Title: A mechanism linking dopamine's roles in reinforcement, movement and motivation.

Authors:
Allison E Hamilos*[a,b], Isabella C Wijsman[b], Qinxin Ding[a,b], Pichamon Assawaphadungsit[a,b], Zeynep Ozcan[a] and John A Assad*[b,c]

Affiliations:
[a]Whitehead Institute, Cambridge, Massachusetts, 02142, USA
[b]Department of Neurobiology, Harvard Medical School, Boston, 02115, USA
[c]Istituto Italiano di Tecnologia, Genova, Italy
*For correspondence: A.E.H. (ahamilos@wi.mit.edu) or J.A.A. (jassad@hms.harvard.edu)

Summary:
Dopamine neurons (DANs) play seemingly distinct roles in reinforcement,[1-3] motivation,[4,5] and movement,[6,7] and DA-modulating therapies relieve symptoms across a puzzling spectrum of neurologic and psychiatric symptoms.[8] Yet, the mechanistic relationship among these roles is unknown. Here, we show DA's tripartite roles are causally linked by a process in which phasic striatal DA rapidly and persistently recalibrates the propensity to move, a measure of vigor. Using a self-timed movement task, we found that single exposures to reward-related DA transients (both endogenous and exogenously-induced) exerted one-shot updates to movement timing—but in a surprising fashion. Rather than reinforce specific movement times, DA transients quantitatively *changed* movement timing on the next trial, with larger transients leading to earlier movements (and smaller to later), consistent with a stochastic search process that calibrates the frequency of movement. Both abrupt and gradual changes in external and internal contingencies—such as timing criterion, reward content, and satiety state—caused changes to the amplitude of DA transients that causally altered movement timing. The rapidity and bidirectionality of the one-shot effects are difficult to reconcile with gradual synaptic plasticity, and instead point to more flexible cellular mechanisms, such as DA-dependent modulation of neuronal excitability. Our findings shed light on how natural reinforcement, as well as DA-related disorders such as Parkinson's disease, could affect behavioral vigor.

MAIN TEXT:
Both electrical[9] and optogenetic activation[10-12] of dopamine neurons (DANs) can gradually reinforce actions.[10-15] It is widely believed DANs drive gradual learning via phasic dopamine (DA) transients, with ongoing debate over what these transients represent in reinforcement learning frameworks (e.g., causal association,[16] learning rate,[3] value,[4,5] reward/performance-prediction errors[1,2,17-22], etc.). Yet, multiple lines of evidence suggest DA also plays "real-time" roles in motor control[6,7,23-25] and motivation,[5,25] which are thought to be distinct from DA's role in gradual learning.[6,22,26] For example, tonic DA levels are associated with movement vigor, with both tonic DA and vigor increasing in reward-rich environments.[6,26] Exogenous DA/DAN manipulations also causally modulate the probability,[7,23] number,[23] acceleration[23] and variability[11] of movements, and DA-modulating drugs are used ubiquitously in clinical settings as a "tuning dial" to adjust patients' propensity to move.[8,27] DA repletion/agonists increase movement propensity (e.g., to

treat Parkinson's disease), whereas DA antagonists have the opposite effect (e.g., to treat mania, Huntington's disease, delirium, etc.). Yet, it remains a mystery whether *endogenous*, reward-related DA transients are related to DA's role in movement.[23,26,28-30]

Because DA drugs bidirectionally influence the propensity to move, we suspected DA's roles in reinforcement and movement might be mechanistically linked by a tuning process, whereby endogenous DA transients calibrate the propensity to move based on feedback from recent reward history. Individual DANs encode both reward- and movement-related information,[30,31] suggesting these functions may be intertwined or somehow co-dependent.

We recently showed DANs play a causal, real-time role in movement initiation while mice perform a self-timed movement task.[7] Given this task also shows outcome-related DA signaling and quantitative behavioral changes related to motivational state, we realized this paradigm would allow us to study DA's tripartite roles in the same context, allowing us to test the calibration hypothesis.

Task for quantifying propensity to move

We trained head-fixed, water-deprived mice to perform a self-timed movement task[7] (**Fig 1a**). On each trial, mice had to estimate elapsed time to decide when to lick after a brief cue (tone/LED). Mice received a juice reward if (and only if) the first-lick occurred during a hidden reward window (3.3-7s after the cue), which had to be discovered through trial-and-error. Only the first lick determined the trial outcome (unrewarded/rewarded). Mice learned to self-time licks, with the timing distribution stabilizing within 10-20 training sessions (**Fig 1b**, **Extended Data Fig 1**).

Timing behavior is naturally variable,[32,33] with both humans and well-trained animals making many relatively early and late movements during the same session.[7,34-36] We exploited this variability to ask how DAN signaling related to trial-by-trial updates to movement timing. DA has long been implicated in timing behavior. For example, Parkinsonism/DA antagonists cause systematically late timing,[33,37-39] whereas DA agonists cause early timing.[40,41] These findings indicate that movement timing can serve as a quantitative, operational definition of what we mean by "propensity to move" (earlier timing=higher propensity; later timing=lower). We thus hypothesized that reward-related DAN signaling might adjust timing based on task reward/error feedback.

Outcome-related DA signals

We used fiber photometry to monitor intracellular Ca^{++} in genetically-defined DANs expressing GCaMP6f, recording from both cell bodies (substantia nigra, SNc: n=4 DAT-cre mice; ventral tegmental area, VTA: n=3 mice) and axon terminals (dorsolateral striatum, DLS, n=4 mice). DANs exhibited RPE-like transients in the timing task: unrewarded first-licks were followed by small dips in intracellular Ca^{++}; rewarded trials were followed by large increases (**Fig 1c**). These signals were surprisingly persistent, decaying for >10s after the reward, much longer than GCaMP6f's off-kinetics (t½ <300 ms at 37°).[42]

The same basic RPE-like features were present across DAN recording sites (SNc, VTA, DLS) and in striatal DA release (**Fig 1d**), which we monitored with the fluorescent DA sensors dLight or DA2m, expressed in either dorsolateral (DLS) or ventrolateral dorsal striatum (VLS[43,44]; **Extended Data Fig 2**).

These persistent RPE-like signals bled into the next trial (**Fig 1e**), becoming continuous with differences in baseline DAN calcium signals that were correlated with timing on the next trial (**Fig 1f**; Pearson's r: –0.79, 95% CI: [-0.84, -0.73]), suggesting outcome-related DA might influence timing based on task feedback from the previous trial.

"Sloshing" strategy for timing updates

There were three possible trial outcomes: 1) early (no reward), 2) rewarded, or 3) no lick. Because mice only rarely did not lick, we focused analyses on early versus rewarded outcomes. Intuitively, an early trial should prompt slower timing on the next trial. Unrewarded DA dips persisted for seconds and were associated with lower pre-cue DAN signals—which were indeed systematically correlated with later timing on the next trial. Conversely, rewarded trials were associated with disproportionate DA increases. Because exogenous DA can gradually reinforce actions,[10-15] we expected animals would move at about the same time on the next trial. Surprisingly, this was not the case. After rewards, DA signals remained elevated for >10s, and elevated baseline was systematically correlated with *earlier* timing on the next trial. Rather than repeating the rewarded time, mice *sped up*. This "sloshing" strategy may seem unintuitive, but is not necessarily irrational: A thirsty mouse may want to move at the earliest possible rewarded time, but this time is unknown—mice had to discover it through experience. By slowing down after early trials and speeding up after rewarded trials, this sloshing effect can cause the timing distribution to build up around the earliest possible rewarded time.

A potential confound is that, on average, relatively late trials might—by chance—tend to be followed by earlier trials (and relatively early followed by later). This "regression to the median" phenomenon could occur even in the absence of goal-directed updates to timing. We thus compared the change in the movement time between consecutive trials, $\Delta T_{(n+1)-n}$, to the change in movement time expected by chance from regression to the median (n=17 mice, 155 sessions; **Supplemental Methods**). $\Delta T_{(n+1)-n.}$ was in excess of chance: as suggested by the DAN signals, animals indeed moved significantly earlier after rewarded trials and significantly later after unrewarded trials (update after rewards: –317ms, 95% CI: [–406ms, –223ms]; update after early trials: +74ms, 95% CI: [+41ms, +106ms]; 10,000x bootstrap procedure; **Fig 1g**).

DA transients explain *changes* in timing

To determine if DA explained sloshing updates, we derived a generalized linear model (GLM) predicting the *change* in the movement time on the next trial, $\Delta T_{(n+1)-n}$, from the amplitude of outcome-related DA transients (**Fig 2a**). This amplitude was encoded as the DA signal averaged over 0-500ms following the previous trial's first-lick ("DA", **Fig 2b**). We controlled for optical artifacts and ongoing body movements by including the tdTomato ("tdt") photometry channel, accelerometer and/or neck EMG signals as predictors.

Across all recording sites (and for both GCaMP and DA sensors), DA was independently predictive of $\Delta T_{(n+1)-n}$ (**Fig 2c**; **Extended Data Fig 3**), in excess of optical artifacts and ongoing body movements. The larger the DA transient amplitude, the larger the timing update on the next trial (highest DA: largest early-shifts; lowest DA: largest late-shifts).

The amplitude of the DA transient was strongly correlated with trial outcome (which was also a strong predictor of $\Delta T_{(n+1)-n}$). Although this would not preclude a causal role for DA, we tested if additional information was present in DA by fitting separate models for early and rewarded trials. Both early and rewarded DA transient amplitudes showed considerable trial-by-trial variability, which had additional predictive power beyond correlation with trial outcome (**Fig 2d**, **Extended Data Fig 3**). These effects were also not explained away by a tendency to revert toward the median movement time (DA coefficient significantly larger than expected by chance in shuffled trial order model (p=7.2e-9, Wilcoxon rank sum test; **Fig 2e**; **Extended Data Fig 4**).

Outcome-related DA was thus positioned to make "one-shot" learning updates to timing. We use the machine learning term "one-shot learning" here because a *single exposure* to an endogenous DA transient was associated with a quantitative update to behavior. To our knowledge, DA learning effects based on variability in endogenous DA transients have not previously been observed.

DA causes one-shot timing updates

To test if outcome-related DA could cause one-shot timing updates, we virally expressed channelrhodopsin (ChR2) in genetically defined DANs for optogenetic manipulations (n=4 DAT-cre mice, **Fig 3a**). To modulate outcome-related DAN signaling, we triggered stimulation in the SNc at the time of the first-lick on 30% of randomly interleaved trials (for inhibition: see **Supplemental Note 1**). Stimulation was calibrated[7,30] outside the context of the task to recapitulate endogenous rewarded DA transients (**Fig 3b**), which matched the amplitude and width of unstimulated rewarded transients during the task (**Fig 3c; Extended Data Fig 5a**).

Stimulation on rewarded trials slightly increased DA transient amplitude, but on average, this remained within the limits of unstimulated reward transients on single trials measured during the same session (**Extended Data Fig 5c**). Control stimulations in mice not expressing opsin indicated that rdLight1 recordings were not contaminated by blue light-induced photoswitching[45] artifacts ("No Opsin" mice, **Fig 3d; Extended Data Fig 5d**).

DAN activation caused mice to move significantly earlier than expected on the next trial, with 4/4 mice showing early-shifting (**Fig 3e-f**; see **Methods**; mean $\Delta\Delta T_{(n+1)-n}$: -88ms, p=0.0019, 10,000x bootstrap test, n=4 mice, 40 sessions). By contrast, stimulation in No Opsin mice had no significant effect on $\Delta T_{(n+1)-n}$ (mean $\Delta\Delta T_{(n+1)-n}$: -6ms, p=0.40, 10,000x bootstrap test, n=4 mice, 40 sessions). These results were consistent across 3 additional complementary statistical analyses (**Extended Data Fig 6**).

Activation also caused subtle early-shifting of the *entire* timing distribution across the session compared to interleaved sham sessions (**Fig 3g**; p<0.025, 1000x bootstrap test; **Extended Data**

Fig 5b,e). This raised the possibility that outcome-related DA may normally control the "set-point" of the movement timing distribution, which we operationally define as the median first-lick time.

DA explains the timing set-point

Surprisingly, across timing paradigms, both rodents and non-human primates frequently tend to move too early, at the expense of reward[7,34-36,46] (**Extended Data Fig 7a**). In our mice, this early-bias developed over days of experience (**Extended Data Fig 1**) and persisted even in expert mice with >70 sessions' experience.[7] Outcome-related DA also appeared highly skewed in this task, with disproportionately large reward-related transients compared to small unrewarded dips (**Extended Data Fig 7b**). We observed that high amplitude (rewarded) DA transients caused much larger early-shifts compared to relatively small late-shifts after DA dips, and overall, mice made larger updates to timing after rewarded trials compared to unrewarded trials (median –317ms vs 74ms; **Fig 1g**). Together, this suggests bias in DA signaling may explain early-biased timing behavior.

But behavior was not early skewed for the entire session—all mice showed marked timing changes *within* each session, initially moving too early, but progressively shifting to later times (**Fig 4a-b; Extended Data Fig 7c**). If allowed to work *ad libitum*, the timing set-point ultimately shifted into the reward window—suggestive of timing calibration. (Importantly, the sloshing model GLMs were robust to this non-stationarity, **Extended Data Fig 4c**).

The Sloshing Model suggested that changes in outcome-related DA could explain changes in the timing set-point. Across each session, as the set-point shifted later, 1) rewarded trials were followed by progressively smaller timing updates (median timing change (in excess of chance), **beginning half of session: -400ms**, 95% CI: [-519ms, -270ms]; **ending: -271ms**, 95% CI: [-403ms, -143ms]; n=17 mice, 155 sessions; **Extended Data Fig 7d**); and, 2) rewarded DA transient amplitudes progressively declined (**Fig 4c; Extended Data Fig 7e-f, Extended Data Fig 8**; *for unrewarded transient effects: see* ***Supplemental Note 2***)**.** Importantly, this DA decline was not explained-away by bleaching or optical artifacts (**Extended Data Fig 7g**, **Supplemental Methods**).

Rewarded DA amplitude was tightly correlated with the timing set-point (median Pearson's r: –0.81; **Fig 4d**). This was not the case for movement-related/optical artifacts (median Pearson's r: +0.19, **Extended Data Fig 7h**). Strikingly, shifts were not always gradual—rewarded DA amplitude decreased suddenly during abrupt shifts of the set-point into the reward window (**Fig 4e**; R^2=0.65, p=0.0003), which was also not explained away by optical/movement related artifacts (R^2=0.19, p=0.10; **Extended Data Fig 7i**).

DA calibrates the propensity to move

Altogether, across sessions and recording sites/modalities, rewarded DA transients decreased in amplitude as timing behavior became more accurate. DA transients also controlled trial-by-trial timing updates, and these updates became progressively smaller as behavior became more accurate (**Extended Data Fig 7d**). Together, these findings suggest a calibration process for motor timing. This adaptive decrease in stochastic learning is reminiscent of adaptive stochastic gradient descent algorithms.[47] However, because rewarded DA and set-point tended to change monotonically across the session, we tested several complementary manipulations to "break the

monotonicity" between these co-variates. Each manipulation produced marked changes in the animals' timing strategy, including late-to-early timing shifts. Importantly, in all settings, rewarded DA transients explained bidirectional changes in the timing set-point, as follows:

Different reward identities. First, we serendipitously discovered that well-trained mice continued to participate even when juice rewards were replaced by the usual "correct" tone (**"No Juice" Task**; solenoid valve disconnected)—but set-point dynamics were strikingly different (**Fig 4f-g**; **Extended Data Fig 7j**). Mice still showed early bias for the first ~150 trials, but unexpectedly, the set-point suddenly and dramatically shifted into the reward window—significantly earlier in the session than in the +Juice task (p=0.001, Wilcoxon rank sum test; 5 mice, 18 paired sessions, **Fig 4h**). Mice continued to work for as many as 1000 trials under these conditions with correct timing on most trials, exhibiting much more "accurate" behavior than the same animals performing the +Juice task on interleaved sessions (same mice as **Figs 1b, 4b-d**).

Surprisingly, mice still showed "rewarded" DA transients in the No Juice task (in response to the correct tone, **Fig 4i**), but these were narrower and smaller in amplitude than reward transients in the +Juice task (**Fig 4j**). Moreover, reward-transient amplitude declined significantly earlier in the session during No Juice sessions (n=5 mice, 18 paired sessions; p=0.0073, Wilcoxon rank sum test, **Fig 4k, Extended Data Fig 7k-m**), and the timing of this abrupt, early decline explained the rapid shift in timing set-point (median Pearson's r: –0.78; **Fig 4l; Extended Data Fig 7o**). These effects were not explained by changes in ongoing body movements or optical artifacts (**Extended Data Fig 7l,n**).

Interestingly, the persistence of high amplitude rewarded transients in the +Juice task suggests that juice skewed normal timing updates: despite apparent motivation, mice received many fewer rewards/time compared to the No Juice task. The resulting early timing bias is reminiscent of the effects of DA-modulating drugs of abuse on timing behavior, such as amphetamines.[40]

Abrupt changes in reward identity. Second, we trained mice to alternate between blocks of trials with either No Juice rewards (correct tone only) or +Juice. During No Juice blocks, the timing set-point rapidly late-shifted into the reward window. This was rapidly reversed in +Juice blocks, in which the timing set-point rapidly shifted significantly earlier (n=3 mice, 5 block changes; p=0.032, Wilcoxon rank sum test; **Fig 5a**). Again, the amplitude of rewarded DA transients explained these set-point shifts (median Pearson's r: –0.89; **Fig 5b-c**).

The rapid changes in DA-transient amplitude might be correlated—but not causal—to abrupt set-point shifts. Therefore, we also had mice alternate between blocks of trials with either No Juice or No Juice+Stimulation, in which DANs were optogenetically activated at the time of first-licks that occurred either within the reward window or on all trials (n=3 mice, 20 block changes; **Extended Data Fig 9a-e**). Stimulation was calibrated to recapitulate usual juice-related DA transients without delivering liquid rewards (as in **Fig 3b**). The set-point rapidly late-shifted during No Juice blocks, but then rapidly and significantly early-shifted during +Stimulation blocks (off-on: p=0.001, on-off: p=0.007; Wilcoxon rank sum test; **Extended Data Fig 9b**), even when stimulation was only delivered on rewarded trials (off-on: p=0.009, on-off: p=0.002; n=2 mice, 12 block

changes, **Fig 5d**). Stimulation did not change the set-point in control mice expressing No Opsin (n=4 mice, 20 block changes; **Fig 5d; Extended Data Fig 9b-e**). Again, the amplitude of rewarded DA transients explained the set-point shifts (median stim: –0.66, no opsin: –0.56; **Fig 5e-f**, **Extended Data Fig 9d-f**). These results emphasize that rewarded DA transients induced *shifts* to earlier times—rather than reinforcing the rewarded movement time.

Rapid satiation. Third, in addition to sudden shifts, the timing set-point also showed gradual late-drift across sessions, potentially related to fatigue or satiation. The sloshing model suggested DA could also explain state/motivation-dependent shifts. We thus tested whether *rapid* satiation could change rewarded DA and set-point. After completing >=100 trials of the +Juice task and receiving >=10 rewards, the task was paused, and mice were allowed to drink *ad libitum* until they stopped licking available juice (n=3 mice, 15 sessions; range: 0.75-1.8mL (daily water consumption for a water-deprived mouse is ~1mL); **Fig 5g-h**). The task was then resumed. Rewarded DA and timing set-point remained highly anticorrelated throughout the session (median Pearson's r: –0.70, **Fig 5i**), but rewarded DA decreased rapidly after the pause (p=8.5e-5, Wilcoxon rank sum test; n=11 photometry sessions; **Extended Data Fig 9g**), and the timing set-point showed significant late-shifting (p=1.7e-5, Wilcoxon rank sum test; **Extended Data Fig 9h**). In contrast, in interleaved control sessions (empty syringe; no juice given during the pause), rewarded DA did not show this precipitous decrease (p=0.0026; **Fig 5h**), and there was no significant change in timing set-point (p=0.18; **Fig 5g**). Altogether, rapid juice consumption caused subsequent rewarded DA transients to rapidly decrease in amplitude, which was associated with late-shifting.

Changing timing constraints. Finally, fourth, to test if endogenous changes in rewarded DA transients could re-calibrate the timing set-point in response to different environmental timing contingencies, we changed the rewarded time window in blocks of trials ("Timeshift Task;" e.g., from 7-12s to 2-3.3s; n=7 mice; **Fig 5j**). The new reward window was not signaled and had to be discovered through trial-and-error. We hypothesized increases in rewarded DA would cause early-shifting to match the constraints of the task (and decreases, late-shifting). Because fatigue/satiety should increase monotonically over the course of the session, we reasoned early-shifts would disentangle goal-directed processes from fatigue/satiation.

The Timeshift Task produced pronounced shifts in the animals' timing behavior, including directed changes in timing in relation to the new reward window (**Fig 5j**, **Extended Data Fig 9i**) and exploratory behavior that appeared to anticipate block changes (**Extended Data Fig 9j**). The amplitude of rewarded DA transients continued to explain the timing set-point (median Pearson's r: –0.64; 7 mice, 54 sessions; **Fig 5k-l**), and intriguingly, *unrewarded* DA transients also became more correlated with the set-point (median r: –0.36; **Fig 5l**). Early-shifts were accompanied by increases in DA transient amplitude (and late-shifts by decreases), and these effects were not explained away by optical artifacts (tdt: Rewarded trials median r: –0.20; unrewarded r: 0.10; **Extended Data Fig 9k**).

Together, these results show that changes in reward-related DA cause shifts in the behavioral timing set-point, both in relation to changing environmental conditions (such as changes in

environmental timing contingencies for reward or the presence/absence of juice rewards) and internal state (such as gradual or rapid satiation). These findings demonstrate that DA quantitatively and adaptively calibrates the propensity to move in response to both external and internal constraints.

DISCUSSION

A cornerstone hypothesis of biological reinforcement learning is that reward-related DA transients (whatever their nature[3,5,16,19,48,49]) lead to changes in future behavior. However, to our knowledge, variability in *endogenous* reward-related DA transients has not been shown to relate to variability in subsequent behavior, despite serious effort to detect such effects.[3,30,50,51] Here, we show for the first time that the amplitude of native DA reward transients is quantitatively related to one-shot, goal-directed updates to movement timing on the *next* trial, >10s later. This suggests a functional link between reward and the calibration of the propensity to move (an aspect of vigor), unifying two seemingly distinct aspects of DA function.[6] In addition, we found that asymmetry in the relative amplitude of rewarded/unrewarded DA transients explained asymmetry in one-shot updates, and that changes in the relative amplitude of these transients could causally shift the timing set-point. This process, reminiscent of adaptive stochastic gradient descent,[52] allowed the animals to adjust their propensity to move in response to changing experimental contingencies (e.g., timing constraints). Changes in DA transients and behavior could also be caused by changes in internal/motivational state (reward identity, satiety/fatigue), demonstrating that flexibility in reward-related DA signaling mediates both strategic and motivationally driven changes to movement timing. Together, DA's tripartite roles in reinforcement, movement and motivation are mechanistically linked via this calibration process.

Relationship to reported rapid DA effects. Others have observed rapid, "real-time" effects of DA on movement vigor—but at much shorter latency than the reward-related effects we report here (within 100s of milliseconds before movement), and those studies specifically excluded reward-related DA/DAN transients from analyses.[23,25] We previously found that ongoing ramping DAN/DA dynamics *during* the self-timed interval (between the cue and the first-lick) predicted self-timed movement initiation (**Fig 1f, right**).[7] Moreover, physiologically calibrated optogenetic DAN activation throughout the self-timed interval could increase the moment-by-moment probability of initiating the self-timed movement (and inhibition, decrease it)[7]—a real-time DAN effect on movement timing. However, even before the trial started, differences in baseline DAN calcium signals were systematically related to the timing of the upcoming movement,[7] suggesting that baseline might reflect a putative "stochastic set-point" that modulates timing by controlling the start-point and slope of the ramping process.[53-55]

Here, we show that this variability in DAN baseline can be traced back *tens of seconds* to variability in the amplitude of the previous trial's outcome-related DA transient—and that variability in this transient can causally update timing on the next trial, 10-25s in the future. The effects of the outcome-related transient were in the same direction as prior real-time stimulation effects (higher DA, earlier movement); thus, parsimony would suggest both outcome-related and baseline/ramping DA exert the same mechanistic effect on timing. But, importantly, the results

presented here demonstrate that, in addition to real-time effects (sub-second to a few seconds), even transient DAN signaling causes effects on movement timing that persist over tens of seconds. This time-course has important mechanistic implications. In particular, the rapid, one-shot nature is difficult to square with conventional views of incremental DA-dependent synaptic plasticity.[56]

Relationship to reported gradual DA effects. Whereas we found one-shot effects of endogenous reward-related DA signals on the next-trial, other studies have found that *exogenous* DA manipulations can reinforce actions, but typically only *gradually*, over many exposures.[10,12,14,57] Such gradual effects could be viewed as consistent with incremental synaptic plasticity. In some cases, DA-mediated reinforcement has been specifically reported as imperceptible on the order of tens of consecutive trials.[2] One study found that optogenetic DAN inhibition throughout the previous trial of a reversal-learning task could affect choice probability on the subsequent trial, but this inhibition had no impact when limited to the rewarded period and did not affect the next trial's movement latency.[31] Other recent work showed an increased propensity to repeat behavioral syllables associated with spontaneous DA transients in freely behaving mice,[11] but this was outside a goal-directed context and did not explore the effects of variation in DA transients. Endogenous DA transient amplitude can also gradually decrease over the course of learning in a variety of other settings.[2,58,59] For example, in birds, endogenous DA signals related to song performance prediction errors have been associated with slow changes in syllable pitch over juvenile learning.[21,22] But again, these effects were gradual—and not linked to changes in movement vigor.[22]

Others have proposed that real-time DA effects on movement are instead the domain of tonic or "quasi-phasic[60]" DA signals rather than reward-related DA transients.[6,22,26,28] Here, we demonstrate that these two modes of DA signaling (phasic and tonic/quasi-phasic) are actually linked, potentially via a novel mechanism.

Potential mechanisms. Why could we detect one-shot learning effects of endogenous DA transients when previous studies did not? One possibility is that our time-to-move metric provided a more sensitive "analog" behavioral readout. Behavioral updates to binary behavioral metrics or in discrete reinforcement-learning tasks, such as bandit tasks, may be difficult to detect on a trial-by-trial basis,[3,50,57] instead requiring multiple repetitions before becoming manifest.

Alternatively, we suspect that one-shot effects on vigor arise via a fundamentally different biological mechanism than synaptic plasticity.[56] First, a common assumption in reinforcement-learning (RL) models is that specific stimuli or actions (or other states of an agent) are reinforced due to temporal association with reward-related DA transients.[56] Importantly, we did not find that reward reinforced a *specific* time-to-move; rather, DA transients induced a bidirectional *update* in timing on the next trial. For example, if an animal was rewarded for moving at 4s, it did not repeat that 4s movement on the next trial—it tended to move *earlier*. Second, RL models often assume incremental learning is mediated by gradual DA-dependent structural synaptic plasticity in cortico-striatal synapses.[56] In contrast, we found that reward-related DA transients changed timing on a trial-to-trial basis, and (in our previous study) DAN stimulation during the

timing interval affected movement time within only 100s of milliseconds to a few seconds.[7] This dynamism seems unlikely to arise via incremental DA-dependent synaptic plasticity.

What is the cellular/circuit mechanism for reward-related DA transient updates to vigor? Theoretical work has suggested frameworks for understanding DA's effects on vigor, but the biological implementations were not clear.[6,61] However, any proposed mechanism should have the following properties: 1) it should support *bidirectional* effects on timing; 2) it should have a *rapid onset* (sub-second to a few seconds); and 3) it should *persist* for at least tens of seconds (the duration of the ITI in our experiments). DA-dependent synaptic plasticity is likely too slow and incremental to meet these constraints—and would also seem metabolically inefficient to support back-and-forth, trial-by-trial updates.

By contrast, DA also modulates the excitability of striatal projection neurons (SPNs) via metabotropic receptors.[62,63] A salient example is a careful striatal slice study, which found that calibrated, transient optogenetic stimulation of DAN axons caused rapid (sub-second) yet long-lasting (minutes) increases in the firing rate of direct pathway dSPNs in response to depolarizing current injections.[63] This excitability effect was blocked by a specific inhibitor of Protein Kinase A (PKA). PKA is an intriguing candidate effector, because DA transients *in vivo* have been shown to drive increased PKA activity in dSPNs, whereas DA dips cause increased PKA activity in indirect pathway iSPNs, with a latency of a few seconds and persistence for tens of seconds,[64] similar to the dynamics of the behavioral effects we observe. It is possible that the relative balance of dSPN:iSPN excitability could scale dynamics of the cortico-basal ganglia-thalamic circuit[65,66] in a manner that allows for bidirectional changes in timing—and modulation of vigor more broadly. Outcome-related DA transients could drive these PKA changes. (The persistence of outcome-related DAN signals into the baseline of the next trial are suggestive of changes in tonic DA, which could also drive and/or contribute to these effects, although we are cautious to make arguments about tonic DA due to our dF/F-limited measurement window. This idea merits future investigation with techniques more suitable to measuring tonic dopamine, such as with quantitative fluorescence lifetime photometry measurements, which are still in development.[67]) This scaled-dynamics model provides a framework for understanding vigor-related symptoms of DA-related disorders of movement, such as bradykinesia and akinesia in Parkinson's disease.

As a final note, we are not arguing against DA's role in synaptic plasticity. On the contrary, rapid DA-mediated effects on striatal excitability could operate in parallel—or even be synergistic—with slower changes in synaptic strength. For example, DA-induced changes in excitability could modulate the probability of spike-timing-dependent synaptic plasticity.[56] Increased vigor could also enhance reinforcement learning simply by allowing an agent to explore its environment more briskly, resulting in more exposures to the reinforced action or association. When an animal encounters a reward-rich environment, the ensuing increase in reward-related DA transients would increase the propensity to move, which in turn could allow the animal more opportunity to discover specific causal structure in its environment, potentially through synaptic plasticity mechanisms. This interpretation could explain DA's association with exploratory behavior,[6] as well as the longstanding observation that animals increase the frequency and speed of all movements—not just instrumental, reward-associated movements—in reward-rich

environments.[6] Our findings here reveal a mechanistic link between reward and vigor that provides a platform for further dissecting the relationship between reinforcement, motivation and movement.

Supplemental Notes

Supplemental Note 1: Optogenetic inhibition

We also attempted to inhibit DANs (stGtACR2 or JAWS), but despite serious effort (5 mice), we could not identify stimulation parameters that would prevent reward-related DA transients, despite finding parameters that caused DA dips outside the context of the task. Furthermore, more aggressive inhibition at higher light powers (e.g., 5-30 mW) in the context of the task—but not outside the context of the task—caused *seconds-long* rebounds in baseline DA after the offset of inhibition (i.e., this was not the typical transient "rebound spike" sometimes seen after inhibition, but a more sustained increase in DA). Thus—paradoxically—aggressive inhibition tended to *raise* DA overall on stimulated trials. We suspect inhibition was ill-posed in this preparation: we could not detect the lick and initiate stimulation fast enough to prevent reward-related action potentials from triggering DA release, and more powerful inhibition caused paradoxical rebounds. Others have also been unable to effectively block reward-triggered DA with inhibition and have proposed similar rationales (e.g., Ref [31]).

Because we could not prevent first-lick related transients, we did not proceed with the Inhibition cohort. Instead, in lieu of optogenetic inhibition experiments, we caused reductions to outcome-related DA through non-optogenetic manipulations (**Fig 4f-l, Fig 5**), whose effects were consistent with the sloshing model.

Supplemental Note 2: Unrewarded DA transient changes across the session (+Juice Task)

By contrast to rewarded DA transients, unrewarded transients showed only a slight decrease in amplitude in the usual +Juice Task (**Extended Data Fig 7e**), and unrewarded timing updates showed relatively little change (**beginning: +64ms**, 95% CI: [+22ms, +103ms]; **end: +94ms**, 95% CI: [+35ms, +153ms]). Unrewarded transients did however show set-point related changes in the Timeshift Task (**Fig 5k**).

METHODS

Animals

We used adult male and female hemizygous DAT-cre mice[68] (B6.SJL-*Slc6a3*$^{tm1.1(cre)Bkmm}$/J, RRID:IMSR_JAX:020080; The Jackson Laboratory, ME) or wild-type C57BL/6 mice (>2mo old). Mice were housed in a temperature and humidity-controlled facility on a reversed night/day cycle (12h dark/12h light). Behavioral sessions occurred during the dark cycle. All protocols were approved by the Harvard Institutional Animal Care and Use Committee and MIT Committee on Animal Care (Harvard IACUC protocol #1343, Animal Welfare Assurance Number #A3431-01; MIT CAC protocol #2306000543). All work was in accordance with the NIH Guide for the Care and Use of Laboratory Animals. **Surgeries** were conducted under aseptic conditions with every effort to minimize suffering, as we described previously.[7] Mice were **water deprived** as we described previously[7] (≥80% initial body weight) with *ad libitum* feeding. **Histology** to confirm viral expression and fiber placement was conducted as we reported previously.[7]

Stereotaxic coordinates (from bregma and brain surface)

Viral Injection:

- SNc[7]: 3.16mm posterior, +/-1.4mm lateral, 4.2mm ventral
- VTA[7]: 3.1mm posterior, +/-0.6mm lateral, 4.2mm ventral
- DLS[7]: 0mm anterior, +/-2.6mm lateral, 2.5mm ventral.
- VLS[43]: 0.34mm anterior, +/- 2.75mm lateral, 3.7mm ventral (rdLight1/dLight1.3), 3.2mm ventral (D1/A2A-cre; GCaMP)

Fiber Optic Tips (ventral position): 200µm 0.53NA or 0.55NA blunt fiber optic cannulae (Doric Lenses, Quebec, Canada) or tapered fiber optic cannulae (200 µm, 0.60 NA, 1.5, 2 or 2.5mm tapered shank, OptogeniX, Lecce, Italy) were positioned at SNc, VTA, DLS or VLS:

- SNc/VTA[7]: 4.0mm (photometry) or 3.9mm (optogenetics).
- DLS[7]: 2.311mm (blunt) or 4.0mm (tapered)
- VLS (rdLight1/dLight1.3): 3.4mm (blunt)
- VLS[43] (D1/A2A-cre; GCaMP): 3.0mm (blunt or midpoint of 2.5mm tapered)

Viruses

Photometry:

- tdTomato ("tdt"): AAV1-CAG-FLEX-tdT ($5.3*10^{10}$ to $5.3*10^{12}$gc/mL; UNC Vector Core, NC), 500nL alone or mixed with other fluorophores (below).
- mCherry (D1/A2A-cre experiments): AAV8-Syn-DIO-mCherry ($2*10^{12}$gc/mL; Datta Lab), mixed with GCaMP6f.
- gCaMP6f (at SNc or VTA): 300nL AAV1.Syn.Flex.GCaMP6f.WPRE.SV40 ($2.5*10^{13}$gc/mL; Penn Vector Core, PA) mixed 3:1 with tdt or mCherry.
- DA_{2m} (at DLS): 200-300nL AAV9-hSyn-DA4.4(DA2m) ($3*10^{12}$gc/mL; Vigene, MD) mixed 2:1 or 3:1 with tdt
- dLight1.1 (at DLS): 300nL AAV9.hSyn.dLight1.1.wPRE ($9.6*10^{12}$gc/mL, Children's Hospital Boston, MA) mixed 3:1 with AAV1.CB7.CI.TurboRFP.WPRE.rBG (10^{12}gc/mL, Penn Vector Core).

513 rdLight1 (at VLS): 500nL AAV2/9-Syn-RdLight1.0 ($2*10^{12}$gc/mL; #construct-1646-aav2-9,
514 Canadian Neurophotonics Platform (CNP), Quebec, Canada).
515 dLight1.3 (at VLS): 500nL AAV2/DJ-CAG-dLight1.3b ($2*10^{12}$gc/mL, CNP)
516 Optogenetic stimulation/inhibition (bilateral at SNc):
517 ChR2: 1000nL AAV5-EF1a-DIO-hChR2(H134R)-EYFP-WPRE-pA ($3.2*10^{13}$gc/mL, UNC
518 Vector Core, NC)
519 stGtACR2: 300nL AAV2/8-hSyn1-SIO-stGtACR2-FusionRed ($4.7*10^{11}$gc/mL,
520 Addgene/Janelia Viral Core, VA)
521
522
523 **Behavior**
524 Neurophysiology and behavioral data was acquired as we described previously.[7]
525
526 **Self-timed movement task (original version):** Mice were head-fixed with a juice tube positioned
527 in front of the tongue. The spout was placed ~1-1.5mm ventral and ~1-1.5mm anterior to the
528 mouth. During periods when rewards were not available, a houselamp was illuminated. At trial
529 start, the houselamp turned off, and a random delay ensued (0.4-1.5s) before a cue (simultaneous
530 LED flash and 3300Hz tone, 100ms) indicated start of the timing interval. The timing interval was
531 divided into two windows, early (0-3.333s) and reward (3.333-7s), followed by the intertrial
532 interval (ITI, 7-17s). The window in which the mouse first licked determined the trial outcome
533 (early, reward, or no-lick). An early first-lick caused an error tone (440Hz, 200ms) and houselamp
534 illumination, and the mouse had to wait until the full trial interval had elapsed before beginning
535 the ITI. Thus, there was no advantage of licking early. A first-lick during the reward window caused
536 a reward tone (5050Hz, 200ms) and juice delivery (~4µL apple juice or Gatorade; **+Juice Task**),
537 and the houselamp remained off until the end of the trial. If the timing interval elapsed with no
538 lick, a time-out error tone played (131Hz, 2s), the houselamp turned on, and ITI commenced.
539 During the ITI and pre-cue delay ("lamp-off interval"), there was no penalty or feedback for licking.
540 Mice learned the task in 3 stages, as described previously.[7] All mice learned the task and worked
541 for ~400-1,800 trials/session.
542
543 **Pre-training/post-training.** To encourage participation, each session (except for SNc/VTA/DLS
544 GCaMP mice) commenced with 10 trials of "pre-training" during which juice was delivered
545 triggered on the first-lick within the reward window, but during which early licks were not
546 penalized. Pretraining was always conducted with juice rewards.
547
548 **One-shot Stimulation (+Juice Task, Fig 3)**. Closed-loop stimulation was delivered at SNc and
549 triggered on first-lick on 30% of randomly interleaved trials in the context of the +Juice Task. Mice
550 were trained without stimulation for 13-16 sessions before starting stimulation sessions. Each
551 mouse completed 10 pairs (20 sessions) of interleaved stimulation and sham sessions, where
552 sham sessions were identical to stim sessions (patch cables attached, etc.), except the stimulation
553 LED was turned off and no light was delivered. Photometry was usually not collected during sham
554 sessions to reduce bleaching from recording on consecutive days. See **Optogenetics** section for
555 stimulation parameters, calibration and analysis.
556

No Juice Task (Fig 4f-l). In the usual (+Juice) task, feedback was either an error tone (440Hz, 200ms) or correct tone (5050 Hz, 200 ms) plus juice reward. The No Juice Task was identical to the Juice+ Task and used the same feedback tones, but no juice was delivered (dispensing solenoid valve disconnected). In a subset of animals trained on an open rig with a loud juice solenoid, a decoy solenoid was triggered, but no juice was dispensed. After >=35 days of +Juice Task, expert mice successfully performed the No Juice Task, some for >1000 trials/session (one mouse tested on days 23 and 25 showed less engagement). Anecdotally, we observed that consecutive No Juice sessions could reduce participation in subsequent sessions in some animals, and so No Juice sessions were typically interleaved with +Juice sessions to promote participation. Together, these factors limited the number of No Juice sessions that we could collect.

Alternating Juice/No Juice Task (Fig 5a-c). Mice alternated between unsignaled blocks of usual +Juice or No Juice Task.

Alternating Stimulation Blocks (Fig 5d-f). Mice were trained in a No Juice Task variant in which closed-loop SNc DAN stimulation was delivered in blocks of trials. During stimulation blocks, stimulation was delivered on either 30%, 60% or 100% of first-licks on all trials (separate blocks, **Extended Data Fig 9a**) *or* only on 100% of rewarded first-licks (**Fig 5d**). See **Optogenetics** section for stimulation parameters, calibration and analysis.

Rapid Satiation Task (Fig 5g-i). At >20d of expert task training, Rapid Satiation sessions were collected, where the ongoing task was paused once behavior met two criteria: ≥100 trials completed and ≥10 juice rewards received. Video and photometry recording were continued during the pause (mice remained headfixed on the behavior rig). The spout was retracted, and the mice were hand-fed undiluted Cool Blue Gatorade from a 1mL syringe until they stopped initiating licks to the offered syringe. The spout was then moved back into position and the task resumed. In Control sessions, the same procedure was followed, except an empty syringe was offered.

Both Rapid Satiation and Control sessions were collected on a background of the +Juice task with undiluted Cool Blue Gatorade rewards. We used a newer version of the +Juice task, in which there was also an unsignaled late window after the reward window (7-14s). In this version of the task, the trial duration was always 14s, with the late window occurring between the unsignaled end of the reward window and the start of a 3s-long ITI (total trial+ITI duration 17s, as in the original +Juice Task). First-licks during the late window triggered a distinct, late error tone (132Hz, 500ms) and illumination of the houselamp (similar to early trials). Mice still had to wait until 14s to start the ITI.

In analyses, to account for missing time while the task was paused, the total duration of the pause was divided by 17 (the trial length was 17s) and rounded to recover the number of missing trials. Rewarded DA transient amplitudes ("rewDA") were then interpolated linearly and smoothed with a 20-trial gaussian kernel. Edges (trials before the first reward and after the last reward in the session) were omitted.

Rapid Satiation sessions were collected in pairs with interleaved control sessions. However, some sessions had to be excluded either because the mouse received no rewards before the pause (1 session), the mouse stopped participating after the pause (1 session), or there were issues with lick detection due to a broken spout (3 sessions). Although including these sessions did not change the result of analyses, they were omitted from the final rewDA analysis. Statistical comparisons of rewDA before and after the pause were performed on the average rewDA signal across 50 trials before and after the maximum duration pause (36 trials) and compared by Wilcoxon rank sum test.

Timeshift Task (Fig 5j-l). In the usual +Juice and No Juice tasks, there was only one rewarded time window (3.333-7s). In the Timeshift Task, this window changed dynamically in blocks of trials. The Timeshift Task was introduced after >28 days of training on the usual +Juice Task. Each block used one of five reward windows: a) 1-2s, b) 2-3.333s, c) 3.333-7s, d) 7-12s, e) 9-14s. There was no set schedule for changing between blocks. Block changes were signaled by a ~90s timeout (task paused). The new reward window was not indicated and had to be discovered through trial-and-error. In a subset of sessions, a "late" window was added (as in the Rapid Satiation Task) such that mice were not immediately alerted when the reward window had closed. Behavior was grossly similar on both versions of the Timeshift Task, and thus, for all analyses, data from the standard Timeshift Task and late window Timeshift Task was pooled. Mice were also trained on a No Juice version in some Timeshift sessions (7 mice, 9 sessions), and these data were pooled with +Juice Timeshift sessions.

Photometry

Signal collection. Fiber optics were illuminated with 475nm blue LED light (Plexon, TX) (SNc/VTA: 50µW, DLS: 35µW, VLS: 30µW) measured at patch cable tip with a light-power meter (Thorlabs, NJ). Green fluorescence was collected via a custom dichroic mirror (Doric Lenses, Quebec, Canada) and detected with a Newport 1401 Photodiode (Newport Corporation, CA) or integrated Doric cube. For red florescence, 550nm lime LEDs were used (Plexon, TX). Fluorescence was allowed to recover ≥1d between recording sessions. To avoid optical crosstalk, the red control channel was recorded at one of the other implanted sites (SNc, VTA, DLS or VLS, 550nm lime LED, Plexon, TX) while GCaMP6f, dLight1.1/1.3/rdLight1 or DA_{2m} were recorded simultaneously only at the other implanted sites.

dF/F. Raw fluorescence for each session was pre-processed by removing rare singularities (single points >15 STD from the mean) by interpolation to obtain F(t). To correct photometry signals for bleaching, dF/F was calculated as:

$$\frac{dF}{F}(t) = \frac{F(t) - F_0(t)}{F_0(t)}$$

where $F_0(t)$ is the mean fluorescence during the 5s preceding the current trial, as well as 5 trials before the current trial and 5 trials after the current trial (as we characterized and validated previously[7]).

Correlation of baseline photometry signals with first-lick time. Baseline signals were de-noised by pooling based on first-lick time into 0.25s pools from 0.75 to 4s after the cue, as in **Fig 1d**. Pools with first-lick times after 4s were not included due to insufficient numbers of trials in each pool to de-noise the measurement (inclusion criteria >=175 trials). Pearson's r was calculated from each pool's average dF/F signal calculated within windows of time relative to lamp-off (left of **Fig 1d**, r: –0.63; –10.5 to 10s before lamp-off) or the cue (right of **Fig 1d**, r: –0.79; –0.4 to 0s before the cue (excludes lamp-off event)). These signals were related to the earliest first-lick time in each pool.

Sloshing Model. To test the independent contribution of each task-related input to the photometry signal and select the best model, we employed nested fitting, in which each dataset was fit multiple times (in "nests"), with models becoming progressively more complex in subsequent nests. The nests fit to the photometry data employed the inputs $X^{(j)}$ at each jth nest:

Null Model:	$X^{(0)} = x_0$
Nest 1:	$X^{(1)} = X^{(0)}$ + EMG and/or accelerometer
Nest 2:	$X^{(2)} = X^{(1)}$ + red control fluorophore ("tdt")
Nest 3:	$X^{(3)} = X^{(2)}$ + outcome-related photometry signal

The outcome-related photometry signal was taken as the mean dF/F signal within the first 500ms after the first-lick (peak or median dF/F produced similar results). All d_j design matrix predictors were 0,1 normalized.

The GLM for each nest took the form:

$$Y = \Theta X^{(j)}$$

Where Y is the *1*x*N* vector of the change in the movement time between the current trial, n, and the next trial, n+1 ($\Delta T_{n-(n+1)}$), where *N* is the total number of trials in the session; $X^{(j)}$ is the *d*x*N* design matrix for nest *j*, where the rows correspond to the d_j predictors for nest *j* and the columns correspond to each of the *N* trials of Y; and Θ is the *d*x*1* vector of fit weights. Because of day-to-day/mouse-to-mouse variation (ascribable to many possible sources, *e.g.*, different neural subpopulations, expression levels, behavioral states, *etc*.), each session was fit separately.
To guard against a trivial relationship between the trial outcome and DA signal, separate models were fit for early or rewarded trials.

Among the nests, the best model for each session was selected by BIC (AIC and AICc were typically in agreement with BIC), which was typically the model including the photometry signal. All reported data is for the full model.

Residual sloshing models (Extended Data Fig 4a). To determine whether outcome-related DA was predictive of ΔT in excess of chance (regression to the median effects), the next trial's movement time was shuffled, and the sloshing model was refit on the shuffled data (10,000 shuffled datasets). The coefficients from the 50th percentile shuffled model (50th percentile

shuffled DA coefficient) were used to predict the true data (y-fit shuffle) and calculate shuffled-R^2. Y-fit shuffle was subtracted from the true dataset to obtain a residual dataset, and the residual model was derived by refitting the sloshing model on the residual dataset.

Shuffled models/regression to the median (Fig 2e, Extended Data Fig 4b-c). To guard against spurious relationships between DA and ΔT arising from behavioral non-stationarity, each session was divided into four quartiles of trials (same number of trials with participation in each quartile) to isolate more stationary periods of behavior. The sloshing model was fit on each quartile (including movement/optical artifact control signals and mean DA photometry signal during the first 500ms after the first-lick). The trial order (for each quartile) was then shuffled 10,000x and the sloshing model fit on each shuffled dataset to produce the distribution of DA coefficients for each quartile assuming no trial-order dependence. To determine if the true fit coefficient was in excess of the shuffle, a Wilcoxon rank sum test was applied to compare the set of true coefficients across quartiles (and sessions) to the set of mean shuffled DA coefficients.

Optogenetics

Opsin expression. 3 cohorts of mice were tested: Activation (ChR2, 4 mice), Inhibition (stGtACR2, 4 mice or JAWS, 1 mouse), and No Opsin (no SNc injection, 4 mice). All mice co-expressed rdLight1 at VLS for co-monitoring of stimulation effects on DA release. All stimulations were delivered via fiber optic cannulae implanted over SNc.

Co-monitored photometry. We selected ChR2 (blue-activated opsin)/rdLight1 (green-excited fluorophore) as our DA co-monitoring strategy because we had previously observed issues with off-target excitation of red light activated opsins (e.g., ChrimsonR) during prolonged co-recording of dLight1.1 (a blue light-excited fluorophore) in DLS (red opsins are also activated by blue light, albeit with lower efficiency, which becomes significant during hours-long timecourses of photometry illumination in our behavioral sessions).[7] This off-target excitation would have created a confound, because we had previously shown that DAN activation limited to the self-timing interval increases the probability of early first-licks.[7] ChR2, by contrast, is not appreciably activated by green light (used for rdLight1 excitation), and was thus a cleaner preparation. One challenge of blue stimulation/green-excited photometry is that stimulation light causes a DC-offset artifact in the red photometry channel while the stimulation LED is turned on. To avoid crosstalk between the stimulation LED and the photometry recording, two complimentary correction methods were used, with similar results: 1) the brief stimulation up-times were omitted from the photometry signal (missing points filled by interpolation), or 2) the signal was decontaminated by measuring the instantaneous increase in rdLight1 signal at the beginning of each pulse of blue light, which was subtracted from the voltage trace during each pulse. Both corrections were taken before dF/F correction and yielded similar photometry signals, however the second method had the advantage of compatibility with continuous illuminations used during stGtACR2 inhibition experiments (see **Stimulation Parameter Calibration**, below).

Stimulation parameter calibration. Activation parameters were calibrated outside the context of the task. Mice sat on the typical behavioral platform and received 10 unexpected juice rewards

(ca. 4µL each), spaced at least 15s apart. A grid screen of stimulation parameters and light powers identified conditions that best recapitulated both the amplitude and width of the typical reward-related transient. Mice received 10 stimulations with each stimulation protocol, and DA release in VLS was monitored throughout the calibration procedure with rdLight1. Light intensity and pulse width were varied until the resulting rdLight transient matched the amplitude and width typically seen with unexpected juice rewards.

The final parameters used in all experiments were as follows: Optogenetic stimulation was delivered bilaterally via blunt fiber optics at SNc (5mW 465nm LED, 10Hz, 250ms, 35% duty cycle). One ChR2 animal had partial response to 5mW blue light, and thus received 10mW blue light, with a larger but still partial effect (Mouse X4-3-1M). We could not identify a set of inhibition parameters that prevented first-lick related transients in the task and did not proceed with the Inhibition cohort (**Supplemental Note 1**).

One-shot stimulation analyses. Initial power simulations indicated that ~5,000 trials (10x 500 trial sessions) would be necessary to achieve 80% power to detect a 20% change in first-lick time. Thus, at least 10 sessions of stimulation (~500 trials each) were collected from each animal.

Three complementary analyses examined single-trial stimulation effects (**Extended Data Fig 6**):

1. **"deldel,"** difference in the timing update (del1) between stimulated vs unstimulated trials (del2):

$$\Delta(\Delta T_{(n+1)-n})_{stim\text{-}unstim} = \Delta T_{(n+1)-n \mid \text{trial n stimulated}} - \Delta T_{(n+1)-n \mid \text{trial n unstimulated}}$$

2. **"dAUC,"** difference in the area under the cdf of the next trial (n+1) timing distribution for stimulated vs unstimulated trials (as described previously[7]).
3. **"cdf-rank,"** percentage of points in the n+1th trial cdf after stimulated trials that were less than matching points in the n+1th trial cdf after unstimulated trials. cdfs were interpolated such that each point was matched between the stimulated and unstimulated curves.

Analyses were performed at 3 levels, with consistent results across analyses and levels:

<u>By Session</u>

Within single session analyses: deldel, dAUC and cdf-rank analyses compared the distributions or first-lick times on the next trial after stimulated vs unstimulated trials. As expected from preliminary power calculations, these statistics were noisy (**Extended Data Fig 6 top row, single dots**).

Between groups single session comparisons: The test statistic for each single session was bootstrapped 10,000x, and the mean of the single session test statistics for each group were subtracted to derive a between groups comparison test statistic (activation – no opsin), which was then compared by Wilcoxon rank sum test (**Extended Data Fig 6 top row, between group statistic**).

By Mouse

Within "Mega" session comparisons: Because initial power simulations suggested 5000 trials were needed for 80% power to detect a 20% change in first-lick time, data were pooled across stimulated sessions from individual animals to generate "mega" sessions, which had ~5000 trials each. The test statistic for each mega session was reported as dots in **Extended Data Fig 6**, **second row**.

Between groups mega session comparisons: The test statistics for each group of animals' mega sessions were compared by Wilcoxon rank sum test. 2-sided unpaired t-test found similar results.

All Mice

Within "super mega session" comparisons: All data from all sessions in each cohort were pooled to generate a "super mega session" to compare all stimulated and unstimulated trials by 10,000x bootstrap procedure (same as used on within-single session analyses). Test statistic and p-value reported in the Main Text and **Extended Data Fig 6, third row**.

Whole session behavioral analyses

"Set point" calculations: Running median first-lick time. For visualization, the timing set-point was defined as the 50-trial running median centered at the current trial, excluding missing trials (e.g., no-lick) and reactions (first-lick <0.5s). This captured the peak of the timing distribution (whereas including reactions tended to underestimate the peak). For quantitative comparison to rewarded DA transients (which were less abundant) and in tasks with rapidly changing set-point dynamics, the number of trials included in the running median was adaptively reduced to avoid over-smoothing (10-50 trials used, depending on analysis: e.g., Timeshift No Juice task used 15 trials). Data provenance text files are included with each figure and specify the smoothing used in each analysis.

+Juice/No Juice Task Comparisons. For animals with No Juice Task sessions (n=5 VLS rdLight mice, 11 sessions), the nearest +Juice session with photometry was matched for comparisons of behavior (timing set-point) and outcome-related DA signals. For set-point analysis, 15 sessions were available for comparison (e.g., **Figs 1b, 4b-d**). Some flanking sessions were collected without photometry (we allow signals to recover from bleaching for at least 1d between recording sessions)—in these situations, the most recent +Juice photometry session was used for DA analyses. Together, this produced 18 paired photometry sessions for No Juice vs Juice+ comparisons (**Fig 4i-k**; n=5 mice, 9 sessions each task). Except for one session (Randy_VLS_47), paired +Juice photometry sessions happened before the paired No Juice session. Because the +Juice sessions mostly had 500 trials total, for consistency, we used only the first 500 trials from each No Juice session for comparisons, although some No Juice sessions were recorded for >1000 trials. The Q1 and Q4 lick-triggered averages shown in **Fig 4** were thus taken on trials 1-125 and 375-500, respectively, irrespective of session duration.

Set-point cross-over analysis (Fig 4e). Sudden shifts of the set-point into the reward window were eventually observed in all animals when allowed to work ad libitum, typically between trials 400-1100. However, because many sessions were only recorded for 500 trials, only a subset of sessions were recorded long enough to observe cross-overs (we validated this in a large behavioral dataset from animals allowed to work until disengagement, typically between trial 1100-1800, 3 mice, 120 sessions). Sessions with set-point cross-over were identified in the SNc GCaMP or VLS rdlight1 datasets as those sessions with set-point non-stationarity of >=+1.5s (16 sessions total). If the set-point started within the reward window at the beginning of the session, the set-point had to first exit and then re-enter the reward window (1/16 sessions). Evolution in first-lick time and rewarded DA transient amplitude ("rewDA") was taken as a running median over 75 trials (37-38 either side of each trial), where missing data was interpolated linearly (edges excluded as above to avoid artifacts). To de-noise the derivative of the rewDA signal, rewDA was first downsampled (every 15 trials taken). The downsampled trial with maximum derivative of the DA signal (–1*rewDA') was compared to the "set-point cross-over trial," defined as the trial on which the median of the timing distribution first crossed into the reward window (>3.333s). The cross-over trial number and –rewDA' were 0,1-normalized.

Code Availability. All custom behavioral software and analysis tools are available with sample datasets at https://github.com/harvardschoolofmouse.

Data Availability. All datasets supporting the findings of this study will be made publicly available at the time of publication via DANDI, with source data files provided for all figures.

Acknowledgements

We thank L. Hou, E. Sayed, J.G. Mikhael, and S. Hrvatin for discussions; M. Yang, S.L. Jacobs-Skolik, N. Hunter, V. Berezovskii, J. LeBlanc, T. LaFratta, O. Mazor, and P. Gorelik for technical assistance; and B. Sabatini, C. Harvey, S.R. Datta, M. Andermann, A. Mohebi and W. Regehr for reagents. The work was supported by NIH grants UF-NS108177 and U19 NS113201; NIH core grant EY-12196; the Valhalla Foundation and the Whitehead Fellows Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

A.E.H. and J.A.A. conceived the study, designed the experiments, acquired funding, supervised the project, and wrote and edited the manuscript. **A.E.H.** Performed experiments, wrote the code for analysis, analyzed the data, and wrote the first draft. **I.C.W., Q.D., Z.O. and P.A.** assisted with experimental design, performed experiments, processed data, and assisted with analysis. All authors reviewed the final manuscript.

Author ORCIDs

Allison E Hamilos https://orcid.org/0000-0001-9486-0017
Isabella C Wijsman https://orcid.org/0009-0000-5107-7823
Qinxin Ding https://orcid.org/0009-0009-6571-2678
Pichamon Assawaphadungsit https://orcid.org/0009-0003-0343-5368
Zeynep Ozcan https://orcid.org/0000-0002-8311-1958
John A Assad https://orcid.org/0000-0002-7689-5336

Competing Interests Declaration

J.A.A. is a co-founder of OptogeniX, which produces the tapered optical fibers used in some experiments.

Additional Information

Corresponding Author: Allison E. Hamilos and John A. Assad

Supplementary information:

Extended Data Figures
Supplemental Notes 1-2
Supplemental Methods

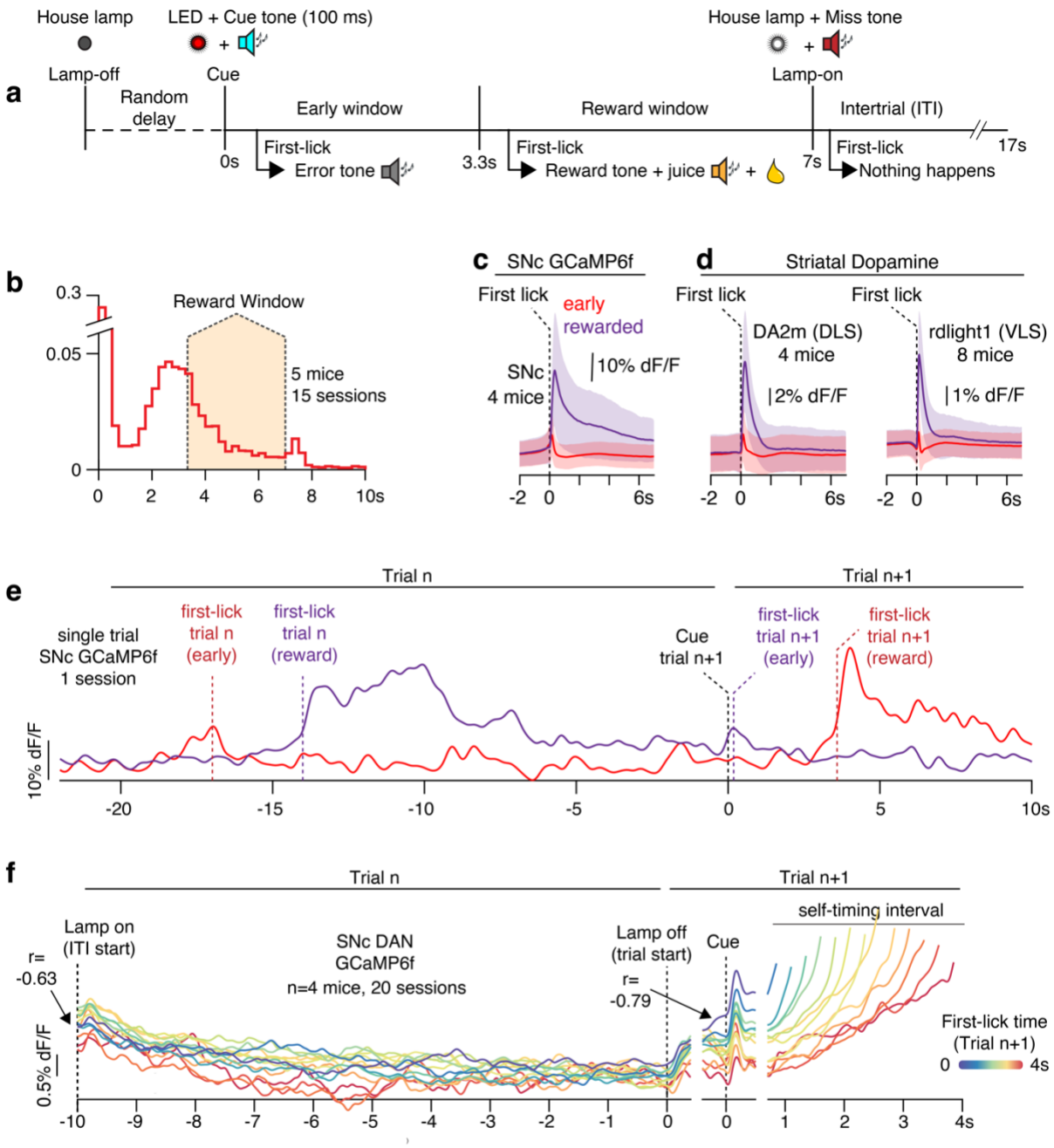


Fig 1. Outcome-related DA is related to timing on the next trial. a, Task. **b,** Whole-session timing distribution, same 5 mice as in **Fig 4**. **c-d,** Outcome-related signals; shading: std. **e-f,** Timing-related DAN signals: **e,** Single-trial. **f,** DAN signals pooled by movement time and plotted up to 150ms before spout contact. X-axis breaks indicate change of alignment. **g,** Change in timing set-point on next trial (ΔT); shading: 95% CI (10,000x bootstrap).

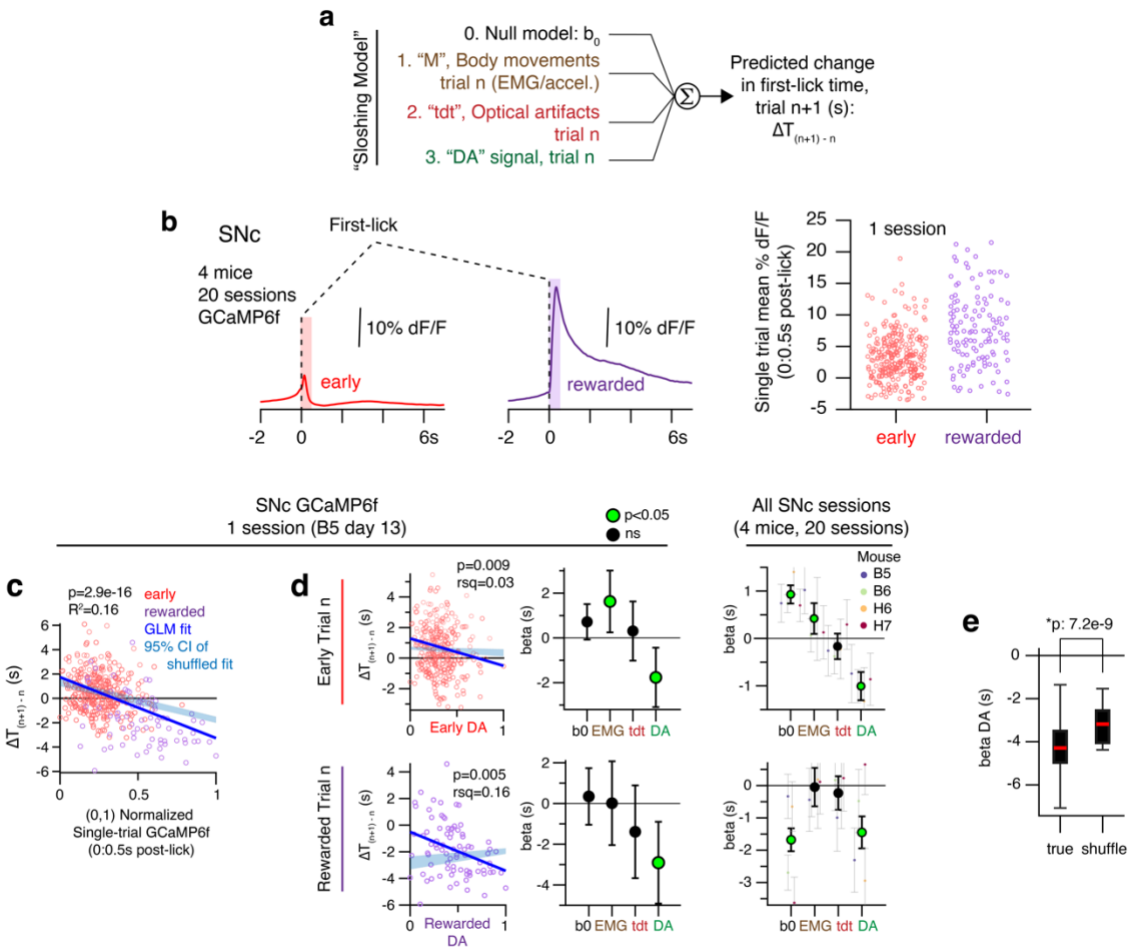


Fig 2. Outcome-related DAN signals predict changes in timing on the next trial. a, Sloshing GLM. **b,** Trial-by-trial variability in outcome-related DAN signals. Left: average. Right: single-trial mean DAN signals within 0.5s post-first-lick, jittered for clarity. **c-d,** Sloshing Model fit. Coefficient error bars: 95% CI. **e,** Model DA coefficients, true data vs 10,000x shuffle fit, all sessions. Standard box plot (red line: median). *p: Wilcoxon rank sum test. Models for all signals: **Extended Data Figs 3-4**.

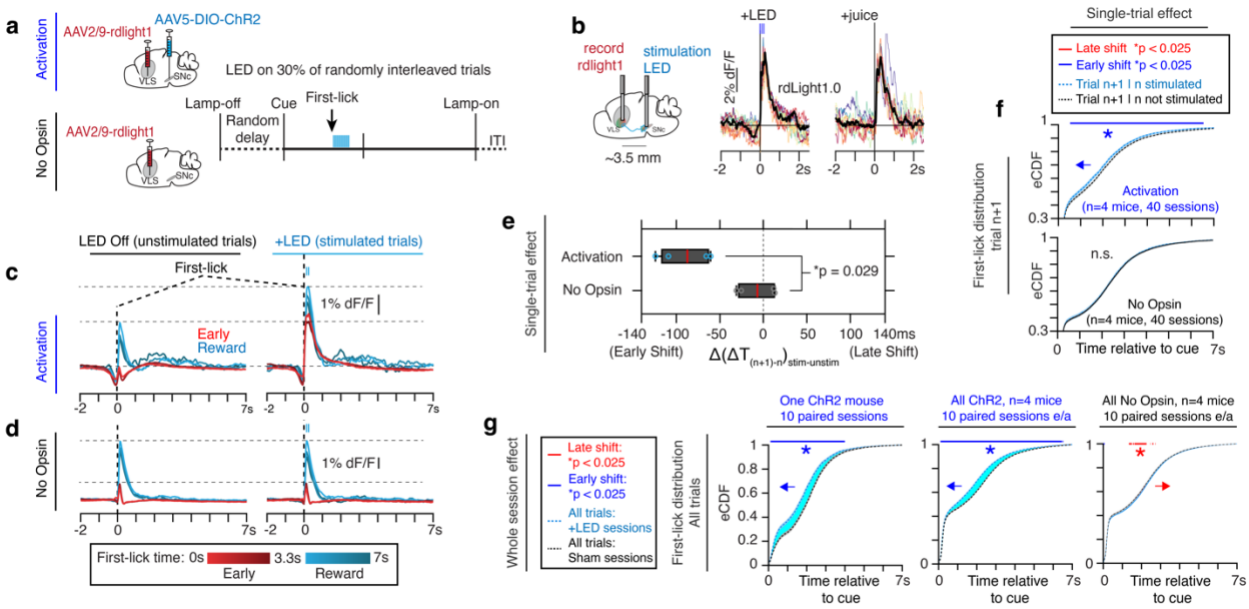


Fig 3. Outcome-related DA causes one-shot timing updates. a, Optogenetics strategy. **b,** Stimulation parameter calibration, outside context of task. Colors: single trials, black: ave. Left: 10x stimulations; Right: 10x unexpected juice rewards (~4µL each). **c-d,** Task DA. **e-f,** Single trial ("one-shot") stimulation effects. **e:** Standard box plot, red bar: median. n=4 Activation mice; n=4 No Opsin mice, 10 sessions/mouse. Wilcoxon rank sum test, *p<0.05. **f:** First-lick distribution on next trial (n+1). Arrow: direction of effect. Shading: 95% CI; Bar: *p<0.025 (1000x bootstrap). **g,** Whole session effects (stimulated vs interleaved sham sessions). Shading: 95% CI; Bar: *p<0.025 (1000x bootstrap).

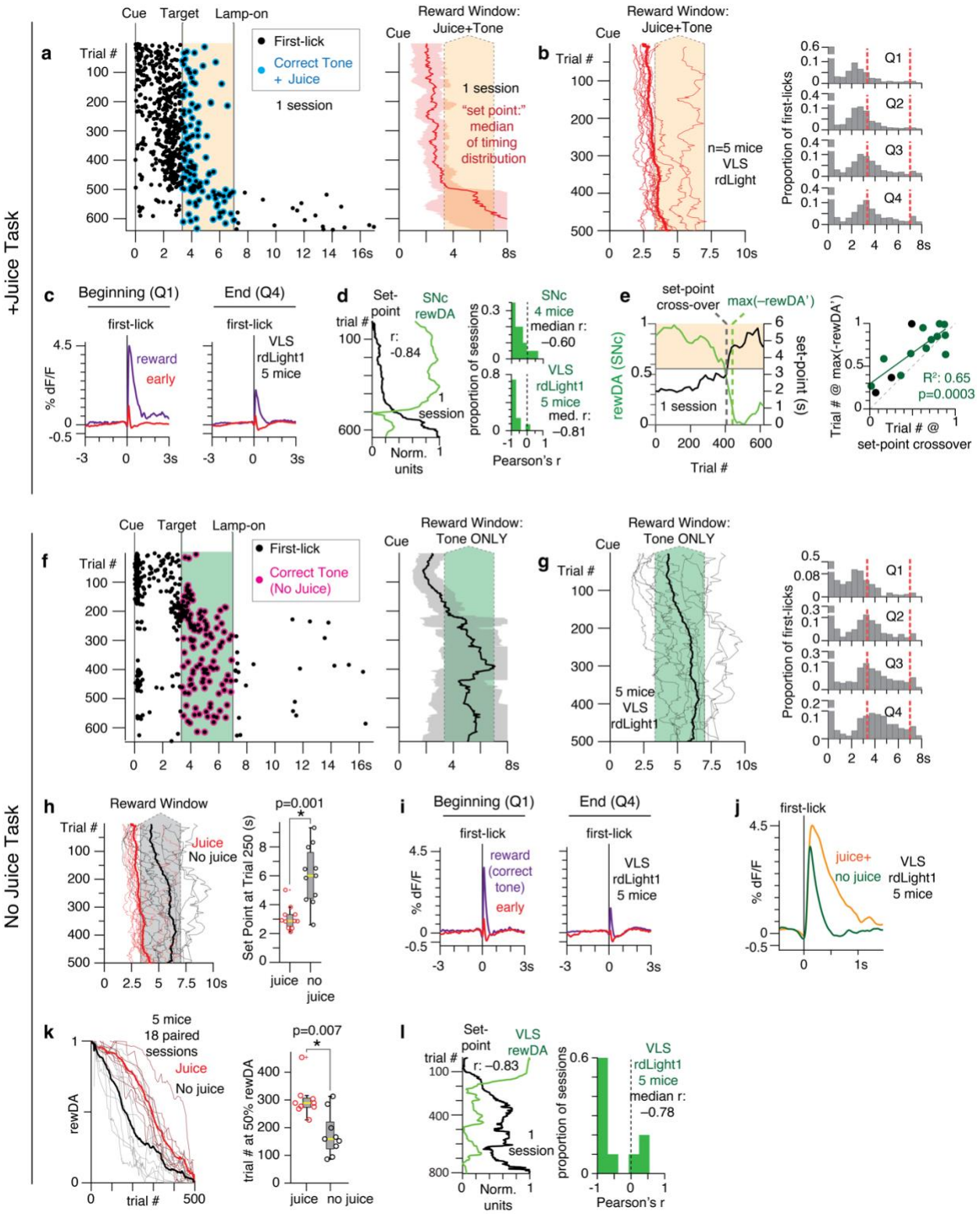


Figure 4. Timing changes across behavioral sessions. a-b, Whole-session behavior, +Juice Task. "Set-point": running median first-lick time. **b,** Sessions shown for comparison to matched No Juice Task sessions (**Fig 4f-l**). Q1-Q4: quartiles of trials in session. **c,** Outcome-related DA. **d,** Normalized rewarded DA transient amplitude ("rewDA") vs timing set-

point. **e,** max(-rewDA'): trial at maximum negative derivative of rewDA. Cross-over: trial at which set-point crossed into reward window. Right: Circles—single sessions. Black circles—No Juice sessions. **f-g:** No Juice Task whole session behavior, as in panels a-b. **h,** +Juice vs No Juice behavior. Boxplot, yellow line: median; Wilcoxon rank sum test. **i,** No Juice Task outcome-related DA. **j-k:** rewDA, +Juice vs. No Juice Task. **j** shows Q1 signals. **l,** No Juice rewDA vs timing set-point.

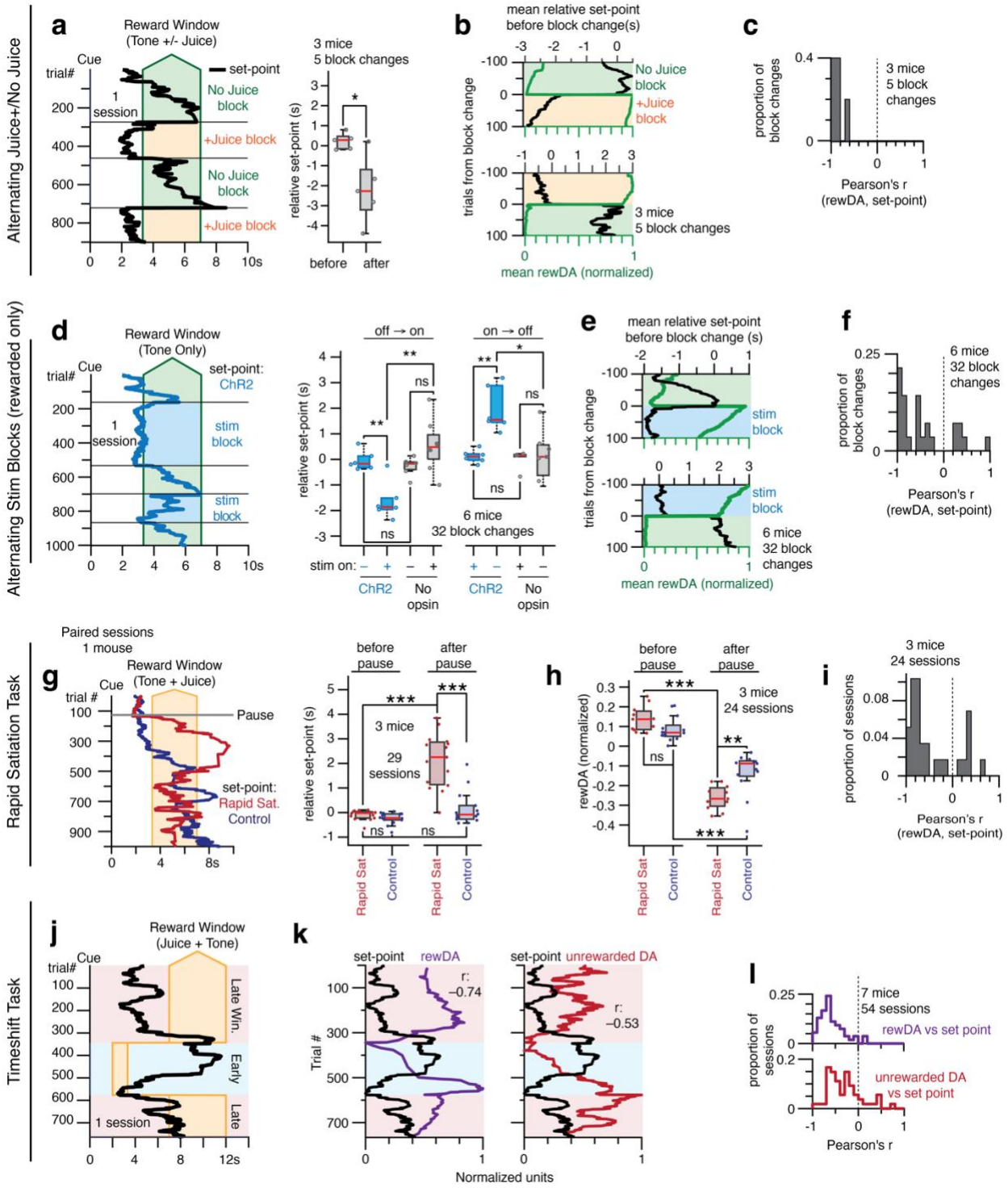
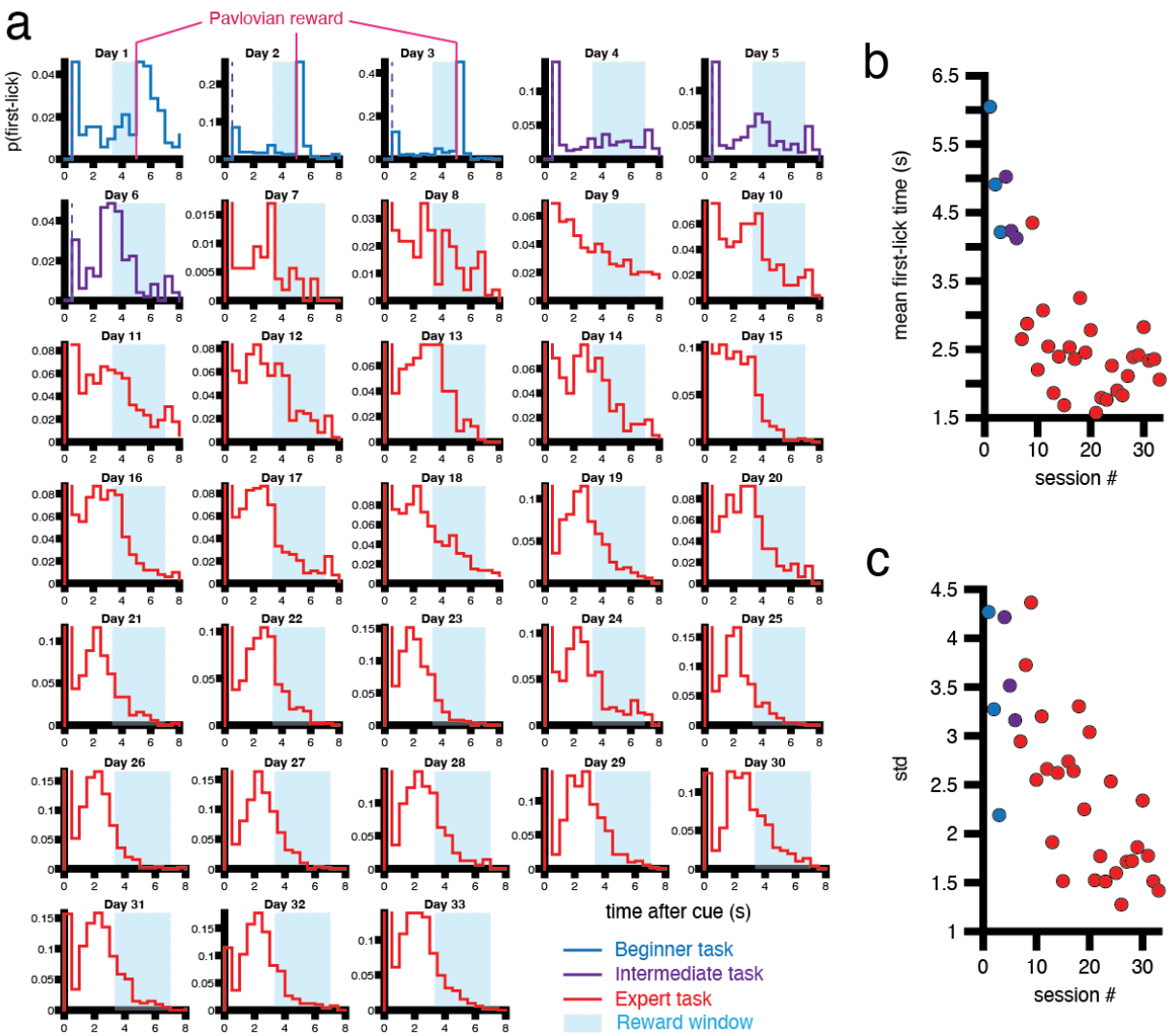
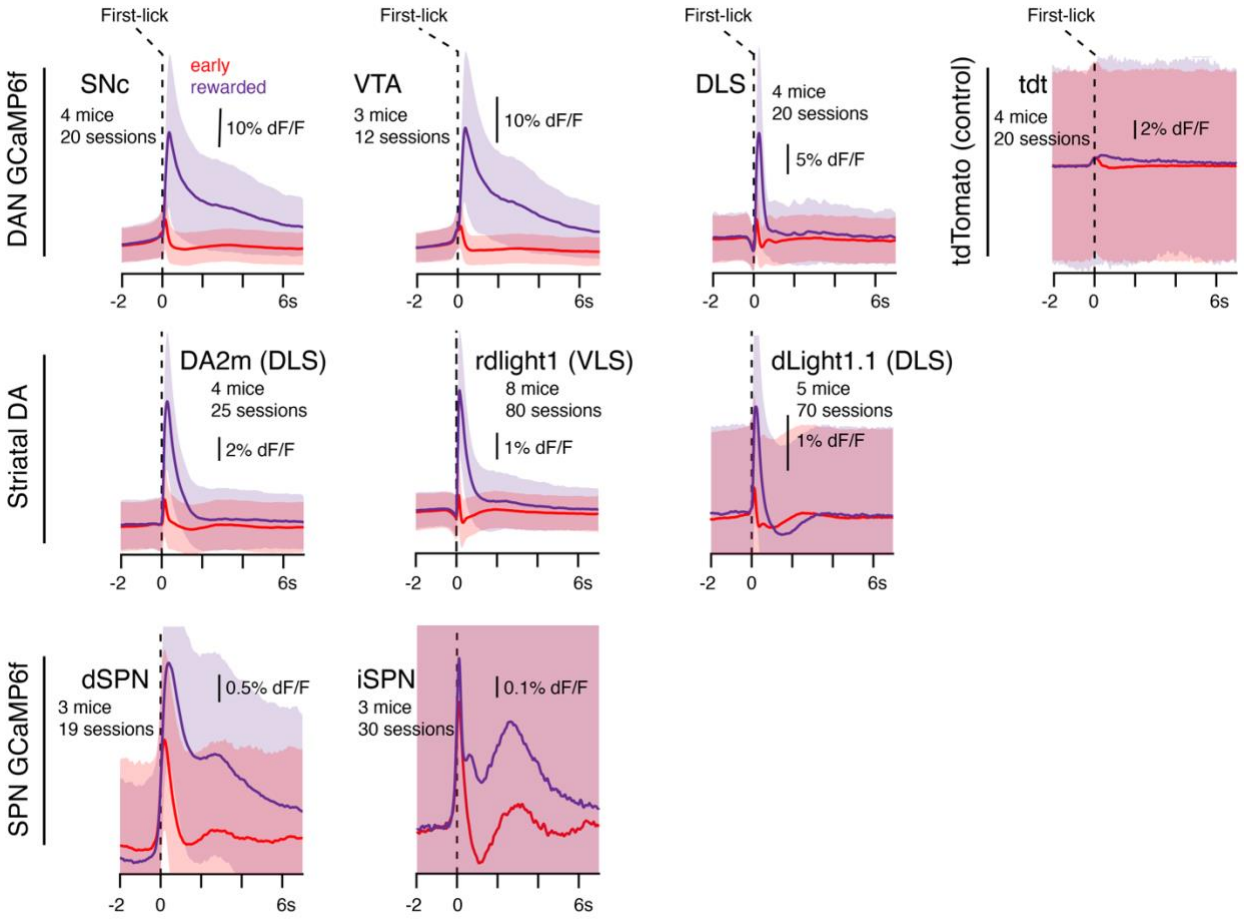


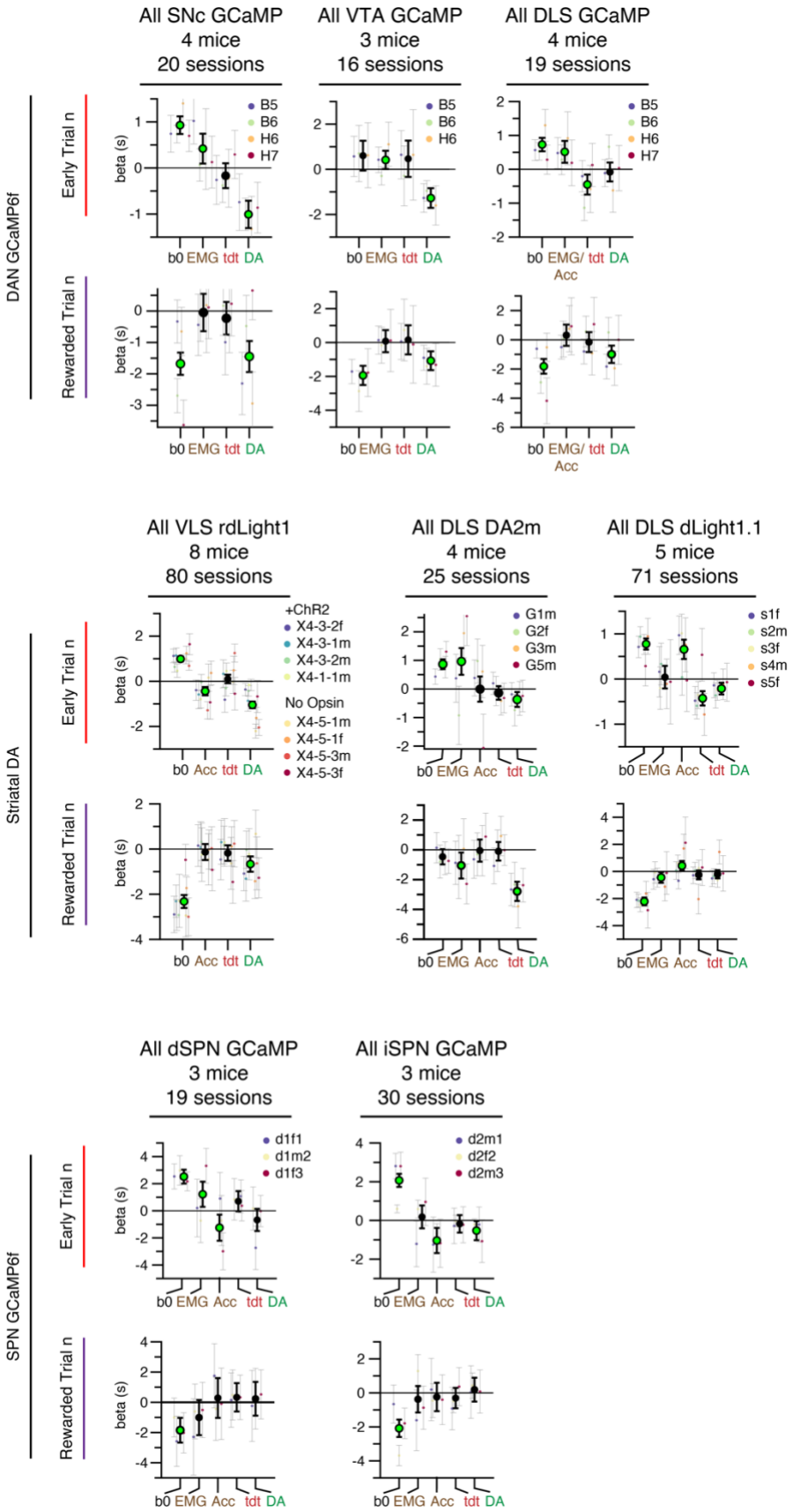
Figure 5. DA calibrates propensity to move. a-c, Alternating Juice/No Juice Task. Standard box plot, Wilxocon rank sum, p*=0.032. **b-c,** Correlation of reward-related DA ("rewDA") and set-point across block changes. **d-f,** Alternating Stimulation Blocks, Rewarded Trials only. *p<0.05, **p<0.01, ***p<0.001. **g-i,** Rapid Satiation task: causes abrupt decline in rewDA amplitude accompanied by rapid late-shifting. **j-l,** Timeshift Task. **k,** Set-point remains strongly correlated with rewDA; unrewarded DA peak amplitude (within 500ms after first-lick) shows stronger correlation with set-point than in the standard +Juice/No Juice tasks.



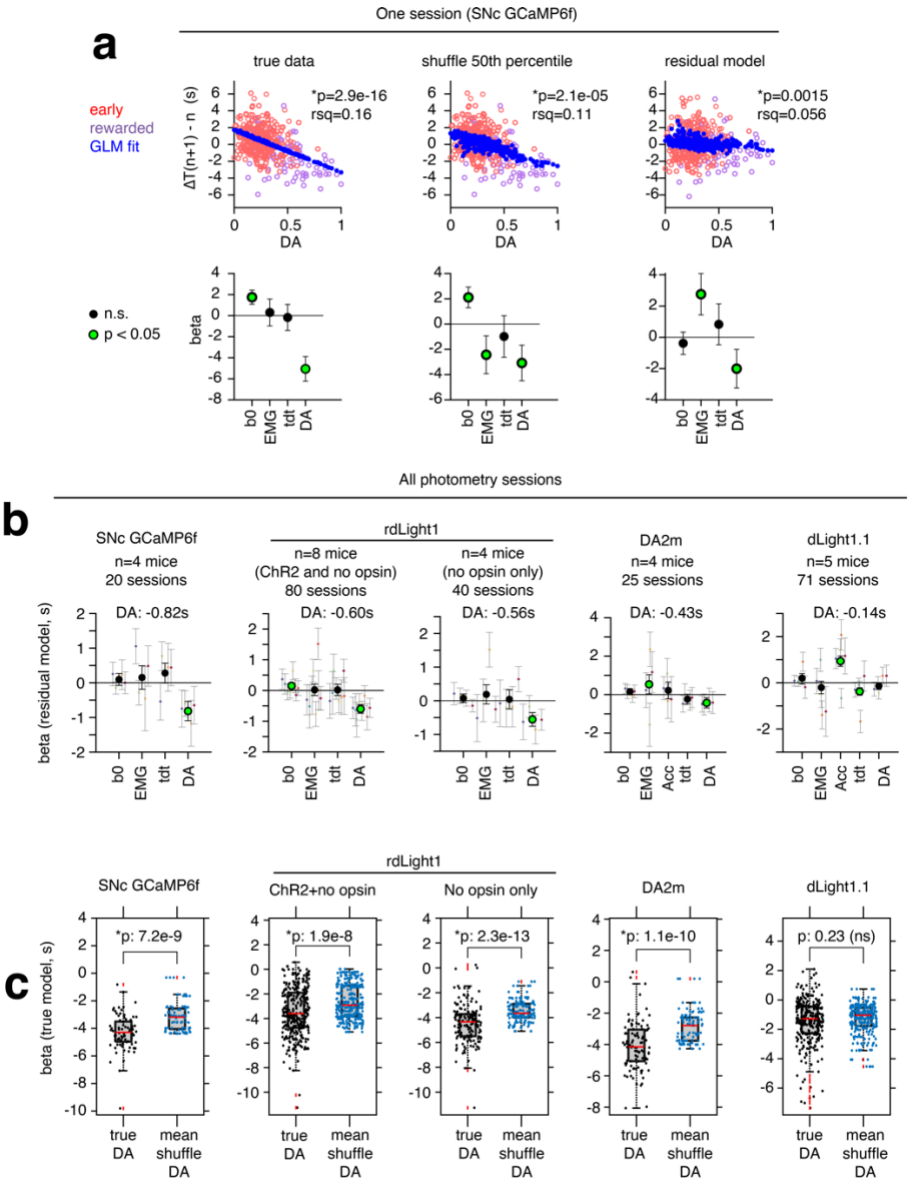
Extended Data Fig 1. Timing behavior over the course of learning. Animals trained with Juice rewards (*but not No-Juice rewards, see **Fig 4***) show early-biased timing behavior—we will address this in detail later in the paper. All mice start by learning the +Juice Task, shown here. Timing peaks emerge over the course of learning, stabilize within the first 20 sessions of training, and early bias in the timing peak persists throughout +Juice Task training. We will come back to this early bias later in the paper. **a)** Histograms show the distribution of first-licks after the cue on each day of +Juice Task training for one representative mouse. This mouse was trained for 33 consecutive sessions with 500 trials/session, with no training days on weekends. The mouse was transitioned from the Beginner task (blue traces, Pavlovian reward at 5s, 0.5s reaction window) to the Intermediate task (purple traces; no Pavlovian rewards, 0.5s reaction window) on day 4. Transition to the Expert task (red, no Pavlovian rewards and no reaction window) occurred on day 7. Y-range of histograms was maximized to clearly show the timing distribution peak; reaction peak (<0.5s) is truncated in some plots and indicated by open trace. +Juice Task distributions become stable by day 20 of training, and early timing bias persists throughout training when juice rewards are used. **b)** Mean first-lick time for each day of training. Early timing bias (mean first-lick time between 2-3s) is firmly established by day 10 and persists throughout +Juice training. **c)** Standard deviation of first-lick time for each day of training declines as timing distribution becomes more stereotyped over the course of training.



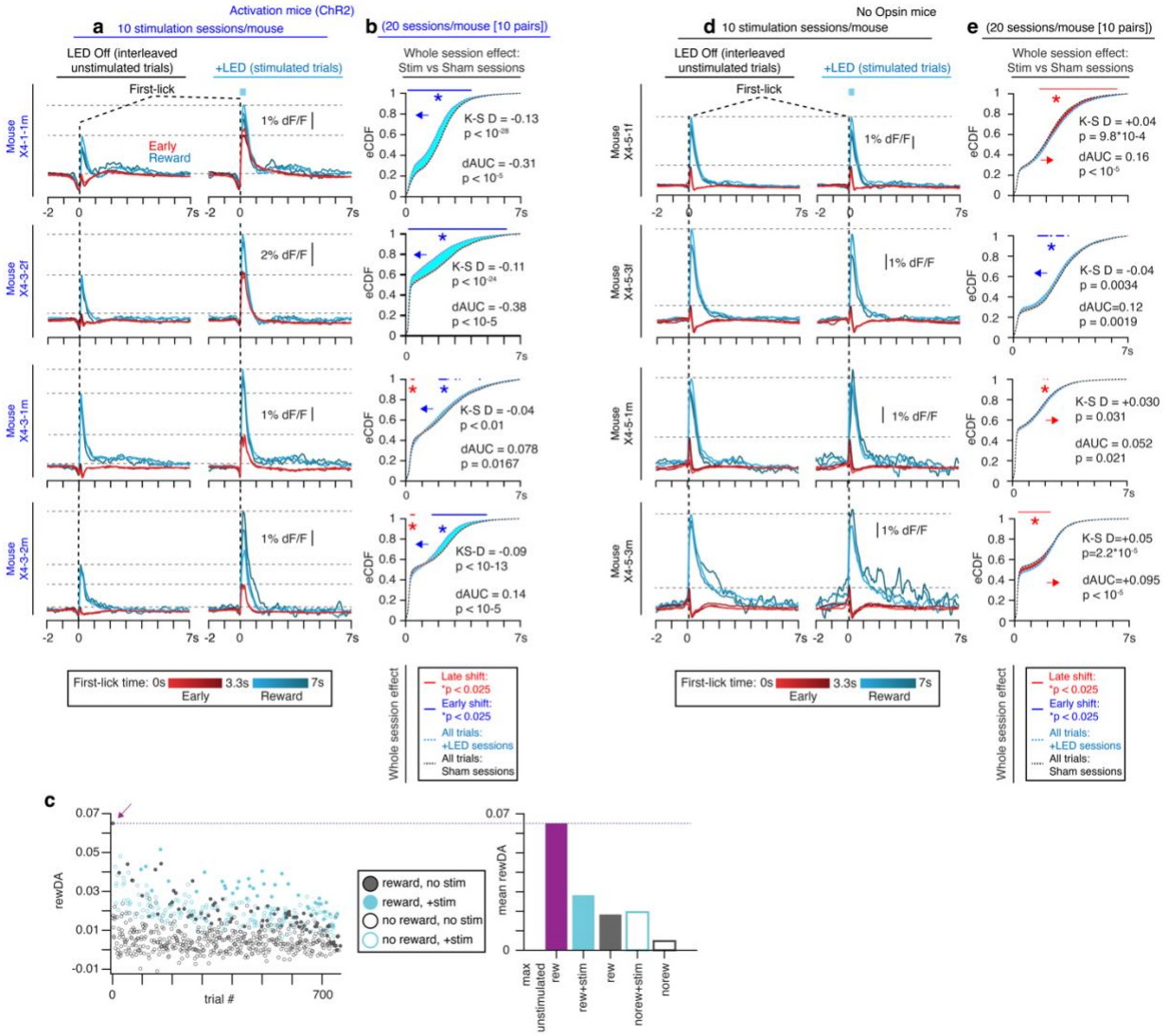
Extended Data Fig 2. Outcome related dopamine and striatal projection neuron (SPN) signals. Shading: standard deviation.



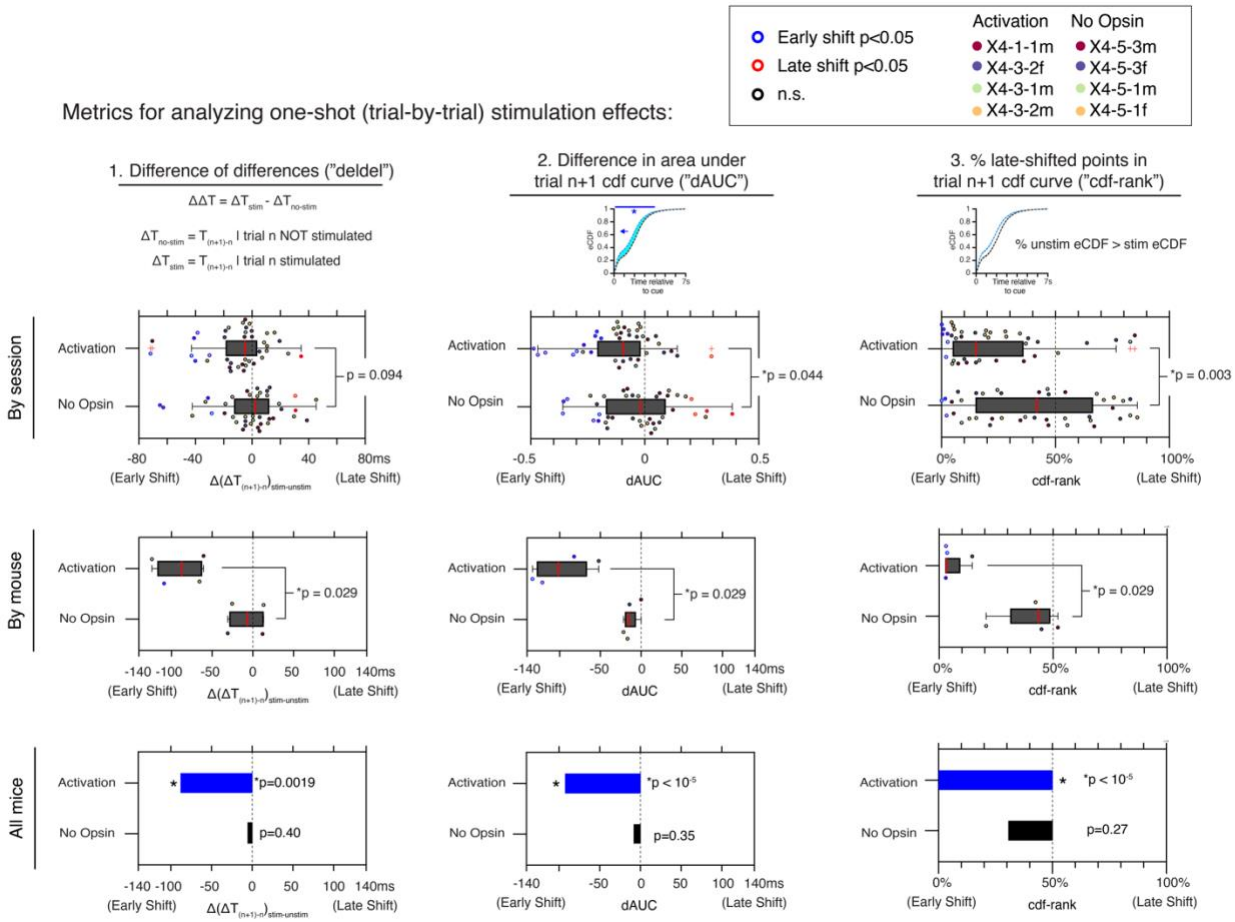
Extended Data Fig 3. Sloshing model coefficients, all signals, all sessions. Small colored dots indicate coefficients from individual mice.



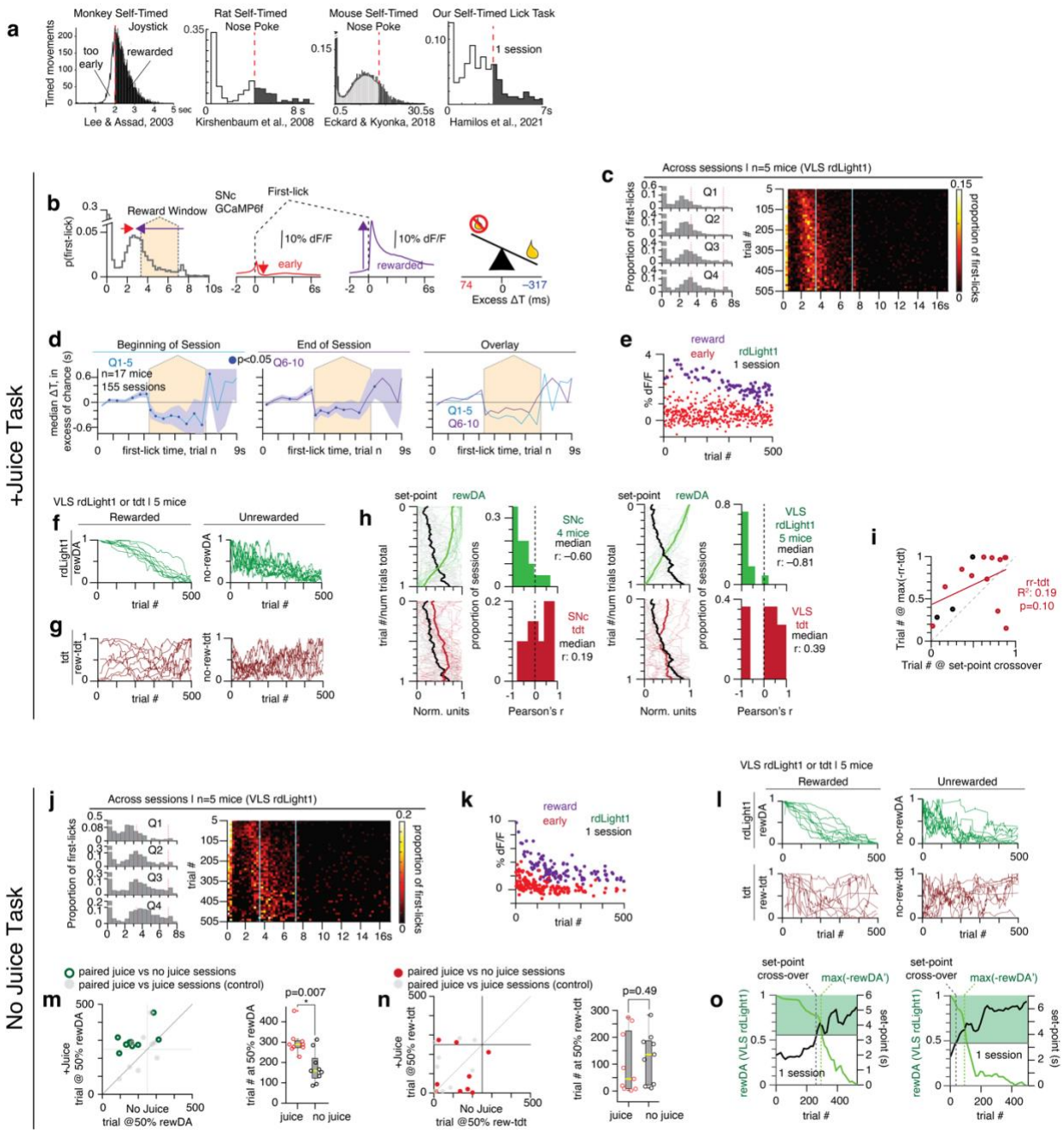
Extended Data Fig 4. Outcome-related DA explains ΔT in excess of chance. **a-b,** Shuffle models. The next trial's movement time was shuffled, and the sloshing model (**Fig 2a**) was refit on the shuffled data (10,000 shuffled datasets total). The coefficients from the 50th percentile shuffled model (50th percentile shuffled DA coefficient) were used to predict the true data (y-fit shuffle) and calculate R^2 (**a, middle panel**). Y-fit shuffle was subtracted from the true dataset to obtain a residual dataset, and the residual model was derived by refitting the sloshing model on the residual dataset (**a, right panel**). Outcome-related DA remained a significant predictor to the residual model, both in single sessions (**a, right-bottom panel**) and across sessions for SNc GCaMP, VLS rdlight1, and DLS DA2m. DLS dlight1.1 DA signals did not remain significant after the shuffle procedure, likely due to relatively poor S:N quality in the dlight1.1 dataset. **c.** Comparison of true vs shuffled model DA coefficients fit on quartiles of trials in the session. To guard against spuriously significant DA coefficients arising from non-stationarity in timing behavior across the session, true and shuffle models were fit on quartiles of trials in each session (e.g., first quarter of trials in the session, second quarter, etc.). Consistent with the residual models in **b**, the population of true model DA coefficients was significantly larger in magnitude (more negative) than the population of mean shuffled model DA coefficients (except for DLS dlight1.1; Wilcoxon rank sum test, *p<0.05). Outcome-related DA is thus predictive of ΔT in excess of regression to the median effects (see **Supplemental Methods**) and unlikely to be explained by behavioral non-stationarity.



Extended Data Fig 5. One-shot optogenetic manipulations—all mice. LED triggered on 30% of randomly interleaved first-licks. **a,d.** VLS rdlight1, all stimulation sessions, 10 sessions/mouse. Each row is a different mouse. **b,e.** Comparison of trial n+1 first-lick distributions (all trials) from stimulation sessions (blue trace) vs interleaved sham sessions (black trace). Blue shading: early-shift effect of LED; red shading: late-shift effect. Arrow: indicates overall direction of effect. Top bars (blue/red): denotes significant difference between eCDF curves (*p < 0.025, 95% CI, 1000x bootstrap test). K-S D: signed Kolmolgorov-Smirnov distance metric (negative indicates early shift effect). dAUC: difference in area under eCDF curves, p: 10,000x permutation test. **c,** Peak rewarded DA transient amplitude ("rewDA"), 1 session. Purple arrow and dashed line indicate the first rewarded (unstimulated) trial, whose amplitude is ~0.065 dF/F—corresponding to the purple bar in the bar plot. Bar plot also shows mean peak rewDA amplitude for each stimulation condition and trial outcome (reward/no reward). Stimulated trials remained below the naturally occurring peak rewDA amplitude throughout the session and on average.

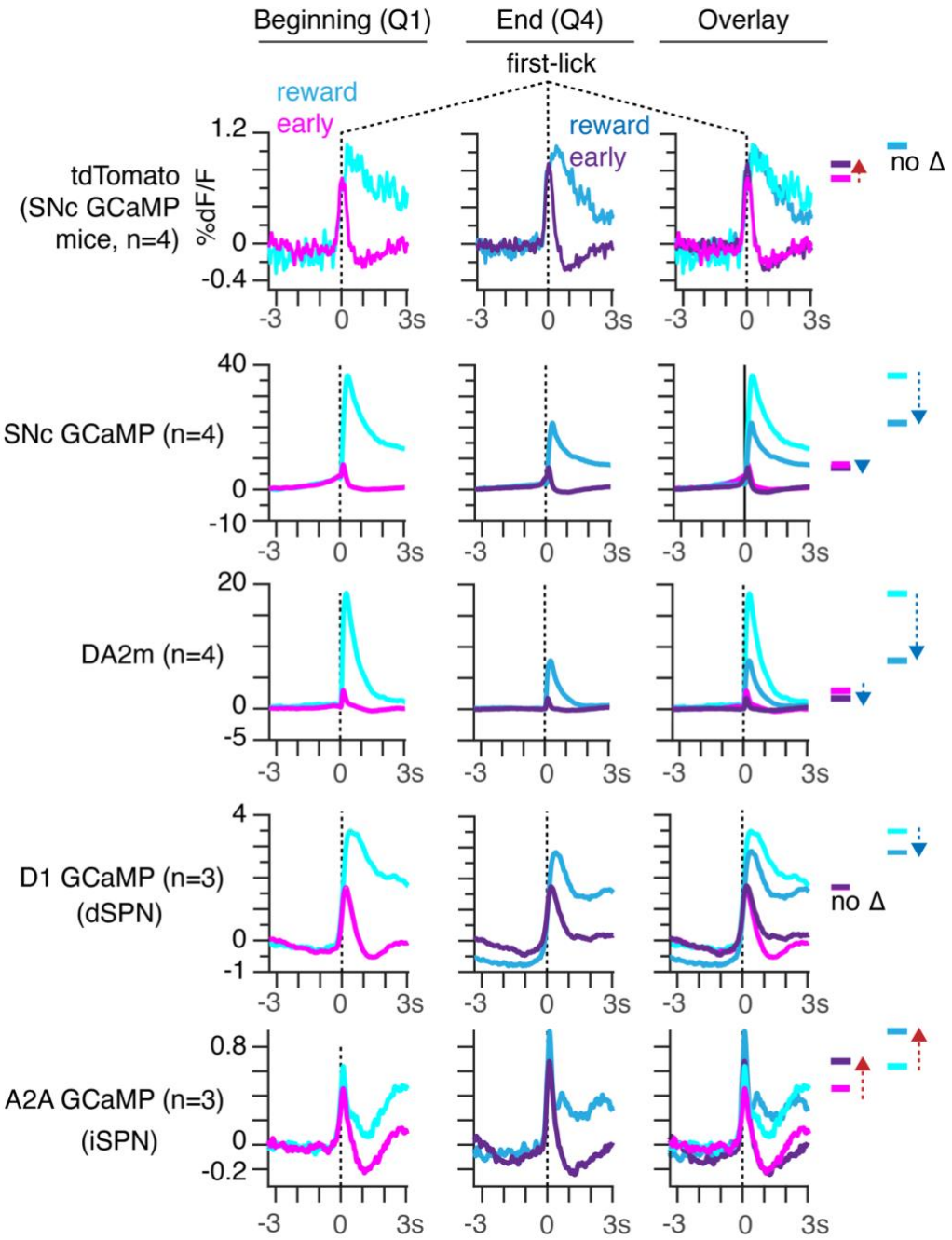


Extended Data Fig 6. One-shot optogenetic manipulations – complementary analysis methods, all sessions and mice. Comparisons: stimulated vs unstimulated trials from the same (stimulated) session. Dot colors indicate whether the test statistic was significant at the individual session level (p<0.05). Note these metrics can also be used on whole session comparisons (stim sessions vs interleaved sham sessions), as in **Fig 3g**.

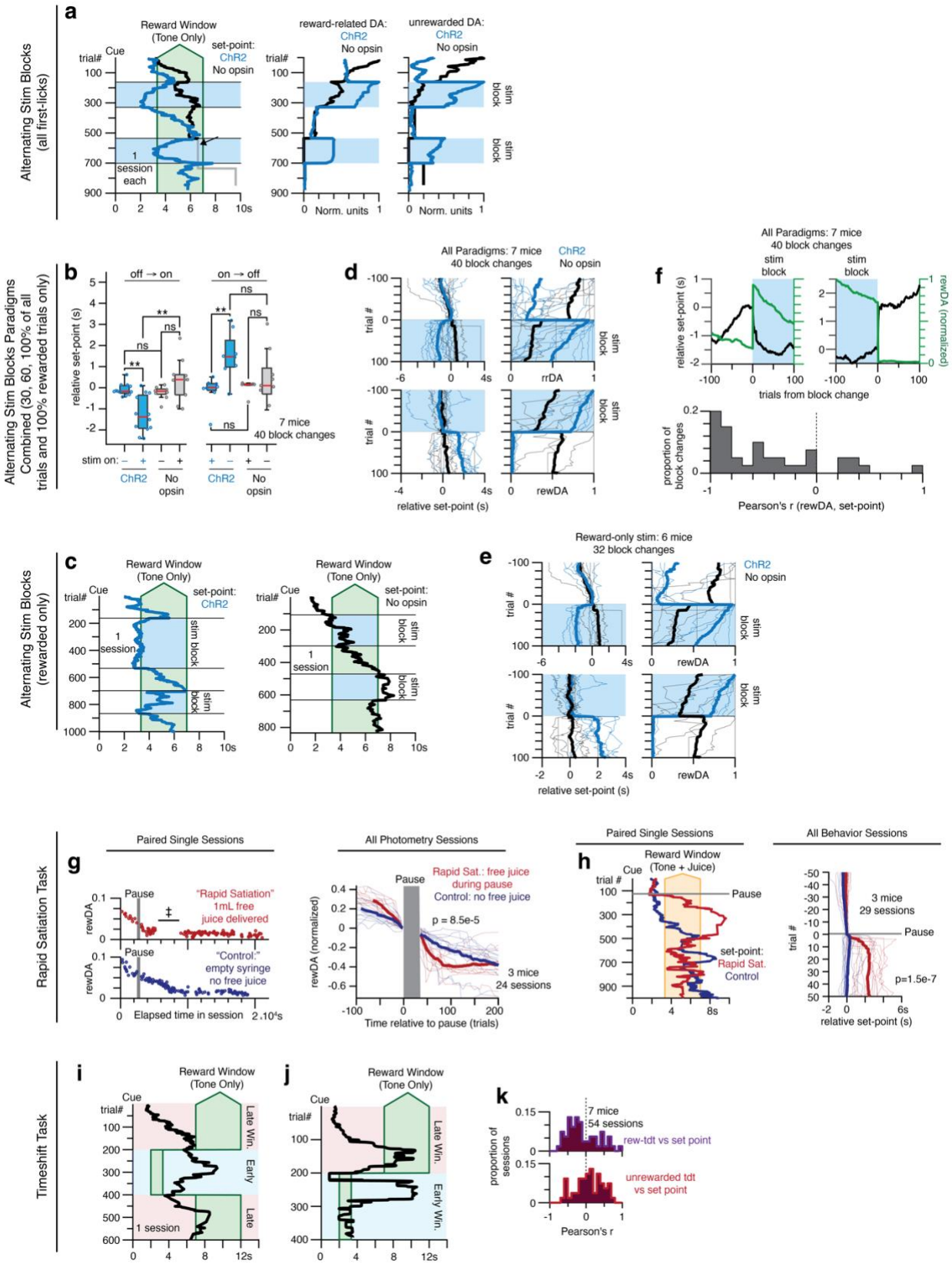


Extended Data Fig 7. Whole session timing behavior. **a,** Timing distributions from monkeys, rats, and mice show early-bias relative to the reward window. Figure panels/data adapted and used with permission. **b,** Skew in outcome-related DA responses appears similar to skew in the timing distribution with respect to the reward window. **c&j,** Heatmap view of first-lick distribution. Same mice shown in panels c and j, interleaved sessions. "Q:" quartile of trials in session with a first-lick. **d,** The effect of trial outcome on timing updates becomes less skewed toward the end of the session (rewarded trials followed by smaller updates). **e&k,** Single trial peak amplitude DA signal within 500ms after the first-lick. **f&l,** smoothed single trial data plotted for each session. Each line is one session. "rewDA:" reward-related DA peak amplitude; "no-rewDA:" unrewarded peak. **g,** Peak outcome-related tdTomato, same sessions as panel f. **h,** Correlation of all single-session outcome-related DA (top) and tdt (bottom) transients with timing set-point. Thin traces: single sessions; thick: average. Left: SNc GCaMP6f; Right: VLS rdLight1 signals. Control rew-tdt signals are

not consistently correlated with the timing set-point, arguing against a trivial correlation with ongoing body movements/optical artifacts. **i,** max(-rew-tdt'): trial at maximum negative derivative of rewarded tdt signal amplitude. Cross-over: trial at which set-point crossed into reward window. Circles: single sessions. Black circles indicate No Juice sessions. The strong correlation of set-point cross-over with rewarded DA amplitude (**Fig 4e**) is not present in control tdt signals, arguing against a trivial correlation with ongoing body movements/optical artifacts. **m-n,** Comparison of Juice+ and No Juice Task rewDA shows that rewDA declines significantly earlier in the session in the No Juice Task compared to the Juice+ Task, but there was no difference in the 50% max point for tdt data collected on the same sessions. Standard box plot with Wilcoxon rank sum test. The scatterplot points are the trial at which rewDA (panel m) or rew-tdt (panel n) reached 50% max peak amplitude for paired Juice+/No Juice session. Green/red dots: the data for either rdLight1 (green) or tdt (red); light grey dots: control data obtained from randomly pairing Juice+ sessions from the same animal, which shows 50%-max tends to occur around the same trial on sessions of the same task-variant within the same mouse. **o,** Additional single-session examples of cross-over analysis in No Juice task (analogous to **Fig 4e**).



Extended Data Fig 8. Reward-related photometry across the session. Q1: Beginning of session (first quartile of consecutive trials). Q4: End of session (fourth quartile). All DA/DAN recordings show decline in rewarded DA transient amplitude across the session, with comparably little change in unrewarded (early) transient amplitude. Optical artifacts (tdTomato) show the opposite direction of effect on unrewarded trials, whereas rewarded trial amplitude is similar across the session. Recordings from direct pathway striatal projection neurons (dSPNs) show similar patterns to DA/DAN recordings. Indirect pathway SPNs (iSPNs) show the *opposite* effect on average: both early- and reward-related iSPN transient amplitude *increase* toward the end of the session. This difference is consistent with DA's effect on iSPN excitability and provides supporting evidence against bleaching-related artifacts explaining the reward-related DA downtrend (**Supplemental Methods**).



Extended Data Fig 9. Set-point manipulations: more detail. **a,** Alternating Stimulation Blocks task, 100% of trials. *Left*: Example single sessions with same block schedule (1 ChR2 mouse, 1 No Opsin mouse, 1 session each). Exposure to

blue light initially caused a small early-shift in the No Opsin mouse, but this was much smaller than in the ChR2 mouse and did not recur during the second stimulation block. The No Opsin set-point trace is greyed after the arrow because the mouse did not participate on sufficient trials during blocks 4-5 to meaningfully calculate the running median first-lick time. *Right*: DA transient peak amplitude ("rewDA") in the example ChR2 session. Interestingly, a slow late-shift dynamic was present even with ChR2 stimulation, and stimulated DA transients declined progressively over the course of the session. **b,d,f,** Combined Alternating Stimulation Blocks paradigm results (30%, 60%, 100% and rewarded-trial only versions combined). ChR2: n=3 mice, 20 block changes; No Opsin: n=4 mice, 20 block changes. **b,** Same layout as **Fig 5d**, Wilcoxon rank sum, *p<0.05, **p<0.01, ***p<0.001. **c,** Comparison of ChR2 and No Opsin set-point dynamics, single sessions (2 mice, 1 session each). Despite exposure to blue LED, No Opsin mouse does not show early-shift during stimulation blocks. **d-e,** Single session block change set-point and dynamics. Thick trace: mean, thin traces: individual sessions. Blue: ChR2 mice, Black: No Opsin Mice. **f,** *Top:* Overlay of mean set-point and rewDA dynamics at block change. *Bottom:* Pearson's correlation of set-point and rewDA across all Alternating Stimulation Block sessions. **g-h,** Rapid Satiation Task: comparison of set-point and rewDA dynamics for single sessions and across all sessions. **g**: Left: Single session examples shown for paired, interleaved sessions from the same mouse. Black line/dagger(‡) in the top single-session plot indicates that the interval without red dots had no rewarded trials during this particular session (the set-point was highly late-shifted with reduced participation during this time, see panel *h*). Right: All photometry sessions. Thin traces: single sessions; thick: average. **h**: Left: Behavior from the same paired single sessions as in panel *g*, Right: all behavior sessions. **i-j,** Timeshift Task: example No Juice session set-point dynamics. **k,** Unlike DA, the co-monitored tdt control channel was not consistently correlated with set-point changes and did not explain away the correlation between outcome-related DA and set-point (compare to **Fig 5l**).

Supplemental Methods

Regression to the median analyses.

Timing behavior may arise from a stochastic generative process, in which timing on any trial can be modeled as stochastic "draws" from a latent distribution, *T*. This presents potential confounds to detecting trial-by-trial learning. In this framework, the expected median of the next trial's movement time would be the median of the latent distribution on the next trial, $T^{(n+1)}$. If the latent distribution is stationary trial-by-trial (i.e., $T^{(n)} = T^{(n+1)}$), we would still see changes in movement timing on the next trial (Δt_n) that arise by chance, but which are not necessarily in response to learning. Thus, even though some deterministic process (say, outcome-related DA signaling) might control this tendency to "regress to the median" of $T^{(n)}$, we could not rule out the possibility that regression to the median of $T^{(n)}$ happens by chance (i.e., due to a different process). Thus, to detect learning effects in excess of chance would require that Δt_n be in excess of the regression to the median effect. In a stochastic generative model of timing, we would interpret this kind of learning as an update to the latent distribution on the next trial, such that $T^{(n)} \neq T^{(n+1)}$. We define regression to the median on each trial, n, as:

$$R2M_n = \text{median}(T_n) - t_n$$

Where t_n is the movement time on trial n and T_n is the vector of all movement times in consideration. We say "in consideration" because we cannot simply use all movement times in the session due to slow, non-stationarity in the timing distribution across the session, which might arise from processes other than learning (e.g., satiation/fatigue) and which necessitates a more local calculation of median(T_n) to isolate "micro-nonstationarity" arising from trial-by-trial updates. Thus, the median(T_n) is an uncertain quantity. To address this, we calculated median(T_n) in 3 complementary ways:

A. We broke each session into deciles of trials and calculated the median of the distribution within each decile.
B. We broke the session into deciles as in A, but we bootstrapped the next trial's movement time by resampling the distribution of movement times within the decile 10,000 times.
C. We calculated a local, running median movement time for each trial (100 trials, centered at current trial).

$R2M_n$ for each method was then compared to the true Δt_n on each trial, defined as in the rest of the paper ($\Delta t_n = t_{n+1} - t_n$; where late-shifting is positive by convention). Because of the high level of trial-by-trial variability, we resolved Δt in excess of R2M as a significant difference between the distributions of $R2M_n$ vs Δt_n across all trials (R2M and ΔT), calculated by Wilcoxon rank sum test. The median of each distribution was then used to calculate "ΔT in excess of chance" (**Fig 1g**) defined as:

$$\Delta T \text{ in excess of chance} = \text{median}(\Delta T\text{-}R2M)$$

Assuming timing arises by a stochastic generative process, this excess term is interpreted as the change in the median of $T^{(n+1)}$ versus $T^{(n)}$.

For the analysis dataset, we used all +Juice photometry sessions that recorded DAN/DA signals from mice expressing no opsin (17 mice, 155 sessions). To ascribe excess effects to trials with different outcomes, we pooled trials either within 0.5s bins of time with respect to the cue on trial n (as in **Fig 1g**). We also tried more broadly pooling early and rewarded trials (statistics reported in-line in the text). "Early trials" included all trials with a first-lick time falling within a "notch" of the timing distribution (between the local median and opening of the reward window at 3.333s). We looked at notch trials because we expected learning updates (late-shift) and R2M effects (early-shift) for these trials to be in opposite directions, presenting the most challenging conditions to detect learning effects. To put pooled early and rewarded trial analyses on the same footing, the rewarded trial "notch" included trials pooled within an equivalent interval in the reward window (e.g., if the notch were between 3 to 3.333s for a local measurement, the rewarded notch interval was taken as 3.333-3.666s).

All three R2M approximation methods (A, B and C) returned similar results in all analyses. R2M effects were present in the dataset, indicating the mice did indeed have a tendency to "regress to the median" movement time, consistent with a stochastic generative process underlying timing. However, ΔT was in excess of R2M by all metrics and depended on the trial outcome: early trials (>0.5 to 3.333s) showed late-shifting in excess of chance, whereas rewarded trials showed early-shifting—consistent with learning effects. Interestingly, "reaction" trials, within 500ms of the cue, also showed small but significant early-shifting, which is likely a multi-trial effect (i.e., rewarded trials are often followed by reactions, and the subsequent trial (n+2) also appears to show some residual early-shift). Although there was no opportunity for mice to be "late" in this version of the timing task, rare trials in which the first-lick fell within the ITI (unrewarded) showed subsequent late-shifting, consistent with DA dips observed on these trials. It is important to note that these late-shifts represent shifts in the *median* of the putative latent distribution on trial n+1, which were small (100s of milliseconds) compared to the first-lick time (>7s)—which means that the first-lick time on trial n+1 was still more likely to be much earlier than that on trial n. Finally, we note that the size of the excess ΔT effects was on the same order of magnitude as one-shot optogenetic effects (10s-100s of milliseconds, **Fig 3**), suggesting that outcome-related DA controls updates to the stochastic "set-point" of the latent distribution on trial n+1.

Bleaching analyses

As is standard for photometry measurements, we corrected for bleaching using dF/F. Nevertheless, there remains some possibility that bleaching is either not fully corrected by this processing *or* that dF/F itself introduces some insidious artifact that could explain the downtrend in rewarded DA transients ("rewDA") across the session. *A priori*, this is unlikely:

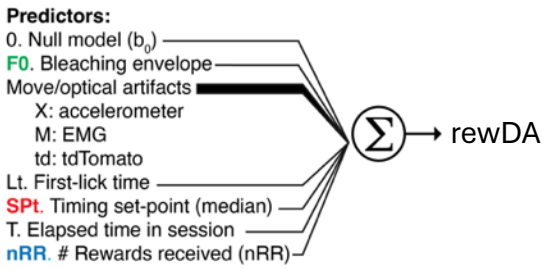
1. The amplitude of unrewarded transients showed relatively little change across the session (**Extended Data Fig 7e,f,k**)—bleaching artifacts would likely affect signals similarly for all trial outcomes.

2. Indirect pathway striatal projection neuron (iSPN) GCaMP signals showed the *opposite trend* across the session on average: rewarded iSPN transient amplitude *increased*, on average (n=3 A2A-cre mice, 30 sessions; **Extended Data Fig 8 bottom row**). iSPNs receive negative input from DANs, thus these findings are consistent with decreased DAN signaling.
3. The timing set-point showed abrupt shifts that occurred at the moments of the greatest decline in rewDA (**Fig 4e**), these changes occurred at different times in the session (**Extended Data Fig 7o**) and occurred significantly earlier in the No Juice Task (**Fig 4k**). Bleaching would likely cause similar decay rates across all sessions from the same mouse.
4. Reward-related tdTomato transients (rew-tdt) did not show the same pattern of decline across the session (**Extended Data Fig 7g,l, Extended Data Fig 8 top row**) and did not explain abrupt shifts in set-point (**Extended Data Fig 7i**).
5. Causal effects of rewDA on set-point cannot be explained away by bleaching (**Fig 5**, **Extended Data Fig 9**).
6. Rapid Satiation (**Fig 5g-i**) caused a sudden, precipitous decline in rewDA that was significantly different from that observed in control sessions in the same mice, which cannot be explained by bleaching.

Nevertheless, to directly assess the contribution of bleaching to the decline in rewDA, we derived a GLM predicting the peak amplitude of rewDA from bleaching-related and behavioral predictors. To derive a bleaching predictor, we applied the raw fluorescence, F0, from the dF/F correction. F0 was measured for each trial as the average fluorescence signal during the 5s preceding the Lamp-off event, averaged across the 10 nearest neighboring trials.[7]

We first fit the model with F0 alone. As expected, because F0 decreases monotonically, F0 was positively correlated with set-point and the decline in rewDA, and thus was a significant positive coefficient in the "bleaching-only model." This verified that bleaching was encoded properly.

Subsequently, we introduced behavioral predictors to derive nested models containing permutations of the following predictors:



Across DAN/DA recordings, predictors related to ongoing body movements, optical artifacts, and first-lick time were relatively poor predictors of rewDA, as evaluated by variance explained and 4 complementary model criterion metrics (AIC, BIC, AICc, and CAIC). For the +Juice Task, the best models always included the number of rewards received (nRR) as a significant predictor, which weakened F0's coefficient, indicating that variance in rewDA was independently explained by

variables related to satiation and/or behavioral accuracy, in excess of bleaching. Supporting this, when the models were fit on data from the Timeshift Task, in which the set-point no longer trivially correlated with nRR, set-point became a stronger predictor in the model, and all 4 information criterion metrics agreed that the most parsimonious model of rewDA required both set-point and nRR as significant predictors, in excess of bleaching. Together, these data indicate that rewDA signals contained information related to propensity to move +/- satiation, in excess of bleaching. Thus, bleaching did not explain away the relationship of rewDA to changes in the timing set-point.

References

1 Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997). https://doi.org/10.1126/science.275.5306.1593

2 Schultz, W., Apicella, P. & Ljungberg, T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J Neurosci* **13**, 900-913 (1993). https://doi.org/10.1523/JNEUROSCI.13-03-00900.1993

3 Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. *Nature* **614**, 294-302 (2023). https://doi.org/10.1038/s41586-022-05614-z

4 Mohebi, A. *et al.* Dissociable dopamine dynamics for learning and motivation. *Nature* **570**, 65-70 (2019). https://doi.org/10.1038/s41586-019-1235-y

5 Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nat Neurosci* **19**, 117-126 (2016). https://doi.org/10.1038/nn.4173

6 Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)* **191**, 507-520 (2007). https://doi.org/10.1007/s00213-006-0502-4

7 Hamilos, A. E. *et al.* Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements. *Elife* **10** (2021). https://doi.org/10.7554/eLife.62583

8 Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci* **12**, 366-375 (1989). https://doi.org/10.1016/0166-2236(89)90074-x

9 Olds, J. & Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *J Comp Physiol Psychol* **47**, 419-427 (1954). https://doi.org/10.1037/h0058775

10 Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* **16**, 966-973 (2013). https://doi.org/10.1038/nn.3413

11 Markowitz, J. E. *et al.* Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature* **614**, 108-117 (2023). https://doi.org/10.1038/s41586-022-05611-2

12 Xiao, L. *et al.* A Basal Ganglia Circuit Sufficient to Guide Birdsong Learning. *Neuron* **98**, 208-221 e205 (2018). https://doi.org/10.1016/j.neuron.2018.02.020

13 Adamantidis, A. R. *et al.* Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *J Neurosci* **31**, 10829-10835 (2011). https://doi.org/10.1523/JNEUROSCI.2246-11.2011

14 Kim, K. M. *et al.* Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. *PLoS One* **7**, e33612 (2012). https://doi.org/10.1371/journal.pone.0033612

15 Tsai, H. C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080-1084 (2009). https://doi.org/10.1126/science.1168878

16 Jeong, H. *et al.* Mesolimbic dopamine release conveys causal associations. *Science* **378**, eabq6740 (2022). https://doi.org/10.1126/science.abq6740

17 Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**, 581-589 (2017). https://doi.org/10.1038/nn.4520

18 Mikhael, J. G., Kim, H. R., Uchida, N. & Gershman, S. J. Ramping and State Uncertainty in the Dopamine Signal. *bioRxiv*, 805366 (2019). https://doi.org/10.1101/805366

19 Kim, H. R. *et al.* A unified framework for dopamine signals across timescales. *bioRxiv*, 803437 (2019). https://doi.org/10.1101/803437

20 Qian, L. *et al.* Prospective contingency explains behavior and dopamine signals during associative learning. *Nat Neurosci* (2025). https://doi.org/10.1038/s41593-025-01915-4

21 Gadagkar, V. *et al.* Dopamine neurons encode performance error in singing birds. *Science* **354**, 1278-1282 (2016). https://doi.org/10.1126/science.aah6837

22 Kasdin, J. *et al.* Natural behaviour is learned through dopamine-mediated reinforcement. *Nature* (2025). https://doi.org/10.1038/s41586-025-08729-1

23 da Silva, J. A., Tecuapetla, F., Paixão, V. & Costa, R. M. Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244-248 (2018). https://doi.org/10.1038/nature25457

24 Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505-510 (2016). https://doi.org/10.1038/nature18942

25 Ko, D. & Wanat, M. J. Phasic Dopamine Transmission Reflects Initiation Vigor and Exerted Effort in an Action- and Region-Specific Manner. *J Neurosci* **36**, 2202-2211 (2016). https://doi.org/10.1523/JNEUROSCI.1279-15.2016

26 Cai, X. *et al.* Dopamine dynamics are dispensable for movement but promote reward responses. *Nature* **635**, 406-414 (2024). https://doi.org/10.1038/s41586-024-08038-z

27 Cepeda, C., Murphy, K. P., Parent, M. & Levine, M. S. The role of dopamine in Huntington's disease. *Prog Brain Res* **211**, 235-254 (2014). https://doi.org/10.1016/B978-0-444-63425-2.00010-6

28 Niv, Y. Neuroscience: Dopamine ramps up. *Nature* **500**, 533-535 (2013). https://doi.org/10.1038/500533a

29 Salamone, J. D. & Correa, M. The Neurobiology of Activational Aspects of Motivation: Exertion of Effort, Effort-Based Decision Making, and the Role of Dopamine. *Annu Rev Psychol* **75**, 1-32 (2024). https://doi.org/10.1146/annurev-psych-020223-012208

30 Coddington, L. T. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat Neurosci* **21**, 1563-1573 (2018). https://doi.org/10.1038/s41593-018-0245-7

31 Parker, N. F. *et al.* Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat Neurosci* **19**, 845-854 (2016). https://doi.org/10.1038/nn.4287

32 Gallistel, C. R. & Gibbon, J. Time, rate, and conditioning. *Psychol Rev* **107**, 289-344 (2000). https://doi.org/10.1037/0033-295x.107.2.289

33 Merchant, H., Harrington, D. L. & Meck, W. H. Neural basis of the perception and estimation of time. *Annu Rev Neurosci* **36**, 313-336 (2013). https://doi.org/10.1146/annurev-neuro-062012-170349

34 Lee, I. H. & Assad, J. A. Putaminal activity for simple reactions or self-timed movements. *J Neurophysiol* **89**, 2528-2537 (2003). https://doi.org/10.1152/jn.01055.2002

35 Kirshenbaum, A. P., Brown, S. J., Hughes, D. M. & Doughty, A. H. Differential-reinforcement-of-low-rate-schedule performance and nicotine administration: a systematic investigation of dose, dose-regimen, and schedule requirement. *Behav Pharmacol* **19**, 683-697 (2008). https://doi.org/10.1097/FBP.0b013e328315ecbb

36 Eckard, M. L. & Kyonka, E. G. E. Differential reinforcement of low rates differentially decreased timing precision. *Behav Processes* **151**, 111-118 (2018). https://doi.org/10.1016/j.beproc.2018.02.022

37 Malapani, C. *et al.* Coupled temporal memories in Parkinson's disease: a dopamine-related dysfunction. *J Cogn Neurosci* **10**, 316-331 (1998). https://doi.org/10.1162/089892998562762

38 Meck, W. H. Affinity for the dopamine D2 receptor predicts neuroleptic potency in decreasing the speed of an internal clock. *Pharmacol Biochem Behav* **25**, 1185-1189 (1986). https://doi.org/10.1016/0091-3057(86)90109-7

39 Meck, W. H. Neuroanatomical localization of an internal clock: a functional link between mesolimbic, nigrostriatal, and mesocortical dopaminergic systems. *Brain Res* **1109**, 93-107 (2006). https://doi.org/10.1016/j.brainres.2006.06.031

40 Dews, P. B. & Morse, W. H. Some observations on an operant in human subjects and its modification by dextro amphetamine. *J Exp Anal Behav* **1**, 359-364 (1958). https://doi.org/10.1901/jeab.1958.1-359

41 Schuster, C. R. & Zimmerman, J. Timing behavior during prolonged treatment with dl-amphetamine. *J Exp Anal Behav* **4**, 327-330 (1961). https://doi.org/10.1901/jeab.1961.4-327

42 Helassa, N., Podor, B., Fine, A. & Torok, K. Design and mechanistic insight into ultrafast calcium indicators for monitoring intracellular calcium dynamics. *Sci Rep* **6**, 38276 (2016). https://doi.org/10.1038/srep38276

43 Chen, Z. *et al.* Direct and indirect pathway neurons in ventrolateral striatum differentially regulate licking movement and nigral responses. *Cell Rep* **37**, 109847 (2021). https://doi.org/10.1016/j.celrep.2021.109847

44 Lee, J., Wang, W. & Sabatini, B. L. Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nat Neurosci* **23**, 1388-1398 (2020). https://doi.org/10.1038/s41593-020-00712-5

45 Taniguchi, J. *et al.* Comment on 'Accumbens cholinergic interneurons dynamically promote dopamine release and enable motivation'. *bioRxiv*, 2023.2012.2027.573485 (2024). https://doi.org/10.1101/2023.12.27.573485

46 Yang, Z., Inagaki, M., Gerfen, C. R., Fontolan, L. & Inagaki, H. K. Integrator dynamics in the cortico-basal ganglia loop underlie flexible motor timing. *bioRxiv* (2024). https://doi.org/10.1101/2024.06.29.601348

47 Ward, R., Wu, X. & Bottou, L. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *J Mach Learn Res* **21**, 1-30 (2020).

48 Berke, J. D. What does dopamine mean? *Nat Neurosci* **21**, 787-793 (2018). https://doi.org/10.1038/s41593-018-0152-y

49 Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc Biol Sci* **285** (2018). https://doi.org/10.1098/rspb.2018.1645

50 Beron, C. C., Neufeld, S. Q., Linderman, S. W. & Sabatini, B. L. Mice exhibit stochastic and efficient action switching during probabilistic decision making. *Proc Natl Acad Sci U S A* **119**, e2113961119 (2022). https://doi.org/10.1073/pnas.2113961119

51 Long, C. *et al.* Physiological constraints on the rapid dopaminergic modulation of striatal reward activity. *bioRxiv* (2022). https://doi.org/10.1101/2022.09.16.508310

52 Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* **12(Jul)**, 2121-2159 (2011).

53 Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron* **98**, 1005-1019 e1005 (2018). https://doi.org/10.1016/j.neuron.2018.05.020

54 Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian Computation through Cortical Latent Dynamics. *Neuron* **103**, 934-947 e935 (2019). https://doi.org/10.1016/j.neuron.2019.06.012

55 Maimon, G. & Assad, J. A. A cognitive signal for the proactive timing of action in macaque LIP. *Nat Neurosci* **9**, 948-955 (2006). https://doi.org/10.1038/nn1716

56 Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616-1620 (2014). https://doi.org/10.1126/science.1255514

57 Lee, K. *et al.* Temporally restricted dopaminergic control of reward-conditioned movements. *Nat Neurosci* **23**, 209-216 (2020). https://doi.org/10.1038/s41593-019-0567-0

58 Salinas-Hernandez, X. I., Zafiri, D., Sigurdsson, T. & Duvarci, S. Functional architecture of dopamine neurons driving fear extinction learning. *Neuron* **111**, 3854-3870 e3855 (2023). https://doi.org/10.1016/j.neuron.2023.08.025

59 Salinas-Hernandez, X. I. *et al.* Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. *Elife* **7** (2018). https://doi.org/10.7554/eLife.38818

60 Floeder, J. R., Jeong, H., Mohebi, A. & Namboodiri, V. M. K. Mesolimbic dopamine ramps reflect environmental timescales. *bioRxiv* (2024). https://doi.org/10.1101/2024.03.27.587103

61 McClure, S. M., Daw, N. D. & Montague, P. R. A computational substrate for incentive salience. *Trends Neurosci* **26**, 423-428 (2003). https://doi.org/10.1016/s0166-2236(03)00177-2

62 Tritsch, N. X. & Sabatini, B. L. Dopaminergic modulation of synaptic transmission in cortex and striatum. *Neuron* **76**, 33-50 (2012). https://doi.org/10.1016/j.neuron.2012.09.023

63 Lahiri, A. K. & Bevan, M. D. Dopaminergic Transmission Rapidly and Persistently Enhances Excitability of D1 Receptor-Expressing Striatal Projection Neurons. *Neuron* **106**, 277-290 e276 (2020). https://doi.org/10.1016/j.neuron.2020.01.028

64 Lee, S. J. *et al.* Cell-type-specific asynchronous modulation of PKA by dopamine in learning. *Nature* **590**, 451-456 (2021). https://doi.org/10.1038/s41586-020-03050-5

65 Mello, G. B., Soares, S. & Paton, J. J. A scalable population code for time in the striatum. *Curr Biol* **25**, 1113-1122 (2015). https://doi.org/10.1016/j.cub.2015.02.036

66 Costacurta, J. C., Bhandarkar, S., Zoltowski, D. M. & Linderman, S. W. Structured flexibility in recurrent neural networks via neuromodulation. *bioRxiv* (2024). https://doi.org/10.1101/2024.07.26.605315

67 Lodder, B. *et al.* Absolute measurement of fast and slow neuronal signals with fluorescence lifetime photometry at high temporal resolution. *bioRxiv* (2025). https://doi.org/10.1101/2025.01.10.632162

68 Backman, C. M. *et al.* Characterization of a mouse strain expressing Cre recombinase from the 3' untranslated region of the dopamine transporter locus. *Genesis* **44**, 383-390 (2006). https://doi.org/10.1002/dvg.20228