Acetylcholine demixes heterogeneous dopamine signals for learning and moving

Hee Jae Jang[1], Royall McMahon Ward[1], Carla E. M. Golden[1], Christine M. Constantinople[1*]

[1] Center for Neural Science, New York University; New York, NY 10003.

*To whom correspondence should be addressed; E-mail: constantinople@nyu.edu.

Midbrain dopamine neurons promote reinforcement learning and movement vigor. A major outstanding question is how dopamine-recipient neurons in the striatum parse these heterogeneous signals. Here we characterized dopamine and acetylcholine release in the dorsomedial striatum (DMS) of rats performing a decision-making task. We found that dopamine acted as a reward prediction error (RPE), modulating behavior and DMS spiking on subsequent trials when coincident with pauses in cholinergic release. In contrast, at task events that elicited coincident bursts of acetylcholine and dopamine, dopamine preceded contralateral movements and predicted movement vigor without inducing plastic changes in DMS firing rates. Our findings provide a circuit-level mechanism by which cholinergic modulation allows the same dopamine signals to be used for either movement or learning depending on instantaneous behavioral context.

In reinforcement learning, agents learn the value of taking actions in different states; those state-action values are updated based on RPEs, or the difference between received and expected rewards (*1*). Evidence suggests that dopaminergic neurons instantiate RPEs in the brain (*2–9*). One target for dopamine neurons is the DMS, a striatal region necessary for learning action-outcome associations (*10*). Dopamine is thought to drive plasticity at synaptic inputs onto

medium spiny neurons such that animals will be more (or less) likely to take actions in states that produced positive (or negative) RPEs. However, RPEs measured during behavior have not been directly linked to synaptic plasticity *in vivo*.

Dopamine is also implicated in motor control (*11*). Degeneration of dopamine neurons is a hallmark of Parkinson's Disease (*12*). Perturbations of dopamine in the DMS affect movement initiation, kinematics, and vigor (*13–15*). Studies have found dopamine responses that reflect movements and cannot be easily interpreted as RPEs (*5*, *14*, *16*, *17*).

Here, we tested the hypothesis that striatal acetylcholine dynamically determines whether dopamine drives learning or movement vigor in rats performing a decision-making task. While previous work has suggested that cholinergic interneurons may gate dopamine-dependent plasticity (*18–21*), this has not been tested in behaving animals. We developed a task with salient reward cues that were separable in time from movement-related events, which allowed us to independently examine neuromodulatory signatures of RPE and movement.

Temporal wagering task provides behavioral readouts of rats' reward expectations and movement We trained rats on a self-paced temporal wagering task that manipulated reward expectations by varying reward statistics in blocks of trials (*22*) (Fig. 1, A and B). Rats initiated trials by poking into a central nose port. This triggered an auditory cue, the frequency of which indicated the volume of water reward offered on that trial (5, 10, 20, 40, or 80 μL). Rats were required to maintain fixation in the central nose poke for $\sim$1 second while the auditory cue played; leaving early triggered a white noise sound and terminated the trial. At the offset of the auditory cue, the reward was assigned randomly to one of two side ports, indicated by an LED. The rat could wait for an unpredictable delay to obtain the reward, or at any time could opt out by poking in the other side port to start a new trial. On 15-25% of trials, rewards were withheld to assess how long rats were willing to wait for them (catch trials), providing a behavioral readout of how much they valued the reward offer.

To manipulate reward expectations, rats experienced blocks of trials offering low (5, 10, or 20 μL) or high (20, 40, or 80 μL) volumes of water. These were interleaved with mixed blocks offering all rewards (Fig. 1B). Rats initiated trials more quickly in high blocks compared to low blocks, providing a behavioral readout of their subjective value of the environment (Fig. 1C).

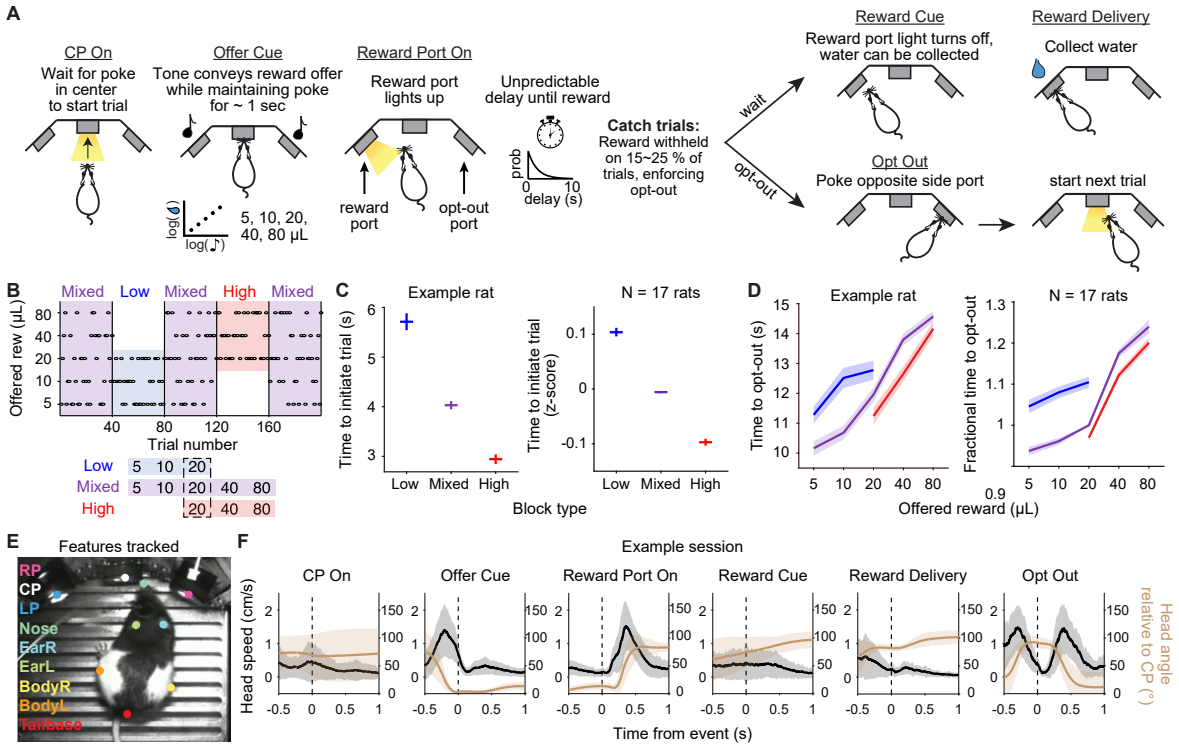

Fig. 1. Temporal wagering task provides behavioral readouts of rats' reward expectations and movement. (**A**) Schematic of behavioral paradigm. (**B**) Schematic of the block structure of the task. All sessions start with a mixed block and alternate with low or high blocks. Blocks are 40 completed trials. (**C**) Time to initiate trial by block type averaged across sessions for an example rat (mean ± s.e.m.). Right, Z-scored trial initiation time by block across rats (N = 17, mean ± s.e.m.). (**D**) Left, Time to opt-out by offered reward volume in each block averaged across sessions for an example rat (mean ± s.e.m.). Right, Fractional time to opt-out by offered reward in each block across rats (N = 17, mean ± s.e.m.). Time to opt-out is divided by the time for 20 µL in mixed block for each rat before averaging. (**E**) Features tracked with DeepLabCut using videography. (**F**) Head speed (black) and angle relative to the central nose port (brown) aligned to different task events in an example session (mean ± 1 s.d. across trials).

We also found that rats modulated how long they were willing to wait for rewards based on the average reward in each block, but wait times reflected a distinct strategy from the decision to initiate a trial (*22*) (Fig. 1D). Given our previous findings that initiation times rely on a

dopamine-dependent reinforcement learning algorithm (*22*, *23*), we focused on that aspect of behavior in this study.

Deep pose estimation (*24*) showed that rats made stereotypical orienting movements at certain task events (Fig. 1, E and F). When the side LED lit up indicating the reward port on that trial, rats oriented towards it, and after opting out, they oriented back to the center port to initiate a new trial. Thus, our task includes reward cues that are separable in time from orienting movements, which require DMS (*25*).

Dopamine and acetylcholine show distinct dynamics at task events We next measured dopamine and acetylcholine release in the DMS in this task. To measure dopamine, we used fiber photometry of an optical dopamine sensor, GRAB_{DA} (Fig. 2A). We used a static fluorophore mCherry and/or isosbestic control to correct for brain motion (*26*), with similar results for both methods (fig. S1).

At the offer cue, we observed phasic dopamine release that scaled with the offered reward volume, consistent with an RPE (Fig. 2B). Additionally, because the timing of the rewards was unpredictable and they were omitted on catch trials, we reasoned that when the rat learns reward is available, there should be larger prediction errors for long compared to short delays (*27*, *28*). Binning trials with different reward delays revealed that dopamine release at the reward cue scaled with delay, again consistent with an RPE (Fig. 2, C and D). These RPE representations did not depend on the spatial location of the reward port (fig. S2).

However, at other task events when cues elicited stereotypical orienting movements, dopamine release showed a larger phasic response for orienting movements contralateral to the recording hemisphere (Fig. 2, E and F). This was true when the reward port was assigned and rats oriented toward that port, and following opt-outs, when rats oriented back toward the center port to initiate a new trial. Contralateral signals were DMS-specific as dopamine in the ventral striatum showed similar encoding of RPEs, but not contralateral movement selectivity (*5*) (fig. S3). Computing the discriminability index of dopamine responses for different task variables confirmed that dopamine encoded reward volume, delay, and contralateral movements at distinct task events (fig. S4). Therefore, dopamine in the DMS multiplexes RPEs and contralateral movement signals at distinct time points on single trials.

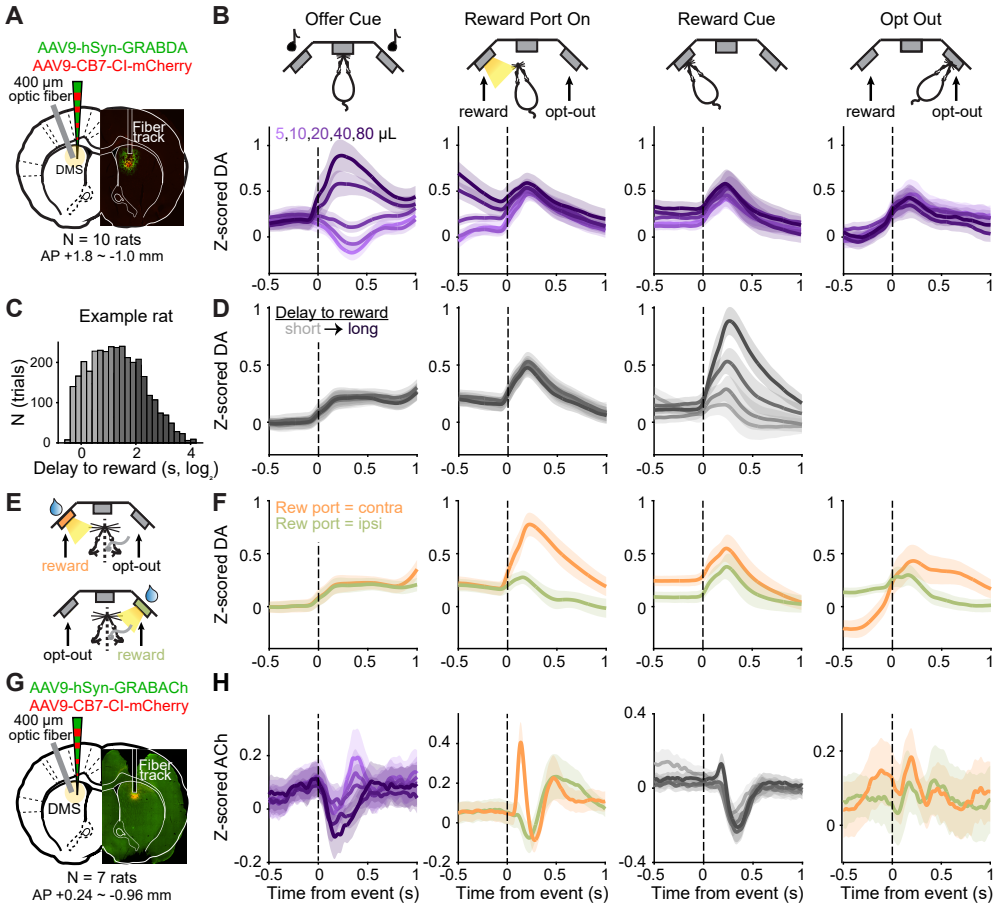

Fig. 2. Dopamine and acetylcholine in the DMS show distinct dynamics at reward- and movement-associated events. (**A**) Viral strategy to measure dopamine release in the DMS and representative histology. (**B**) Average event-aligned dopamine release during mixed blocks for different reward offers (N = 10 rats, mean ± s.e.m.). (**C**) Distribution of delay to reward, drawn from an exponential distribution, for an example rat. colors indicate quartiles. (**D**) Average event-aligned dopamine release split by delay to reward quartiles across all blocks (N = 10 rats, mean ± s.e.m.). (**E**) Schematic of contralateral and ipsilateral trials shown for a rat implanted in the right hemisphere. The side of implant is counterbalanced across rats. (**F**) Average event-aligned dopamine release split by contralateral (orange) and ipsilateral (green) trials (N = 10 rats, mean ± s.e.m.). (**G**) Viral strategy to measure acetylcholine release in the DMS and representative histology. (**H**) Average event-aligned acetylcholine release split by the task variables most strongly encoded by dopamine at each task event (N = 7 rats, mean ± s.e.m.).

We measured acetylcholine release in the DMS using fiber photometry of the GRAB$_{ACh}$ sensor (Fig. 2G). We found prominent dips in acetylcholine release at cues that elicited dopamine RPEs (Fig. 2H). In contrast, when the side LED indicated the reward port, we observed a strong burst in acetylcholine release on contralateral trials (Fig. 2H), when a phasic increase in dopamine release was also observed (Fig. 2F). At opt-out, we observed a similar but weaker rise in acetylcholine on contralateral trials. The difference in the strength of the acetylcholine burst at the two movement events could reflect more temporal jitter between when rats opt-out and then orient to initiate subsequent trials (fig. S5). Therefore, the relative phase relationship between dopamine and acetylcholine is different when dopamine encodes RPEs versus reflects contralateral movements.

Coincident dips in acetylcholine are required for dopamine RPEs to drive learning We next sought to determine whether dopamine at putative RPE events implemented reinforcement learning. We modeled rats' initiation times as inversely proportional to the value of the environment, which they estimated with the standard reinforcement learning algorithm $V_{t+1} = V_t + \alpha(R_t - V_t)$. Our behavioral model reproduces rats' behavior, including that trial initiation times are sensitive to the block type (*22*) (Fig. 3A).

The model predicted that RPEs should scale with offered rewards, which we observed in the data (Fig. 3B). The model also predicted that RPEs should be inversely related to expectations, or the average reward in each block (Fig. 3C). Indeed, the dopamine response to 20 µL was the largest in low blocks and negative in high blocks, consistent with an RPE (Fig. 3C). The learning algorithm in our model predicts that RPEs and behavior will be influenced by previous rewards, with the most recent rewards having the greatest influence and decaying exponentially further in the past (Fig. 3D). Regressing dopamine area under the curve (AUC) and initiation times at the time of the offer cue against previous rewards revealed negative, exponentially decaying influences of previous rewards on phasic dopamine responses and rats' initiation times, consistent with an RPE (Fig. 3D). Finally, there was a strong relationship between the magnitude of the phasic dopamine signal at the offer cue and the change in initiation times on the subsequent trial (Fig. 3E). Collectively, these results argue that dopamine at the offer cue acts as an RPE that updates the value of the environment for subsequent initiation times.

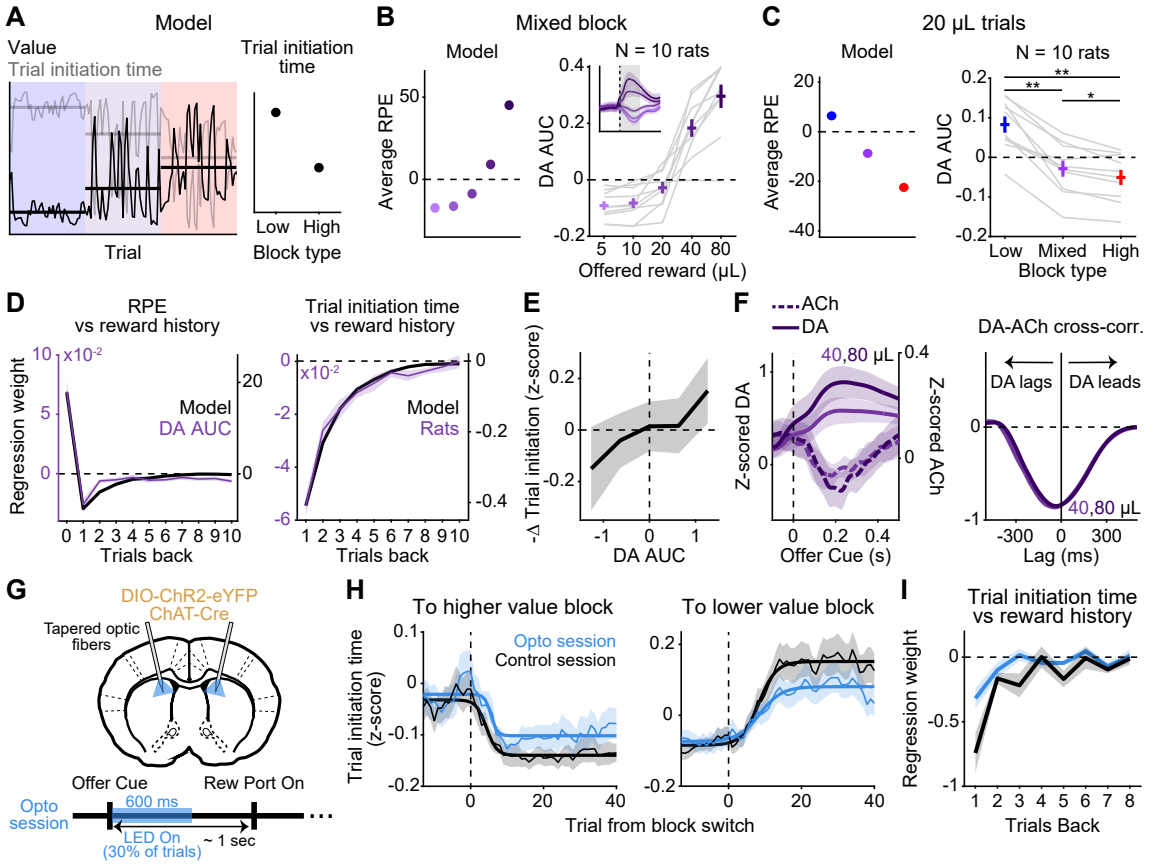

Fig. 3. Coincident dip in acetylcholine is required for dopamine RPE to drive learning. (**A**) Left, Model predicted value (black) and simulated initiation time (grey). Shaded colors indicate block type (blue: low, purple: mixed, red: high). Solid lines indicate average block values. Right, Model predicted initiation times by block. (**B**) Left, Model predicted RPE by offered reward in mixed blocks. Right, Dopamine AUC at the offer cue (0 $\sim$ 0.5 s) in mixed blocks (N = 10 rats, mean $\pm$ s.e.m.). Grey lines show individual rats. (**C**) Same as in **B** but on 20 μL offer trials by block type (N = 10 rats, mean $\pm$ s.e.m., two-sided Wilcoxon signed rank: low vs mixed, p = 0.002; low vs high, p = 0.002; mixed vs high, p = 0.013). (**D**) Left, Dopamine AUC (purple) at the onset of the offer cue (0 $\sim$ 0.35 s) regressed against reward history in mixed blocks (N = 10 rats, mean $\pm$ s.e.m.). Black line shows model predicted RPE regressed against reward using trial statistics from example session in **A-C**. Right, Z-scored initiation time (purple) regressed against reward history in mixed blocks (N = 10 rats, mean $\pm$ s.e.m.). Black line shows model predicted initiation time regressed against reward. (**E**) Dopamine AUC at the offer cue (0 $\sim$ 0.35

s) versus -Δ z-scored trial initiation time, averaged across rats (N = 10, mean $\pm$ s.e.m.). For each rat, dopamine AUC is binned in 5 linearly spaced intervals and then z-scored before averaging across rats. (**F**) Left, Average dopamine (solid purple) overlaid with acetylcholine (dashed purple) release at the offer cue on 40and 80 µL offer trials (N = 10 GRAB$_{DA}$ rats, 7 GRAB$_{ACh}$ rats, mean $\pm$ s.e.m.). Right, Cross-correlation of average dopamine and acetylcholine signals at offer cue (0-0.5 s). Negative lag indicates dopamine following acetylcholine signal. (**G**) Top, Viral strategy to photoactivate ChAT-positive cells in the DMS. Bottom, Stimulation protocol. (**H**) Z-scored trial initiation time around block transitions averaged across control (black) and optogenetic (blue) sessions (N = 5 rats, mean $\pm$ s.e.m.). Transitions to a higher value block include low to mixed and mixed to high transitions. Transitions to a lower value block include high to mixed and mixed to low transitions. (**I**) Trial initiation time regressed against reward history averaged across rats (N = 5, mean $\pm$ s.e.m.) in optogenetic (blue) versus control (black) sessions. Across panels, *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001.

At the offer cue, dopamine RPEs coincide with dips in acetylcholine. For visualization, we highlight the temporal correlation between large positive RPEs and cholinergic dips, in which dopamine peaks slightly follow cholinergic dips (Fig. 3F, fig. S6). We hypothesized that cholinergic dips might help medium spiny neurons interpret dopamine as an RPE for learning. To test this hypothesis, we photoactivated cholinergic interneurons by injecting a ChAT-Cre virus and a Cre-dependent channelrhodopsin (ChR2) conjugated to YFP and delivering continuous 465 nm light via tapered optical fibers implanted bilaterally in the DMS (Fig. 3G). The LED was triggered on 30% of randomly selected trials at the onset of the offer cue, and delivered for 600 ms at 1.9-2.3 mW.

On optogenetic sessions, rats showed weaker behavioral changes at block transitions, consistent with cholinergic activation preventing learning on a subset of trials (Fig. 3H). Further, regression revealed a weaker influence of previous rewards on trial initiation times from optogenetic sessions (Fig. 3I). We did not observe these effects in sham rats (fig. S7).

Dopamine at RPE events promotes plasticity Since dopamine at the offer cue is consistent with an RPE, we hypothesized that it might produce plasticity in recipient DMS cells (*29–31*).

Dopamine receptor binding can influence the excitability of medium spiny neurons (*32*, *33*), or modify the strength of convergent synaptic inputs (*29–31*). However, it has remained unclear how and when RPE-like updating of DMS activity actually occurs *in vivo* during behavior. For RPE-mediated updating to be useful, it should produce rapid induction and persistent expression of synaptic plasticity, but it is not clear how the timescales of conventional short-term and long-term plasticity allow behaviorally useful changes to occur in this manner.

We recorded single unit action potentials using Neuropixels probes implanted in the DMS (Fig. 4B). We used the trial-by-trial change in initiation times as a behavioral proxy for RPEs on each trial (Fig. 3E). We observed trial-by-trial changes in firing rate following changes in initiation times at the offer cue, but not at subsequent events (Fig. 4C). To quantify this effect, we regressed trial-by-trial changes in firing rates for 300 ms after each event against putative RPEs (-Δ initiation times). The highest proportion of cells had significant slope parameters at the offer cue (51.6%, 130/252 units) with 53.1% (69/130 units) having positive slopes (e.g., example cell in Fig. 4D) and 46.9% (61/130 units) having negative slopes (Fig. 4, E and F). The relationship between putative RPEs and trial-by-trial changes in firing rate decreased at events that were further from the offer cue, suggesting that plasticity was temporally specific to the time of the RPE. We confirmed that the results in Fig. 4C-F reflect RPEs from the previous trial rather than the initiation time on the current trial (fig. S8).

To characterize the timescale of these effects, we identified sequences of trials in which large positive/negative RPEs (top/bottom 15%) on trial N were followed by negligible RPEs ($|\text{RPE}| < 0.2$) on trial N+1. If dopamine effects are long-lasting, the change in firing rate from trial N to trial N+1 should persist to trial N+2, since there was a negligible RPE on the intervening trial. Fig. 4G shows the firing rates of two neurons on such a trial sequence. We analyzed firing rates for all such sequences separately for cells with positive and negative firing rate correlations with RPEs. In both groups, positive RPEs on trial N changed firing rates on trial N+1, and that change persisted to trial N+2 (Fig. 4H). These data are consistent with RPEs producing lasting changes in task-evoked firing rates that endured over trials. This result generally held for other thresholds used to select RPEs on trial N (fig. S9A). Negative RPEs produced more variable effects on trial-by-trial spiking for these specific sequences of trials (fig. S9B), possibly reflecting different reliability of plasticity mechanisms following increases ver-

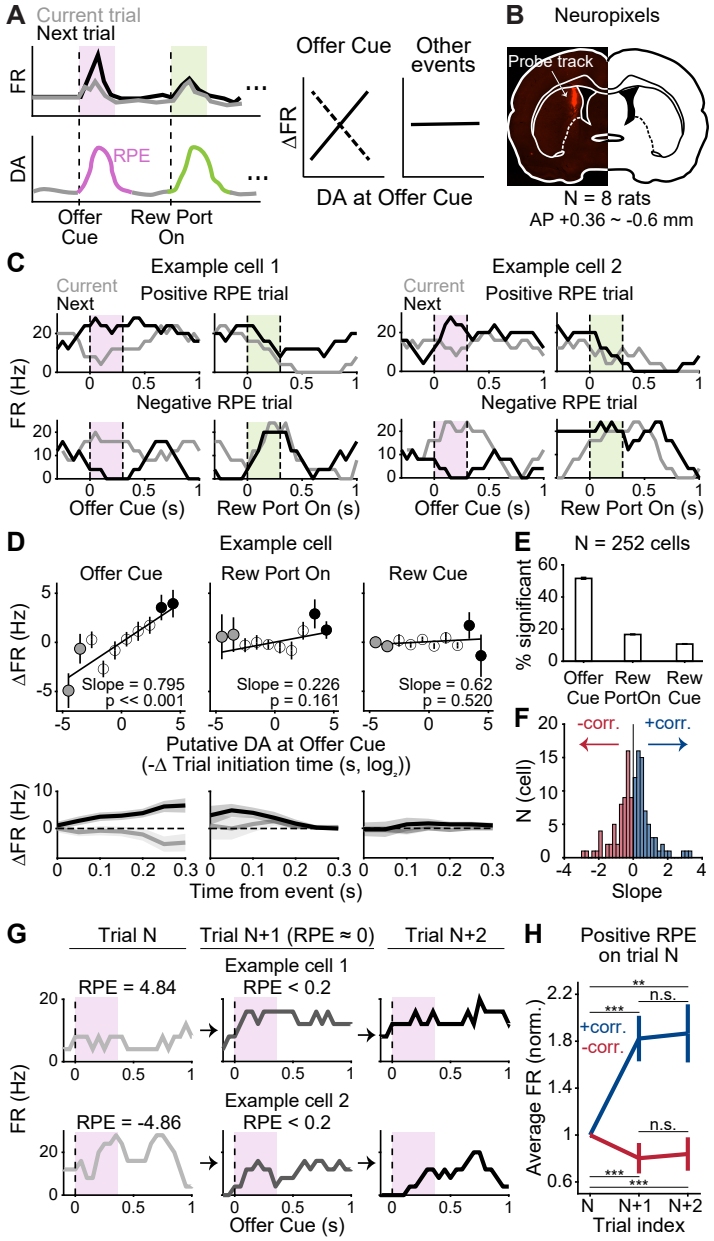

Fig. 4. Dopamine RPE predicts DMS firing rate on subsequent trials. (**A**) Schematic of the prediction for state- or input-specific dopamine-dependent synaptic plasticity where RPE at the time of the offer cue produces trial-by-trial changes in firing rates at that specific event but not at other states. (**B**) Example histology of Neuropixels probe track in the DMS. (**C**) Firing rate of two example cells on current (grey) and the subsequent trial (black) following a positive and negative RPE. Shaded boxes indicate the time window used to compute average firing rate (0 ∼

0.3 s). (**D**) Top, Trial-by-trial changes in average firing rate surrounding each task event ($0 \sim 0.3$ s) versus putative dopamine at offer cue (-Δlog_2 trial initiation time) for an example cell. Solid lines show the best fit line (t-test on slope: Offer Cue, $p << 0.001$; Rew Port On, $p = 0.161$; Rew Cue,$p = 0.520$) and the filled black/grey circles indicate trials with large positive/negative RPEs, respectively. Trials are binned in 15 linearly spaced intervals for visualization (mean $\pm$ s.e.m.). Bottom, Change in average firing rate on trials with large positive (black) and negative RPEs (grey) for the same example cell in the top panel (mean $\pm$ s.e.m. across trials). (**E**) Fraction of cells with significant slopes at different task events. Error bars are binomial confidence intervals. (**F**) Histogram of slopes for significantly modulated cells at the offer cue (red: positively modulated by RPE, blue: negatively modulated by RPE). (**G**) Firing rate of two example cells aligned to the onset of offer cue on 3 consecutive trials where a big positive/negative RPE on the first trial is followed by a negligible RPE on the second trial. (**H**) Change in average z-scored firing rate at offer cue ($0 \sim 0.3$ s) of positively (blue) and negatively (red) modulated cells on 3 consecutive trials for which a large positive RPE (top 15 %) on trial N is followed by a negligible RPE on trial N+1 (N = 145/164 trial sequences for +corr./-corr., mean $\pm$ s.e.m., two-sided Wilcoxon signed rank: positively modulated cells, trial N vs N+1, $p = 0.0004$; trial N vs N+2, $p = 0.0065$; trial N+1 vs N+2, $p = 0.4169$; negatively modulated cells, trial N vs N+1, $p << 0.0001$; trial N vs N+2, $p << 0.0001$; trial N+1 vs N+2, $p = 0.1898$). Across panels, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

sus decreases of dopamine. Nonetheless, positive RPEs in the DMS modulate evoked firing on subsequent trials at the time of the RPE but not at neighboring task events (Fig. 4, C to E), and those changes persist over trials (Fig. 4, G and H).

Coincident dopamine and acetylcholine bursts drive movement Next, we sought to characterize the contralateral selective dopamine signal at events accompanying orienting movements. The time to peak of the dopamine signal preceded that in head speed by $\sim$100 ms (Fig. 5, A and C). Moreover, the amplitude of the dopamine signal at these events predicted the vigor of the subsequent orienting movement. When the reward port is assigned, rats quickly orient towards the side port once the LED turns on. The AUC of the dopamine signal at this event

was significantly larger when reaction times were in the fastest versus slowest quartile (Fig. 5B). Similarly, after rats opt out, they orient back to the center port to start a new trial, and the amplitude of the dopamine signal at opt-out was also larger when rats were faster to initiate the next trial (Fig. 5D). Therefore, the phasic dopamine signal at these movement-associated events precedes and predicts the vigor of upcoming contralateral orienting movements.

To test whether dopamine at this event promoted plasticity, we tested for a significant correlation between trial-by-trial changes in the firing rates of DMS cells and dopamine magnitude (using the reaction time on the previous contralateral trial as a proxy for dopamine). We found that for the overwhelming majority of cells, there was no significant correlation (fig. S10). In contrast to RPE events, when dopamine influenced trial-by-trial firing rates, we found that DMS cells encoded movement vigor at this timepoint. We trained a logistic regression decoder to predict whether the reaction time on each trial was in the fastest or slowest quartile, based on the firing rates of simultaneously recorded neurons. Cross-validated classification accuracy was high and above chance in 17 of 18 sessions (Fig. 5E), suggesting that dopamine at movement-associated events dictates the vigor of upcoming contralateral orienting movements but does not result in persistent changes in task-evoked spiking of DMS cells.

Since phasic dopamine responses were coincident with bursts of acetylcholine release at movement events as opposed to dips at the RPE events (Fig. 5, F and G), we hypothesized that coincident activation of dopamine and acetylcholine might invigorate movements (Fig. 5H). Indeed, photoactivating cholinergic interneurons at the offer cue, when there was normally a cholinergic dip, reliably elicited movements (Fig. 5I). Optogenetic stimulation in the ChAT-ChR2 group, but not the sham group, elicited movements and caused rats to break center fixation prematurely ("violate") on almost every stimulation trial (Fig. 5J). To test whether the magnitude of the cholinergic signal affected vigor, we separately analyzed effects of optogenetic stimulation on trials with large reward offers (40/80 μL), which produced more negative cholinergic dips, and small reward offers (5/10 μL), which produced shallower dips (Fig. 5K). Rats exhibited faster movements on trials with small reward offers, in which we elicited cholinergic stimulation on top of a higher baseline (Fig. 5K). These data suggest that cholinergic bursts permit dopamine to invigorate movements, and the magnitude of the burst has graded effects on vigor.

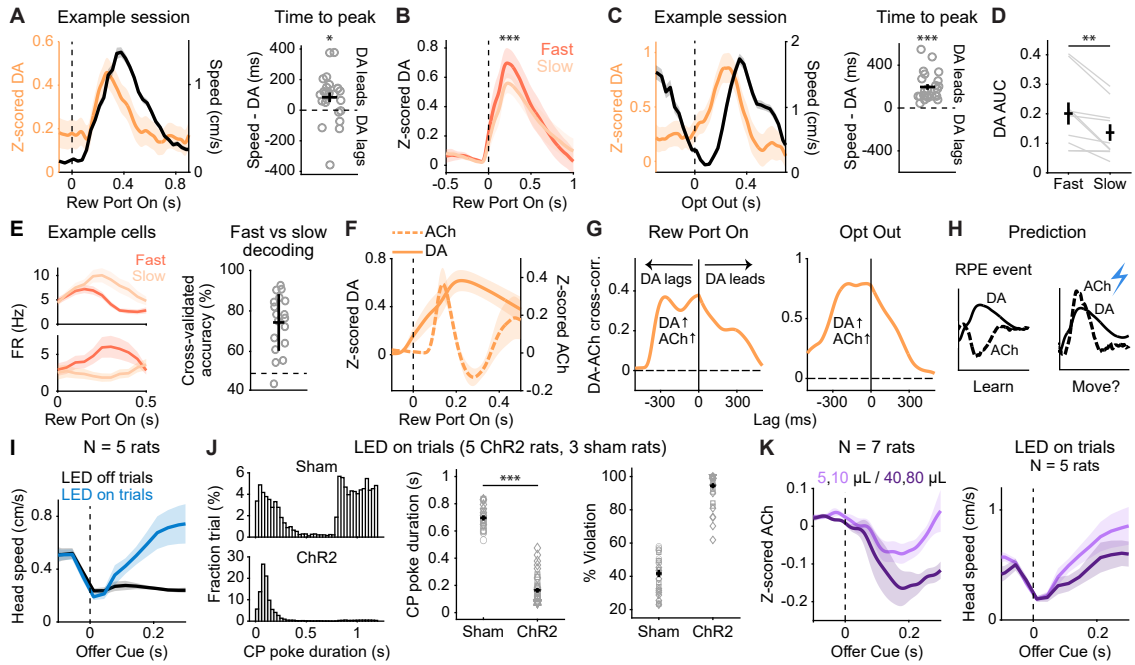

Fig. 5. Coincident dopamine and acetylcholine bursts in DMS drive movement. (**A**) Left, Dopamine release on contralateral trials (orange) when the side LED turns on, overlaid with head speed (black) in an example session (mean ± s.e.m. across trials). Right, Difference in time to peak in dopamine compared to head speed (N = 24 sessions from 5 rats, mean ± s.e.m. across sessions, t-test: p = 0.0135). Empty circles indicate individual sessions. (**B**) Average dopamine release on contralateral trials aligned to when the side LED turns on, split by reaction time (fast: bottom quartile, slow: top quartile; N = 164 sessions from 10 rats, mean ± s.e.m., mean difference in AUC = 0.0475, two-sided Wilcoxon signed rank: p << 0.0001). (**C**) Same as in **A** but for opt-out (t-test: p << 0.0001). (**D**) Dopamine AUC at opt-out split by the speed of the following trial initiation time (fast: bottom quartile, slow: top quartile; N = 10 rats, mean ± s.e.m., two-sided Wilcoxon signed rank: p = 0.0020). (**E**) Left, Average firing rate of two simultaneously recorded example cells on fast (dark orange) versus slow (light orange) reaction time trials (mean ± s.e.m. across trials). Right, Average cross-validated classification accuracy of a logistic regression decoder. Dashed horizontal line indicates chance performance (N = 18 sessions from 8 rats, mean ± s.e.m. across sessions). Grey circles indicate individual sessions. (**F**) Average z-scored dopamine (solid orange) and acetylcholine (dashed orange) release (N =

10 GRAB$_{DA}$ rats, 7 GRAB$_{ACh}$ rats, mean $\pm$ s.e.m.). (**G**) Cross-correlation of average dopamine and acetylcholine signals at movement events (0-0.5 s). Negative lag indicates dopamine following acetylcholine. (**H**) Schematic of artificially inducing a burst in acetylcholine at an RPE event to mimic the endogenous dynamics at movement events. (**I**) Head speed aligned to the offer cue on optogenetically stimulated (blue) and control (black) trials (N = 5 ChR2 rats, mean $\pm$ s.e.m.). (**J**) Left, Histogram of center port nose poke duration in sham (top, N = 3) and ChR2 (bottom, N = 5) rats on stimulation trials. Middle, Average center port nose poke duration in sham and ChR2 rats on stimulation trials. Grey markers indicate individual sessions with different marker shapes for each rat (mean $\pm$ s.e.m. across sessions, two-sample t-test: $p << 0.0001$). Right, Fraction of violation trials out of stimulation trials in sham and ChR2 rats (mean $\pm$ s.e.m. across sessions). (**K**) Left, Average z-scored acetylcholine release at offer cue on small (5 and 10 µL, light purple) versus large (40 and 80 µL, dark purple) reward offer trials (N = 7 rats, mean $\pm$ s.e.m.). Right, Head speed aligned to offer cue on optogenetically stimulated trials split by small versus large reward offers (N = 5 ChR2 rats, mean $\pm$ s.e.m.). Across panels, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Discussion Although cholinergic interneurons only comprise 1-3% of neurons in the striatum (*34*), they are thought to be critical for basal ganglia function. Tonically active neurons (putative cholinergic interneurons) develop pauses in firing following cues that predict appetitive or aversive outcomes (*20*, *35–40*), suggesting that cholinergic pauses are important for or reflect learning (*41*, *42*), possibly through local modulation of dopamine release (*43–45*). While some studies reported weak relationships between cholinergic firing and movement (*20*, *35*, *36*, *39*), others have found signals that reflect motor responses (*40*, *46*, *47*). Our data suggest that the relative phase relationship between dopamine and acetylcholine (*48*, *49*) is key for determining dopamine's action on recipient neurons. Recent work suggests that temporal dynamics of dopamine release can vary depending on the behavioral engagement of striatal subregions (*50*). We speculate that modulation of the phase relationship between dopamine and acetylcholine provides a circuit mechanism for arbitration between behavioral controllers originating from different striatal subregions, and when striatal subregions do not support behavior, dopamine

and acetylcholine dynamics will be slightly out of phase (*48*). Moreover, dopamine signals that seem inconsistent with RPE (for instance, phasic bursts to cues that predict aversive outcomes (*51*)), may be more interpretable in the context of concurrent cholinergic dynamics.

Classic studies suggested that corticostriatal plasticity is governed by a three-factor rule requiring coincident pre- and post-synaptic spiking and dopamine receptor binding (*31*, *52*, *53*). Recent work has revised this view, arguing that a fourth factor - pauses in cholinergic interneurons - is required for induction of synaptic plasticity (*18*). Our findings support this hypothesis. Cholinergic pauses might gate plasticity by relieving M1-receptor-mediated blockade of L-type calcium channels, thereby promoting endocannabinoid release (*53–55*), or by relieving M4-receptor mediated suppression of adenylyl cyclase and subsequent PKA activation, particularly in D1-receptor expressing neurons (*56*, *57*).

Despite the prominence of the dopamine RPE hypothesis, it has remained unclear how and when RPE-like updating of striatal activity actually occurs during behavior. While reinforcement learning models of behavior assume that learning is reflected in trial-by-trial changes in state-action values (Q-values), represented in corticostriatal synaptic weights (*58*), studies in slices typically evaluate effects of plasticity protocols on the order of minutes after tetanic stimulation or pre/post spike pairing. Our data suggest that induction of plasticity can occur within the seconds-scale inter-trial intervals in our task, while expression of plasticity via changes in striatal firing rates is persistent over trials. Moreover, the temporal specificity of effects at RPE but not movement events further suggests that trial-by-trial changes in firing rates did not simply reflect dopaminergic modulation of excitability (*32*, *33*). These results add to recent findings in other brain regions of rapid *in vivo* plasticity that operates on behavioral timescales (*59*), and validate key assumptions of reinforcement learning accounts of behavior.

References

1. R. S. Sutton, A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
2. J. Olds, Self-stimulation of the brain: Its use to study local effects of hunger, sex, and drugs. *Science* **127**, 315–324 (1958).
3. D. Corbett, R. A. Wise, Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbrain: a moveable electrode mapping study. *Brain research* **185**, 1–15 (1980).
4. E. E. Steinberg *et al.*, A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience* **16**, 966–973 (2013).
5. N. F. Parker *et al.*, Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature neuroscience* **19**, 845–854 (2016).
6. M. J. Sharpe *et al.*, Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience* **20**, 735–742 (2017).
7. H.-C. Tsai *et al.*, Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080–1084 (2009).
8. A. A. Hamid *et al.*, Mesolimbic dopamine signals the value of work. *Nature neuroscience* **19**, 117–126 (2016).
9. A. R. Adamantidis *et al.*, Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *Journal of neuroscience* **31**, 10829–10835 (2011).
10. J. Cox, I. B. Witten, Striatal circuits for reward learning and decision-making. *Nature Reviews Neuroscience* **20**, 482–494 (2019).
11. J. D. Berke, What does dopamine mean? *Nature neuroscience* **21**, 787–793 (2018).
12. P. Redgrave *et al.*, Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience* **11**, 760–772 (2010).
13. J. A. Da Silva, F. Tecuapetla, V. Paixão, R. M. Costa, Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244–248 (2018).
14. M. W. Howe, D. A. Dombeck, Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).
15. J. W. Barter *et al.*, Beyond reward prediction errors: the role of dopamine in movement kinematics. *Frontiers in integrative neuroscience* **9**, 39 (2015).
16. R. S. Lee, M. G. Mattar, N. F. Parker, I. B. Witten, N. D. Daw, Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. *Elife* **8**, e42992 (2019).

17. P. D. Dodson *et al.*, Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism. *Proceedings of the National Academy of Sciences* **113**, E2180–E2188 (2016).

18. J. N. Reynolds *et al.*, Coincidence of cholinergic pauses, dopaminergic activation and depolarisation of spiny projection neurons drives synaptic plasticity in the striatum. *Nature Communications* **13**, 1296 (2022).

19. S. J. Cragg, Meaningful silences: how dopamine listens to the ACh pause. *Trends in neurosciences* **29**, 125–131 (2006).

20. G. Morris, D. Arkadir, A. Nevet, E. Vaadia, H. Bergman, Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron* **43**, 133–143 (2004).

21. P. Calabresi, B. Picconi, A. Tozzi, M. Di Filippo, Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in neurosciences* **30**, 211–219 (2007).

22. A. Mah, S. S. Schiereck, V. Bossio, C. M. Constantinople, Distinct value computations support rapid sequential decisions. *Nature communications* **14**, 7573 (2023).

23. C. E. Golden *et al.*, Estrogenic control of reward prediction errors and reinforcement learning. *bioRxiv*, 2023–12 (2023).

24. A. Mathis *et al.*, DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience* **21**, 1281–1289 (2018).

25. J. Lee, W. Wang, B. L. Sabatini, Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nature neuroscience* **23**, 1388–1398 (2020).

26. M. S. Creamer, K. S. Chen, A. M. Leifer, J. W. Pillow, Correcting motion induced fluorescence artifacts in two-channel neural imaging. *PLoS computational biology* **18**, e1010421 (2022).

27. S. Kobayashi, W. Schultz, Influence of reward delays on responses of dopamine neurons. *Journal of neuroscience* **28**, 7837–7846 (2008).

28. C. K. Starkweather, B. M. Babayan, N. Uchida, S. J. Gershman, Dopamine reward prediction errors reflect hidden-state inference across time. *Nature neuroscience* **20**, 581–589 (2017).

29. J. N. Kerr, J. Wickens, Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. *Journal of neurophysiology* **85**, 117–124 (2001).

30. D. Centonze *et al.*, Unilateral dopamine denervation blocks corticostriatal LTP. *Journal of neurophysiology* **82**, 3575–3579 (1999).

31. W. Shen, M. Flajolet, P. Greengard, D. J. Surmeier, Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848–851 (2008).

32. A. K. Lahiri, M. D. Bevan, Dopaminergic transmission rapidly and persistently enhances excitability of D1 receptor-expressing striatal projection neurons. *Neuron* **106**, 277–290 (2020).

33. J. Lee, B. L. Sabatini, Striatal indirect pathway mediates exploration via collicular competition. *Nature* **599**, 645–649 (2021).

34. H. Kimura, P. L. McGeer, F. Peng, E. G. McGeer, Choline acetyltransferase-containing neurons in rodent brain demonstrated by immunohistochemistry. *Science* **208**, 1057–1059 (1980).

35. M. Kimura, J. Rajkowski, E. Evarts, Tonically discharging putamen neurons exhibit set-dependent responses. *Proceedings of the National Academy of Sciences* **81**, 4998–5001 (1984).

36. P Apicella, E. Scarnati, W Schultz, Tonically discharging neurons of monkey striatum respond to preparatory and rewarding stimuli. *Experimental brain research* **84**, 672–675 (1991).

37. P. Apicella, S. Ravel, M. Deffains, E. Legallet, The role of striatal tonically active neurons in reward prediction error signaling during instrumental task performance. *Journal of Neuroscience* **31**, 1507–1515 (2011).

38. T. Aosaki, M. Miura, T. Suzuki, K. Nishimura, M. Masuda, Acetylcholine–dopamine balance hypothesis in the striatum: An update. *Geriatrics & gerontology international* **10**, S148–S157 (2010).

39. S. Ravel, E. Legallet, P. Apicella, Responses of tonically active neurons in the monkey striatum discriminate between motivationally opposing stimuli. *Journal of Neuroscience* **23**, 8489–8497 (2003).

40. P. M. Blazquez, N. Fujii, J. Kojima, A. M. Graybiel, A network representation of response probability in the striatum. *Neuron* **33**, 973–982 (2002).

41. T. Aosaki, A. M. Graybiel, M. Kimura, Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science* **265**, 412–415 (1994).

42. N. T. Franklin, M. J. Frank, A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. *Elife* **4**, e12029 (2015).

43. R. Cachope *et al.*, Selective activation of cholinergic interneurons enhances accumbal phasic dopamine release: setting the tone for reward processing. *Cell reports* **2**, 33–41 (2012).

44. S. Threlfell *et al.*, Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. *Neuron* **75**, 58–64 (2012).

45. C. Liu *et al.*, An action potential initiation mechanism in distal axons for the control of dopamine release. *Science* **375**, 1378–1385 (2022).

46. L. Benhamou, O. Kehat, D. Cohen, Firing pattern characteristics of tonically active neurons in rat striatum: context dependent or species divergent? *Journal of Neuroscience* **34**, 2299–2304 (2014).

47. M. Howe *et al.*, Coordination of rapid cholinergic and dopaminergic signaling in striatum during spontaneous movement. *Elife* **8**, e44903 (2019).

48. A. C. Krok *et al.*, Intrinsic dopamine and acetylcholine dynamics in the striatum of mice. *Nature* **621**, 543–549 (2023).

49. L. Chantranupong *et al.*, Dopamine and glutamate regulate striatal acetylcholine in decision-making. *Nature* **621**, 577–585 (2023).

50. A. A. Hamid, M. J. Frank, C. I. Moore, Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733–2749 (2021).

51. W. Schultz, Dopamine reward prediction-error signalling: a two-component response. *Nature reviews neuroscience* **17**, 183–195 (2016).

52. J. N. Reynolds, B. I. Hyland, J. R. Wickens, A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).

53. A. C. Kreitzer, R. C. Malenka, Endocannabinoid-mediated rescue of striatal LTD and motor deficits in Parkinson's disease models. *Nature* **445**, 643–647 (2007).

54. Z. Wang *et al.*, Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. *Neuron* **50**, 443–452 (2006).

55. A. C. Kreitzer, R. C. Malenka, Striatal plasticity and basal ganglia circuit function. *Neuron* **60**, 543–554 (2008).

56. E. Ince, B. J. Ciliax, A. I. Levey, Differential expression of D1 and D2 dopamine and m4 muscarinic acetylcholine receptor proteins in identified striatonigral neurons. *Synapse* **27**, 357–366 (1997).

57. J. A. Goldberg, J. B. Ding, D. J. Surmeier, Muscarinic modulation of striatal function and circuitry. *Muscarinic receptors*, 223–241 (2012).

58. R. Chen, J. H. Goldberg, Actor-critic reinforcement learning in the songbird. *Current opinion in neurobiology* **65**, 1–9 (2020).

59. K. C. Bittner, A. D. Milstein, C. Grienberger, S. Romani, J. C. Magee, Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).

Acknowledgments

We thank Tanya Sippy, Cristina Savin, Robert Froemke, Richard Tsien, and members of the Constantinople lab for feedback. We thank research technicians in the Constantinople lab for animal training.

Funding

DP2MH126376

R01MH136272

Alfred P. Sloan Research Fellowship

Author contributions

Conceptualization: CMC, HJJ

Data acquisition: HJJ, RMW, CEG

Data analysis: HJJ

Visualization: HJJ

Funding acquisition: CMC

Supervision: CMC

Writing: CMC, HJJ

Competing interests

Authors declare that they have no competing interests.

Supplementary materials

Materials and Methods

Figs. S1 to S10

Supplementary materials

Materials and Methods

Rats 32 male and 7 female Long-evans rats between the ages of 6 and 24 months were used for this study. Out of 8 rats used for electrophysiology experiment, 3 rats were TH-Cre and 1 rat was Adora2a transgenics. No behavioral difference in task performance was found between the transgenics and wild-type rats. Rats were water restricted to motivate them to perform the task. At the end of the day, rats were given 20 minutes of access to ad libitum water. Following behavioral sessions on Friday until mid-day Sunday, rats received ad libitum water. Animal use procedures were approved by the New York University Animal Welfare Committee (UAWC #2021-1120) and carried out in accordance with National Institute of Health standards.

Behavior training and task logic Rats were trained in incremental stages of task complexity in a high-throughput behavioral facility using a computerized training protocol as described previously (*1*).

LED illumination from the central nose port indicated that the animal could initiate a trial by poking its nose in that port ("CP On"). Upon trial initiation the center LED turned off. While in the central nose port, rats needed to maintain center fixation for a duration drawn uniformly from [0.8, 1.2] seconds. During the fixation period, a tone played from both speakers, the frequency of which indicated the volume of the offered water reward for that trial [1, 2, 4, 8, 16 kHz, indicating 5, 10, 20, 40, 80 μL rewards] ("Offer Cue"). Following the fixation period, one of the two side LEDs was illuminated, indicating that the reward might be delivered at that port ("Rew Port On"); the side was randomly chosen on each trial. This event also initiated a variable and unpredictable delay period, which was randomly drawn from an exponential distribution with mean = 2.5 seconds. The reward port LED remained illuminated for the duration of the delay period, and rats were freely behaving during this period. When reward was available, the reward port LED turned off ("Rew Cue"), and rats could collect the offered reward by nose poking in that port ("Rew Delivery"). The rat could also choose to terminate the trial at any time by nose poking in the opposite, un-illuminated side port ("Opt Out"), after which a new trial would immediately begin. On a proportion of trials (15-25%), the delay period would only end if the

rat opted out (catch trials). If rats did not opt-out within 100 seconds on catch trials, the trial would terminate.

The trials were self-paced: after receiving their reward or opting out, rats were free to initiate another trial immediately. However, if rats terminated center fixation prematurely, they were penalized with a white noise sound and a time out penalty, typically lasting 2 seconds ("Violation"). The trial following a premature fixation break offered the rats the same amount of reward in order to dis-incentivize premature terminations for small volume offers.

We introduced semi-observable, hidden-states in the task by including uncued blocks of trials with varying reward statistics (*2*): high and low blocks, which offered the highest three or lowest three rewards respectively, were interspersed with mixed blocks, which offered all volumes. There was a hierarchical structure to the blocks, such that high and low blocks alternated after mixed blocks (e.g., mixed-high-mixed-low, or mixed-low-mixed-high). Each session always started with a mixed block. Blocks transitioned after 40 successfully completed trials. Because rats prematurely broke fixation on a subset of trials, in practice, blocks varied in number of trials.

Rats were considered to have achieved task proficiency when their time to opt-out exhibited linear sensitivity to the offered reward volume and time to opt-out for the 20 µL offer in low blocks was longer than in high blocks for at least 3 consecutive days. Rats that satisfied these criteria underwent surgeries specific to experiments. After recovering from surgeries, rats were retrained until they achieved the same pre-surgery performance which typically took one session. Rats were then habituated to being tethered for 2-3 days before proceeding with recordings and/or manipulations.

Surgery Rats were anaesthetized (isofluorane at 4% for induction and 1.5-2.5% for maintenance) and placed in a stereotaxic frame. Skull screws were inserted to secure implants before proceeding with surgeries specific to experiments.

Fiber photometry: For fiber photometry recordings, rats were injected with 300-480 nL of GRAB$_{DA}$ (AAV9-hSyn-GRAB_DA2h, AddGene #140554) or GRAB$_{ACh}$ (AAV9-hSyn-ACh4.3, WZ Biosciences #YL10002) in the DMS (AP: -0.4 mm, ML: $\pm$3.5 mm, DV: -4.6 mm or AP: 1.7 mm, ML: $\pm$2.8 mm, DV: -4.3 mm from the skull surface at bregma at a 10° angle). 40 nL of

virus was injected at a time at a rate of 10 nL sec^{-1} and the syringe was lowered stochastically to inject across 0.3 mm in DV around the target depth. For rats with mCherry (AAV9-CB7-CI-mCherry-WPRE-RBG, AddGene #105544) as a control fluorophore, the GRAB sensors were mixed with mCherry in 1.5:1 $\sim$ 2.5:1 ratio prior to injection. For rats with isosbestic control, only GRAB sensor was injected. Rats were then bilaterally implanted with 400 µM core, 0.5 NA optical fibers (Thorlabs) above the injection area. Craniotomies were sealed with Kwiksil and the fibers were fixed to the skull with dentin and Metabond. Rats were allowed to recover for a minimum of 7 days.

Electrophysiology: The day before surgery, Neuropixels probe assembly was prepared by mounting the probe (Neuropixels probe 1.0) onto a custom-designed 3D printed piece and staining with DiI. Open Ephys built-in self tests were run on the probe assembly prior to surgery. Rats were implanted with a probe assembly in the DMS (AP: -0.24 mm, ML: $\pm$ 2.5 mm, DV: -4.4 mm from the skull surface at bregma; 6 rats in the left hemisphere, 2 rats in the right hemisphere). A ground wire was inserted in the cerebellum and fixed with Dentin. Craniotomies were covered with Kwik-Cast and gaps between the probe assembly and the skull were sealed with sterile paralube vet ointment. The probe was then fixed to the skull with Metabond and dental acrylic. Rats were allowed to recover for a minimum of 5 days.

Optogenetics: For optogenetic manipulations, rats were injected with 550-700 nL of ChAT-Cre (rAAV-ChAT-CRE-WPRE-hGH, Biohippo #BHV12400208-6) and ChR2 (AAV9-EF1a-double floxed-hChR2, AddGene #20298) mixed in 1:1 ratio in the DMS (AP: 0.24 mm, ML: $\pm$2.5 mm, DV: -4.0 mm from the skull surface at bregma at a 6° angle). 50 nL of virus was injected at a time at a rate of 10 nL sec^{-1} and the syringe was lowered stochastically as described previously. Rats were then bilaterally implanted with 200 µM core tapered optical fibers (0.66 NA, 1 mm active length, 6.5 mm implant length, Optogenix) in the injection area such that the emitting tip of the fiber is in the target region. For the sham group, rats were bilaterally implanted with tapered optical fibers in the DMS without virus infusion.

Markerless pose tracking Videos and synchronization pulses were acquired using methods described below for each experiment type. Deeplabcut software (v.2.2.3) was used for tracking the behavior ports (right port, central port, and left port) and 6 body parts (right ear, nose, left

ear, mid-point along the right torso, mid-point along the left torso, and base of the tail). Frames to be annotated and used as a training set were selected using k-means clustering. For each training cycle, 100-200 frames were extracted and annotated. The training was initialized with ResNet-50 and run for 500,000 iterations. Upon completion of training, the model-predicted positions of the features of interest were visually inspected. Outlier frames were then extracted manually and re-annotated. Frame annotation and training was repeated until no significant outliers existed, which typically took 4-5 training cycles with a total of 400-600 annotated frames.

The tracked features were smoothed in time with a moving median filter of 0.167 s. To obtain head speed and angle relative to the behavior ports for analysis, the centroid of the head was calculated by averaging 3 body parts on the head (earR, earL, and nose). Extreme outliers in head speed and angle were removed by truncating the 99th percentile. Head speed and angle were aligned to task events by aligning the synchronization pulses with Bpod behavior events.

Fiber photometry experiment Video and fiber photometry data acquisition: Video data were acquired using cameras attached to the ceiling of behavior rigs to capture the top-down view of the arena (Doric USB3 behavior camera, Sony IMX290, recorded with Doric Neuroscience Studio v6 software). To synchronize with neural recordings and Bpod task events, the camera was connected to a digital I/O channel of a fiber photometry console (FPC) and triggered at 30 Hz via a TTL pulse train. The remaining 3 digital I/O channels of the FPC were connected to the 3 behavior ports to obtain TTL pulses from port LEDs turning on. An analog channel of the FPC was connected to the patch cords to record fluorescence. Fluorescence from activity-dependent (GRAB$_{DA}$ and GRAB$_{ACh}$) and activity-independent (isosbestic or mCherry) signals was acquired simultaneously via demodulation and downsampled on-the-fly by a factor of 25 to $\sim$ 481.9 Hz.

Fiber photometry data preprocessing: The recorded demodulated fluorescence was corrected for photobleaching and motion using Two-channel motion artifact correction (*3*) with mCherry or isosbestic signal as the activity-independent channel. The motion-corrected signal was z-scored and then aligned to task events using synchronization TTL pulses from the behavior ports.

Electrophysiology experiment Neuropixels data acquisition: Spike data were acquired from 384 channels at 30 kHz using Neuropix-PXI hardware and OpenEphys. To synchronize with behavior, task event TTL pulses were recorded simultaneously using NI-DAQmx.

Spike sorting: Spike data were preprocessed using Kilosort 2.5. After preprocessing, clusters that were identified from Kilosort as single-units were manually inspected using the Phy2 Python package. Clusters with unreliable waveform shape, autocorrelogram, or spike amplitude timecourse were manually classified as multi-unit or noise and removed from further analysis. For the identified single-units, spike data were aligned to behavioral events and binned into 5 ms bins using custom Matlab codes.

Optogenetics experiment After recovering from surgery, rats were habituated to being tethered until they achieved pre-surgery performance which typically took 1-2 sessions. Following habituation, control sessions (no LED) and stimulation sessions (LED on on 30% of trials) were alternated every 3 days for a minimum of 9 sessions total per rat. Blue LED modules (465 nm, 1.9-2.3 mW, PlexBright LEDs, Plexon) were connected to a commutator and 200 µM core patch cords that were tethered to the fiber implants of rats. The connections between the LEDs, patch cords, and fiber implants were sealed with black electric tape or black shrink tubes to block light leak. Continuous stimulation was delivered randomly on 30% of trials for 600 ms at the onset of central nose pokes. If rats retracted from the central nose port before 600 ms was up, stimulation was also terminated.

Video and synchronization pulses to align with behavioral events were acquired with a Doric USB3 behavior camera and Doric Neuroscience Software with methods described above for fiber photometry recordings.

Behavioral model We modeled the trial initiation times as inversely proportional to the value of the environment, which is updated according to the following recursive equation: $V_{t+1} = V_t + \alpha(R_t - V_t)$, where $R_t - V_t$ is the RPE, and α is a learning rate ranging from 0 to 1 that dictates the dynamics with which values are updated over trials. We used trial statistics from an example rat to simulate the model. The learning rate parameter used in the example model simulation was 0.4.

Trial-by-trial spiking analysis To account for biases in statistical test due to imbalanced dataset across task events, we performed random upsampling of the minority dataset. If the ratio of the number of data points between the majority and the minority datasets exceeded 2.5, the minority dataset was upsampled via random sampling with gaussian noise (sigma = 1). Only sessions that had at least 20 trials were included for analysis.

Logistic regression model To quantify encoding of movement vigor in DMS spikes, we trained a logistic regression decoder. For each session, the predictor tensor matrix (Trial x Neuron x Time) was built by concatenating spikes from 0 to 0.5 seconds after the onset of reward port assignment. For each contralateral trial in a given session, it was classified as fast/slow if it were in the bottom/top quartile of the reaction time distribution. Only sessions with at least 5 simultaneously recorded DMS units were included for analysis. We used the Matlab function fitclinear to train the decoder to predict 1 if a given trial is fast, or 0 if slow. To cross-validate, we implemented 10-fold cross-validation with ridge regression. The regularization strength was determined by training the decoder sequentially on a set of 11 logarithmically-spaced regularization strengths ranging from 10^{-4} to $10^{-0.2}$ and selecting the value that minimizes classification error.

Histology Following completion of each experiment, rats were deeply anesthetized and transcardially perfused with phosphate buffer saline (PBS) and formalin for histological verification. Brains were post-fixed with formalin for 3 days and sectioned at 50 μm coronal slices on a Leica VT1000 vibratome. For electrophysiology experiment, sections were imaged with an Olympus VS120 Virtual Slide Microscope to verify probe location. For fiber photometry experiment, sections were stained to enhance GFP and imaged to verify virus expression and fiber implant location.

Immunohistochemistry: For brain tissues acquired for fiber photometry experiment, sections were treated with 1% hydrogen peroxide in 0.01 M PBS for 30 min, washed with 0.01 M PBS, and blocked in 0.01 M PBS containing 0.05% sodium azide and 1% bovine serum albumin (PBS-BSA-Azide) for 30 minutes prior to treating with primary antibodies. GFP was amplified with Rabbit anti-GFP (Thermo Fisher Scientific A11122, 1:2,000, lot #2083201) as

primary antibody and Goat anti-Rabbit AlexaFluor 488 IgG (Thermo Fisher Scientific A11008, 1:200, lot #2147635) as secondary antibody. Primary and secondary antibodies were made in PBS-BSA-Azide and primary antibodies additionally contained 0.2% triton-x. Sections were incubated in primary antibody overnight at room temperature, washed with PBS for 15 min, and treated with secondary antibody for 1 hour in dark. After washing with PBS for 15 min, sections were mounted on glass slides using Prolong Gold (Vector Laboratories).

Figs. S1 to S10

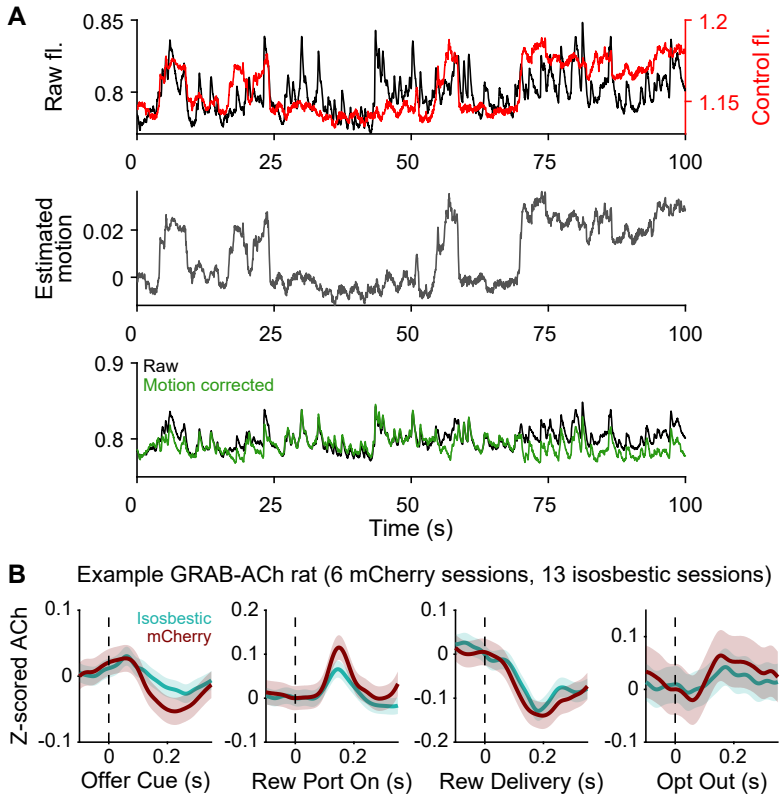

fig. S1. Motion correction using TMAC produces similar event-aligned responses for both mCherry and isosbestic control. (**A**) Top, Example raw fluorescence (black) overlaid with simultaneously recorded mCherry fluorescence (red). Middle, Estimated motion calculated with TMAC (grey). Bottom, Raw fluorescence overlaid with TMAC motion corrected fluorescence (green). (**B**) Average motion-corrected acetylcholine signal using isosbestic (green, N = 13 sessions) or mCherry (red, N = 6 sessions) as the activity-independent channel for an example rat.

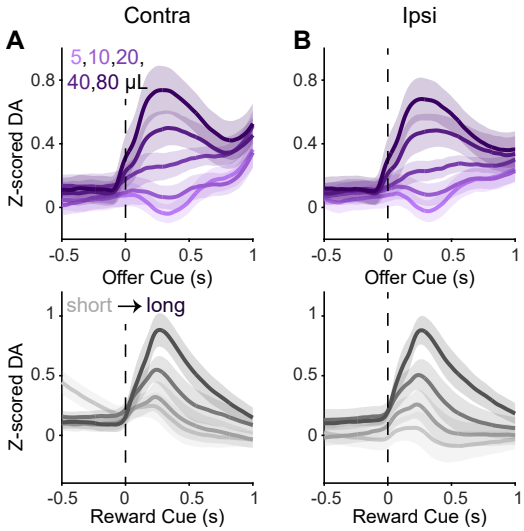

fig. S2. Dopamine release is not side selective at RPE-associated events. (**A**) Average dpoamine release split by offered reward volume in mixed blocks at Offer Cue (top) and by reward delay at Rew Cue (bottom) on contralateral trials. (**B**) Same as in **A** but on ipsilateral trials.

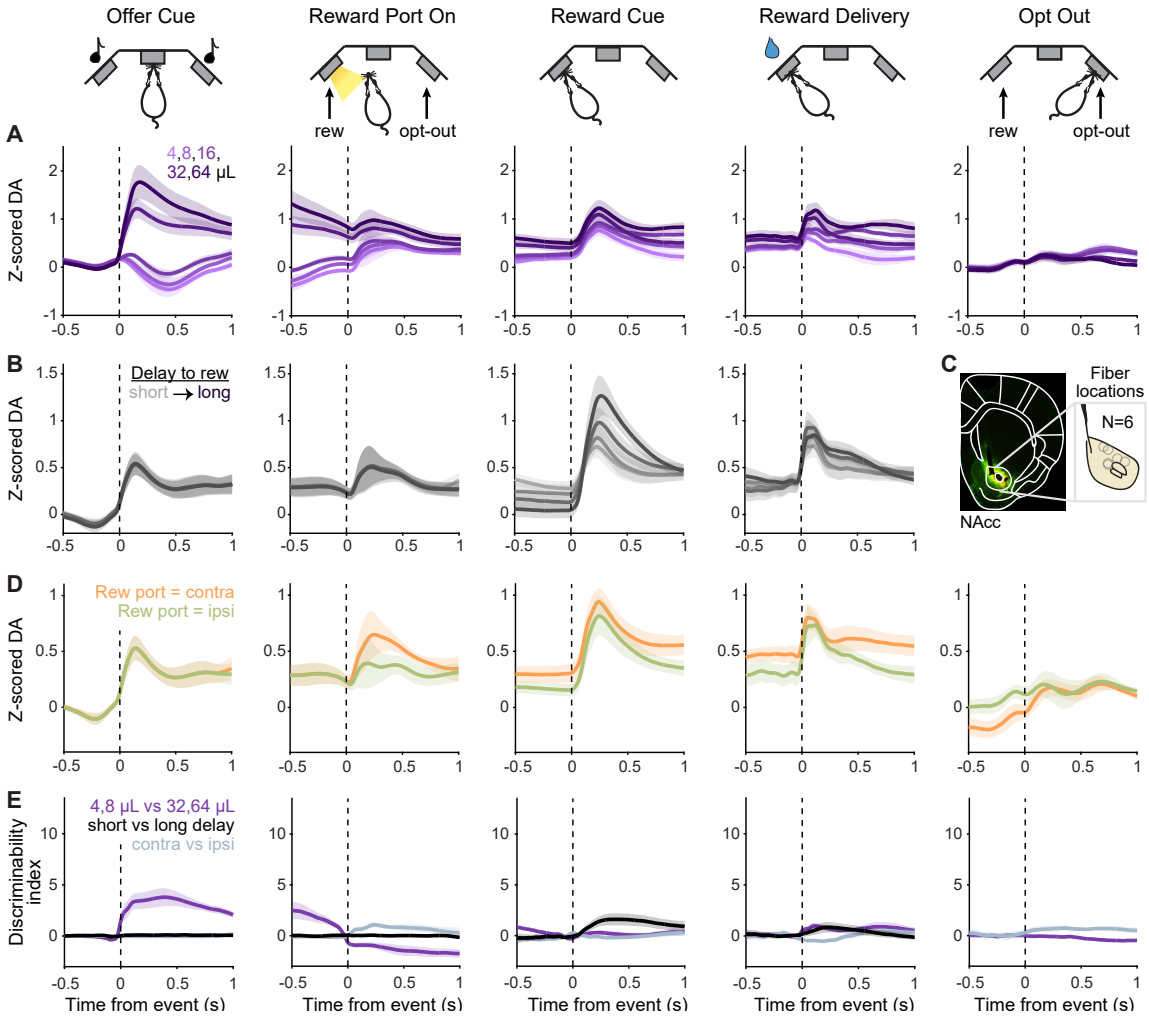

fig. S3. Dopamine release in the NAcc does not show contralateral selectivity. (**A**) Average event-aligned dopamine release in the NAcc split by offered reward volume in mixed blocks. (**B**) Average event-aligned dopamine release in the NAcc split by reward delay quartiles across all blocks. (**C**) Representative histology image of viral expression of GRAB$_{DA}$ and mCherry and recovered fiber placements. (**D**) Average event-aligned dopamine release in the NAcc split by contralateral and ipsilateral trials. (**E**) Average discriminability index of different task variables (purple: reward volume, black: reward delay, light blue: contra/ipsi) aligned to all task events. N = 6 rats, mean ± s.e.m. across panels.

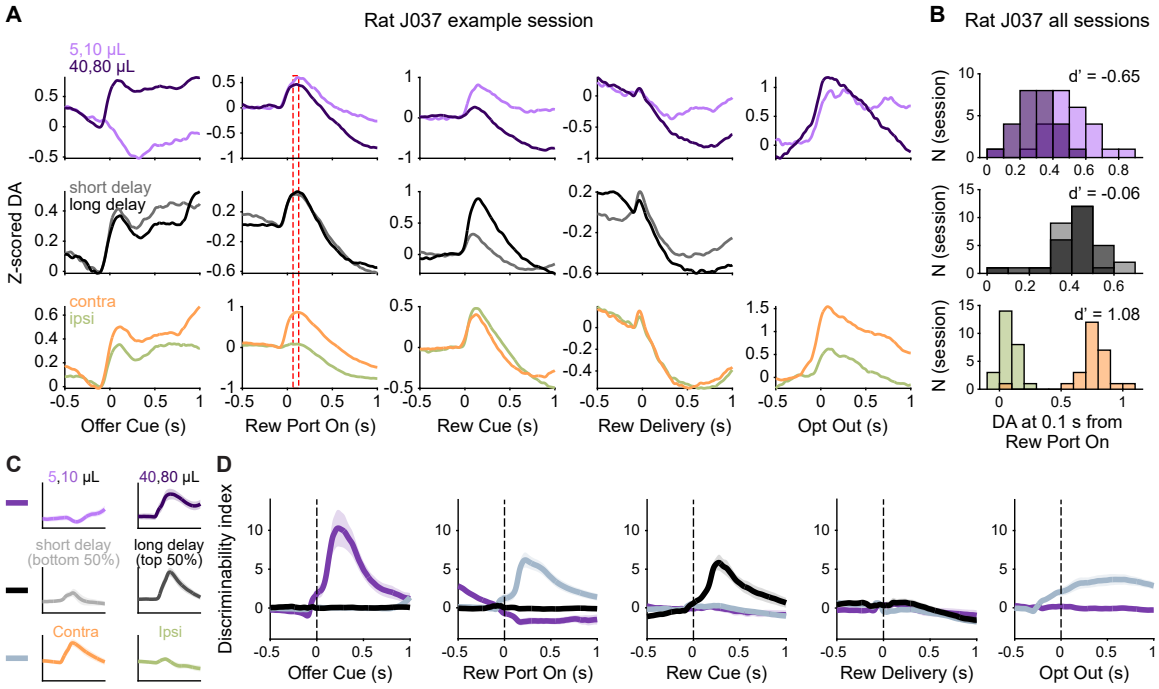

fig. S4. Discriminability index quantifies encoding of task variables at different events. (**A**) Top, Average z-scored dopamine release on small (5, 10 µL) versus big (40, 80 µL) reward offer trials aligned to all task events in an example session. Middle, Average z-scored dopamine release on short (bottom 50% of reward delay distribution) vs long (top 50%) reward delay trials. Bottom, Average z-scored dopamine release on contralateral versus ipsilateral trials. Across panels, the red vertical line indicates the time point the histograms in **B** were generated from (t = 0.1 s from Rew Port On onset). (**B**) Distribution of z-scored dopamine at 0.1 s from Rew Port On (red vertical line in **A**) across 26 sessions from an example rat. (**C**) Schematic of how task variables are binarized to compute discriminability index. (**D**) Average discriminability index of different task variables (purple: reward volume, black: reward delay, light blue: contra/ipsi) aligned to all task events (N = 10 rats, mean ± s.e.m.).

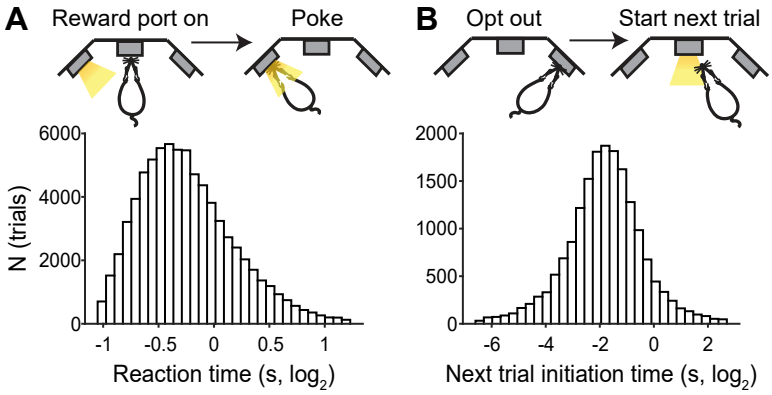

fig. S5. Behavioral variability is higher at opt-out compared to reward port assignment. (**A**) Histogram of rats' reaction time to the side LED lighting up to indicate the side of the reward port. (**B**) Histogram of rats' time to initiate trial following opt-outs. N = 17 rats across panels.

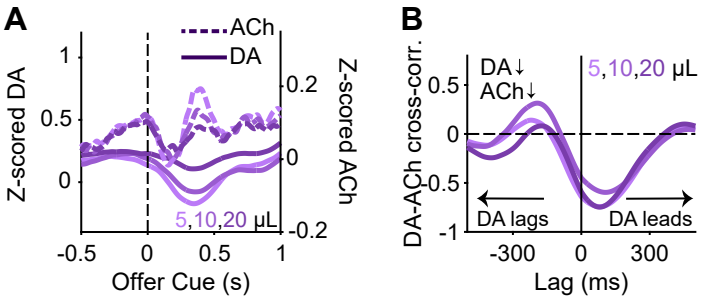

fig. S6. Acetylcholine dips slightly lead dips in dopamine. (**A**) Average z-scored dopamine (solid, N = 10 rats) and acetylcholine (dashed, N = 7 rats) signals for small reward offer trials aligned to the onset of offer cue. (**B**) Cross-correlation between average dopamine and acetylcholine signals around offer cue ($0 \sim 0.5$ s) for small reward offer trials. Negative lag indicates dopamine lagging acetylcholine.

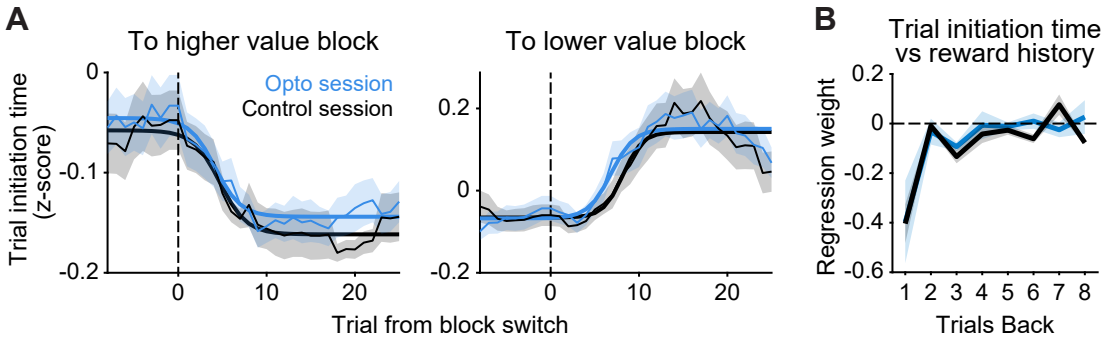

fig. S7. LED stimulation does not impair learning in sham rats. (**A**) Z-scored trial initiation time around block transitions averaged across control (black) and optogenetic (blue) sessions. (**B**) Trial initiation time regressed against reward history in optogenetic (blue) versus control (black) sessions. N = 3 rats, mean ± s.e.m. across panels.

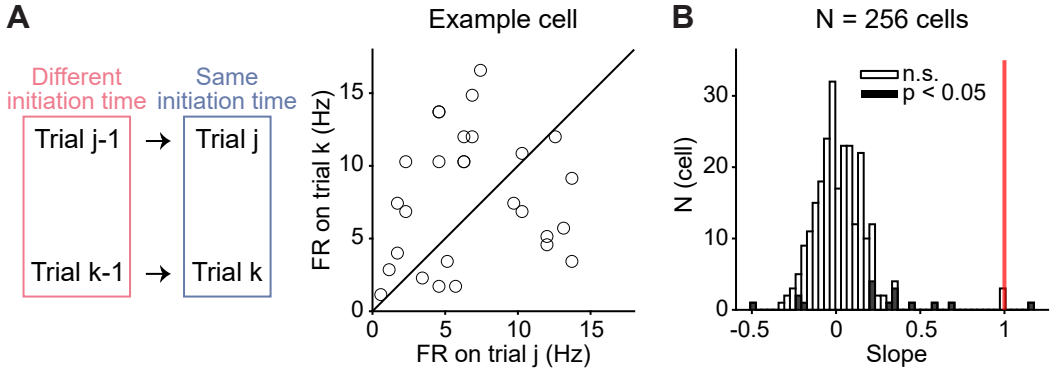

fig. S8. Trial initiation speed on current trial alone does not explain DMS firing rate. (**A**) Left, Trials j and k have similar trial initiation times (difference < 0.01 s) but different initiation times on the previous trial (difference > 2 s). Right, Average firing rate at offer cue ($0 \sim 0.3$ s) on trials j and k for an example cell. Solid line is the unity line. (**B**) Histogram of best fit line slopes for all cells. Red vertical line indicates slope of 1.

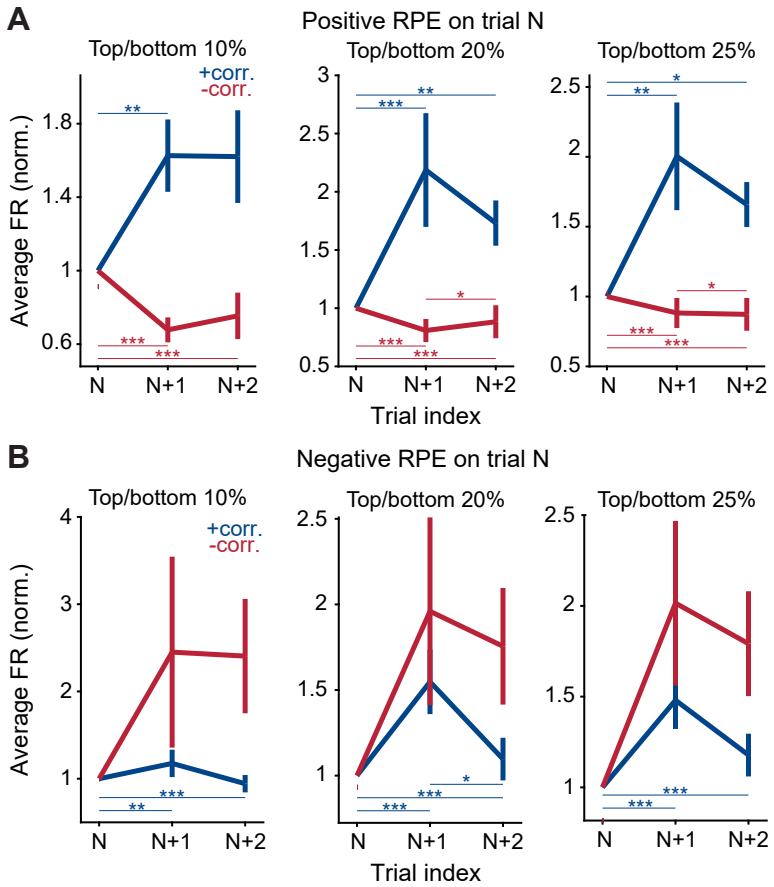

fig. S9. RPE produces lasting changes in task-evoked firing rates for a range of RPE thresholds. (**A**) Change in average z-scored firing rate at offer cue ($0 \sim 0.3$ s) on 3 consecutive trials where a big positive RPE on trial N is followed by a negligible RPE on trial N+1. Different panels are for different thresholds used to select RPE on trial N. Two-sided Wilcoxon signed rank: Left, 10% threshold: +corr. cells, N = 100 trial sequences, trial N vs N+1, p = 0.0064; trial N vs N+2, p = 0.0620; trial N+1 vs N+2, p = 0.4269; -corr. cells, N = 115 trial sequences, trial N vs N+1, p $<<$ 0.0001; trial N vs N+2, p $<<$ 0.0001; trial N+1 vs N+2, p = 0.5312. Middle, 20% threshold: +corr. cells, N = 191 trial sequences, trial N vs N+1, p = 0.0001; trial N vs N+2, p = 0.0395; trial N+1 vs N+2, p = 0.1138; - corr. cells, N = 223 trial sequences, trial N vs N+1, p $<<$ 0.0001; trial N vs N+2, p $<<$ 0.0001; trial N+1 vs N+2, p = 0.0182. Right, 25% threshold: +corr. cells, N = 245 trial sequences, trial N vs trial N+1, p = 0.0025; trial N vs trial N+2, p = 0.0434; trial N+1 vs trial N+2, p = 0.3320; -corr. cells, N

= 280 trial sequences, trial N vs trial N+1, p << 0.0001; trial N vs trial N+2, p << 0.0001; trial N+1 vs trial N+2, p = 0.0339. (**B**) Same as in **A** but for a big negative RPE on trial N. Two-sided Wilcoxon signed rank: Left, +corr. cells, N = 119 trial sequences, trial N vs trial N+1, p = 0.0010, trial N vs trial N+2, p << 0.0001; Middle, +corr. cells, N = 240 trial sequences, trial N vs trial N+1, p = 0.0005; trial N vs trial N+2, p << 0.0001; trial N+1 vs trial N+2, p = 0.0108; Right, +corr. cells, N = 299 trial sequences, trial N vs trial N+1, p << 0.0001; trial N vs trial N+2, p << 0.0001. Across panels, *p < 0.05, **p < 0.01, ***p < 0.001.

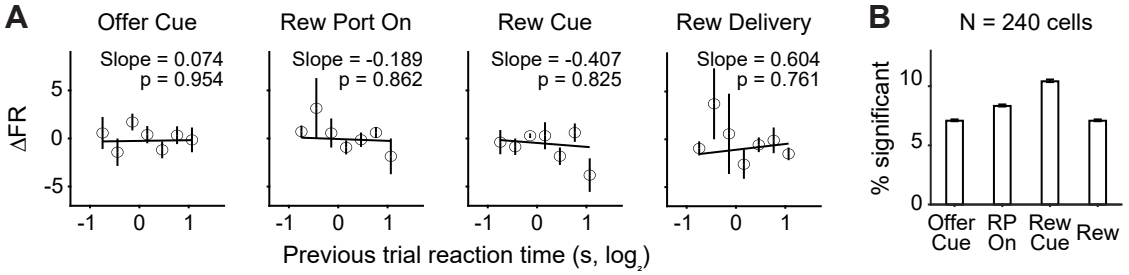

fig. S10. Speed of contralateral movement does not modulate DMS firing rate on subsequent trial. (**A**) Trial-by-trial changes in average firing rate surrounding each task event (0 ∼ 0.3 s) versus putative dopamine at Rew Port On (i.e., vigor of contralateral movement) for an example cell. Solid lines show the best fit line (t-test on the slope: Offer Cue, p = 0.954; Rew Port On, p = 0.862; Rew Cue, p = 0.825; Rew Delivery = 0.761). Trials are binned in 7 linearly spaced intervals for visualization (mean ± s.e.m.). (**B**) Fraction of cells with significant slopes at different task events. Error bars are binomial confidence intervals.