Habit formation coincides with shifts in reinforcement representations in the sensorimotor striatum

Citation: Smith, Kyle S., and Graybiel, Ann M. "Habit Formation Coincides with Shifts in Reinforcement Representations in the Sensorimotor Striatum." Journal of Neurophysiology 115, 3 (March 2016): 1487–1498 © 2016 American Physiological Society

As Published: https://doi.org/10.1152/jn.00925.2015

Publisher: American Physiological Society

Persistent URL: http://hdl.handle.net/1721.1/113332

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Habit Formation Coincides with Shifts in Reinforcement Representations in the Sensorimotor Striatum

Kyle S. Smith[1] and Ann M. Graybiel[2]

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire 03755

[2]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Corresponding Authors:

Dr. Kyle Smith

6207 Moore Hall

Hanover, NH 03755

kyle.s.smith@dartmouth.edu

603-646-1486

Dr. Ann Graybiel

43 Vassar Street

Building 46, Room 6133

Cambridge, MA 02139

graybiel@mit.edu

617-253-5785

Running title: Reward dynamics in sensorimotor striatum

Keywords: electrophysiology; learning; reinforcement; basal ganglia

Abstract

Evaluating outcomes of behavior is a central function of the striatum. In circuits engaging the dorsomedial striatum, sensitivity to goal value is accentuated during learning, whereas outcome sensitivity is thought to be minimal in the dorsolateral striatum and its habit-related corticostriatal circuits. However, a distinct population of projection neurons in the dorsolateral striatum exhibits selective sensitivity to rewards. Here, we evaluated the outcome-related signaling in such neurons as rats performed an instructional T-maze task for two rewards. As the rats formed maze-running habits and then changed behavior after reward devaluation, we detected outcome-related spike activity in 116 units out of 1479 recorded units. During initial training, nearly equal numbers of these units fired preferentially either after rewarded runs or after unrewarded runs, and the majority were responsive at only one of two reward locations. With over-training, as habits formed, firing in non-rewarded trials almost disappeared, and reward place preference declined. Thus error-related signaling was lost, and reward signaling became generalized. Following reward devaluation, in an extinction test, post-goal activity was nearly undetectable, despite accurate running. Strikingly, when rewards were then returned, post-goal activity reappeared and recapitulated the original early response pattern, with nearly equal numbers responding to rewarded and unrewarded runs and to single rewards. These findings demonstrate that outcome evaluation in the dorsolateral striatum is highly plastic and tracks stages of behavioral exploration and exploitation. These signals could be a new target for understanding compulsive behaviors that involve changes to dorsal striatum function.

Introduction

Outcome evaluation is essential for behavioral flexibility and learning. Behaviors that become well engrained habits characteristically lack sensitivity to changes in the acquired estimates of outcome value (Dickinson 1985). Also characteristic of habits is a loss of sensitivity to consequences (Graybiel 2008), and this insensitivity is evident in compulsions such as addiction (Everitt and Robbins 2015; Jonkman et al. 2012). The transition into such habitual states requires plasticity in neural networks responsible for coordinating and automating actions (Balleine et al. 2009; Graybiel 2008; Packard 2009; Smith and Graybiel 2013a; Yin and Knowlton 2006; Yin et al. 2008). The ventromedial region of the striatum and the medial part of the dorsal striatum are known to be critical for behaviors requiring value representations leading to learning, but neurons sensitive to reward and task outcome have also been found in the dorsolateral part of the striatum (DLS), including its sensorimotor regions (Apicella et al. 1991; Barnes et al. 2005; Cromwell et al. 2005; Fujii and Graybiel 2003; Hikosaka et al. 1989; Hollerman et al. 1998; Jin and Costa 2010; Jog et al. 1999). Prominent responses of striatal projection neurons (SPNs) at goal-reaching have been shown to evolve through learning (Barnes et al. 2005; Brigman et al. 2013; Jog et al. 1999; Rueda-Orozco and Robbe 2015; Smith and Graybiel 2013a; Thorn et al. 2010). Moreover, a sharp distinction has been made between reward-sensitive DLS neurons and larger classes of DLS neurons active prior to goal-reaching but not active at reward delivery (Schmitzer-Torbert and Redish 2004).

These findings present a paradox. Lesion studies have confirmed the DLS and its corticostriatal circuits as necessary for habitual modes of behavior, but generally not for behavioral flexibility directed by goals and related reinforcement signaling (Daw et al. 2005; Dolan and Dayan 2013; Packard 2009; Yin and Knowlton 2006). Given that action-related spike activity in the DLS has been found to undergo dynamic changes in activity that coincides with habit learning (Barnes et al. 2005; Gremel and Costa 2013; Howe et al. 2011; Jog et al. 1999; Smith and Graybiel 2013a; Tang et al. 2007; Tang et al. 2009; Thorn et al. 2010; Tricomi et al.

2009), we searched for potential plasticity of outcome-related spike responses of DLS SPNs as rats were trained and over-trained on a maze task, and then as they performed after reward-devaluation.

Materials and Methods

Subjects. Six adult male Sprague-Dawley rats (Taconic) were individually housed, maintained on a reverse light-dark cycle, and food restricted to within 85% of pre-surgical weight during the experiments. Experiments were run during their dark (active) cycle. All procedures were approved by the Committee on Animal Care at the Massachusetts Institute of Technology. Data related to the behavior of these animals and the activity of separate task-related striatal units have been reported (Smith and Graybiel 2013a). Here we focused on the activity of a distinct population of units that we found to be active shortly after the cessation of maze runs. Behavioral results (Fig. 1C) here were for six of the rats labeled as the over-trained group in the previous report (Smith and Graybiel 2013a)**.**

T-Maze task. Rats were given training on a T-maze task (Smith and Graybiel 2013a; Smith et al. 2012) in which they waited at a start platform, heard an auditory warning cue, and then after a gate was lowered, ran down the maze (Fig. 1, *A* and *B*). Just prior to the T-junction of the maze, one of two tone cues (1 or 8 kHz) was sounded. This cue instructed the rat to turn left (e.g., 1 kHz) or right (e.g., 8 kHz) to receive ~0.2-ml liquid reward at the goal site. The rewards were chocolate milk and 30% sucrose in water, and each was paired with one of the goal arms for a given rat but pseudorandomly assigned across rats. Entry into the incorrect end-arm resulted in no reward. Rats were left alone to consume the reward after correct runs, or to rest after incorrect runs, before being guided back to the starting platform to await the next trial. Rats did not display any overt preference for one reward over the other; each was consumed fully after correct runs.

Training proceeded in daily ca. 40-trial sessions until a 72.5% accuracy criterion was reached, followed by 10+ days of over-training. Then, the value of one reward was reduced using a devaluation procedure. For this, each rat received 3 pairings of free home-cage consumption of one reward with subsequent injection of 0.3 M lithium chloride (10 ml/kg, i.p.; 3 pairings spaced 48 hr apart). Aversion to the paired reward was confirmed by reduced home-cage intake. Specifically, rats consumed on average close to 30 ml of the reward prior to the first lithium injection, and then consumed ~1 ml in the final test after the lithium pairings (Smith and Graybiel 2013a). Forty-eight hours after the final pairing, the rats were returned to the maze task in an unrewarded probe session followed by a series of normal sessions in which reward was given after correct runs. Typically the devalued reward was not fully consumed after a correct run to it, and in such cases it was removed before the next trial. The identity of the devalued reward was pseudorandomly assigned between rats.

Tetrode implantation and single-unit recordings. Prior to T-maze training, rats were surgically fitted with head-stages containing 12-24 independently drivable tetrodes targeting the infralimbic cortex and the DLS. For the DLS recordings analyzed here, during the first post-operative week, tetrodes were lowered to the target recording location. Final recording positions spanned: −0.1-2.0 mm anterior-posterior relative to bregma; 2.5-4.6 mm lateral to the midline; 3.8-5.8 mm ventral from skull. Electrical signals were amplified at 100-10000, sampled at 32 kHz, band-pass filtered for 600-6000 Hz, and recorded by a Cheetah data acquisition system (Neuralynx, Bozeman, MT) (Smith and Graybiel 2013a). An overhead CCD camera tracked LEDs on the head-stage preamplifiers (30 Hz sampling rate), and photobeams were placed every ca. 17.5 cm. Task-control was provided by a MED-PC program (Med Associates, Inc., St. Albans, VT) or MATLAB (Mathworks, Natick, MA).

Histology. Two days prior to perfusion, the rats were anesthetized, and electrolytic lesions were then made by passing current through the tetrode tips (25 μA, 10 s). For perfusion, the rats were anesthetized with a lethal dose of sodium pentobarbital (100-145 mg/kg) and were perfused through the heart with 0.9% saline followed by 4% buffered paraformaldehyde in 0.1M $NaKPO_4$ buffer. Brains were post-fixed in the paraformaldehyde solution and the placed in cryoprotectant solution (1:3 glycerol in 0.1 M phosphate buffer with sodium-azide), and cut at 30 μm. Transverse sections were stained with cresylecht violet. Some sections were immunostained for activated microglia (CD11b) to mark tetrode tracks and aid their localization under a microscope (Polikov et al. 2005; Smith and Graybiel 2013a).

Data Analysis. Single units were manually isolated (Offline Sorter; Plexon, Inc., Dallas, TX). DLS units classified as putative SPNs were analyzed. By manual classification, these SPN units exhibited characteristic <5 Hz baseline firing rates, bursty firing profiles in autocorrelograms, and medium waveform width, which were features distinct from those of fast-spiking neurons (which exhibited high rates, narrow waveforms, short interspike intervals) and tonically active neurons (which exhibited medium rates, wide waveforms, long interspike intervals) (Barnes et al. 2005; Berke et al. 2004; Kubota et al. 2009; Schmitzer-Torbert and Redish 2004; Smith and Graybiel 2013a).

SPN unit activity was analyzed with respect to co-registered task events: warning cue onset, gate opening, locomotion onset, instruction cue onset, turn start, turn end, and goal arrival (Barnes et al. 2005; Berke et al. 2004; Kubota et al. 2009; Schmitzer-Torbert and Redish 2004; Smith and Graybiel 2013a). Units were divided into three categories: (1) task-responsive, if their spike activity exceeded 2 SDs above a pre-run baseline period for three consecutive 30-ms bins within ±200 ms of a task event as determined by activity averaged over a session (n = 858/1479); (2) post-goal-responsive if their spike activity during a 0.2-3.2 s after goal arrival similarly exceeded 2 SDs above a pre-run baseline period for three consecutive 30-ms bins

within ±200 ms of a task event (n = 116/1479); and (3) non-task-responsive. The 3-s post-goal window was selected so as not to overlap with the goal-arrival period but to avoid the subsequent part of the intertrial interval, when there was greater variability in behavior and recording signal quality. We did not attempt to track single SPNs across sessions; they were treated independently, session by session, as members of the population of units recorded during that session. As a consequence, we were unable to detect potential learning-related plasticity of individual SPNs.

To analyze post-goal activity, firing rates in moving 50-ms windows were normalized for each responsive unit by subtracting activity in each 50-ms window by a baseline average of that unit. The 3-s period preceding the warning cue served as a pre-run baseline period; baseline activity for single sessions was subtracted for each unit from its average event-related activity of interest for that session. For isolating activity on different trial types (e.g., correct runs), the baseline for those trials only was used and subtracted from event-related activity on those trials per unit. Baseline-subtracted 100-ms windows were also used for plotting the data. Response latencies were defined as the first 50-ms window of at least two successive windows after goal arrival in which a responsive post-goal unit exceeded 2 SDs of baseline firing. For comparisons of overall post-goal activity on goal response types and learning stages, the baseline-subtracted average of the 50-ms windows was calculated for the 0.2-3.2 s post-goal time window (i.e., averaging the sixty 50-ms bins). Using these data, T-tests ($P < 0.05$ significance) were applied to differentiate among units with significantly greater post-goal activity to correct, incorrect, left maze reward or right maze reward outcomes, respectively, or units with similar activity between these reward end-arm variables. These analyses were done after averaging activity to each outcome (rewarded or unrewarded) for all such trials in a session in order to minimize the influence of differential trial numbers. ANOVA and Bonferroni-corrected *post-hoc* tests were used to compare activity for run outcomes for each response class, learning stages, response latencies, and magnitudes of firing increases or decreases (i.e., relative to zero after per-unit

baseline-subtraction). The main outcome comparisons of interest were for units responding more to correct than to incorrect runs ('correct related'), units responding more to incorrect than to correct runs ('incorrect related'), and units responding similarly to both ('correct-and-incorrect related'), in direct *t*-test comparisons of their firing rates in the averaged **0.2-3.2 s** post-goal period per session. Due to the more stable trial number of left and right goal entries, the main goal identity comparison was for units responding to the left goal only or to the right goal only ('single-goal responsive'), or left-and-right ('both-goal responsive'), as judged by activity in the 0.2-3.2 s post-goal period relative to baseline activity. Population size changes with learning were compared with a Kolmogorov-Smirnov *z*-test. Linear regression was used to correlate percent correct behavior with population distributions and firing rates. To do this analysis, neural signals were averaged for each session. Those averages included: (1) the proportion of spike activity on correct runs averaged during the 0.2-3.2 s period of all post-goal units $[(Correct_{rate}-Incorrect_{rate}) / (Correct_{rate}+Incorrect_{rate})]$; (2) the proportion of correct-related units of the population $[(Correct_{pop}-Incorrect_{pop}) / (Correct_{pop}+Incorrect_{pop})]$; (3) the summation of these two proportion measures $[(Correct_{pop}-Incorrect_{pop}) / (Correct_{pop}+Incorrect_{pop})] + [(Correct_{rate}-Incorrect_{rate}) / (Correct_{rate}+Incorrect_{rate})]$, and (4) the percentage of single goal location related units of the population [(single goal count/both goal count)*100].

Training sessions were staged in order to maintain consistent levels of performance accuracy and recorded unit counts for analysis: stages 1-2 (first two sessions), stages 3-4 (pairs of sessions ≥60% correct), stage 5 (first pair of sessions ≥72.5%), stages 6-13 (subsequent pairs of sessions ≥72.5%, the criterion for task acquisition). Stages after devaluations were: Probe (unrewarded probe session), stages PP1-2 (first two post-probe rewarded sessions), stages PP3-6 (subsequent pairs of rewarded sessions) (Smith and Graybiel 2013a).

We examined licking behavior in three additional rats on the T-maze task in order to assess in detail the onset of licking after arrival at the goal location and the duration of reward consumption, by means of close-up video recordings of the reward port and frame-by-frame

analysis for three sessions of runs per rat. Goal arrival was marked by a photobeam 5 cm proximal to the reward trough.

We also analyzed the relative distributions of post-goal units and other recorded units based on histological assessment of tetrode tips and *post-hoc* reconstruction of recording sites based on tetrode lowering history. The location of recorded post-goal units were compared to the location of non-task-related and non-goal-related units on dorsal-ventral (DV), anterior-posterior (AP), and medial-lateral (ML) axes using ANOVA.

Results

Behavior. Here we specifically analyzed the activity of outcome-responsive units recorded in 6 rats that were followed through initial training, over-training to induce habit formation (Smith and Graybiel 2013a), reward devaluation, and reinstatement of reward as reported in our prior work (Fig. 1, *A*-*C*) (Smith and Graybiel 2013a). That prior study included 7 rats, but one lacked DLS recordings; thus 6 were included here. These 6 rats reached criterion performance (stage 5), rapidly reached stable performance asymptote by stage 6 (Fig. 1*C*), and then during the unrewarded probe test after devaluation, continued to run accurately as though by habit. During post-probe rewarded sessions, as reported (Smith and Graybiel 2013a), the 6 over-trained rats developed three performance routines: running to the non-devalued goal when so instructed (100% of those trials), running to the devalued goal when so instructed (~40% of trials instructed to the devalued goal), and running the wrong-way to the non-devalued goal (~60% of trials instructed to the devalued goal). Throughout these post-devaluation sessions, rats consistently drank only ~50% of the time after the few correct runs to the devalued goal (corresponding to ~0.4 ml), and drank all non-devalued rewards they received. This behavior suggests stability and specificity of the aversion. The in-task run speed and occurrence of vicarious trial-and-error head-turn movements at the choice point declined over the course of

training, and then rose in post-devaluation sessions (speed and head turns were higher on runs instructed to the devalued compared to non-devalued goal) (Smith and Graybiel 2013a).

In the analysis of licking behavior, we found that behavior after correct runs always included a continuous period of reward licking and swallowing. The mean time from goal arrival to the first lick of reward was 0.55 s (range: 0.40-1.23 s), and the mean licking duration was 12.28 s (range: 7.06-21.21 s). Behavior after incorrect trials involved one or two sniffs at the empty reward trough, followed by quiet sitting with occasional head movement. Only one lick was detected several seconds after an incorrect run in this analysis; otherwise licking was not observed after incorrect trials.

Post-goal unit characterization. Out of the 1479 recorded DLS SPNs, 116 (8%) responded exclusively after goal arrival ('post-goal-related'). This population was distinct from the simultaneously recorded population (~58% of recorded units) that respond phasically during maze running in the same experiments (Smith and Graybiel 2013a). The post-goal-related units did not respond during maze runs at all, but then, after arrival at a goal location, exhibited a sharp rise in spiking at, on average, 359 ms (± 20.51 SEM; Fig. 1, *E* and *F*), slightly preceding the onset of reward licking after correct runs. As noted further below, most units did not respond to both rewards, even though licking occurred for both, so that the unit responses were not obligatorily related to licking. The post-goal response typically began with a phasic peak that ramped to a maximum followed by a mid-level plateau of activation that tapered down to near baseline levels after several seconds (Fig. 1, *E* and *F*). The post-goal units were distinct from a separate population of units that were phasically responsive around the actual goal-arrival time, when the animals stopped running (Smith and Graybiel 2013a); the activity of these goal-arrival units subsided as the post-goal activity emerged (Fig. 1*F*). Thus, the end of maze runs was marked by the breaking of a goal photobeam, a phasic peak of goal-arrival units, and a following distinct excitation of post-goal units. Relative to the total number of recorded DLS

units, the population size of post-goal units remained stable throughout training time (Kolmogorov-Smirnov $z = 1.22$, $P = 0.10$; Fig. 1*D*), except during the probe test (see below).

These post-goal units were widely distributed across recording sites, and could be found on the same tetrode as units exhibiting the task-bracketing pattern that we have reported to be related to habit learning (Barnes et al. 2005; Jog et al. 1999; Regier et al. 2015; Smith and Graybiel 2013a; Thorn et al. 2010) (Figs. 1*F* and 2*A*). Post-goal unit locations overlapped with those of other recorded DLS units, including those responsive or not responsive to goal arrival (Fig. 2*A*). However, the center of gravity of post-goal units was shifted (Fig. 2*B*). Compared to the positions of non-goal related units, post-goal units were more lateral [$F(1,1740) = 11.81$, $P = 0.001$], posterior [$F(1,1740) = 13.85$, $P = 0.000$] and dorsal [$F(1,1740) = 4.28$, $P = 0.039$]. Nevertheless, post-goal units, like SPNs with task-related responses, could be found throughout the recorded DLS region. This histological assessment also showed that for the post-goal units, the range of their DV coordinates was similar across training stages, due to similar variations in tetrode lowering and non-lowering across training stages [Stages 1-4, DV range in mm: 4.0-5.5; Stages 5-8, range: 3.8-5.5, Probe day, 4.5 (one unit); post-probe, range: 3.8-5.6].

Plasticity of spike activity related to habit emergence. During early training, post-goal unit activity occurred after both rewarded and unrewarded trials, as though the post-goal units as a population tracked the outcomes of both correct (rewarded) and incorrect (unrewarded) outcomes of the runs. The majority of individual post-goal units nevertheless exhibited a clear-cut firing bias toward responding either for rewarding outcomes or for unrewarded outcomes after erroneous behavior (Fig. 3, *A* and *B*). A minority of units encoded both possible outcomes similarly.

The proportions of these classes of post-goal units markedly changed as training continued into the over-training period, demarcating the transition from learning to habit (Kolmogorov-Smirnov $z = 2.31$, $P < 0.001$; Fig. 3*C*). During task acquisition (stages 1-4), relatively similar

numbers of post-goal units responded after correct runs (42% of recorded; 8/19) and after incorrect runs (32%; 6/19), with some responding after both correct and incorrect runs (26%; 5/19). During the over-training period, however, the post-goal population became progressively more heavily weighted toward firing at rewarding outcomes (Figs. 3*C* and 4*A*). During early over-training (stages 5-8), the distribution shifted to favor correct-only units (42%; 10/24) or correct-and-incorrect units (38%; 9/24) over incorrect-only units (21%; 5/24). By the end of over-training (stages 9-13), there was a greater than 12:1 ratio of units firing more after correct runs (68%; 25/37) to units responding more after incorrect runs (5%; 2/37) (Fig. 4*A*), a sharp change from the 1.3:1 ratio during initial acquisition. The proportion of correct-and-incorrect responding units also fell (27%; 10/37). This change in outcome representation during habit formation occurred without a change in the total proportion of post-goal responsive units recorded (Fig. 1*D*), suggesting that there was a shift in the preferred outcome-related firing of a post-goal ensemble of units of stable ensemble size.

The proportion of reward-related units compared to error-related units was positively correlated with percent correct maze runs (R^2 = 0.25, F = 5.05, P = 0.04; Fig. 4*B*). The firing rates of reward-related units relative to error-related units trended toward a positive correlation with run accuracy as well, but did not reach statistical significance (R^2 = 0.14, F = 2.05, P = 0.18; Fig. 4*B*). This index adding these populations and firing rate proportions to capture both changes together was highly correlated with run accuracy (R^2 = 0.44, F = 10.06, P = 0.007; Fig. 4*C*). Thus, accuracy increased as the disparity between correct signaling and error signaling increased.

We next examined the firing dynamics of the neuronal subpopulations in outcome response categories (correct-related, incorrect-related, or both). A three-factor ANOVA detected modulation of post-goal activity by learning stage [F(3,219) = 2.3, P = 0.002], but not by response category (P = 0.29) or by trial outcome (correct or incorrect; p = 0.31). Significant interactions were found between response category and learning [F(6,219) = 2.2, P < 0.001] as

well as response category and trial outcome [$F(2,219) = 17.50$, $P < 0.001$]. Thus post-goal firing rates changed according to stages of habit expression, and subpopulations with particular outcome responses could change their response preferences according both to the level of habit expression exhibited by the animals and to the actual outcome received after runs (see Fig. 4).

Units responding more to rewarded (correct-related) than to unrewarded (incorrect-related) outcomes maintained a consistent elevation of firing rate to reward relative to their baseline rates and to firing after incorrect runs across training stages (Fig. 4*D*). Firing after incorrect runs was not different from baseline at any stage in these correct-related units. By contrast, the activity of units responding more after incorrect runs followed a similar trend as did their population size, with outcome-specific signaling declining as habitual runs emerged during over-training. During initial training, their responses after incorrect runs exceeded their baseline rates and activity after correct runs (Fig. 4*E*). However, during early over-training, the activity of these units became variable and, on average, not significantly elevated above baseline (Fig. 4*E*). The lack of consistent firing above baseline continued during late-overtraining (Fig. 4*E*), when the population size of error-related units had considerably diminished. Despite meeting the criterion set for responsive to errors more than to rewards, the average firing rates of these units were variable enough to fail to exceed their firing to rewards at this late training phase (Fig. 4*E*). Thus, the outcome specificity of these incorrect-outcome units as an ensemble was lost with over-training. Interestingly, the post-goal units with equivalent responses after both correct and incorrect runs responded similarly across the time of habit formation (Fig. 4*F*), and thus their activity appeared unrelated to learning measures.

To assess aggregate strength of reward and error signaling, we compared the firing rates of the post-goal population in correct and incorrect trials. We combined all post-goal units together, calculated an average firing rate for the 0.2-3.2 s post-goal period for each, and used a three-way ANOVA to compare their activity as a function of trial outcome and learning stage.

Firing rates were significantly and differentially modulated by outcome during learning [outcome: $F(1,157) = 8.42$, $P = 0.004$; stage: $F(2,157) = 2.64$, $P = 0.075$; interaction: $F(2,157) = 3.13$, $P = 0.046$]. During training, the average spike rate was similar after correctly and incorrectly performed runs ($P > 0.05$) (Fig. 5*A*, left). A trend toward increased firing rates occurred after correct runs during early over-training but was not statistically different from similarly variable activity after incorrect runs ($P > 0.05$) (Fig.5*A*, middle). During late over-training and habit emergence, this trend became significant ($P < 0.05$) (Fig. 5*A*, right). Thus with full over-training, firing to rewards eclipsed firing to errors, just as the shift away from error representation in population makeup occurred.

The latencies of the spike responses of the post-goal units also changed, becoming shorter after correct runs than after incorrect runs, regardless of training stage [learning: $F(2,167) = 0.13$, $P = 0.88$; outcome: $F(1,167) = 11.31$, $P = 0.001$; interaction: $F(2,167) = 0.31$, $P = 0.73$] (see Fig. 5, *B* and *C*). These data suggested a slight delay in error signaling compared to reward signaling, with little relation of this latency change to stages of habit formation. This difference could partly reflect the weak firing after incorrect runs in late over-training, though the lack of interaction with training suggests that it could be due to a factor not changing with learning, such as covert reward or error detection timing (Fig. 5*C*).

Finally, we compared the session-averaged pre-run baseline firing rates of the post-goal units across learning stages. They did not change. Nor did the baseline firing distinguish trial-to-trial accuracy [learning stage: $F(3,220) = 2.14$, $P = 0.096$; accuracy: $F(1,220) = 0.20$, $P = 0.66$; interaction $F(3,220) = 0.51$, $P = 0.68$].

Our analyses make it unlikely that the occurrence of fewer incorrect trials during late over-training accounted for these differences across training. First, the analyses were averaged over trials, resulting in identical numbers of data points for correct and incorrect runs per session for each unit. Further, the correct-and-incorrect-related units maintained incorrect responsivity during over-training, and the simultaneously recorded units active around run cessation fired

equally on correct and incorrect runs, even during very late over-training with few incorrect trials, as found also in earlier work (Smith and Graybiel 2013a; Thorn et al. 2010) (Fig. 6*A*). We earlier reported that the task-related SPNs active during the maze runs, including those active at goal arrival, fired similarly during correct and incorrect runs throughout learning (Smith and Graybiel 2013a). Moreover, the spike activity of these task-related units (n = 858), if anything, exhibited an opposite pattern of responding after correct or incorrect runs to that shown by the post-goal units (Fig. 6*B*), further underscoring the distinctive response profiles of the post-goal population.

Post-devaluation training: DLS outcome signaling plasticity as the habit breaks. The insensitivity of over-trained behavior to reward devaluation, measured in an unrewarded probe test, offered an opportunity to test whether the post-goal activity that had characterized firing during habitual performance would be retained during such runs, or would instead change according to expected or received outcome. Strikingly, there was a near total absence of post-goal activity during this probe test (Fig. 1*E*). Only 1/54 recorded units exhibited a post-goal response. This sudden reduction suggested that the activity after correct runs during over-training likely required the presence of a reward, and that the absence of activity after unrewarded runs coincided with habit expression in both late over-training sessions and the probe test.

We then evaluated outcome responses during the subsequent post-probe rewarded sessions in order to assess dynamics when the acquired behavior changed, as well as to compare the responses to valued versus devalued outcomes.

Strikingly, significant post-goal activity returned during post-devaluation training (Figs. 1*E*, 4 and 7), and similar numbers of post-goal units responded after correct runs (13/36 units), incorrect runs (10/36), and both correct and incorrect runs (13/36). Again, post-goal activity was significantly elevated in rewarded trials in correct-related units (P = 0.001 vs. baseline; P =

0.003 vs. error trials), and in incorrect trials to incorrect-related units (P = 0.021 vs. baseline; P = 0.002 vs. reward trials) (Fig. 4, *D*-*F*). Thus the response profile after return of the rewards recapitulated the profile found during the initial training phase, despite the fact that the animals had gone through over-training

The spike rates of post-goal units differentiated among the three types of post-devaluation runs [$F(2,91) = 3.83$, $P = 0.025$]: correct runs to the devalued goal (a reward that was known but newly aversive), correct runs to the non-devalued goal (a known and valuable reward), and wrong-way runs to the non-devalued goal (no reward) (Fig. 7). There were no incorrect runs when the rats were instructed to go to the non-devalued goal. Activity remained pronounced after rewarded runs to the non-devalued goal ($P < 0.001$), but was not significantly elevated either after wrong-way (unrewarded) runs to the same goal location (P = 0.18) or after runs to the devalued goal (P = 0.16). In direct comparison, spiking was higher after the rewarded runs to the non-devalued goal relative to spiking after wrong-way runs to this goal (P = 0.031), and spiking was not quite significantly higher relative to runs to the devalued reward (P = 0.086), suggesting a value component to the post-goal response.

Shift from specific to general goal signaling with habit learning. In addition to being sensitive to reinforcement outcome, post-goal units were sensitive to goal position in the maze, irrespective of recording hemisphere (Fig. 8*A*). Consistent with results from prior maze experiments with single rewards given at different locations (Schmitzer-Torbert and Redish 2004), about one-third of post-goal units responded at the left goal, one-third at the right goal, and one-third at both goals, and the average firing rates at the two goals were similar on average [$F(1,229) = 0.26$, $P = 0.61$]. This pattern mirrors the similarly mixed distribution of task-related DLS units that favor left versus right turns on the T-maze (Smith and Graybiel 2013a; Thorn et al. 2010).

However, as with firing patterns related to run accuracy, the distribution of units with specific goal responses changed markedly with learning (Fig. 8, *B*-*D*), significantly so as indicated by an ANOVA for distribution of goal responsivity (left, right, both) by learning stage [$F(4,115) = 3.49$, $P = 0.01$]. During task acquisition (stages 1-4), the majority (68%) of post-goal units responded selectively to one goal (e.g., left) but not to the other (e.g., right), while fewer responded similarly to both goals (Fig. 8*B*). As a result, post-goal representations during this time of trial-and-error learning were highly heterogeneous and covered the factorial combinations of correct/incorrect and left/right possibilities with nearly equal proportion (Fig. 8*D*, left). Thus, some units responded to only correct runs to the left goal, others incorrect runs to the left goal, others correct or incorrect runs to the right goal, and others correct and/or incorrect runs to either goal.

With over-training (stages 5-13), the distribution of reward-specific responses shifted to one in which the majority of post-goal units responded similarly to both left and right goals (Fig. 8*B*); 67% of post-goal units responded to both rewards in early overtraining, and 65% in late overtraining, in comparison to 32% that did so during initial training. This shift to a more generalized relation to reward occurred in parallel to the accentuation of firing related to correct outcomes over firing related to incorrect outcomes. These findings suggest that, as the representation of error outcomes was reduced in favor of reward outcomes during over-training and habit formation, so too were reward-specific representations, which were increasingly replaced by reward-general representations. During the end of over-training (stages 9-13), by far the most common (49%) representation in post-goal units was a response to reward outcomes at either goal location after correct runs (Fig. 8*D*).

In training sessions after reward devaluation (PP1-4), the post-goal responses shifted again, now selective more often to single goals rather than to both goals. The proportions of single-goal and both-goal responses bore close resemblance to the initial proportions detected during the task acquisition phase. In *post hoc* analysis from the ANOVA described above, early

training and post-devaluation sessions did not differ from one another in the proportion of single-goal-responsive units ($P > 0.05$), and both differed significantly from the early and late over-training periods (each, $P < 0.05$) — which themselves did not differ from one another ($P > 0.05$; Fig. 8*B*).

This pattern of change from reward-specific responses during period of initial training and early over-training to a more homogenous goal representation during the period of habitual behavior during later over-training occurred simultaneously with the learning-related shifts in correct vs. incorrect outcome responding. Thus, after the devaluation, the post-goal units regained the fairly uniform distribution of different responses to combinations of goal location and reward outcome that was noted during initial training (Fig. 8*D*, right). Finally, the greater proportion of post-goal units responding to both goals correlated positively with increasing run accuracy, as uncovered by a linear regression test on goal responsivity and percent correct performance per session [$R^2 = 0.13$; $F(1,115) = 16.81$, $P < 0.001$; Fig. 8C], suggest a similar relevance to behavior as that of the decline in error-related signaling.

Discussion

Signals related to the success or failure of a goal-directed behavior can drive learning and can help to stamp in the successful behavior as a habit. Such signals are typically the domain of associative-limbic corticostriatal loops and their midbrain dopamine-containing inputs. Models of these loops suggest shifting of control across associative and sensorimotor corticostriatal loops as behavior shifts between being flexible and being relatively automatic. A central finding in the field, confirmed here, is that outcome-related signaling occurs in the sensorimotor part of the striatum (Apicella et al. 1991; Barnes et al. 2005; Cromwell et al. 2005; Desrochers et al. 2015; Fujii and Graybiel 2003; 2005; Hikosaka et al. 1989; Jin et al. 2009; Schmitzer-Torbert and Redish 2004; Thorn et al. 2010). Of great interest, however, is how trial-

to-trial outcome information is incorporated in action plans through these semi-segregated systems as habits are acquired (Desrochers et al. 2015).

Here we demonstrate that there is a specialized population of neurons in the sensorimotor striatum that change their post-performance responses during learning, shifting from nearly equal responsivity after both successful and unsuccessful behavioral episodes early on to highly biased signaling of rewarded outcomes with only rare signaling of errors. Thus error signaling is a property of the sensorimotor striatum during early learning, but fades if habits are established. The firing patterns of these post-goal neurons also shift from being selective to particular goals — here to one of two possible positions — to a more global response signaling correct performance regardless of the goal position attained. Yet these apparently stable, acquired firing patterns related to outcome can be abruptly reset if behavioral flexibility is again required, a situation that we here imposed experimentally by devaluing one reward. Thus, this population of sensorimotor striatal neurons can track success and failure during early learning, then reduce the error signaling once the behavior being learned becomes habitual, but revert to signaling both success and failure if conditions change. The devaluation experiments also uncovered clear modulation of DLS outcome signaling by reward value. These findings suggest that the sensorimotor striatum can contribute not only to action encoding but also, through outcome signaling, to the mechanisms underlying transitions between goal-directed and habitual behavior.

Many behavioral studies have supported the notion that transitions to habitual performance are marked by a loss of sensory-specific representations of outcomes (e.g., chocolate milk) in favor of general outcome expectations (e.g., something tasty), and as well, a loss of the representation of specific response-outcome contingencies (Balleine and Dickinson 1998; Smith and Graybiel 2014; Yin and Knowlton 2006). Habitual behaviors characteristically are insensitive to changes in reward value or action-reward contingencies that tap into their sensory specific qualities (e.g., conditioned taste aversion) (Dickinson 1985; Holland and Wheeler 2009).

Evidence for corresponding shifts in neural activity is much less extensive. However, behavioral studies have clearly demonstrated that behavior becomes more sensitive to specific outcomes after disruption of habit-related brain regions, including the sensorimotor striatum, and recording studies have shown that spike activity in these regions develops strong task-related patterns that could plausibly override outcome expectation signals operating in parallel in the limbic system (Balleine et al. 2009; Barnes et al. 2005; Gremel and Costa 2013; Hitchcott et al. 2007; Killcross and Coutureau 2003; Smith and Graybiel 2014; Tricomi et al. 2009; Yin and Knowlton 2006). Some dampening of outcome representations during the course of over-training on tasks has also been reported in limbic regions (Atallah et al. 2014; Gremel and Costa 2013; McDannald et al. 2012; Thorn et al. 2010), which may help to produce a limbic-sensorimotor imbalance favoring habits. Our finding here of the loss of error signaling in the DLS as habits are formed could be part of the mechanism by which habitual behaviors become resistant to change despite negative feedback.

Our findings challenge a purely parallel systems view of sensorimotor and limbic-associative loop function during habit learning by demonstrating that the shifts in outcome signaling can in part occur in the sensorimotor circuit itself. This signaling, as well as reported action-related plasticity occurring at the level of limbic medial prefrontal cortex (Hitchcott et al. 2007; Killcross and Coutureau 2003; Smith and Graybiel 2013a), might also help to resolve the issue of how these systems are selected for influence over behavior (Daw et al. (2005). If plasticity related to both behavior and outcome is occurring within both limbic and sensorimotor systems, as proposed previously (Corbit and Janak 2007), a between-system arbitrator may not be essential. This possibility is raised by the finding that local field potential activity in the dorsomedial ('goal-related') striatum and dorsolateral ('habit-related') striatum becomes synchronized at different sub-bands within the theta-band as a result of learning (Thorn and Graybiel 2014). The shift in firing patterns reported here reflects activity around received outcomes, rather than around outcome expectations occurring during behavior. Habits are

behaviors that are maintained from trial to trial, and signaling in both time periods could play a role in that maintenance.

This notion speaks directly to the paradox that we highlighted between models that propose a functional dichotomy between DLS and ventral/medial striatum, wherein DLS functions to learn and control habitual action, and the results of neural recordings in the DLS demonstrating a diversity of activity patterns during habit formation and maintenance (Graybiel 2008; Smith and Graybiel 2014). Our favored interpretation is that such models, which have been sculpted in part from elegant loss-of-function studies, should be challenged to build a more nuanced story incorporating such diverse neural representations of behavior. We propose that habits might arise from multiple different signaling processes within the DLS and elsewhere, including outcome-related signaling as shown here and potentially additional as-yet unknown signaling processes (Smith and Graybiel 2013a; 2014).

An additional speculation about the function of the outcome encoding of post-goal DLS activity is that it could indicate surprise. Stimuli with surprise value, that is, stimuli that are not fully predicted, can guide associative learning (Pearce and Hall 1980). The consistent reward-related representations might reflect a process of updating and maintaining habit-related associations acquired through training, and the loss of error-related signaling with habit formation could contribute to an associative updating in response to negative feedback. In combination with a stable DLS representation of the task structure, these outcome-signaling patterns could contribute to the maintenance of habits as well as to the loss of error-corrective sensitivity that can be characteristic of habitual behavior, and particularly characteristic of addictions. Outcome reporting has been singled out as a prime neural correlate of reinforcement learning for some neurons in the ventromedial striatum, including the tonically active interneurons thought to correspond to cholinergic interneurons (Atallah et al. 2014), suggesting that this process may be striatum-wide. However, we do not know whether outcome-related signals in other regions of the striatum would follow similar learning-related

changes as those we observed here for habit learning. If not, these signals could have a special function in DLS processing, but if so, similar outcome processing could be more widely distributed than was recognized. The learning-related plasticity of the SPNs recorded here is not likely related to differences in recording depth across learning stages: (1) the range of DLS locations from which post-goal units were recorded was similar across learning stages; (2) after reward devaluation, we observed a reversal of many of the post-goal related activity changes seen during over-training (e.g., return of error-related activity, return of units with responses to single rewards), regardless of recording depth; and (3) there was only one post-goal unit recorded during the unrewarded probe test, suggesting further a relation of activity to task variables rather than to anatomical sites in the DLS.

Plasticity in firing profiles during such exploration-exploitation shifting during habit learning is also characteristic of many DLS units that are active during performance of the task, before the outcome has be achieved (Barnes et al. 2005). DLS activity during performance increases with habit emergence (Gremel and Costa 2013; Jog et al. 1999; Tricomi et al. 2009), yet habits are marked by a temporal restructuring of activity during behavior (Barnes et al. 2005; Jog et al. 1999; Smith and Graybiel 2013a; Tang et al. 2007; Tang et al. 2009; Thorn et al. 2010). The post-performance signaling in the DLS, however, is particularly notable given the strong evidence that this striatal region is responsible for the performance, goal-indifferent phase of behavior (Packard 2009; Yin and Knowlton 2006; Yin et al. 2008). The timing of changes in post-goal activity was not aligned to the changes in task-related activity recorded simultaneously. The post-goal activity changes here were aligned with the emergence (and later loss) of a habitual strategy, whereas task-related activity developed a beginning-and-end habit-related pattern much earlier during learning, and this pattern was maintained after devaluation (Smith and Graybiel 2013a). Determining the relative functional contributions of action-related and post-goal-related signals in the DLS is potentially addressable with

optogenetic tools if the units can be selectively manipulated at fine timescales (Gremel and Costa 2013; Smith and Graybiel 2013a; b; Smith et al. 2012).

Evidence indicates that DLS outcome-related firing occurs in rodents in a range of non-cued tasks, including free maze navigation for rewards (Schmitzer-Torbert and Redish 2004), in free performance of a lever-press sequence (Lumeire and Graybiel 2014), and also in non-human primates developing habitual visual scanning patterns (Desrochers et al. 2015). In the context of literature indicating key roles for DLS in actions sequences (Aldridge et al. 2004; Packard 2009; Root et al. 2010; Smith and Graybiel 2014; Yin and Knowlton 2006; Yin et al. 2008), it is plausible that the outcome signaling changes demonstrated here reflect the evaluation of actions. Other potential features of this evaluation, including signals related to uncertainty or probability, response-outcome associations, integrated cost-benefit signals, and related decision-making computations, could be critical here. Nevertheless, we found clear evidence that these DLS responses can carry information concerning reward identity, reward value, and reward absence, suggesting a role for them in reinforcement learning and decision-making processes. The activity of the post-goal units was not closely linked with oromotor movements, which, in contrast to post-goal activity, were stable across the reward/no-reward conditions and learning stages. Clearly, however, we cannot rule out a potential motor component to the variables that contribute to DLS outcome signals.

It is unclear how the heterogeneity of SPN responses observed here and in prior studies relates to major SPN subtypes as defined chemically and anatomically (e.g., striosome/matrix neurons, D1/D2 receptor-positive neurons) (Barnes et al. 2005; Berke et al. 2009; Kimchi et al. 2009; Schmitzer-Torbert and Redish 2004; Smith and Graybiel 2013a; Tang et al. 2007; Thorn et al. 2010). Even in this task, we detected SPNs responding to maze run events (and subclasses responsive to specific events), SPNs responding in these various forms in the post-goal period, and SPNs not responsive to task events. It is possible that inhibitory responses were undetected in our analyses due to the generally low basal firing rates of SPNs, but they

could add further heterogeneity to DLS representations of the maze behavior. It has been found that both D1-containing and D2-containing SPNs convey overlapping and complementary information relating to action sequence performance (Jin et al. 2014), but their relevance to the post-goal population here remains unclear. Loss-of-function studies not discriminating task-time periods have uncovered roles for indirect pathway striatopallidal SPNs expressing D2 and adenosine 2A receptors in the devaluation insensitivity characteristic of habitual reward seeking, suggesting at least this pathway could be involved (Corbit et al. 2014)Shan et al., 2015). Mice will work for direct, but not indirect, pathway stimulation (Kravitz et al. 2012), raising the potential for those circuits to contain value-related information that could affect the outcome signaling shown here. Dopamine-containing inputs likely contribute to the outcome signaling, given its close relationship to many variables that appear to modulate DLS post-goal activity (Bromberg-Martin et al. 2010; Schultz 2006) and evidence that striatal dopamine release in the DLS can occur after both correct and incorrect goal-seeking behaviors (Howe et al. 2013).

Our findings add to this view of multi-dimensional signaling in the sensorimotor striatum. We demonstrate that even well learned behaviors that have become habitual are subject to monitoring by at least the post-goal population of neurons in the sensorimotor striatum, and that during the acquisition of habitual behavior, this type of monitoring changes from including error feedback during habit formation to excluding error feedback once the habits have been established. This change corresponds in time to a behavioral shift from goal-specific to goal-general responses. An important possibility is that these signals could provide a new potential source of neural dysfunction related to overly fixed behaviors, as exemplified by addictive behavior. Drug craving has in human addicts been linked to the presumed analogue of the DLS in human, within the putamen (Volkow et al. 2006). In rodents, compulsive aspects of drug seeking, as modeled by resistance to a punisher or continued seeking in the presence of drug cues despite lack of drug, also engage and require the DLS (Everitt and Robbins 2015; Jonkman et al. 2012; Willuhn et al. 2012). It is possible that the outcome-related signals we

describe here, with their potential loss of error-related signaling as habits are acquired, could contribute to inflexible behaviors including drug-seeking and drug-taking.

Acknowledgements: It is a pleasure to thank Dordaneh Sugano, Arti Virkud, Christine Keller-McGandy, and Jannifer Lee for help with the experiments and histology, and Dr. Yasuo Kubota for his help with manuscript preparation. This work was supported by National Institutes of Health grants R01 MH060379 (A.M.G.) and F32 MH085454 (K.S.S.); by Office of Naval Research grant N00014-04-1-0208 (A.M.G.); by Nancy Lurie Marks Family Foundation (A.M.G.); by European Union grant 201716 (A.M.G.); and by funding from Mr. R. Pourian and Julia Madadi (A.M.G.). The authors declare no competing interests.

References

Aldridge JW, Berridge KC, and Rosen AR. Basal ganglia neural mechanisms of natural movement sequences. *Can J Physiol Pharmacol* 82: 732-739, 2004.

Apicella P, Ljungberg T, Scarnati E, and Schultz W. Responses to reward in monkey dorsal and ventral striatum. *Exp Brain Research* 85: 491-500, 1991.

Atallah HE, McCool AD, Howe MW, and Graybiel AM. Neurons in the ventral striatum exhibit cell-type-specific representations of outcome during learning. *Neuron* 82: 1145-1156, 2014.

Balleine BW, and Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37: 407-419, 1998.

Balleine BW, Liljeholm M, and Ostlund SB. The integrative function of the basal ganglia in instrumental conditioning. *Behav Brain Res* 199: 43-52, 2009.

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

Berke JD, Breck JT, and Eichenbaum H. Striatal versus hippocampal representations during win-stay maze performance. *J Neurophysiol* 101: 1575-1587, 2009.

Berke JD, Okatan M, Skurski J, and Eichenbaum HB. Oscillatory entrainment of striatal neurons in freely moving rats. *Neuron* 43: 883-896, 2004.

Brigman JL, Daut RA, Wright T, Gunduz-Cinar O, Graybeal C, Davis MI, Jiang Z, Saksida LM, Jinde S, Pease M, Bussey TJ, Lovinger DM, Nakazawa K, and Holmes A. GluN2B in corticostriatal circuits governs choice learning and choice shifting. *Nat Neurosci* 16: 1101-1110, 2013.

Bromberg-Martin ES, Matsumoto M, and Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* 68: 815-834, 2010.

Corbit LH, and Janak PH. Inactivation of the lateral but not medial dorsal striatum eliminates the excitatory impact of Pavlovian stimuli on instrumental responding. *J Neurosci* 27: 13977-13981, 2007.

Corbit LH, Nie H, and Janak PH. Habitual responding for alcohol depends upon both AMPA and D2 receptor signaling in the dorsolateral striatum. *Front Behav Neurosci* 8: 301, 2014.

Cromwell HC, Hassani OK, and Schultz W. Relative reward processing in primate striatum. *Exp Brain Res* 162: 520-525, 2005.

Daw ND, Niv Y, and Dayan P. Actions, Policies, Values, and the Basal Ganglia. In: *Recent Breakthroughs in Basal Ganglia Research*, edited by Bezard E. Hauppauge, NY: Nova Science Publishers, 2005, p. 91–106.

Desrochers TM, Amemori K, and Graybiel AM. Habit Learning by Naive Macaques Is Marked by Response Sharpening of Striatal Neurons Representing the Cost and Outcome of Acquired Action Sequences. *Neuron* 87: 853-868, 2015.

Dickinson A. Actions and habits: the development of behavioral autonomy. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 308: 67–78, 1985.

Dolan RJ, and Dayan P. Goals and habits in the brain. *Neuron* 80: 312-325, 2013.

Everitt BJ, and Robbins TW. Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual review of psychology* 2015.

Fujii N, and Graybiel AM. Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science* 301: 1246-1249, 2003.

Fujii N, and Graybiel AM. Time-varying covariance of neural activities recorded in striatum and frontal cortex as monkeys perform sequential-saccade tasks. *Proc Natl Acad Sci U S A* 102: 9032-9037, 2005.

Graybiel AM. Habits, rituals, and the evaluative brain. *Annual review of neuroscience* 31: 359-387, 2008.

Gremel CM, and Costa RM. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat Commun* 4: 2264, 2013.

Hikosaka O, Sakamoto M, and Usui S. Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J Neurophysiol* 61: 814-832, 1989.

Hitchcott PK, Quinn JJ, and Taylor JR. Bidirectional modulation of goal-directed actions by prefrontal cortical dopamine. *Cereb Cortex* 17: 2820-2827, 2007.

Holland PC, and Wheeler DS. Representation-Mediated Food Aversions. In: *Conditioned Taste Aversion: Behavioral and Neural Processes*, edited by Reilly S, and Schachtman T. Oxford: Oxford University PRess, 2009, p. 196-225.

Hollerman JR, Tremblay L, and Schultz W. Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J Neurophysiol* 80: 947-963, 1998.

Howe MW, Atallah HE, McCool A, Gibson DJ, and Graybiel AM. Habit learning is associated with major shifts in frequencies of oscillatory activity and synchronized spike firing in striatum. *Proc Natl Acad Sci U S A* 108: 16801-16806, 2011.

Howe MW, Tierney PL, Sandberg SG, Phillips PE, and Graybiel AM. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* 500: 575-579, 2013.

Jin DZ, Fujii N, and Graybiel AM. Neural representation of time in cortico-basal ganglia circuits. *Proc Natl Acad Sci U S A* 106: 19156-19161, 2009.

Jin X, and Costa RM. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* 466: 457-462, 2010.

Jin X, Tecuapetla F, and Costa RM. Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat Neurosci* 17: 423-430, 2014.

Jog MS, Kubota Y, Connolly CI, Hillegaart V, and Graybiel AM. Building neural representations of habits. *Science* 286: 1745-1749, 1999.

Jonkman S, Pelloux Y, and Everitt BJ. Differential roles of the dorsolateral and midlateral striatum in punished cocaine seeking. *J Neurosci* 32: 4645-4650, 2012.

Killcross S, and Coutureau E. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb Cortex* 13: 400-408, 2003.

Kimchi EY, Torregrossa MM, Taylor JR, and Laubach M. Neuronal correlates of instrumental learning in the dorsal striatum. *J Neurophysiol* 102: 475-489, 2009.

Kravitz AV, Tye LD, and Kreitzer AC. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat Neurosci* 15: 816-818, 2012.

Kubota Y, Liu J, Hu D, DeCoteau WE, Eden UT, Smith AC, and Graybiel AM. Stable encoding of task structure coexists with flexible coding of task events in sensorimotor striatum. *J Neurophysiol* 102: 2142-2160, 2009.

Lumeire N, and Graybiel AM. Representation of habitual action sequences in the sensorimotor corticostriatal circuit. *Society for Neuroscience Abstracts* 2014.

McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, and Schoenbaum G. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. *Eur J Neurosci* 35: 991-996, 2012.

Packard MG. Exhumed from thought: basal ganglia and response learning in the plus-maze. *Behav Brain Res* 199: 24-31, 2009.

Pearce JM, and Hall G. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* 106: 532-552, 1980.

Polikov VS, Tresco PA, and Reichert WM. Response of brain tissue to chronically implanted neural electrodes. *J Neurosci Methods* 148: 1-18, 2005.

Regier PS, Amemyia S, and Redish AD. Hippocampus and subregions of the dorsal striatum respond differently to a behavioral strategy change on a spatial navigation task. *J Neurophysiol* jn 00189 02015, 2015.

Root DH, Tang CC, Ma S, Pawlak AP, and West MO. Absence of cue-evoked firing in rat dorsolateral striatum neurons. *Behav Brain Res* 211: 23-32, 2010.

Rueda-Orozco PE, and Robbe D. The striatum multiplexes contextual and kinematic information to constrain motor habits execution. *Nat Neurosci* 2015.

Schmitzer-Torbert N, and Redish AD. Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J Neurophysiol* 91: 2259-2272, 2004.

Schultz W. Behavioral theories and the neurophysiology of reward. *Annual review of psychology* 57: 87-115, 2006.

Smith KS, and Graybiel AM. A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron* 79: 361-374, 2013a.

Smith KS, and Graybiel AM. Investigating habits: strategies, technologies and models. *Front Behav Neurosci* 8: 39, 2014.

Smith KS, and Graybiel AM. Using optogenetics to study habits. *Brain Res* 1511: 102-114, 2013b.

Smith KS, Virkud A, Deisseroth K, and Graybiel AM. Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proc Natl Acad Sci U S A* 109: 18932-18937, 2012.

Tang C, Pawlak AP, Prokopenko V, and West MO. Changes in activity of the striatum during formation of a motor habit. *Eur J Neurosci* 25: 1212-1227, 2007.

Tang CC, Root DH, Duke DC, Zhu Y, Teixeria K, Ma S, Barker DJ, and West MO. Decreased firing of striatal neurons related to licking during acquisition and overtraining of a licking task. *J Neurosci* 29: 13952-13961, 2009.

Thorn CA, Atallah HE, Howe MW, and Graybiel AM. Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron* 66: 781-795, 2010.

Thorn CA, and Graybiel AM. Differential entrainment and learning-related dynamics of spike and local field potential activity in the sensorimotor and associative striatum. *J Neurosci* 34: 2845-2859, 2014.

Tricomi E, Balleine BW, and O'Doherty JP. A specific role for posterior dorsolateral striatum in human habit learning. *Eur J Neurosci* 29: 2225-2232, 2009.

Volkow ND, Wang GJ, Telang F, Fowler JS, Logan J, Childress AR, Jayne M, Ma Y, and Wong C. Cocaine cues and dopamine in dorsal striatum: mechanism of craving in cocaine addiction. *J Neurosci* 26: 6583-6588, 2006.

Willuhn I, Burgeno LM, Everitt BJ, and Phillips PE. Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proc Natl Acad Sci U S A* 109: 20703-20708, 2012.

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

Yin HH, Ostlund SB, and Balleine BW. Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur J Neurosci* 28: 1437-1448, 2008.

Figure Legends

Fig. 1. T-maze task and representative DLS post-goal activity. *A*: T-maze task design. *B*: Training timeline. Task phases demarcate periods of task acquisition (stages 1-4), early habit formation (stages 5-8), late habit formation (stages 9-13), unrewarded probe tests after reward devaluation, and post-probe (PP) rewarded sessions. Accuracy criterion was 72.5% correct ($P < 0.05$, χ^2 test compared to chance). *C*: Maze run accuracy over training stages. *D*: Mean percentage of post-goal units relative to total recorded units (black). *E*: Raster plots (top) and histograms (bottom) showing session-averaged activity of two representative post-goal DLS units with firing rate increases only during the period after goal arrival (50-ms bins: 1 s before task initiation, ±200 ms around run start (S) and goal arrival (GA), 3 s after goal arrival). Learning stages in which units were recorded are noted inside histograms (white number). Gray lines denote run start and goal arrival. *F*: Activity of two units recorded simultaneously on the same tetrode during early over-training with different response patterns during maze runs (top) and during post-goal time (bottom). Peri-event windows marked by vertical gray lines show middle half of median peri-event time between the prior and next events, averaged across trials. W, warning cue; G, gate opening; S, run start; I, instruction cue; TS, turn start; TE, turn end; GA, goal arrival.

Fig. 2. Anatomical distribution of post-goal units in DLS relative to non-goal task-responsive units. *A*: Functional maps of recording locations. Recorded units shown by squares on coronal section, shaded to reflect number of post-goal units (red), goal-arrival units (green), and units not responsive to the goal (blue). Gray lines show sections at different AP levels relative to bregma (−0.1 to 2.0 mm). *B*: DV, AP, and ML locations (mean ± SEM), relative to bregma, of post-goal units (red), and units not active around the goal (blue). Goal arrival units (green) shown for comparison.

Fig. 3. Segmentation of post-goal units biased for correct (reward) versus incorrect (error) outcomes. *A*: Histograms of two DLS units recorded during task acquisition that fired more after correct (black) than incorrect (red) maze runs (±3 s around goal arrival). Units showed preferential (top) or exclusive (bottom) firing to one outcome. *B*: Two units firing preferentially (top) or exclusively (bottom) after incorrect runs, recorded during task acquisition. *C*: Percentage of total post-goal units recorded per stage block with greater activity after correct runs (blue), greater activity after incorrect runs (red), or equivalent activity to both outcomes (black), shown for successive training phases. The one post-goal unit in probe session is omitted for clarity as it would exhibit 100%.

Fig. 4. Population and firing activity differentiating outcomes. *A*: Ratio of number of correct-related and incorrect-related post-goal units recorded by training phase. *B* and *C*: Scatterplots and regression lines showing relationship between percent correct maze runs and relative proportion (*B, top*), firing rates (*B, bottom*), and summated population and firing rate proportion (*C*) of correct-related units. Data were averaged for each learning stage (dot). *D-F*: Baseline-subtracted normalized spike activity (mean ± SEM) of correct-related (*D*), incorrect-related (*E*), and correct-and-incorrect-related (*F*) units during the 3-s post-goal period in rewarded (correct; solid line) and error (incorrect; dashed line) trials, shown by training phase. Zero represents pre-trial baseline. *Post-hoc* comparisons: $^{*}P < 0.05$ significant firing elevation compared to zero (baseline); $^{\#}P < 0.05$ correct vs. incorrect trials. All other comparisons to baseline or between trial outcomes, $P > 0.05$.

Fig. 5. Aggregate DLS post-goal activity related to run outcome and learning. *A*: Baseline-subtracted normalized activity (0 = pre-trial baseline) of all post-goal units in correct (black) and incorrect (red) trials, averaged for acquisition (stages 1-4), early over-training (stages 5-8), and late over-training (stages 9-13) phases and plotted for ±200 ms (100-ms bins) around each task

event from baseline to goal arrival, and 3 s following goal arrival (post-goal period). Task events (tick marks): baseline (BL), warning cue, gate, pre run start, post run start, instruction cue, turn on, turn off, goal arrival (GA). Shading show ±SEM. *B*: Baseline-subtracted spike activity after goal arrival, shown in 50-ms bins to illustrate latency of onset and reward/error discrimination. Error bars represent SEM. *C*: Latency for post-goal units to respond (2 SDs above baseline) after goal arrival by trial outcome and learning stage.

Fig. 6. Different learning-related dynamics of post-goal and in-task-responsive DLS units. *A*: A post-goal unit (top) and another unit (bottom) recorded simultaneously on the same tetrode during a late over-training session (stage 10) with 32 correct and 8 incorrect trials (±3 s around goal arrival, 50-ms bins). The top unit fired exclusively after correct runs (black), and not after incorrect runs (red), whereas the bottom unit was phasically active around goal arrival after both correct and incorrect runs. *B*: Baseline-subtracted activity (mean ± SEM) of units with significant responses to maze-run events, averaged for correct (black) and incorrect (red) trials across acquisition and over-training, and shown for ±200 ms around turn end (TE), and 200 ms before and 3 s after goal arrival (GA).

Fig. 7. DLS post-goal spike activity related to run outcome after reward devaluation. *A*: Normalized activity (mean ± SEM) of post-goal units during post-probe (PP) sessions after correct runs to the non-devalued goal (black), correct runs to the devalued goal (blue), and wrong-way runs to the non-devalued goal (red). Baseline-subtracted activity (0 = baseline) around task events, plotted as in Fig. 5*A*. *B*: Average activity during 3 s after goal arrival in PP sessions. $^{*}P < 0.05$ compared to baseline; $^{\#}P < 0.05$ correct compared to incorrect trials; ns, not significantly different from baseline.

Fig. 8. Interaction of post-goal activity with goal location and reward outcome across learning. *A*: Percent of all post-goal units recorded from right and left hemispheres that were responsive above baseline to the ipsilateral goal only (light gray), contralateral goal only (dark gray), or both goals (black). *B*: Percent of post-goal units that were responsive to a single goal location (i.e., left only or right only) by learning stages. Above 50% reflects more single-goal-responsive units; below reflects more units responsive to both maze goals. *C*: Scatterplot and regression fit on post-goal units responsive to a single goal (blue) or to both goals (black) by percent correct maze runs for the session in which it was recorded, collapsed over learning stages. *D*: Pie charts by stages representing breakdown of post-goal units responsive to combinations of one or both goals, and correct-related, incorrect-related, or correct-and-incorrect-related (1 unit in probe test excluded). Percent numbers (rounded) denote proportion of population for each stage block; white numbers denote number of post-goal units recorded.

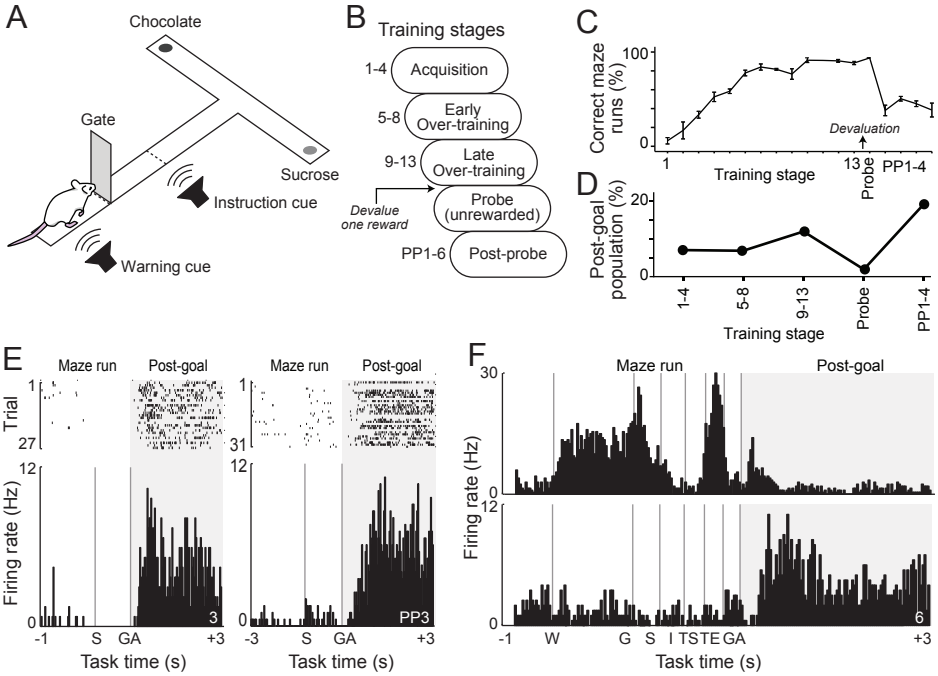
A
Chocolate
Gate
Sucrose
Instruction cue
Warning cue
B
Training stages
1-4
Acquisition
5-8
Early Over-training
9-13
Late Over-training
Devalue one reward
Probe (unrewarded)
PP1-6
Post-probe
C
Correct maze runs (%)
100
0
1
13
Training stage
Devaluation
Probe
PP1-4
D
Post-goal population (%)
20
0
1-4
5-8
9-13
Probe
PP1-4
Training stage
E
Maze run
Post-goal
Trial
1
27
31
Firing rate (Hz)
12
0
3
PP3
-1
S
GA
+3
-3
Task time (s)
F
30
Firing rate (Hz)
W
G
S
I
TS
TE
6

Figure 1

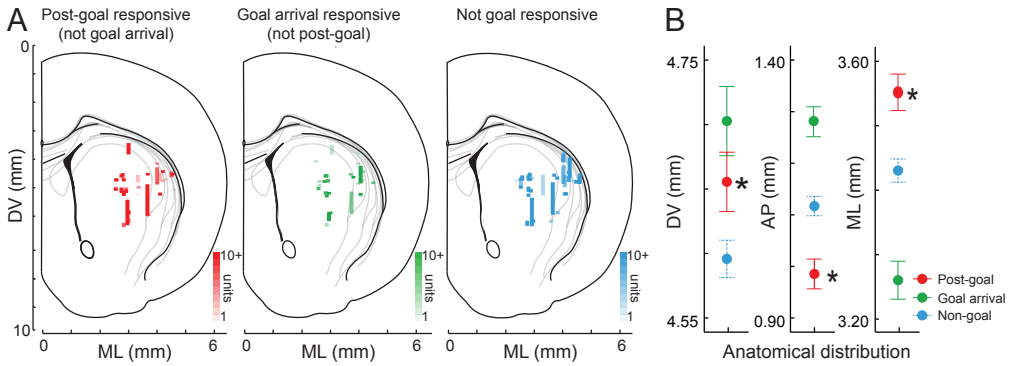
A
Post-goal responsive
(not goal arrival)
Goal arrival responsive
(not post-goal)
Not goal responsive
DV (mm)
ML (mm)
units
10+
1
0
6
10
B
4.75
4.55
1.40
0.90
3.60
3.20
DV (mm)
AP (mm)
ML (mm)
Anatomical distribution
Post-goal
Goal arrival
Non-goal

Figure 2

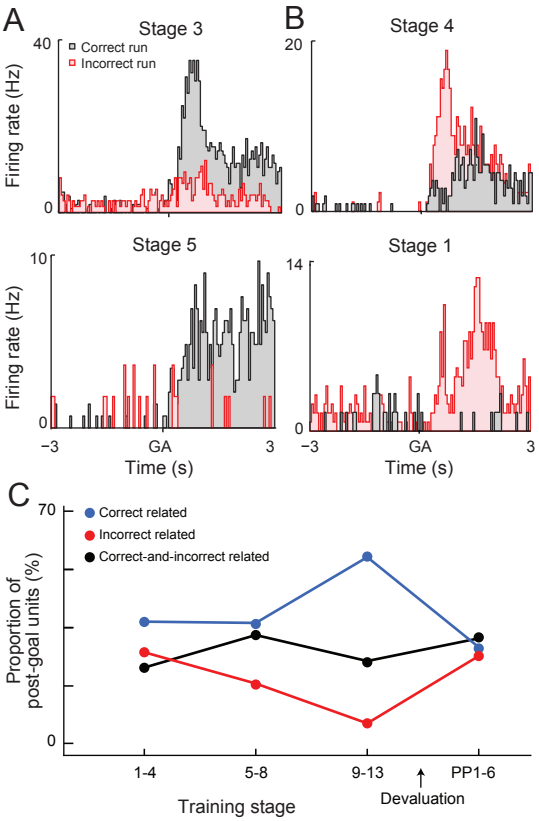
A
Stage 3
Correct run
Incorrect run
40
0
Firing rate (Hz)
B
Stage 4
20
0
Stage 5
10
0
Firing rate (Hz)
−3
GA
3
Time (s)
Stage 1
14
0
−3
GA
3
Time (s)
C
70
0
Proportion of post-goal units (%)
Correct related
Incorrect related
Correct-and-incorrect related
1-4
5-8
9-13
PP1-6
Devaluation
Training stage

Figure 3

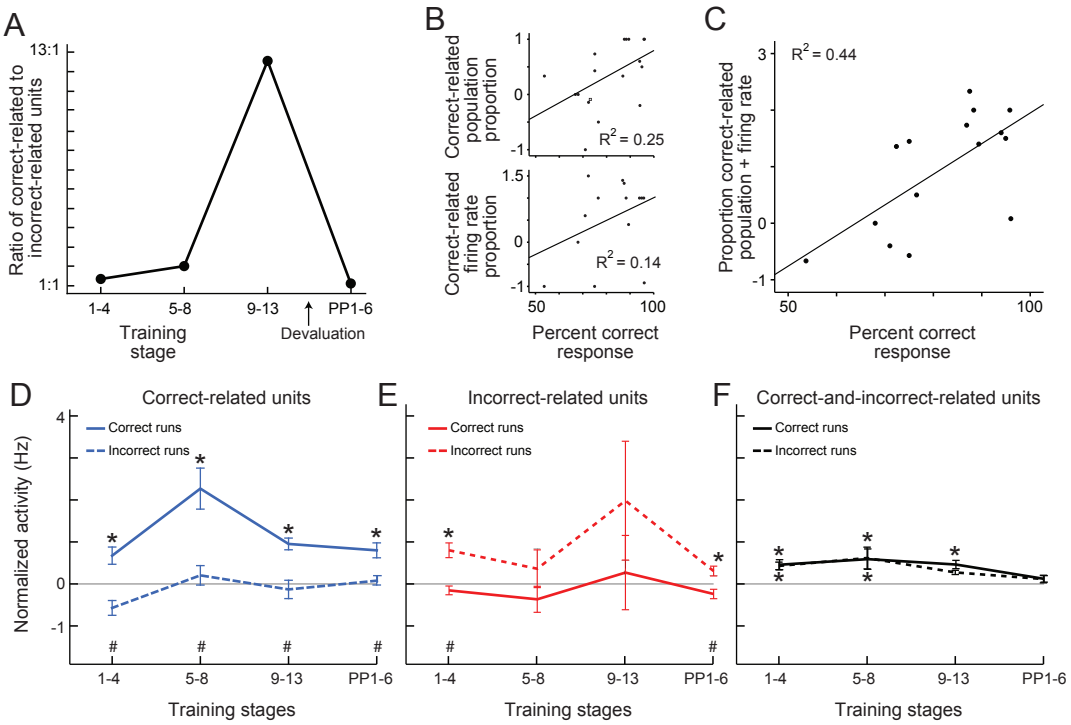
A
13:1
1:1
Ratio of correct-related to incorrect-related units
1-4
5-8
9-13
PP1-6
Training stage
Devaluation
B
Correct-related population proportion
R2 = 0.25
Correct-related firing rate proportion
R2 = 0.14
Percent correct response
C
R2 = 0.44
Proportion correct-related population + firing rate
Percent correct response
D
Correct-related units
Correct runs
Incorrect runs
Normalized activity (Hz)
Training stages
E
Incorrect-related units
F
Correct-and-incorrect-related units

Figure 4

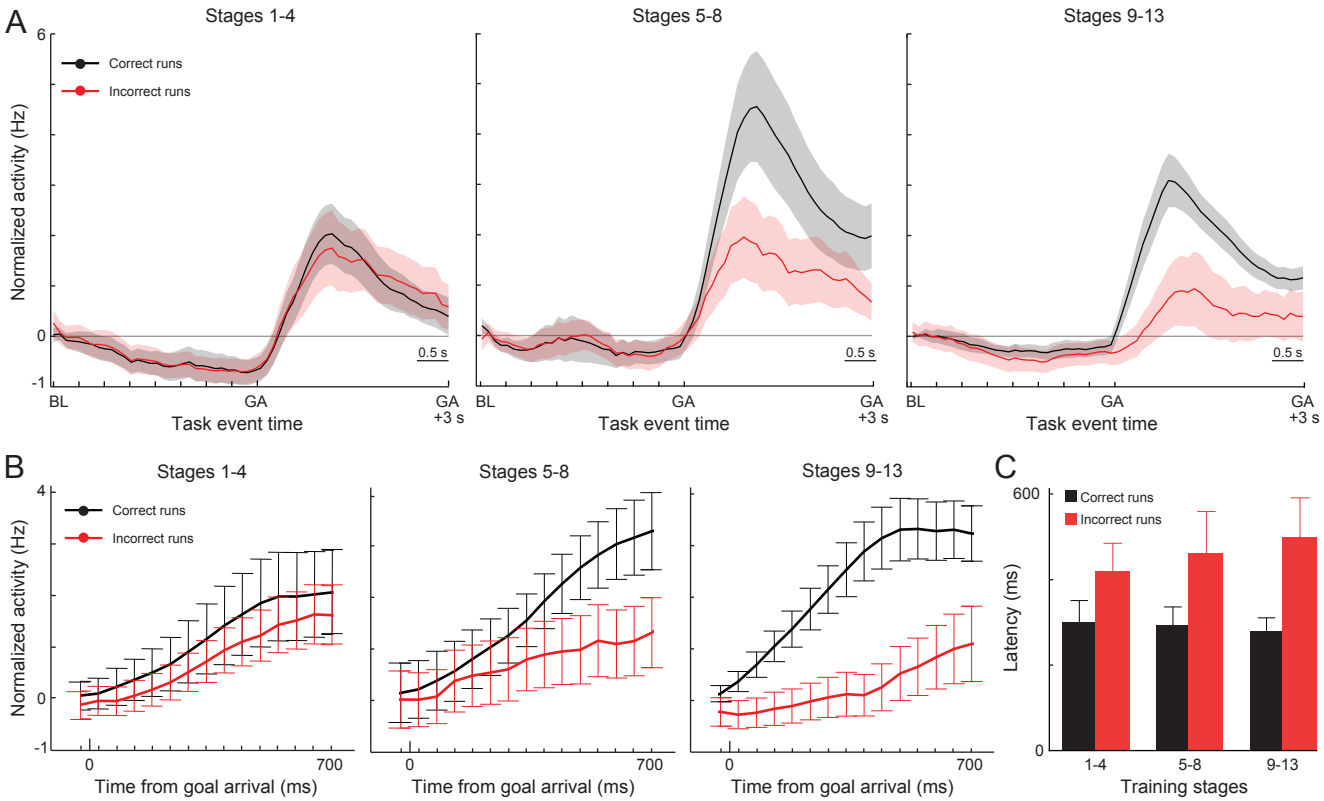
A
Stages 1-4
Stages 5-8
Stages 9-13
Correct runs
Incorrect runs
Normalized activity (Hz)
6
0
-1
0.5 s
BL
GA
GA +3 s
Task event time
B
Stages 1-4
Stages 5-8
Stages 9-13
Correct runs
Incorrect runs
Normalized activity (Hz)
4
0
-1
0
700
Time from goal arrival (ms)
C
Correct runs
Incorrect runs
Latency (ms)
600
0
1-4
5-8
9-13
Training stages

Figure 5

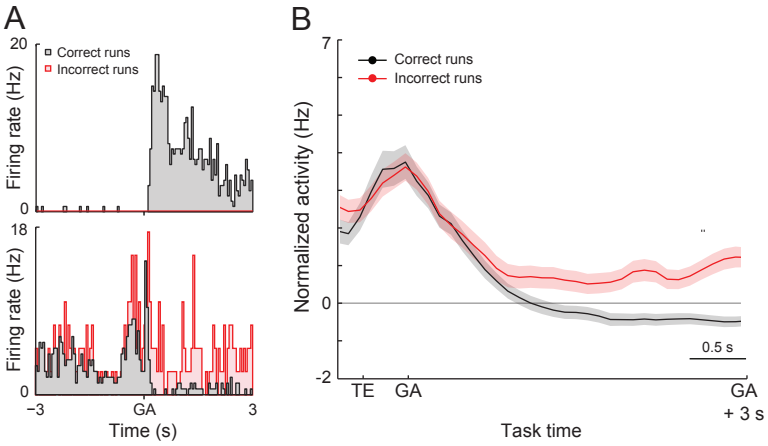
A
B
Correct runs
Incorrect runs
Firing rate (Hz)
20
0
18
0
−3
GA
3
Time (s)
Normalized activity (Hz)
7
0
-2
0.5 s
TE
GA
GA
+ 3 s
Task time

Figure 6

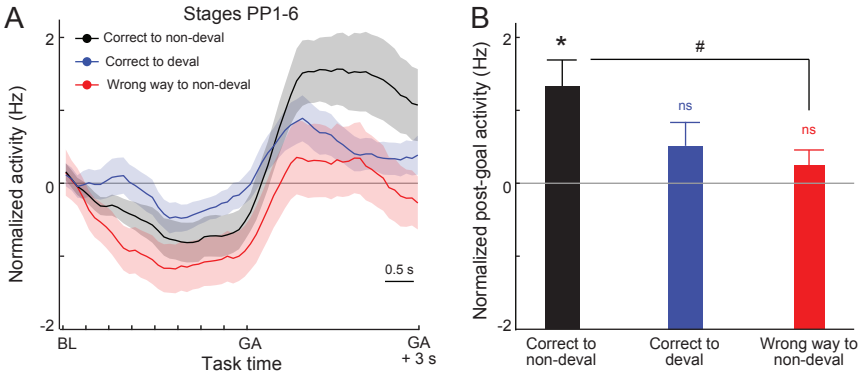
A
Stages PP1-6
Correct to non-deval
Correct to deval
Wrong way to non-deval
Normalized activity (Hz)
2
0
-2
0.5 s
BL
GA
GA
+ 3 s
Task time
B
Normalized post-goal activity (Hz)
2
0
-2
*
#
ns
ns
Correct to
non-deval
Correct to
deval
Wrong way to
non-deval

Figure 7

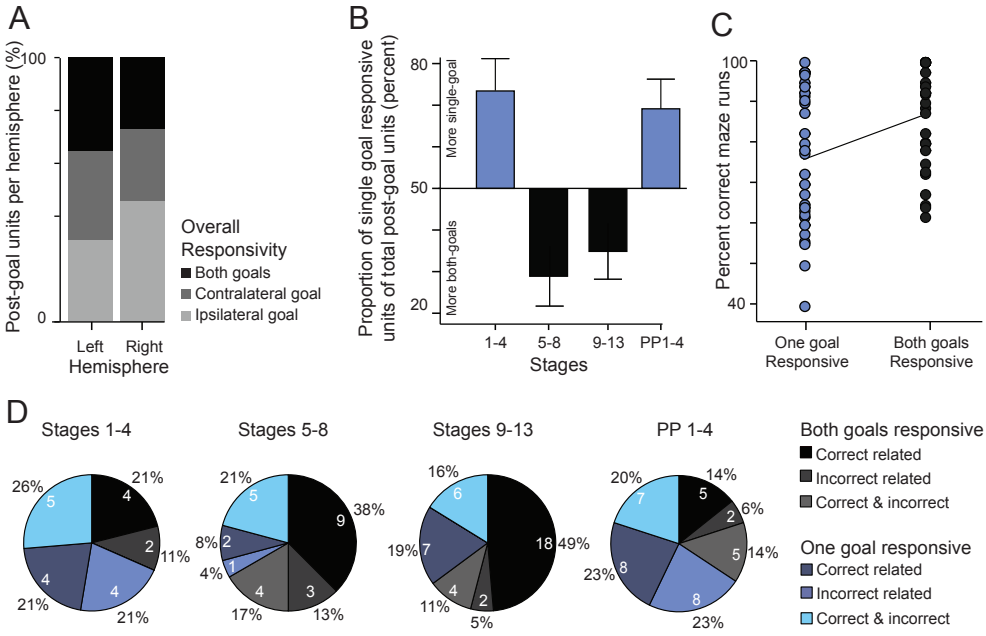
A
Post-goal units per hemisphere (%)
100
0
Left
Right
Hemisphere
Overall
Responsivity
Both goals
Contralateral goal
Ipsilateral goal
B
Proportion of single goal responsive units of total post-goal units (percent)
80
50
20
More single-goal
More both-goals
1-4
5-8
9-13
PP1-4
Stages
C
Percent correct maze runs
100
40
One goal
Responsive
Both goals
Responsive
D
Stages 1-4
Stages 5-8
Stages 9-13
PP 1-4
Both goals responsive
Correct related
Incorrect related
Correct & incorrect
One goal responsive
Correct related
Incorrect related
Correct & incorrect

Figure 8