# Mesolimbic dopamine adapts the rate of learning from errors in performance

Luke T. Coddington[1,*], Sarah E. Lindo[1], and Joshua T. Dudman[1,*]

[1]Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA

***Correspondence:** Luke (coddingtonl@hhmi.org) and/or Josh (dudmanj@janelia.hhmi.org)

## Abstract

Recent success in training artificial agents and robots derives from a combination of direct learning of behavioral policies and indirect learning via value functions. Policy learning and value learning employ distinct algorithms that depend upon evaluation of errors in performance and reward prediction errors, respectively. In animals, behavioral learning and the role of mesolimbic dopamine signaling have been extensively evaluated with respect to reward prediction errors; however, to date there has been little consideration of how direct policy learning might inform our understanding. Here we used a comprehensive dataset of orofacial and body movements to reveal how behavioral policies evolve as naive, head-restrained mice learned a trace conditioning paradigm. Simultaneous multi-regional measurement of dopamine activity revealed that individual differences in initial reward responses robustly predicted behavioral policy hundreds of trials later, but not variation in reward prediction error encoding. These observations were remarkably well matched to the predictions of a neural network based model of behavioral policy learning. This work provides strong evidence that phasic dopamine activity regulates policy learning from performance errors in addition to its roles in value learning and further expands the explanatory power of reinforcement learning models for animal learning.

## Introduction

Biological and artificial agents learn how to adapt behavior to achieve targeted outcomes through experience with an environment. Reinforcement learning (RL) theory describes the algorithms that allow an agent to iteratively improve its success through training [1]. Experience with the environment can be evaluated both with respect to the success of an agent's behavioral 'policy' that directly determines the actions performed as well as an agent's subjective expectations of reward that indirectly guide action. We refer to these distinct aspects of evaluation as 'performance errors' and 'reward prediction errors (RPEs)', respectively. The reinforcement learning algorithms that govern how an agent learns from performance errors and RPEs are distinct and referred to as direct policy learning and (indirect) value learning methods, respectively. Learning methods that combine direct policy learning with value learning have contributed to key breakthroughs in training artificial agents [2,3]. However, the contribution of direct policy learning has been little considered in canonical neuroscience and animal learning paradigms.

Animal learning has commonly been compared to the 'actor-critic' class of algorithms for computational reinforcement learning [4], in which an actor module selects actions based upon the critic module's predictions from value learning. Ascending projections from midbrain dopamine neurons (mDA) convey subjective information about expected outcomes and regulate synaptic and intrinsic plasticity at their targets [5]. mDA activity has thus been hypothesized to implement a 'critic'. Over the last several decades much work has explored how mDA activity matches the predicted update signals (RPEs [6]) for value learning in a critic, resulting in many confirmations [5], but also challenges [7–11]. Moreover, exogenous stimulation of mDA neurons has led to influential observations consistent with some predictions of value learning [12–14]. However recent work has also highlighted that mDA neurons reflect a heterogeneous mix of signals and functions, some of which can be difficult to reconcile with the predictions of value learning models [11,15–23]. That much of this phasic mDA activity is intertwined with the production and monitoring of action [24–27] calls for a better understanding of how dopamine learning signals function in the evaluation of performance errors expected in direct policy learning, but not currently incorporated into actor-critic models of animal learning.

Even though it has received relatively less attention in neuroscience, there are a number of reasons why it is important to explore the potential of direct policy learning to provide "computational and mechanistic primitives" [28] that account for aspects of dopamine function, especially in the context of novel task acquisition by animals. For one, behavioral adaptation during the initial acquisition of tasks is meaningfully variable across individuals [29]. Policy learning methods allow for explicit modeling of these individual behavioral learning trajectories as exploration in the space of policy parameterizations. Performance errors can, by definition, be idiosyncratic to individual animals, whereas RPEs are fixed in the deterministic environment of many laboratory learning tasks. Second, direct policy learning methods have enjoyed substantial success in embodied learning problems in robotics that resemble problems faced by a behaving animal [30]. Third, fundamental observations that mDA stimulation reinforces behavior [11,31–34] could in principle be consistent with a direct effect on behavioral policy, but this perspective requires explicit evaluation of direct policy learning accounts of learning that have not been articulated. Finally, policy learning can be directly driven by behavioral performance error signals, in lieu of or in addition to, RPEs [35,36], connecting them to diverse observations of learning in dopamine-recipient brain areas [20,37–45].

Here we assess whether inferred adaptations of behavioral policy during acquisition of a novel cue-reward association, and corresponding mDA neuron activity, can be reconciled with a direct policy learning perspective (see Logic Outline, Sup. Fig. 1). We first collected a detailed

dataset of multidimensional behavioral changes during the initial acquisition of a classical trace conditioning task to infer putative behavioral policy. We then articulate a novel variant of a policy learning model that quantitatively accounts for the learned behavior and in particular is expressive enough to capture the diverse learning trajectories of individual animals. mDA activity predicted by the model quantitatively matched fiber photometry recordings of mDA activity made continuously throughout learning and explains how individual differences in mDA activity predict learning outcomes. We use the model to identify two novel experimental predictions uniquely consistent with policy learning and tested these predictions using calibrated optogenetic mDA stimulation in closed-loop with behavior. Finally, we show that intense, uncalibrated stimulation can yield effects that appear much more consistent with value learning. Together these results define parallel functions for mesolimbic dopamine in policy and value learning to more fully explain associative learning phenomena.

## Results

### *Characterizing individual learning trajectories over acquisition of a cue-reward pairing*

We tracked multiple features of behavioral responses to classical trace conditioning in mice that had been acclimated to head fixation but had received no other “shaping” or pre-training (Fig. 1a). On “cued” trials an auditory cue (0.5 s, 10 kHz) preceded a ~3 μL sweetened water reward by 1.5 seconds. ”Uncued” probe trials (~10% of trials), in which rewards were delivered in the absence of an auditory cue, were randomly interleaved. Collection latency progressively decreased in both cued and uncued trials across training (Fig. 1b). The presence of a predictive cue showed additional reductions in collection latency (cued: 176 ± 26 ms, uncued: 231 ± 23 ms, p =0.03), indicating that mice learn to use predictive cues to speed reward collection.

We sought to describe idiosyncratic learning across individual mice in addition to the population average by measuring multiple features of behavior. An accelerometer attached to the moveable basket under the mice summarized body movements [11]. High resolution video was used to infer lick rate, whisking state, pupil diameter, and nose motion. We considered two general aspects of behavior that determine how rapidly a mouse can collect a water reward after its delivery. The first is the speed of the reaction to sensory evidence of reward availability that we will refer to as “transient” behavioral responses to reward delivery (Fig. 1e, reflecting the efficiency of behavioral activation to reward delivery). The second is preparation for delivery of water using a preceding stimulus (cue) as a predictor. We refer to these components as “sustained” behavioral responses observed during the delay period after the cue and prior to reward (Fig. 1d, aspects of which are referred to as conditioned responding or anticipatory behavior in other work).

Across all mice the sustained and transient components of learned behavior exhibited noisily monotonic trajectories, with variable magnitudes and time courses across individuals (Fig. 1c-e). To understand how these behavioral adaptations are related to learned improvements in reward collection performance, we built generalized linear models (GLM) to predict reward collection latency across training in each mouse. GLMs (Fig. 1f) using “sustained” and “transient” (non-licking) behavioral measures as predictors captured much of the variance in reward collection efficiency over training ($r^2$ = 0.69 ± 0.11; $r^2$ with shuffled responses = 0.01 ± $3e^{-4}$). We observed a range of consistencies in each predictor’s weighting, with sustained licking having the most consistent relation to reward collection latency (Fig. 1g). However, across all mice each predictor had similar unique predictive power as judged by the partial explained variances of each predictor (1-way ANOVA: p = 0.4). Accordingly, both sustained and transient variables were necessary to accurately predict reward collection latency (Fig. 1h, Friedman's: P = 0.0003; sustained only $r^2$ = 0.51 ± 0.24, vs full model P = 0.004; transient only $r^2$ = 0.46 ± 0.20, vs full model P = 0.002).

Thus, across the population of animals initial learning could be characterized by both gains in transient responding to cues and rewards and the development of sustained behavioral responses that outlasted cue presentation. From the diverse behavioral measures we could infer putative underlying quantities - sustained and transient components of behavioral learning - that manifest as more variable and idiosyncratic measurements across individuals and behavioral modalities (Fig. 1i, see methods for details). From this perspective updates to the behavioral policy over learning for each mouse can be viewed as a trajectory through an abstract 'learning space' defined by sustained and transient dimensions that together improve reward collection performance (Fig. 1j). Across the population, the initial starting points of animals were quite variable and converged towards a more consistent final, learned state.

### ***Mesolimbic dopamine signals predict differences in learning trajectories across mice***

To simultaneously measure and manipulate mDA activity in the same mice, we expressed Cre-dependent jRCaMP1b in DAT-Cre::ai32 mice that transgenically expressed Chr2 under control of the dopamine transporter promoter (Fig. 2a) [11,46]. Optical fibers were implanted bilaterally over the VTA, and unilaterally in the nucleus accumbens core (NAc), and in the dorsal medial striatum (DS) (Fig. 2a). For each mouse we recorded signals from two target areas throughout the first 800 trials of acquisition; NAc in all mice with simultaneous recordings either from contralateral DS (n = 6) or the ipsilateral VTA (n = 3) in subsets.

NAc-DA signals exhibited reward responses at the start of training that became better aligned to reward delivery but did not change significantly in magnitude over training (trials 1-100: 0.82 ± 0.21 z, trials 700-800: 1.16 ± 0.23 z, signed-rank p = 0.13), even as cue responses steadily increased (Fig. 2c-d). In contrast, DS-DA signals were not initially excited by cues or rewards, but developed responses upon further training, consistent with previous reports [21,47,48]. VTA-DA signals more closely resembled NAc-DA signals, and simultaneously recorded signals grew more correlated with training (Fig. 2d-f), consistent with somatic recordings in which mesolimbic DA neurons more synchronously represent cues and rewards as training progressed [11]. By the end of the training period, NAc-DA reward signals exhibited correlates with positive (cued 1.1 ± 0.2 z vs uncued 1.5 ± 0.2 z, p = 0.04) and negative (omitted -0.3 ± 0.1 z, p = 0.002 vs 0) reward prediction errors. VTA-DA signals exhibited similar trends (cued 1.3 ± 0.4 z vs uncued 2.0 ± 0.5 z, p =0.25; omitted -0.3 ± 0.2, p = 0.06 vs 0). DS-DA reward signals did not reflect RPE correlates (cued 0.6 ± 0.3 z vs uncued 0.5 ± 0.3 z, p = 0.7; omitted 0.08 ± 0.14 z, p =0.7).

We next examined whether DA reward responses could explain any of the substantial inter-animal variability in behavioral learning trajectory (Fig. 1j), focusing on NAc-DA given its robust reward responses throughout training. We found substantial inter-animal variance in initial NAc-DA responses in the first 100 trials that was not related to anatomical location of fibers (Fig. 3b; individual axis correlations with NAc-DA reward response trials 1-100: A/P: p=0.5, M/L: p=0.4, D/V: p=0.5; multiple linear regression, p = 0.7). However, low initial NAc-DA reward signals were predictive of more extensive sustained behavior at the end of training (Fig. 3a, C; NAc-DA reward response during trials 1-100 vs sustained behavior during trials 700-800, r = -0.85, p = 0.004), as well as predictive of faster reward collection latencies (Fig. 3d; NAc-DA reward during trials 1-100 vs reward collection latency during trials 700-800, r = 0.81, p = 0.008). Furthermore, estimates of the sustained and transient components were sufficient to accurately reconstruct each mouse's initial dopamine reward signals (Fig. 3e; actual vs predicted from behavior (see Methods) r = 0.99, p < 0.0001).

Responses of mDA neurons to predictive cues emerge alongside cued behaviors, and regularly correlate with the learned value of the cue [5]. We found that individual differences in initial NAc-DA reward signals were not correlated with the learning of NAc-DA cue signals (NAc-DA at rew trials 1-100 vs.NAc-DA at cue trials 700-800, p = 0.5) despite a correlation between NAc-DA cue signal magnitude and the amount of sustained behavior across individual mice at the end of training (Sup Fig. 3a). Thus, individual differences in NAc-DA reward signals correlate with the behaviors that emerge with training, but not with the NAc-DA cue signals that are correlates of value learning.

### *ACTR: a direct policy learning model of classical conditioning*

The above data suggest that it could be useful to formulate novel associative learning as a problem of directly learning a control policy for reward collection that minimizes the time to collect reward. To explore this concept, we first began by specifying a behavioral 'plant' that determines how a control signal produces licking behavior. We used a parsimonious generative model to capture the statistics of licking behavior - a state model that transitions between a quiescent state and a licking state which emits regular licks (at 7Hz as observed behaviorally). A control policy ($\pi(t)$) determines the forward transition rate to licking, and the reverse transition rate reflects an energetic cost to constant licking that decreases in the presence of water such that licking is sustained until collection is complete (Fig. 4a). The control policy for lick transitions was learned by a recurrent neural network (RNN) receiving input at the onset and offset of the cue and the onset of water delivery (Fig. 4b; see Methods for model details). The optimal target policy as identified by a search through thousands of weight initializations minimized performance cost through an enhanced transient component and sustained cued licking that depends upon sustained network dynamics (Sup. Fig. 2) .

During learning, the internal connection strengths between RNN units (equivalent to synaptic weights) were updated in order to minimize a cost function that depended on reward collection latency. This learning rule (ACTR: Adaptive rate, Cost of performance to REINFORCE) was inspired by a recently described, biologically plausible rule for training RNNs [49] that itself drew upon inspiration variants of node perturbation methods [50] and the classic policy optimization methods known as 'REINFORCE' rules [1,35]. As adaptive learning rates are exceptionally useful for optimization in machine learning (e.g., [51]), the ACTR learning rate is not a fixed constant, but rather is an adaptive rate proportional to the activity of a feedback unit (pink output unit in Fig. 4b) that compares behavioral plant output (akin to an efference copy of reward-related action commands [52]) to the policy network's response to reward delivery (akin to reward-predictive sensory evidence). This scheme has a direct and intentional parallel to the phasic activity of midbrain DA neurons which, analogously, is well described by action- and sensory-related components of reward prediction [11,24]. Moreover, mDA neurons receive input from areas involved in determining policy [53,54] and have previously been implicated in modulating learning rate [55,56].

We generated predicted phasic mDA activity in the ACTR model by summing the action- and sensory-related components of feedback unit activity as in previous work with somatic spiking signal [11] and convolving with a temporal kernel matched to the kinetics of the calcium sensor used in Fig. 2-3 [57]. This predicted phasic DA photometry signal from the ACTR model corresponds closely to experimentally measured mDA activity across training and in particular replicated well known RPE correlates (Fig. 4e). Using this mDA-like feedback signal to control ACTR's learning rate allowed us to account for observed learning data and provided novel experimental predictions regarding the expected correlates of mDA activity and function of dopamine-mediated feedback as explained below.

### *Comparison of ACTR dynamics to observed mouse learning*

For repeated ACTR simulations (n=12) from a common initialization, we found that latency to collect rewards declined comparably to observed mouse behavior over training, including a performance gain on cued versus uncued trials (Fig. 4d). Sampling possible ratios of sustained and transient learning produced a range of trajectories (Fig. 4c) comparable to individual observed mice. This can be illustrated by computing the gradient surface from simulations of the plant across a range of policies with varying sustained/transient ratios (Fig. 4f). A direct comparison of the model objective gradient with the gradient inferred from behavioral data (Fig. 4g; see Methods for details) illustrates the similarity between the structure of learning in each case.

We discovered a strong predictor of learning trajectory in experimental data was the initial magnitude of phasic DA responses to reward delivery (Fig. 3). In the ACTR model formulation, the magnitude of the DA response depends on sustained responses to policy output and transient responses to reward-related sensory input. As simulations initialize with no initial sustained policy output (to match the lack of cued licking in naive mice), the predicted DA signal is set by the initial transient response of the RNN to reward input. Thus, we next asked whether increasing reward input, leading to larger initial DA signals, would reproduce the experimentally observed correlation with a reduced sustained component and longer collection latencies at the end of training. Indeed, for 6 distinct network initializations we found that paired comparisons of weaker and stronger reward input strength at initialization of training was sufficient to produce differential predicted DA responses at initiation of training (Fig. 4h) and this difference predicted a delayed collection latency at the end of training (Fig. 4h-I; r=0.73, p=0.007) due to a reduced sustained licking policy (Fig. 4j).

### *Specific predictions of the ACTR model for closed-loop manipulation of mDA reward signals*

There is thus a compelling correspondence between the ACTR model and the detailed structure of experimentally-observed individual differences. But a challenging aspect of understanding dopamine as a feedback signal is that mDA activity both represents the state of learning and functions in the progression of learning. Carefully designed manipulation experiments are required to tease apart representation from function in such feedback systems. We next explain a set of experimental predictions specific to ACTR that provide a distinguishing test of the model.

ACTR (and in general any policy learning model) evaluates whether a given change in parameter values led to better or worse performance. The mDA-like signal ($\beta$) uses policy network output to adaptively determine learning rate, *i.e. how fast* to change (Fig. 4b). The actual *sign* of the change (towards or away from the new policy specification) is determined independently of mDA by the performance error (PE) (Fig. 4b). This dissociation between the rate and the sign of the update in the ACTR learning rule can lead to surprising predictions about the effect of phasic mDA activity on learning. For example, on trials in which collection latencies get longer relative to recent performance, the sign of the policy update directs away from the current parameterization. Enhancing the phasic mDA-like signal at reward (increasing the learning rate) on such a trial would then bias *away* from that policy in the future (resulting in "less" of the associated behavior). Thus depending on the *sign* of the underlying policy gradient, a large mDA-like signal can have an effect that is the opposite of simply 'reinforcing' a preceding action.

In the case of the ACTR model this effect is true by construction; however, it raises the question of how a realizable experiment might be designed to test this prediction of the model. Specifically, the challenge is that an animal's underlying policy is not directly observed by the experimenter. However, we know that the presence or absence of licking on a given trial of behavior is a (noisy) reflection of the underlying behavioral policy. As a result it is possible that using an experimentally accessible observation, licking, could be sufficient. This can be illustrated by classifying trial types: 'lick+' trials in which the agent licked during the delay period, and 'lick−' trials where there was no sustained licking (Fig. 4h).

Reward collection latencies converge to small values as the policy approaches an optimum. However, the stochasticity of the licking plant ensures that some trials with a good policy can result in transiently worse collection latency, *i.e.* a negative performance error. Such trials with a negative performance error are generally lick+ trials that occur relatively late in learning and have sustained licking that happens to terminate prior to reward delivery. Selectively enhancing learning rates on these trials with a negative performance error can have the effect of pushing the model away from a policy with sustained licking especially by reducing transient response components. In contrast, lick− trials are much more likely to be trials in which the sustained policy (generally early in learning) was low and thus only positive signed performance errors can occur. Thus, enhancing learning rates in these trial types generally acts to push away from a low sustained policy (essentially exaggerating the adaptive component of the ACTR learning rule that produces sustained licking) and thus constitute a good matched comparison. As expected, ACTR simulations of trial-type dependent enhancement of learning rate indeed produced opposite signed effects on sustained licking behavior for multiple model initializations (n=3, Fig. 4i). These analyses indicate two experimentally-tractable tests of the relevance of the ACTR model to phasic mDA activity (Fig. 4h) and function (Fig. 4i).

### *Calibrated manipulation of DA reward signals in closed-loop with cued behavior*

The ACTR learning model suggested the surprising possibility that larger NAc-DA reward signals could bias learning away from the sustained component of behavior that preceded reward delivery if (and only if) NAc-DA signals were selectively elevated following robust sustained behavior. This is in stark contrast with both qualitative and formal models in which DA reward signals act as a positive feedback signal to reinforce preceding behavior and thus provides a compelling prediction of the ACTR modelling framework. We first examined whether, during learning in control animals, there was evidence of differential NAc-DA reward signaling on trials with sustained licking (lick+) versus trials without sustained licking (lick−) (Fig. 5a). As mice learned in our task, they gradually increased the probability of licking during the delay period on each trial, with the top half of performers in terms of collection latency (Fig. 5b) reaching a higher probability of licking with less training compared to the bottom half of performers (Fig. 5b). As training proceeded, the top performing mice exhibited larger NAc-DA reward signals during lick− trials than during lick+ trials, as in trials 400-800 there was a significant correlation between mice's lick+/lick− differential and their final reward collection latency (r=0.71, p=0.03, n=9) (Fig5c-e). Thus, as mice became better performers, their NAc-DA reward signals on lick− vs lick+ resembled those predicted by the ACTR model (Fig. 4h).

The ACTR model (Fig. 4i) and the above observational data (Fig. 3, 5) make two further surprising predictions. First, exogenous NAc-DA stimulation contingent on cued licking behavior should bias away from the contingent behavior. Second, augmenting NAc-DA reward signals will not directly translate into larger NAc-DA cue responses, rather DA cue responses will follow the level of sustained cued behavior. We tested these predictions by selectively increasing DA reward signals through optogenetic stimulation in the VTA contingent upon sustained cued behavior. Separate groups of animals experienced each of the following stimulation contingencies: "stimLick+" animals received VTA-DA stimulation at the moment of reward delivery on trials in which we detected licking in the 750 ms preceding reward delivery, while "stimLick−" animals received the same stimulation on trials in which no licking was detected during the delay interval (Fig. 5f-g). Crucially, stimulation was brief (150 ms) and calibrated to endogenous fiber photometry signals in each mouse to approximately double the endogenous NAc-DA reward response (Fig. 5g, see Methods, [11,46]). In order to account for the large discrepancy in stimulated trials that would arise between the two stimulation groups due to eventual predomination of lick+ trials (Fig. 5b), stimLick+ animals were limited to having a max of 50% of total trials stimulated in a given session. This resulted in a comparable number of stimulated trials between the two groups by the end of the training period (Fig. 5g, lower right).

Calibrated enhancement of reward-related activity in VTA→NAc-DA projections in this way had opposite effects on emerging delay-period behavior across the two stimLick contingencies. As in the ACTR model, behavior was biased in opposite directions for each contingency, with stimLick+ animals exhibiting lower and stimLick− animals exhibiting higher sustained licking ((trials 600-800, stimLick+ 1.0 ± 0.7, stimLick- 0.6 ± 0.1, ANOVA $F_{1,72} = 10.5$, $p = 0.002$)). Furthermore, NAc-DA cue signals were biased in matching directions, with the stimLick− group also exhibiting higher NAc-DA cue responses vs stimLick+ (trials 600-800, stimLick+ 0.3 ± 0.1 z, stimLick- 2.6 ± 0.7 z, ANOVA $F_{1,72} = 10.1$, $p = 0.002$). These contingent stimulation experiments confirmed predictions from the ACTR model and provided causal evidence supporting the observational correlations between NAc-DA reward signals and individual differences in learning across mice.

In control animals NAc-DA cue response magnitude was correlated with sustained behavior at the end of training (Sup. Fig. 3a), and the effects of closed-loop stimulation on NAc-DA cue responses matched the direction of the effects on licking behavior. As noted above our modeling suggests that NAc-DA cue responses are correlates that reflect underlying differences in policy but not an immediate cause for the emission of sustained behavior. At the same time, work exploring direct roles of DA in movement [21,58] or motivation [10,59] could suggest that these cue signals function in generating or invigorating sustained behavior. In a subset of mice, at the end of regular training we included an extra session in which we paired reward-calibrated VTA-DA neuron stimulation with cue presentation on a random subset of trials (Sup. Fig. 3b-d). Increasing VTA-DA cue responses in this way had no effect on cued licking in the concurrent trial (control 2.3 ± 1.1 Hz, stim. 2.3 ± 1.0 Hz, $p > 0.99$). Thus within this context and the observed range of NAc-DA cue responses, the magnitude of NAc-DA cue signals is a correlate of learned changes in behavioral policy that does not directly regulate sustained behavior in anticipation of reward delivery [11].

### *Large dopamine manipulations drive value-like learning*

Above we describe a set of results that are difficult to reconcile with pure value or action-value learning accounts, in which DA reward signals should promote the emergence of DA cue signals or directly reinforce contingent behavior. Moreover, we articulate a computational model of biologically plausible, direct policy learning (ACTR) that is consistent with observed data and predicted exogenous stimulation effects. At the same time, seminal results in rodents [12] and monkeys [13] make specific and well-supported arguments for value-like learning effects following exogenous VTA-DA stimulation. It is possible that the novel closed-loop stimulation design used here may explain some differences. However, another technical detail that differentiates the current study is that our VTA-DA stimulation was designed to match the brief (~200 ms) reward-related mDA bursts reported across species [5], and calibrated to the dynamic range of NAc-DA signals measured within the same context. We next explored whether a more intense and sustained stimulation could induce effects more compatible with the predictions of value learning.

We introduced mice used in Figures 1-5 to a novel sensory cue - a 500-ms flash of visible light directed at the chamber wall in front of the mouse. After 10 introductory trials, this visible cue stimulus was paired with exogenous VTA-DA stimulation after 1 s delay for 5 daily sessions (~150 trials per session). One group of randomly selected mice received VTA-DA stimulation (150 ms at 30 Hz and 1-3 mW steady-state power) calibrated to uncued reward responses (stim response = 1.4 ± 0.3 uncued reward response, n=10), while the complement received larger, uncalibrated stimulations (500 ms at 30 Hz and 10 mW steady-state power, stim response = 5.5 ± 0.8 uncued reward response) (Fig. 6b). After 5 sessions, we found that the group receiving calibrated, reward-sized stimulation did not exhibit NAc-DA cue responses above baseline (0.0 ± 0.2 z, p = 0.8), whereas the large, uncalibrated stimulation group exhibited substantial NAc-DA cue responses (0.5 ± 0.2 z, p = 0.02).

The qualitative differences in effects of calibrated and uncalibrated stimulations suggests that uncalibrated stimulation could counteract the suppression of cued licking seen for calibrated stimulation in stimLick+ animals (Fig. 5h, i). To test this possibility we repeated the closed loop stimLick+ experiment (Fig. 5) with a new set of mice, but this time augmented rewards with large, uncalibrated VTA-DA stimulation (500 ms, at 30 Hz and ~10 mW power) on trials in which mice licked during the delay period. Indeed, with this new large exogenous stimulation, the stimLick+ contingency now resulted in increased NAc-DA cue responses within 600 trials (2-way ANOVA, stim group $F_{1,66}$, p = 0.001) as well as increased cued licking (2-way ANOVA, stim group $F_{1,60}$, p =0.01), essentially reversing the sign of the effects of calibrated stimLick+ stimulation (Fig. 5f-i).

## Discussion

It has been proposed that animal learning and its neural basis can be understood in terms of cost functions, circuit architecture, and learning rules - the core computational elements of machine learning [60]. Here we apply this approach to a canonical animal learning paradigm, classical trace conditioning (see Logic Outline, Sup. Fig. 1). We define a cost function which drives learning to minimize the latency to reward collection by exploiting the presence of a predictive cue. We propose a functional circuit architecture in which mDA activity integrates reward-predictive sensory input and current behavioral policy, consistent with known anatomy of forebrain projections and functional data on phasic activity of mDA neurons [24]. Finally, we demonstrate that this circuit architecture, combined with a recurrent neural network and a biologically plausible learning rule (ACTR) is sufficient to optimize reward collection. The formulation of associative learning as a form of direct policy learning has the expressive power to capture individual differences in learning trajectories and replicates canonical neural correlates observed in mDA neurons.

The discovery that the phasic activity of mDA neurons in several species correlated with core predictions of value learning algorithms, in particular TD, has been a dramatic and important advance [5,6,61]. At the same time, reinforcement learning constitutes a large class of methods that include many additional functions for learning about states in the environment and policies for behavioral control. Expanding the space of models for learning about states has yielded recent gains in explanatory power [62–64]. A significant advance of our work is to demonstrate a formulation of policy learning that is consistent not just with signals observed in mDA neuron activity, but also with many aspects of behavior during novel classical conditioning. We speculate that reinforcement learning in other situations in which anticipatory, approach, or operant behavior can be shown to minimize performance errors can also be explained as a descent along a gradient of policy evaluation. Indeed, this space of alternative formulations of classic learning paradigms is beginning to be considered [65,66]. Much evidence shows that implicit predictions and feedback about behavioral performance are fundamental to the central production of action [67]. Further research should clarify how dopamine-recipient circuits 1) represent aspects of an animal's policy [68] and 2) compute performance errors matched to an animal's current objectives [44].

We replicate many previous findings that across learning mesolimbic DA cue responses correlate with inferred value learning and reward responses correlate with RPEs after training (Fig. 2). However, at the level of the individual animals we made several observations that would be surprising in the context of value learning. First, the magnitude of DA signaling at the reward was correlated with behavioral evidence of learning but not the emergence of DA cue signals (as suggested by other recent results [69,70]). Second, the initial magnitude of DA responses to reward was *anti*-correlated with the emergence of sustained cued-behavior (Fig. 3). Lastly, exogenous stimulation of VTA-DA neurons did not necessarily increase DA signals at a predictive cue (Fig. 5), rather such effects depended on DA-independent learning or a magnitude of DA stimulation that far exceeded reward signals measured in our task (Fig. 6).

These results suggest that the dopamine-dependent attribution of motivational value to cues [12,71–73] is at least partially dissociable from the regulation of policy learning within the same mesolimbic circuits, suggesting that they may be complimentary building blocks for animal learning [28].

It may be that in other contexts the threshold activation of DA for inducing value learning is lower due to local circuit conditions, or phasic DA signaling is larger such that physiological dopamine signals are sufficient to support value learning. Such flexibility could support exploration of the policy space when appropriate (as in the naive conditions studied here), while promoting value learning in the conditions where it has been studied predominantly - in well-trained animals deciding between discrete options in the environment. Value effects following higher power, longer simulations may depend on specific receptor recruitment within a circuit [74], and/or recruitment of a wider population of DA circuits, as a spectrum of DA function and signaling has been demonstrated across striatal subregions [15,18,69,75–78]. Future experiments exploring a putative threshold for phasic DA-driven value learning may yield insight into both adaptive and maladaptive functions of dopamine-recipient circuits in the forebrain. Drugs of abuse can enhance DA signaling, recruiting dopaminergic value learning as in our large uncalibrated stimulations (Fig. 6) , or leading to suboptimal policy learning (Fig. 3) -- both intriguing perspectives on the mixtures of adaptive and maladaptive learning observed in the context of addiction [79,80]. Finally, our results underscore the importance of matching exogenous mDA stimulation to measured signals, and support the idea that extended, high-magnitude mDA stimulation is an important model of addiction [81] that is dissociable from natural learning about rewards.

Recent success in artificial intelligence has been achieved by direct learning of policy in parallel to learning about states of the environment [3,82]. The variable optimization path of direct policy methods [30] is consistent with the meaningfully variable learning trajectories of individual animals (Fig. 1). By examining the diverse learning trajectories of individual mice, we discovered that individual variability in the naive response of mesolimbic mDA neurons to reward delivery predicted the variable quality of performance in individual trained mice assessed hundreds of trials later (Fig. 3). This relationship was non-intuitive; enhanced mDA responses were associated with worse learning (slower reward collection and less cued behavior). The ACTR model provided insight into how this surprisingly inverted relation could arise. A key idea in the ACTR model is that the sign of policy updates is determined by performance errors from the behavioral policy. The (unsigned) rate at which the policy is updated is determined by feedback from the (derivative of the) current policy output. We show that mesolimbic mDA activity is quantitatively well predicted by the activity of this feedback unit which controls adaptive learning rate. As a consequence, even in naive animals the reward response of mDA neurons (or the feedback unit of the ACTR model) reflects the initial state of the policy network and is thus predictive of the final policy obtained after learning as observed both experimentally (Fig. 3d) and in the ACTR model (Fig. 4h). In addition to predicting activity of mDA neurons, insights from the model also allowed us to predict two key experimental tests (Fig. 4k, l): the magnitude of mDA activity depends upon the current policy and exogenous stimulation of DA neurons can either enhance or impair learning contingent upon the policy when stimulated; both of which were confirmed (Fig. 5).

There are many opportunities to extend the current model formulation, in particular the recurrent neural network component, to capture more biological reality and evaluate the biologically plausible, but currently incompletely tested, cellular and circuit mechanisms of its learning rule in greater detail. Given that adaptive control over the magnitude of learning rate can be a key determinant of machine learning performance (*e.g.* proximal policy optimization [2]), studying how adaptive control of learning rates are implemented in animals, and especially across diverse tasks, may provide additional algorithmic insights to those developed here. Our work argues that future efforts to explain learning in biological systems that control continuous action, like the basal ganglia [83], should continue to explore a larger space of reinforcement learning algorithms by incorporating direct policy learning driven by performance errors.

## Methods

**Animals.** All procedures and animal handling were performed in strict accordance with protocols (11-39) that were approved by the Institutional Animal Care and Use Committee (IACUC) and consistent with the standards set forth by the Association for Assessment and Accreditation of Laboratory Animal Care (AALAC). For behavior and juxtacellular recordings we used 24 adult DAT-Cre::ai32 mice (3-9 months old) resulting from the cross of DAT$^{IRES\textit{cre}}$ (The Jackson Laboratory stock 006660) and Ai32 (The Jackson Laboratory stock 012569) lines of mice, such that a Chr2/EYFP fusion protein was expressed under control of the endogenous dopamine transporter Slc6a3 locus to specifically label dopaminergic neurons. Animals were housed on a 12-hour dark/light cycle (8am-8pm) and recording sessions were all done between 9am-3pm. Following at least 4 days recovery from headcap implantation surgery, animals' water consumption was restricted to 1.2 mL per day for at least 3 days before training. Mice underwent daily health checks, and water restriction was eased if mice fell below 75% of their original body weight.

**Behavioral training.** Mice were habituated to head fixation in a separate area from the recording rig in multiple sessions of increasing length over >= 3 days. During this time they received some manual water administration through a syringe. Mice were then habituated to head fixation while resting in a spring-suspended basket in the recording rig for at least two 30+ minute sessions before training commenced. No liquid rewards were administered during this recording rig acclimation, thus trial 1 in the data represents the first time naive mice received the liquid water reward in the training environment. The reward consisted of 3 μL of water sweetened with the non-caloric sweetener saccharin delivered through a lick port under control of a solenoid. A 0.5 s, 10 kHz tone preceded reward delivery by 1.5 s on "cued" trials, while 10% of randomly selected rewards were "uncued". Matching our previous training schedule [11], after three sessions, mice also experienced "omission" probe trials, in which the cue was delivered by not followed by reward, on 10% of randomly selected trials. Intertrial intervals were chosen from randomly permuted exponential distribution with a mean of ~25 seconds. Ambient room noise was 50-55 dB, while an audible click of ~53 dB attended solenoid opening upon water delivery and the predictive tone was ~65 dB loud. Mice experienced 100 trials per session and one session per day for 8-10 days. In previous pilot experiments, it was observed that at similar intertrial intervals, behavioral responses to cues and rewards began to decrease in some mice at 150-200 trials. Thus the 100 trial/session limit was chosen to ensure homogeneity in motivated engagement across the dataset.

Some animals received optogenetic stimulation of VTA-DA neurons concurrent with reward delivery, contingent on their behavior during the delay period (see technical details below). Following trace conditioning with or without exogenous DA stimulation, 5 mice experienced an extra session during which VTA-DA neurons were optogenetically stimulated concurrently with cue presentation (Sup. Fig. 3). Mice were then randomly assigned to groups for a new experiment in which a light cue predicted VTA-DA stimulation with no concurrent liquid water reward (5-7 days, 150-200 trials per day). The light cue consisted of a 500 ms flash of a blue LED directed at the wall in front of head fixation. Intertrial intervals were chosen from randomly permuted exponential distributions with a mean of ~13 seconds. Supplementary Table 1 lists the experimental groups each mouse was assigned to in the order in which experiments were experienced.

|  | trace conditioning group |  |  |  | cued VTA-DA stimulation |  |
|---|---|---|---|---|---|---|
|  | (Fig. 1-3, 4a-d) |  |  | (Sup Fig. 3) | (Fig. 6a-d) |  |
| mouse | control | stimLick+ | stimLick- | stim. at cue | reg. stim. | large stim. |
| 1 | x |  |  |  | x |  |
| 2 | x |  |  |  |  |  |
| 3 |  |  | x |  |  | x |
| 4 | x |  |  |  |  | x |
| 5 |  |  | x |  | x |  |
| 6 | x |  |  |  |  | x |
| 7 |  | x |  |  |  | x |
| 8 |  | x |  |  | x |  |
| 9 | x |  |  |  | x |  |
| 10 |  |  | x |  | x |  |
| 11 | x |  |  |  |  | x |
| 12 |  | x |  |  | x |  |
| 13 |  | x |  | x |  | x |
| 14 |  |  | x |  |  | x |
| 15 | x |  |  | x | x |  |
| 16 | x |  |  | x | x |  |
| 17 |  | x |  | x | x |  |
| 18 |  |  | x |  | x |  |
| 19 | x |  |  |  |  |  |
| 20 |  |  | x | x | x |  |

**Supplementary Table 1.** Subsequent experimental groups for each mouse.

**Video and behavioral measurement.** Face video was captured at 100 Hz continuously across each session with a single camera (Flea 3, FLIR) positioned level with the point of head fixation, at a ~30º angle from horizontal. Dim visible light was maintained in the rig so that pupils were not overly dilated, while an infrared LED (model#) trained at the face provided illumination for video capture. Video was post-processed with custom matlab code (available at: www.github.com/).

Briefly, for each session, a rectangular region of interest (ROI) for each measurement was defined from the mean of 500 randomly drawn frames. Pupil diameter was estimated as the mean of the major and minor axis of the object detected with the MATLAB 'regionprops' function, following noise removal by thresholding the image to separate light and dark pixels, then applying a circular averaging filter and then dilating and eroding the image. This noise removal process accounted for frames distorted by passage of whiskers in front of the eye, and slight differences in face illumination between mice. For each session, appropriateness of fit was verified by overlaying the estimated pupil on the actual image for ~20-50 randomly drawn frames. A single variable, the dark/light pixel thresholding value, could be changed to ensure optimal fitting for each session. Nose motion was extracted as the mean of pixel displacement in the ROI Y-axis estimated using an image registration algorithm (MATLAB ‘imregdemons’). Whisker pad motion was estimated as the absolute difference in the whisker pad ROI between frames (MATLAB ‘imabsdiff’; this was sufficiently accurate to define whisking periods, and required much less computing time than ‘imregdemons’). Whisking was determined as the crossing of pad motion above a threshold, and whisking bouts were made continuous by convolving pad motion with a smoothing kernel. Licks were timestamped as the moment pixel intensity in the ROI in between the face and the lick port crossed a threshold.

Body movement was summarized as basket movements recorded by a triple-axis accelerometer (Adafruit, ADXL335) attached to the underside of a custom-designed 3D-printed basket suspended from springs (Century Spring Corp, ZZ3-36). Relative basket position was tracked by low-pass filtering accelerometer data at 2.5 Hz. Stimulations and cue deliveries were coordinated with custom-written software using Arduino Mega hardware (www.arduino.cc). All measurement and control signals were synchronously recorded and digitized (at 1 kHz for behavioral data, 10 kHz for fiber photometry data) with a Cerebus Signal Processor (Blackrock Microsystems). Data was analyzed using Matlab software (Mathworks).

**Sustained and transient measures and abstract learning trajectories.** To describe the relationship between behavioral adaptations and reward collection performance, for each mouse in the control group a generalized linear model (GLM) was created to predict reward collection latency from sustained and transient predictor variables on each trial. Sustained changes in licking, whisking, body movement, and pupil diameter were quantified by measuring the average of each of those signals during the 1 s delay period preceding cued rewards. The nose motion signal was not included as it did not display consistent sustained changes. Transient responses in the whisking, nose motion, and body movement were measured as the latency to the first response following reward delivery. For whisking, this was simply the first moment of whisking following reward delivery. For nose motion, the raw signal was convolved with a smoothing kernel and then the first response was detected as a threshold crossing of the cumulative sum of the signal. For body movement, the response was detected as the first peak in the data following reward delivery. On occasional trials no event was detected within the analysis window. Additionally, discrete blocks of trials were lost due to data collection error for mouse 3-session 7, mouse4-session 5, and mouse9-session 4. In order to fit learning curves through these absent data points, missing trials were filled in using nearest neighbor interpolation.

Trial-by-trial reward collection latencies and predictor variables were median filtered (MATLAB ‘medfilt1(signal,10)’) in order to minimize trial-to-trial variance in favor of variance due to learning across training. After z-scoring the predictor variables, collection latency was then predicted according to the following equation:

β values were fit using MATLAB ‘glmfit’. The unique explained variance of each predictor was calculated as the difference in explained variance between the full model and a partial model in which β values were fit without using that predictor.

Sustained and transient predictor variables were used to define abstract learning trajectories which were plots of collection latency against the inferred transient and sustained variables for each of the first 800 cue-reward trials of training. Transient and sustained variables were calculated as the first principal component of the individual transient and sustained variables used in the GLM fits. For visualization we fit a parametric model to all 3 variables (single exponential for sustained, double exponentials for transient and latency using MATLAB 'fit' function). Quality of fits and choice of model were verified by visual inspection of all data for all mice. An individual mouse's trajectory was then visualized by plotting downsampled versions of the fit functions for latency, transient and sustained. Arrowheads were placed at logarithmically spaced trials.

In order to quantify the total amount of sustained behavior in each mouse at a given point in training ("sustained", Fig. 3c), each sustained measure (pupil, licking, whisking, body movement) was z-scored and combined across mice into a single data matrix. The first principal component of this matrix was calculated and loading onto PC1 was defined as a measure of an inferred underlying 'sustained' component of the behavioral policy. This created an equally weighted, variance-normalized combination of all sustained measures to allow comparisons between individual mice. An analogous method was used to reduce the dimensionality of transient variables down to a single 'transient' dimension that captures the majority of variance in transient behavioral variables across animals. Initial NAc-DA signals were predicted from trained behavior at trials 700-800 by multiple regression (specifically, pseudoinverse of the data matrix of transient and sustained variables at the end of training multiplied by data matrix of physiological signals for all animals).

**Combined fiber photometry and optogenetic stimulation.** In the course of a single surgery session, DAT-Cre::ai32 mice received:
1) Bilateral injections of AAV2/1-CAG-FLEX-jRCaMP1b in the VTA (150 nL at the coordinates -3.1 mm A/P, 1.3 mm M/L from bregma, at depths of 4.6 and 4.3 mm) or in the SNc (100 nL at the coordinates -3.2 mm A/P, 0.5 mm M/L, depth of 4.1,  mm).
2) Custom 0.39 NA, 200 μm fiber cannulas implanted bilaterally above the VTA (-3.2 mm A/P, 0.5 mm M/L, depth of -4.1 mm).
3) Fiber cannula implanted unilaterally in the dorsomedial striatum (DS; 0.9 mm A/P, 1.5 mm M/L, depth of 2.5 mm) and nucleus accumbens core (NAc; 1.2 mm A/P, 0.85 mm M/L, depth of 4.3 mm). Hemisphere choice was counterbalanced across individuals. A detailed description of the methods has been published [46].

Imaging began >20 days post-injections using custom-built fiber photometry systems (Fig. 2a)[46]. Two parallel excitation-emission channels through a 5-port filter cube (FMC5, Doric Lenses) allowed for simultaneous measurement of RCaMP1b and eYFP fluorescence, the latter channel having the purpose of controlling for the presence of movement artifacts. 470 nm and 565 nm fiber-coupled LEDs (M470F3, M565F3, Thorlabs) were connected to excitation ports with acceptance bandwidths of 465-490 nm and 555-570 nm respectively with 200 μm, 0.22 NA fibers (Doric Lenses). Light was conveyed between the sample port of the cube and the animal by a 200 μm core, 0.39 NA fiber (Doric Lenses) terminating in a ceramic ferrule that was connected to the implanted fiber cannula by a ceramic mating sleeve (ADAL1, Thorlabs) using index matching gel to improve coupling efficiency (G608N3, Thorlabs). Light collected from the sample fiber was measured at separate output ports (emission bandwidths 500-540 nm and 600-680 nm) by 600 μm core, 0.48 NA fibers (Doric Lenses) connected to silicon photoreceivers (2151, Newport).

A time-division multiplexing strategy was used in which LEDs were controlled at a frequency of 100 Hz (1 ms on, 10 ms off), offset from each other to avoid crosstalk between channels. A Y-cable split each LED output between the filter cube and a photodetector to measure output power. LED output power was 50-80 μW. This low power combined with the 10% duty cycle used for multiplexing, prevented local ChR2 excitation [46] by 473 nm eYFP excitation. Excitation-specific signals were recovered in post-processing by only keeping data from each channel when its LED output power was high. Data was downsampled to 100 Hz, then band-pass filtered between 0.01 and 40 Hz with a 2nd-order Butterworth filter. Though movement artifacts were negligible when mice were head-fixed in the rig (the moveable basket was designed to minimize brain movement with respect to the skull [11]), according to standard procedure the least squares fit of the eYFP movement artifact signal was subtracted from the jRCaMP1b signal. dF/F was calculated by dividing the raw signal by a baseline defined as the polynomial trend (MATLAB 'detrend') across the entire session. This preserved local slow signal changes while correcting for photobleaching. Comparisons between mice were done using the z-scored dF/F.

Analysis windows were chosen to capture the extent of mean phasic activations following each kind of stimulus. For NAc-DA and VTA-DA, reward responses were quantified from 0-2 s after reward delivery and cue responses from 0-1 s after cue delivery. DS-DA exhibited significantly faster kinetics, and thus reward and cue responses were quantified from 0 to 0.75 s after delivery.

Somatic Chr2 excitation was performed with a 473 nm laser (50mW, OEM Laser Systems) coupled by a branching fiber patch cord (200 µm, Doric Lenses) to the VTA-implanted fibers using ceramic mating sleeves. 30 Hz burst activations (10 ms on, 23 ms off) were delivered with durations of either 150 ms for calibrated stimulation or 500 ms for large stimulations. For calibrated stimulation, laser power was set between 1-3 mW (steady state output) in order to produce a NAc-DA transient of similar amplitude to the largest transients observed during the first several trials of the session. This was confirmed during analysis to have roughly doubled the size of reward-related NAc-DA transients (Fig. 5g). For large stimulations, steady state laser output was set to 10 mW.

## Computational learning model: ACTR

***Behavioral plant.*** An important aspect of this modeling work was to create a generative agent model that would produce core aspects of reward-seeking behavior in mice. To this end we focused on licking, which in the context of this task is the unique aspect of behavior critical for reward collection. A reader may look at the function *dlRNN_Pcheck_transfer.m* within the software repository to appreciate the structure of the plant model. We describe the function of the plant briefly here. It is well known that during consumptive, repetitive licking mice exhibit sustained periods of ~7Hz licking. This we modeled as a simple fixed rate process from a 'lick' state that emitted observed licks at a fixed time interval of 150 ms. The onset of this lick pattern relative to entry into the lick state was modeled as starting at a random phase. We used the simplest possible model in which behavior consisted of two states 'rest' and 'lick' with stochastic transitions between states governed by forward and backward transition rates. The backward transition rate was a constant that depended upon the presence of reward {5e-3 ms without reward, 5e-1 ms with reward}. This change in the backwards rate captured the approximate duration of consumptive licking. The forward rate was governed by the scaled RNN output (see below) and a background tendency to transition to licking as a function of trial time (analogous to an exponential rising hazard function; $\tau$=200ms). The output unit of the RNN was constrained to {-1,1} by the tanh activation function and scaled by $S$=0.02/ms to convert to a transition rate. Behavior of the plant for a range of policies is illustrated in Fig. 4a. A large range of parameterizations were explored with qualitatively similar results. Chosen parameters were arrived at by scanning many different simulations and matching average initial and final latencies for cue-reward pairings across the population of animals. More complicated versions (high-pass filtered, non-linear scaling) of the transition from RNN output to transition rate can be explored in the provided function. However, all transformations were found to produce qualitatively similar results and thus the simplest (scalar) transformation was chosen for reported simulations for clarity of presentation.

***Recurrent neural network (RNN).*** As noted in the main text the RNN component of the model and the learning rules used for training drew upon inspiration from [49] that itself drew upon inspiration variants of node perturbation methods [50] and the classic policy optimization methods known as 'REINFORCE' rules [1,35]. Briefly, [49] demonstrated that a relatively simple learning rule that computed a nonlinear function of the correlation between a change input and change in output multiplied by the change in performance on the objective was sufficiently correlated with the analytic gradient to allow efficient training of the RNN. This uses a version of node perturbation to explore potential weight changes within the network. Below we delve into the learning rule as implemented here or a reader may examine the commented open source code to get further clarification as well. First, we describe the structure of the RNN and some core aspects of its function in the context of the model. The RNN was constructed largely as described in [49] and was very comparable to the structure of a re-implementation of that model in [84]. In our implementation we add a few novel features (the mDA-like feedback component, distinct manner in which to calculate the learning) that we will discuss.

Although we explored a range of parameters governing RNN construction, many examples of which are shown in Sup. Fig. 2, the simulations shown in the main results come from a network with 50 units ($N_u$=50; chosen for simulation efficiency, larger networks were explored extensively as well), densely connected ($P_c$=0.9), spectral scaling to produce sustained dynamics ($\boldsymbol{g}$=1.3), a characteristic time constant ($\boldsymbol{\tau}$=25ms), and a standard tanh activation function for individual units. Initial internal weights of the network ($W_{ij}$) were assigned according to the equation:

$$W_{ij} = \boldsymbol{g} * \mathcal{N}(0,1) * (P_c * N_u)^{-1/2} \qquad \text{(1) line 351 } \textit{RNN-dudlab-master-LearnDA.m}$$

The RNN had a single primary output unit with activity that constituted the continuous time policy (*i.e.* $\pi(t)$) input to the behavior plant (see above), and a 'feedback' unit that did not project back into the network as would be standard, but rather was used to produce adaptive changes in the learning rate (described in more detail in "Learning rules" section below).

***Objective function.*** Evaluation of model performance was calculated according to an objective function that defines the cost as the weighted sum of a performance cost (2, "$\text{cost}_P$") and a network stability cost (3, "$\text{cost}_N$").

$$\text{cost}_P = 1 - e^{-\Delta t/500} \qquad (2)$$

$$\text{cost}_N = \text{sum}(\ |\ \boldsymbol{\delta\pi}(t)/\boldsymbol{\delta}t\ |\ ) \qquad (3)$$

$$R_{obj} = \text{cost}_P + w_N * \text{cost}_N \qquad \text{(4) } \textit{e.g.} \text{ line 209 in } \textit{dlRNN-train\_learnDA.m}$$

$$\langle R(T) \rangle = \boldsymbol{\alpha}_R * R_{obj}(T) + (1-\boldsymbol{\alpha}_R) * R_{obj}(T-1) \qquad \text{(5) } \textit{e.g.} \text{ line 323 in } \textit{dlRNN-train\_learnDA.m}$$

Where $T$ is the trial index. In all presented simulations, $w_N$=0.25. A filtered average cost, **R**, was computed as before [49] with $\boldsymbol{\alpha}_R$=0.75 and used in the update equation for changing network weights via the learning rule described below. For all constants ($\boldsymbol{\alpha}_R$ and $w_N$) a range of values were tried with qualitatively similar results. The performance objective was defined by $\text{cost}_P$ where $\Delta t$ is the latency to collected reward after it is available. The network stability cost ($\text{cost}_N$) penalizes high-frequency oscillatory dynamics that can emerge in some (but not all) simulations. Such oscillations are inconsistent with observed dynamics of neural activity to date.

***Identifying properties of RNN required for optimal performance.*** In order to examine what properties of the RNN were required for optimal performance, we scanned through thousands of simulated network configurations (random initializations of $W_{ij}$) and ranked those networks according to their mean cost ($R_{obj}$) when run through the behavior plant for 50 trials (an illustrative group of such simulations is shown in Sup. Fig. 2). This analysis revealed a few key aspects of the RNN required for optimality. First, a sustained policy that spans time from the detection of the cue through the delivery of water reward minimizes latency cost. Second, while optimal RNNs are relatively indifferent to some parameters (*e.g.* $P_c$) they tend to require a coupling coefficient (g) $\geqq$1.2. This range of values for the coupling coefficient is known to determine the capacity of a RNN to develop sustained dynamics [85]. Consistent with this interpretation we found that optimal policies were observed uniquely in RNNs with large leading eigenvalues (Sup. Fig. 2; *i.e.* long time constant dynamics [86]). These analyses define the optimal policy as one that requires sustained dynamics of output unit activity that span the interval between the cue offset and reward delivery and further reveal that an RNN with long timescale dynamics is required to realize such a policy. Intuitively: sustained anticipatory behavior, or "conditioned responding", optimizes reward collection latency. If an agent is already licking when reward is delivered the latency to collect that reward is minimized.

***RNN Initialization for simulations.*** All mice tested in our experiments began training with no sustained licking to cues and a long latency (~1 second or more) to collect water rewards. This indicates that animal behavior is consistent with an RNN initialization that has a policy $\pi$(t)~0 for the entire trial. As noted above there are many random initializations of the RNN that can produce clear sustained behavior and even optimal performance. Thus, we performed large searches of RNN initializations (random matrices $W_{ij}$) and used only those that had ~0 average activity in the output unit. We used a variety of different initializations across the simulations reported in Fig. 4 and indeed there can be substantial differences in the observed rate of convergence depending upon initial conditions (as there are across mice as well). For simulations of individual differences in Fig. 4h-j 6 distinct network initializations were chosen (as described above) and paired comparisons were made for the control initialization and an initialization in which the weights of the inputs from the reward to the internal RNN units were tripled.

***Learning rules.*** Below we articulate how each aspect of the model acronym, ACTR (<u>A</u>daptive rate <u>C</u>ost of performance <u>t</u>o <u>R</u>EINFORCE), is reflected in the learning rule that governs updates to the RNN. The connections between the variant of node perturbation used here and REINFORCE [35] has been discussed in detail previously [49]. There are two key classes of weight changes governed by distinct learning rules within the ACTR model. First, we will discuss the learning that governs changes in the 'internal' weights of the RNN ($W_{ij}$). The idea of the rule is to use perturbations (1-10Hz rate of perturbations in each unit; simulations reported used 3Hz) to drive fluctuations in activity and corresponding changes in the output unit that could improve or degrade performance. To solve the temporal credit assignment problem we used eligibility traces similar to those described previously [49]. One difference here was that the eligibility trace decayed exponentially with a time constant of 500 ms and it was unclear whether decay was a feature of prior work. The eligibility trace ($\mathscr{e}$) for a given connection *i,j* could be changed at any time point by computing a nonlinear function ($\mathscr{S}$) of the product of the derivative in the input from the *i*th unit ($\mathscr{x}_i$) and the output rate of the *j*th unit ($\mathscr{r}_j$) in the RNN according to the equation:

$$e_{i,j}(t) = e_{i,j}(t-1) + \mathscr{S}[\ \mathscr{r}_j(t-1)\times(\mathscr{x}_i(t)-\langle\mathscr{x}_i\rangle)\ ] \qquad \text{(5) line 82 in } \textit{dlRNN\_engine.m}$$

As noted by Miconi, the function $\mathscr{S}$ need only be a signed, nonlinear function. Similarly, in our simulations we also found that a range of functions could all be used. Typically, we either used $\mathscr{S}(y)=y^3$ or $\mathscr{S}(y)=|y|*y$ and simulations presented were generally the latter which runs more rapidly.

The change in a connection weight ($W_{ij}$) in the RNN in the original formulation [49] is then computed as the product of the eligibility trace and the change in performance error scaled by a learning rate parameter. Our implementation kept this core aspect of the computation, but several critical updates were made and will be described. First, since the eligibility trace is believed to be 'read out' into a plastic change in the synapse by a phasic burst of dopamine firing [87]. Thus, we chose to evaluate the eligibility at the time of the computed burst of DA firing estimated from the activity of the parallel feedback unit (see below for further details). Again, models that do not use this can also converge, but in general converge worse and less similarly to observed mice. The update equation is thus,

$$W_{i,j}(T) = W_{i,j}(T\text{-}1) + \beta_{DA} \times \eta_W \times e_{i,j}(t_{DA}) \times (R_{obj}(T) - \langle R(T) \rangle) \quad \text{(6) } \textit{e.g.} \text{ line 286 in } \textit{dlRNN-train\_learnDA.m}$$

Where $\eta_W$ is the baseline learning rate parameter and is generally used in the range {1e-5,1e-4} (however, a large range was scanned for the data shown in Fig. 4c); and $\beta_{DA}$ is the 'adaptive rate' parameter that equals the derivative of the feedback unit response, equivalent to the proposed role of DA in modulating learning.

As noted in the description of the behavioral data described in Fig. 1 it is clear that animal behavior exhibits learning of both sustained behavioral responses to the cue as well as transient learning that reduces reaction times between sensory input (either cues or rewards) and motor outputs. This is particularly prominent in early training where a dramatic decrease in reward collection latency occurs even in the absence of particularly large changes in the sustained component of behavior. We interpreted this transient component as a 'direct' sensorimotor transformation consistent with the treatment of reaction times in the literature [88] and thus transient learning updates weights between sensory inputs and the output unit (one specific element of the RNN indexed as 'o' below). This transient learning was also updated according to performance errors. In particular the difference between $R_{obj}(T)$ and the activity of the output unit at the time of reward delivery. For the cue updates were proportional to the difference between the derivative in the output unit activity at the cue and the performance error at the reward delivery. These rates were also scaled by the same $\beta_{DA}$ adaptive learning rate parameter:

$$W_{trans,o}(T) = W_{trans,o}(T\text{-}1) + \beta_{DA} \times \eta_I \times (R_{obj}(T) - \pi(t_{reward})) \quad \text{(7) } \textit{e.g.} \text{ line 299}$$

Where $\eta_I$ is the baseline transient learning rate and typical values were 0.033 in presented simulations.

We compared acquisition learning in the complete model to observed mouse behavior, scanning ~2 orders of magnitude for two critical parameters $\eta_I$ and $\eta_W$ or (1) the baseline rate governing changes in the input weights of sensory inputs, analogous to the transient learning component of observed behavior and (2) the baseline rate governing updates to internal weights ($W_{ij}$) of the RNN, analogous to the sustained component of observed behavior, respectively (Fig. 4c). As described above, simulations were initialized from network configurations that had no sustained licking behavior nor any direct licking in response to cue/reward inputs. As these simulations make clear, a broad range of model parameters lead to convergence upon an optimal policy that minimizes performance cost through learned increases in sustained policy and an enhanced transient component as observed in behavioral data (Fig. 4c, analogous to plots in Fig. 1h).

***Visualizing the objective surface.*** In order to visualize the objective surface that governs learning we scanned a range of combinations of transient and sustained components of an analytic policy passed through the behavior plant. The space of transient components covered was [-0.1 0 $10^{[-1:0.1:1]}$] and the space of sustained values covered was 0.33*[0:0.05:1] corresponding to the range of observed licking behavior in experiments. For each pair of values a policy was computed and passed through the behavior plant 1000 times to get an accurate estimate of the mean performance cost. These simulation results are presented in Fig. 4f.

In the case of experimental data the full distribution of individual trial data points across all mice (N=7200 observations) was used to fit a 2nd order, 2d polynomial (MATLAB; 'fit'). Observed trajectories of sustained vs transient were superimposed on this surface by finding the nearest corresponding point on the fit 2d surface for the parametric sustained and transient trajectories. These data are presented in Fig. 4g.

***Simulating closed-loop stimulation of mDA experiments.*** We sought to develop an experimental test of the model that was tractable (as opposed to inferring the unobserved policy for example). The experimenter in principle has access to real-time detection of licking during the cue-reward interval. In simulations this also can easily be observed by monitoring the output of the behavioral plant. Thus, in the model we kept track of individual trials and the number of licks produced in the cue-reward interval. For analysis experiments (Fig. 4h) we tracked these trials and separately calculated the predicted DA responses depending upon trial type classification. For simulations in Fig. 4h we ran simulations from the same initialization in 9 replicates (matched to the number of control mice) and error bars reflect the standard error.

To simulate calibrated stimulation of mDA neurons, we multiplied the adaptive rate parameter, $\beta_{DA}$, by 2 based upon the same trial classification. Specifically, on trials with no licks detected (stim lick−) or on trials with at least one lick detected (stim lick+). For simulations reported in Fig. 4i we used 3 conditions: control, stim lick−, stim lick+. For each of these 3 conditions we ran 9 simulations (3 different initializations, 3 replicates) for 27 total learning simulations (800 trials). This choice was an attempt to estimate the expected experimental variance since trial classification scheme is an imperfect estimate of underlying policy.

***Code availability.*** All code relating to simulating the ACTR model and for a reader to explore both described parameterizations and explore a number of implemented, but unused in this manuscript, features can be found at https://github.com/dudmanj/RNN_learnDA. Specific line numbers are provided within the code for a subset of critical computations in the model.

**Histology.** Mice were killed by anesthetic overdose (isoflurane, >3%) and perfused with ice-cold phosphate-buffered saline (PBS), followed by paraformaldehyde (4% wt/vol in PBS). Brains were post-fixed for 2 h at 4° C and then rinsed in saline. Whole brains were then sectioned (100 µm thickness) using a vibrating microtome (VT-1200, Leica Microsystems). Fiber tip positions were estimated by referencing standard mouse brain coordinates [89].

**Statistical analysis.** Two-sample, unpaired comparisons were made using Wilcoxon's rank sum test (MATLAB 'ranksum'); paired comparisons using Wilcoxon signed-rank test (MATLAB 'signrank'). Multiple comparisons with repeated measures were made using Friedman's test (MATLAB 'friedman'). Comparisons between groups across training were made using 2-way ANOVA. Correlations were quantified using Pearson's correlation coefficient (MATLAB 'corr'). Linear regression to estimate contribution of fiber position to variance in mDA reward signals was fit using MATLAB 'fitlm'. Errors are reported as standard errors of the mean (s.e.m.). All sample sizes refer to the number of mice in the sample.

**Data Availability.** The data and custom code used to generate results supporting the findings of this study are within the [dudman lab github] repo including both modeling code (https://github.com/dudmanj/RNN_learnDA) and analysis code ([dudman lab github]).

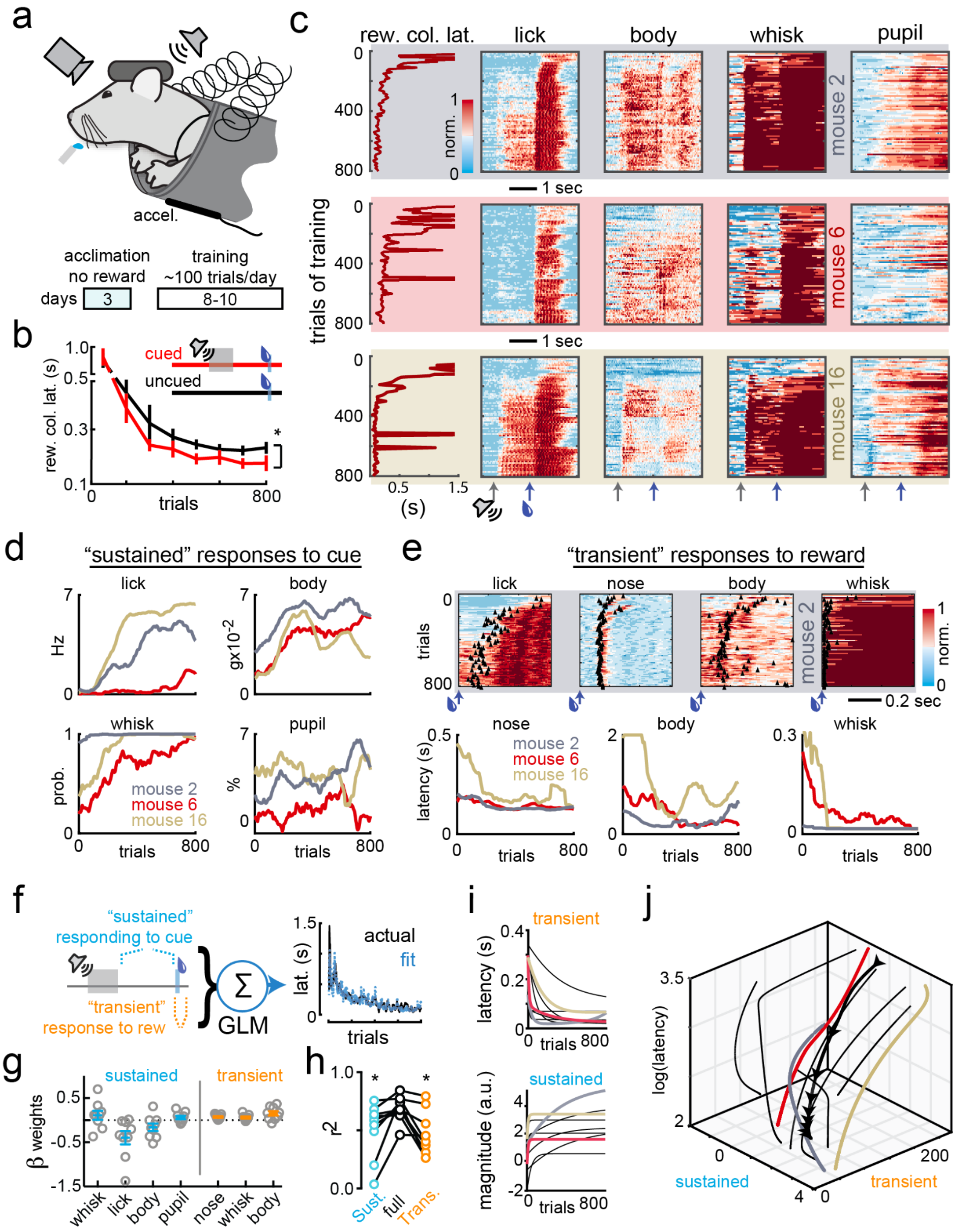

a
accel.
acclimation
no reward
training
~100 trials/day
days
3
8-10
b
rew. col. lat. (s)
1.0
0.5
0.3
0.1
0
trials
800
cued
uncued
*
c
rew. col. lat.
lick
body
whisk
pupil
trials of training
0
400
800
norm.
1
0
1 sec
mouse 2
mouse 6
mouse 16
0.5
1.5
(s)
d
"sustained" responses to cue
lick
body
Hz
gx10^-2
7
0
whisk
pupil
prob.
%
1
mouse 2
mouse 6
mouse 16
trials
800
e
"transient" responses to reward
lick
nose
body
whisk
trials
800
mouse 2
norm.
0.2 sec
nose
body
whisk
latency (s)
0.5
2
0.3
mouse 2
mouse 6
mouse 16
trials
f
"sustained"
responding to cue
"transient"
response to rew
Σ
GLM
1.5
lat. (s)
actual
fit
trials
g
β weights
0.5
-0.5
-1.5
sustained
transient
whisk
lick
body
pupil
nose
whisk
body
h
1.0
r²
0.0
*
*
Sust.
full
Trans.
i
transient
latency (s)
0.4
0.2
0
trials 800
sustained
magnitude (a.u.)
4
2
0
-2
trials 800
j
log(latency)
3.5
2
0
sustained
4
0
200
transient

**Figure 1.Changes to behavioral policy correlate with improved reward collection performance.**
**a)** Naive, thirsty, head-fixed mice underwent classical trace conditioning.
**b)** Learning quantified by a decrease in reward collection latency over training. As training progressed, a predictive cue led to faster reward collection (red) as compared to uncued probe trials (black).
**c)** Reward collection latency (leftmost column) compared to normalized heat maps (right 4 columns) of measures of licking, body movement, whisking probability and pupil diameter, for standard trials in which a 0.5 s auditory cue (grey arrows at cue start) predicted 3 μL sweetened water reward (blue arrows), averaged in 10-trial bins across training. Each row summarizes an individual mouse's learning trajectory,  with a background color that identifies each example mouse in figure panels (d)-(j).
**d)** Moving means of the averaged "sustained" behavioral responses during the 1 s delay between cue end and reward delivery for each example mouse from panel (C).
**e)** (top) Normalized heat maps of licking, nose motion, body movement, and whisking probability in 10 trial bins across training, with mean first response following reward delivery indicated by black triangles. (bottom) Moving means of the "transient" nose, body, and whisking responses, quantified as the latency to the first response following reward delivery, for each example mouse from panel (C).
**f)** Sustained and transient measures as in panels (D) and (E) used as predictors of reward collection latency in a generalized linear model.
**g)** Weights of each predictor in the GLM for each of 9 individual mice.
**h)** Sustained predictors alone (blue) or transient predictors alone (red) provided worse predictions than the full model.
**i)** Abstract learning trajectories were described as exponential fits to the first principal component of transient (top) and sustained (bottom) measurement variables used as predictors in the GLM in panels (F-H).
**j)** Sustained and transient abstract learning trajectories plotted against the latency to collect reward for all mice (example mice from panel C: red, yellow, grey; all other mice: thin black; mean of all mice: thick black line; arrows placed every 100 trials to convey relative speed of updates), providing a full visualization of how inferred policy updates are related to reward collection performance.

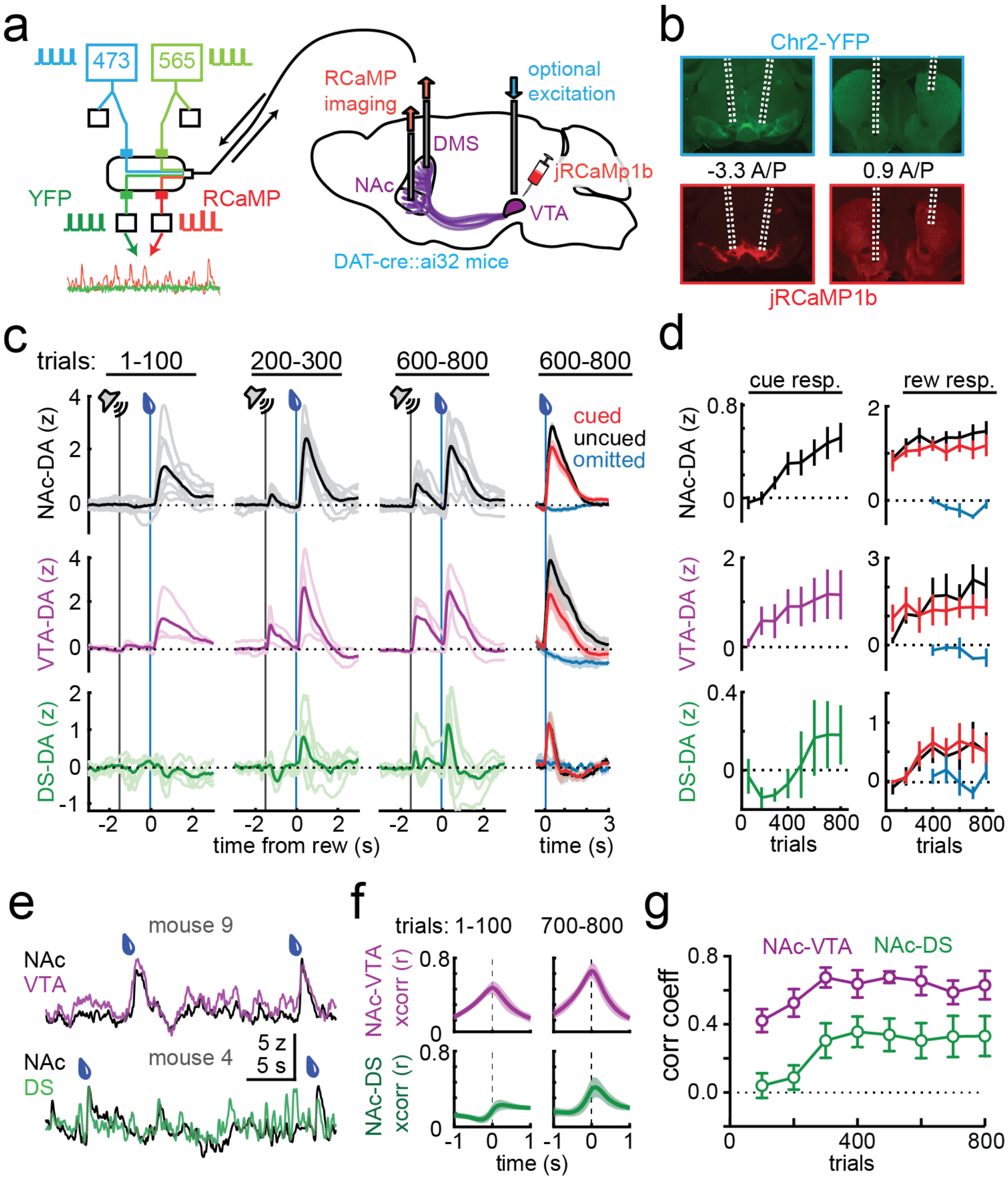


**Figure 2. Mesolimbic DA experiences significant reward signaling throughout acquisition training.**

**a)** (left) Fiber photometry hardware schematic. 10% duty cycle 473 and 656 nm excitations were offset from eachother and split between the main filter cube and photodectors (white squares) that measure output power. Excitation and emission of eYFP and jRCaMP1b fluorescence were conveyed by one cable between the filter cube and the brains of head-fixed animals. YFP and RCaMP emissions were measured at separate filter cube outputs. (right) jRCaMP1b was virally expressed bilaterally in the VTA and SNc of DAT-cre::ai32 mice, allowing measurement and the option for simultaneous manipulation of mesostriatal DA circuits.

**b)** Histology showing example fiber paths and virus expression.
**c)** (left 3 columns) jRCaMP1b DA signals in the nucleus accumbens core (NAc, black, n=9) and simultaneous recordings in the ventral tegmental area (VTA, purple, n = 3) and dorsal striatum (DS, green, n = 6), for cued reward trials in the trial bins indicated across training. (right) Reward or omission signals in NAc (top), VTA (middle), and DS (bottom) in trials 600-800, for cued (red), uncued (black), and cued but omitted (blue) trials.
**d)** Mean signals during the 1 s following cue delivery (left) and 2 s following reward delivery (right) across training for each brain region from panel (C).
**e)** Example simultaneous recordings from NAc-VTA (top) and NAc-DA (bottom).
**f)** Mean cross correlations for simultaneously measured NAc-VTA signals (top row, n =3) and NAc-DS signals (bottom row, n=6) in trials 1-100 (left) and trials 700-800 (right) within trial periods (1 second before cue to 3 seconds after reward).
**g)** Peak cross correlation coefficients between NAc-VTA and NAc-DS signal pairs across training, within trial periods.

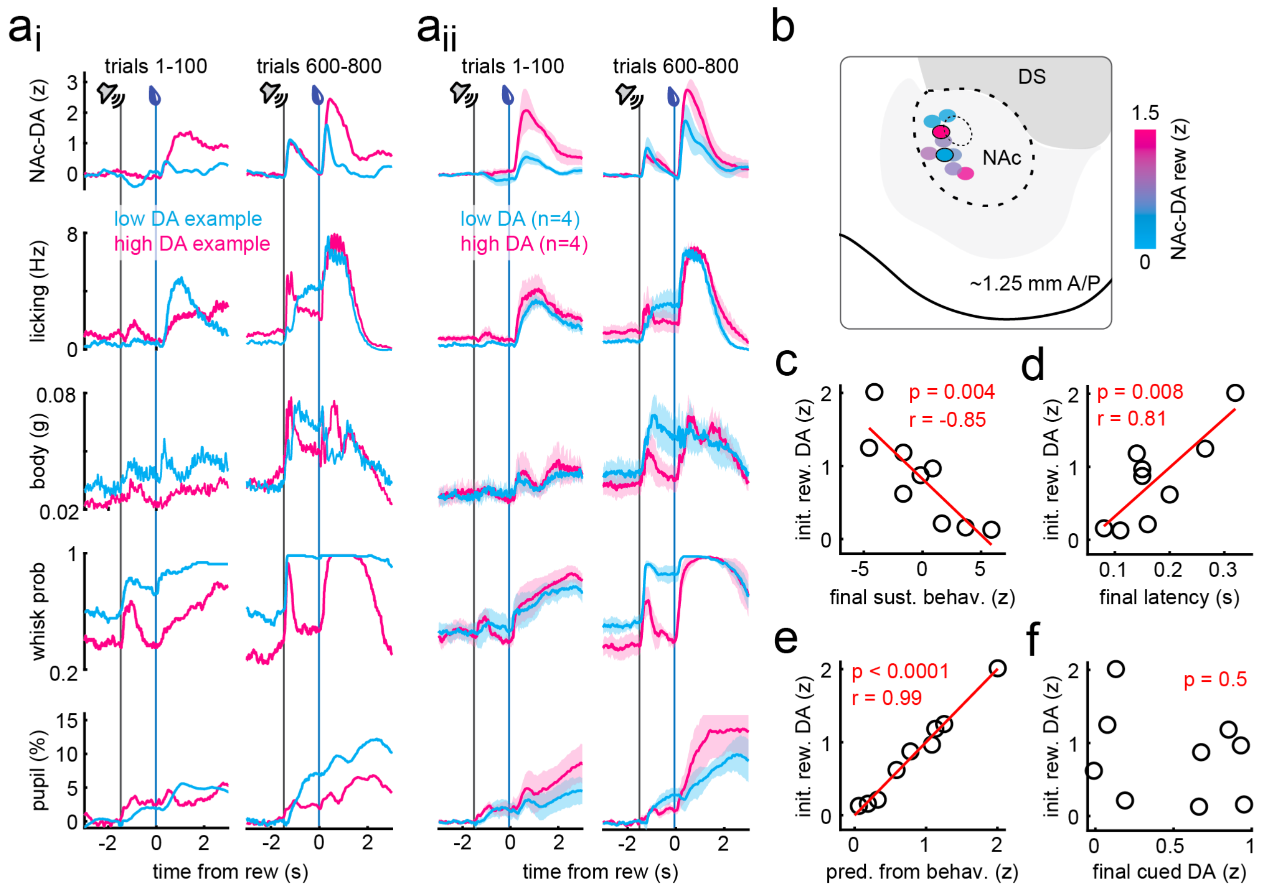


**Figure 3. Individual differences in mesolimbic DA signals correlate with learned behavioral policy.**

**$a_i$)** NAc-DA, licking, body movements, whisking probability, and pupil diameter measurements for example animals with low NAc-DA (blue) and high NAc-DA (pink) initial reward signals.

**$a_{ii}$)** Same as (Ai), except showing the mean of the 4 animals with lowest (blue) and highest (pink) initial NAc-DA reward signals .

**b)** Visualization of fiber locations for each mouse (n=9), color-coded according to the size of their initial NAc-DA reward signals.

**c)** Correlation of total sustained behavior magnitude (see Methods) in trials 700-800 vs initial NAc-DA reward signals trials 1-100 (n=9 mice).

**d)** Correlation of latency to collect reward in trials 700-800 vs initial NAc-DA reward signals trials 1-100 (n=9 mice).

**e)** Correlation of initial NAc-DA reward signals predicted from behavior measures in trials 700-800 to observed initial NAc-DA reward signals.

**f)** Lack of significant correlation between NAc-DA cue signals at trials 700-800 and initial NAc-DA reward signals at trials 1-100.

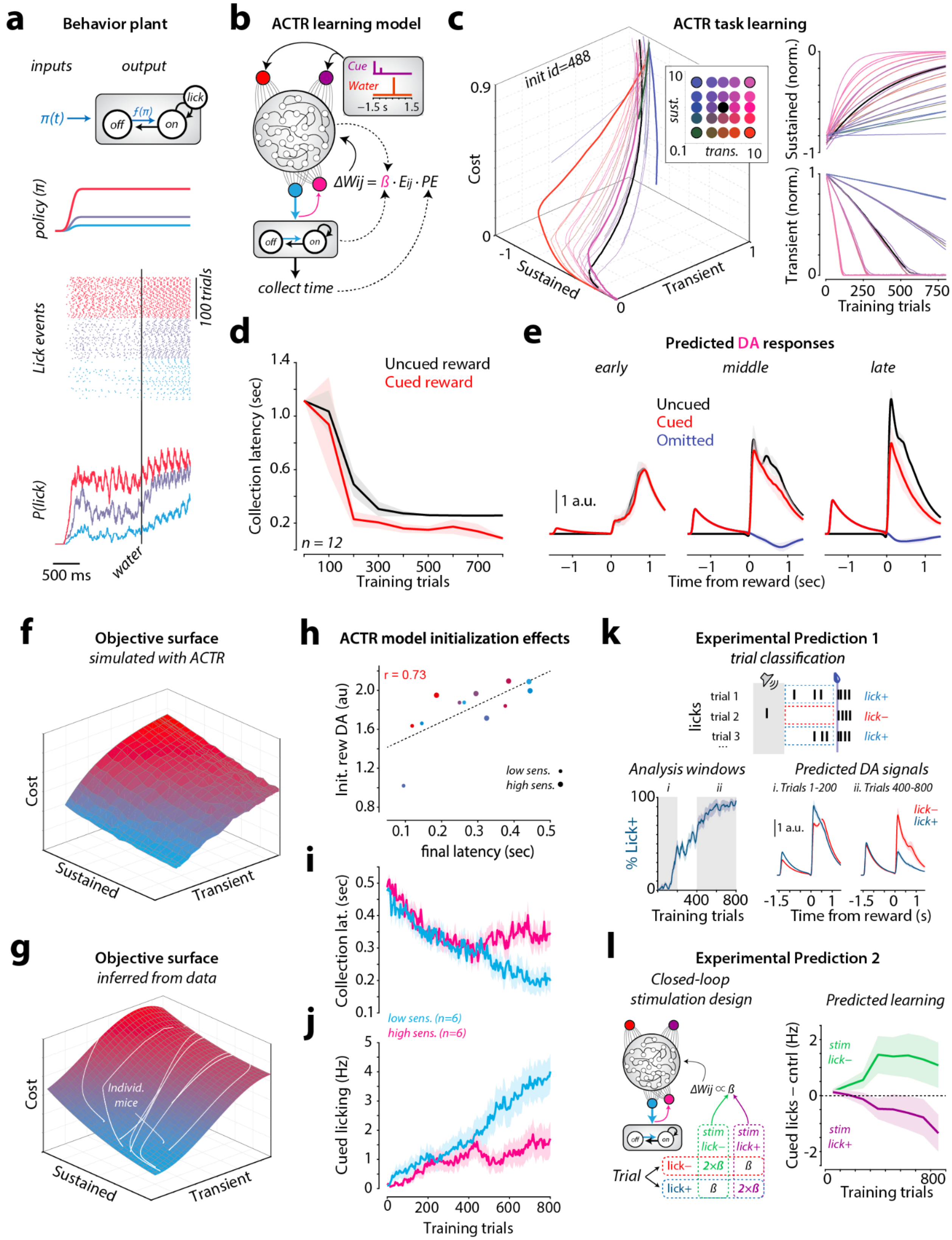


**Figure 4. ACTR model of policy learning during classical conditioning.**

**a)** (top row) Licking behavior was modeled as a two state ({off,on}) plant that emitted 7 Hz lick bouts from the 'on' state. Forward transition rate (off→on) was determined by a policy *π(t)*. Reverse transition rate (on→off) was a constant modulated by the presence of water. Bottom three rows illustrate example licking behavior produced by the plant for three different constant policies (red, purple, blue) before and after water reward delivery (vertical black line) for 100 repetitions of each policy.
**b)** The ACTR model (see methods) learned a control policy for the lick plant as the activity of a single output neuron (cyan) from an RNN that received sensory inputs upon cue onset and offset (purple) and water reward delivery (red). Weights of connections between neurons were updated according to the ACTR learning rule summarized in the equation. $W_{ij}$: weight of the connection between the i-th neuron and the j-th neuron. $\beta$: adaptive learning rate set by a mDA-like signal from a feedback neuron proportional to policy error (pink). $E_{ij}$: eligibility trace for the synapse between the i-th neuron and the j-th neuron. PE: performance error from comparing the latency to collect reward on the current trial to the recent history of collection latencies.
**c)** Different ratios of sustained vs transient learning rates (inset color code) produced a range of trajectories similar to observed trajectories in individual mice (Fig. 1).
**d)** 12 simulations of a common network initialization produced mouse-like decreases in cued and uncued reward collection latency across training.
**e)** Activity of the mDA-like feedback unit, convolved with kernels to match the variance and kinetics of jRCaMP1b measurements of mouse mDA activity, for uncued (black), cued (red), and omitted (blue) reward trials during early (left), middle (middle), and late (right) training periods.
**f)** Objective surface to visualize policy gradient calculated from ACTR model using arbitrary combinations of transient and sustained components (mean of 1000 simulations per point).
**g)** Objective surface fit (2$^{nd}$ order polynomial surface) from observed mouse data. **H)** Simulations with low (small dots, n=6) or high (large dots, n=6) initial reward-related sensory input exhibited a significant correlation between initial predicted mDA reward response and final reward collection latency.
**i)** Reward collection latency and **J)** sustained cued licking for simulations with low (cyan) and high (magenta) initial reward-related sensory input.
**k)** As training progresses (bottom left), ACTR model mDA-like signals (bottom right) display differential signals on trials with sustained cued licking (lick+) versus trials without sustained cued licking (lick-).
**l)** Enhancing the mDA-like adaptive learning rate signal at reward selectively (schematic at left) on either lick- (green) or lick+ (purple) trials biases future licking behavior in opposite directions from the stimulation contingency across training (right): in other words, enhancing mDA-like reward signals on trials with cued licking decreases cued licking in the future.

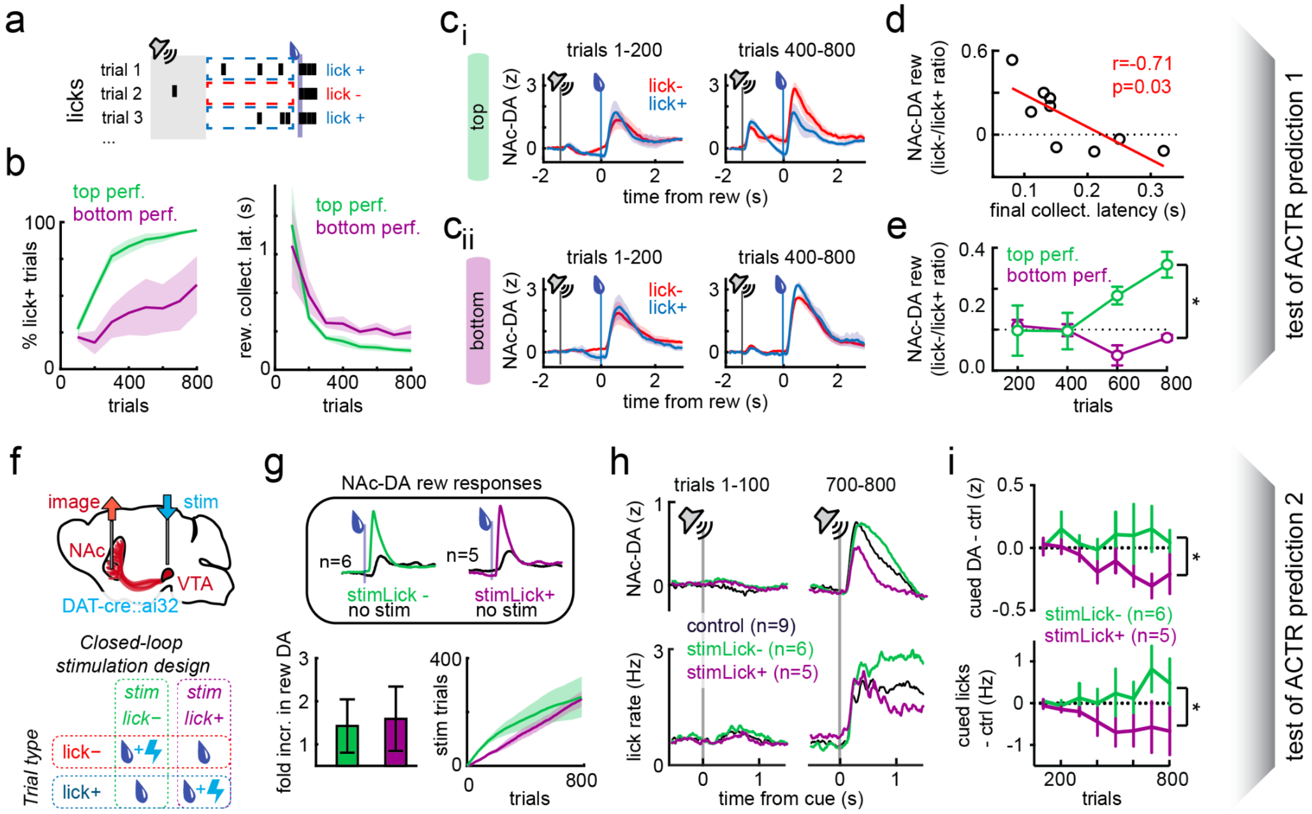


**Figure 5. Role of mesolimbic DA in learning is consistent with an adaptive policy learning rate.**

**a)** Trial types defined as “lick+” trials with at least one lick during the delay between cue and reward and “lick-” trials with no delay period licking.

**b)** Percent of total trials that are “lick+” (left) and reward collection latency (right), for top 4 (green) and bottom 4 (purple) performing mice, as determined by their reward collection latency in trials 700-800.

**c$_i$)** NAc-DA signals early (trials 1-200, left column) and late (trials 400-800, right column) in training for lick- (red) and lick+ (blue) trials, for top 4 performing mice in terms of reward collection latency in trials 700-800.

**c$_{ii}$)** Same as (c$_i$) except for bottom 4 performing mice.

**d)** The ratio of NAc-DA reward signals on lick- vs lick+ trials was correlated with the final reward collection latency.

**e)** The ratio of NAc-DA reward signals on lick- vs lick+ trials increased with training for top performing mice.

**f)** Simultaneous measurement and manipulation of mesolimbic DA signals (top). Closed-loop experiment design, where different groups of mice received VTA stim concurrent with reward delivery on either lick- trials (“stim lick-”) or lick+ trials (“stime lick+).

**g)** (top) Mean NAc-DA reward responses across training with (colored traces) and without (black traces) exogenous stimulation, for stimLick- (green)and stimLick+ (purple) mice. (bottom) Fold increase in NAc-DA reward signals and cumulative sum of stimulated trials for stimLick- (green) and stimLick+ (purple).

**h)** NAc-DA (top) and licking (bottom) during early (left) and late (right) training for control (black), stimLick- (green) and stimLick+ (purple) animals.
**i)** NAc-DA cue responses (top) and cued licking (bottom) for StimLick- (green) and stimLick+ (purple) mice across training, displayed as the difference from control mice.

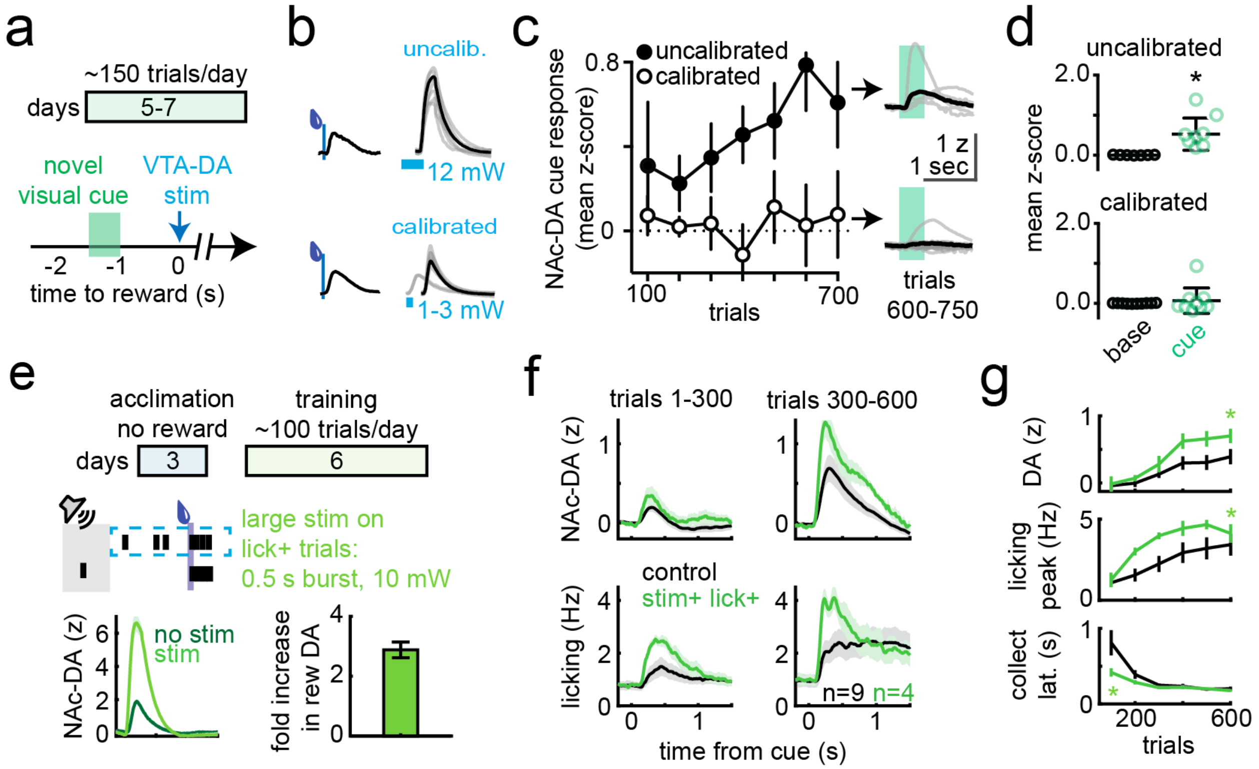


**Figure 6. Large mesolimbic DA manipulations drive value-like learning.**

**a)** Experimental design for VTA-DA stimulation predicted by a 0.5 s light flash at the front of the behavioral chamber.

**b)** Mean uncued NAc-DA reward responses (left) and individual responses to VTA-DA stimulation (right) for mice that either received large, uncalibrated stimulation (top; 30 Hz, 12 mW for 500 ms) or stimulation calibrated to reward responses (bottom; 30 Hz, 1-3 mW for 150 ms).

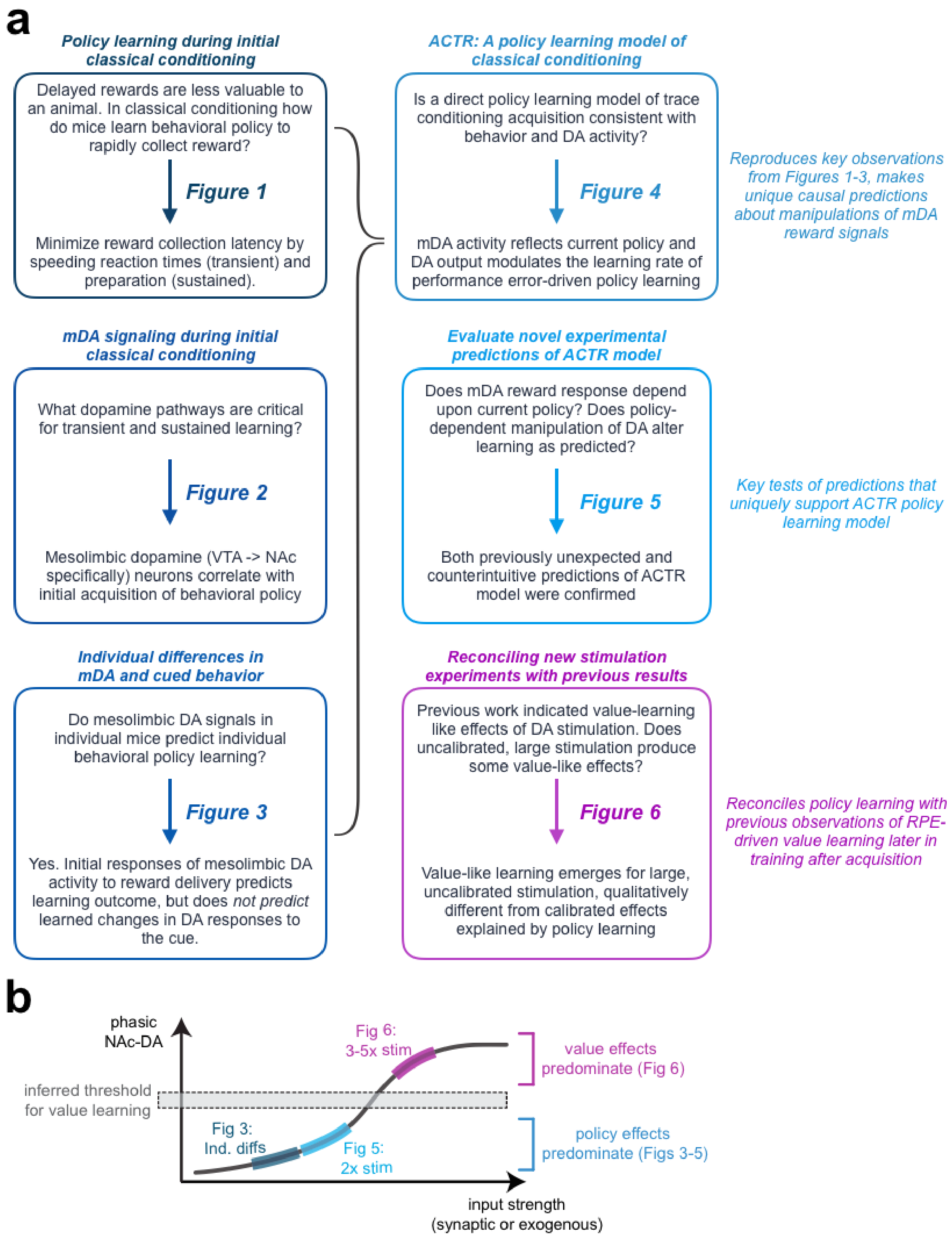
**c)** (left) jRCaMP1b NAc-DA cue responses across training for mice that received large stimulations (5x the size of reward responses; filled circles) or calibrated stimulations (1x the size of reward responses; open circles). (right) Mean NAc-DA traces after 750 training trials.

**d)** Quantification of NAc-DA cue responses at the end of training for large (top) and calibrated (bottom) stimulation.

**e)** (top) Experimental design for new group of mice that experienced a “stim+ lick+” contingency: they received large, uncalibrated VTA-DA stimulation on lick+ trials. (bottom) NAc-DA reward responses on stimulated (light) and unstimulated (dark) trials, and fold increase in NAc-DA reward responses due to stimulation (n=4).

**f)** NAc-DA cue responses (top) and cued licking (bottom) for control (black) and stim+Lick+ (light green) mice.

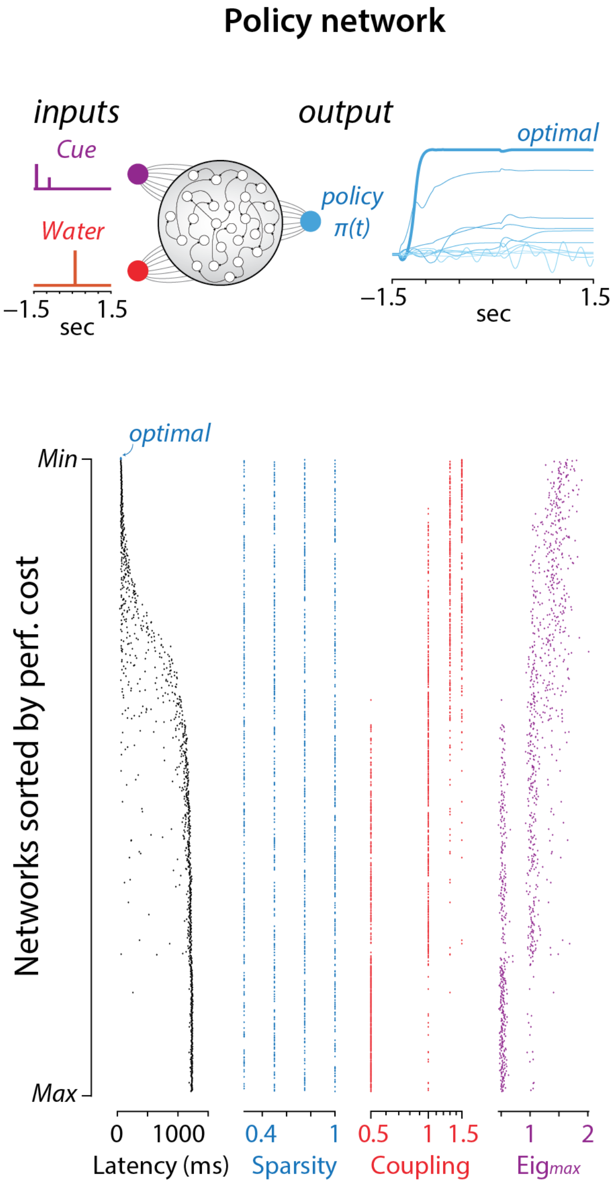
**g)** Quantification across training of NAC-DA cue responses (top), peak cued licking (middle), and latency to collect reward (bottom) for control (black) and stim+lick+ mice (light green).


a
Policy learning during initial classical conditioning
Delayed rewards are less valuable to an animal. In classical conditioning how do mice learn behavioral policy to rapidly collect reward?
Figure 1
Minimize reward collection latency by speeding reaction times (transient) and preparation (sustained).
mDA signaling during initial classical conditioning
What dopamine pathways are critical for transient and sustained learning?
Figure 2
Mesolimbic dopamine (VTA -> NAc specifically) neurons correlate with initial acquisition of behavioral policy
Individual differences in mDA and cued behavior
Do mesolimbic DA signals in individual mice predict individual behavioral policy learning?
Figure 3
Yes. Initial responses of mesolimbic DA activity to reward delivery predicts learning outcome, but does *not predict* learned changes in DA responses to the cue.
ACTR: A policy learning model of classical conditioning
Is a direct policy learning model of trace conditioning acquisition consistent with behavior and DA activity?
Figure 4
mDA activity reflects current policy and DA output modulates the learning rate of performance error-driven policy learning
Reproduces key observations from Figures 1-3, makes unique causal predictions about manipulations of mDA reward signals
Evaluate novel experimental predictions of ACTR model
Does mDA reward response depend upon current policy? Does policy-dependent manipulation of DA alter learning as predicted?
Figure 5
Both previously unexpected and counterintuitive predictions of ACTR model were confirmed
Key tests of predictions that uniquely support ACTR policy learning model
Reconciling new stimulation experiments with previous results
Previous work indicated value-learning like effects of DA stimulation. Does uncalibrated, large stimulation produce some value-like effects?
Figure 6
Value-like learning emerges for large, uncalibrated stimulation, qualitatively different from calibrated effects explained by policy learning
Reconciles policy learning with previous observations of RPE-driven value learning later in training after acquisition
b
phasic NAc-DA
inferred threshold for value learning
Fig 6: 3-5x stim
value effects predominate (Fig 6)
Fig 3: Ind. diffs
Fig 5: 2x stim
policy effects predominate (Figs 3-5)
input strength (synaptic or exogenous)

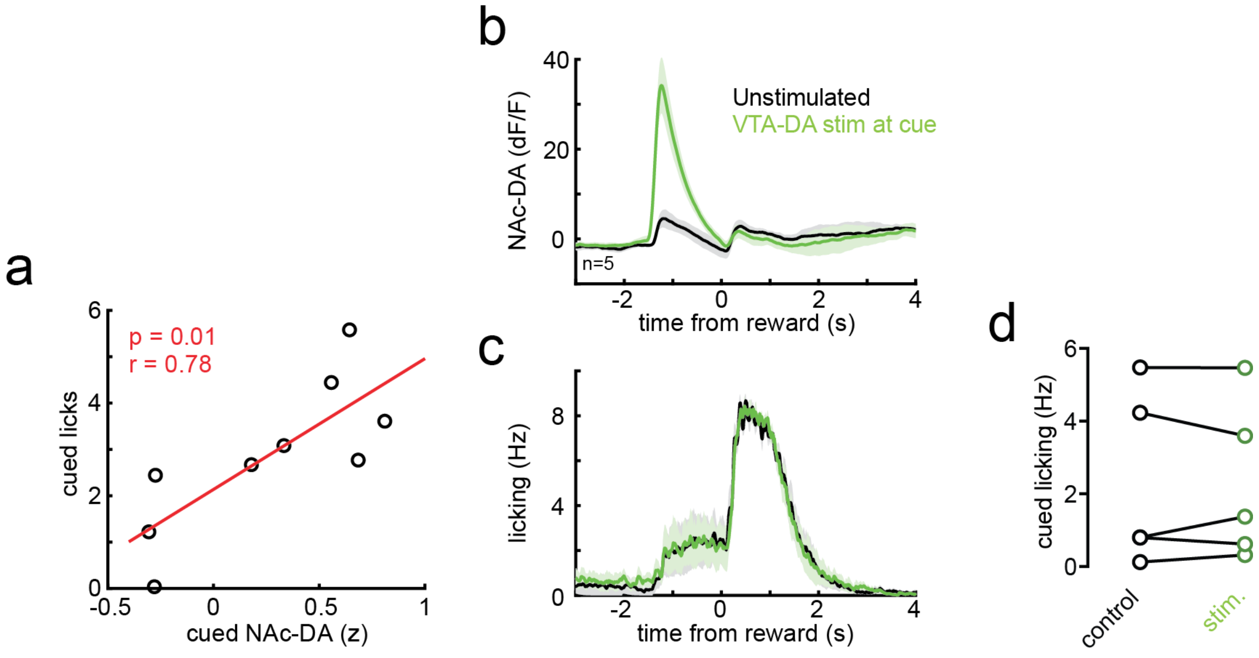
**Supplementary Figure 1.** Logic Outline

**a)** Narrative logic of the manuscript.

**b)** Visualization of proposed dissociation between policy and value learning functions for NAc-DA reward signals



**Supplemental Figure 2. Search through ACTR initialization space.** (top) Schematic of ACTR policy recurrent neural network (RNN). (bottom) We scanned through thousands of simulated network configurations and ranked those networks according to their performance cost (Fig. 4b; cost is a combination of latency cost and a network variance cost, see methods for details). Displayed are the latency to collect reward (black), network sparsity (blue), coupling coefficient (red), leading eigenvalue (purple). This analysis reveals a few key aspects. First, a sustained policy that spans time from the detection of the cue through the delivery of water reward is necessary to minimize latency cost. Second, while optimal RNNs are relatively indifferent to some parameters (sparsity of connectivity) they tend to require a strong coupling coefficient which is known to determine the capacity of a RNN to develop sustained dynamics [85]. Consistent with this interpretation we found that optimal policies were observed uniquely in RNNs with large leading eigenvalues (i.e. long time constant dynamics [86]). These analyses indicate that there are realizable RNN configurations sufficient to produce an optimal policy.



**Supplemental Figure 3. Enhancing mesolimbic cue signals has no immediate effect on cued behavior.**

**a)** Cued licking was correlated with the size of NAc-DA cue responses.

**b)** To test for a causal connection between the size of mesolimbic DA cue responses and cued behavior, in a new session after regular training was complete, we delivered large, uncalibrated VTA-DA stimulation on a random subset of cued reward trials (light green).

**c)** Licking behavior appeared unchanged between unstimulated (black) vs VTA-DA stimulated (light green) trials.

**d)** Quantification of sustained licking during the delay period for unstimulated (black) vs stimulated (green) trials.

## References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
2. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal Policy Optimization Algorithms. *arXiv [cs.LG]* (2017).
3. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
4. Barto, A. G. Adaptive critics and the basal ganglia. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1018&context=cs_faculty_pubs .
5. Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol. Rev.* **95**, 853–951 (2015).
6. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
7. Redgrave, P. & Gurney, K. The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* **7**, 967–975 (2006).
8. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).
9. Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M. & Graybiel, A. M. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575–579 (2013).
10. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
11. 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).
12. Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).

13. Stauffer, W. R. *et al.* Dopamine Neuron-Specific Optogenetic Stimulation in Rhesus Macaques. *Cell* **166**, 1564–1571.e6 (2016).

14. Chang, C. Y. *et al.* Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nat. Neurosci.* **19**, 111–116 (2016).

15. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* (2019) doi:10.1038/s41586-019-1261-9.

16. Menegas, W., Akiti, K., Amo, R., Uchida, N. & Watabe-Uchida, M. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nat. Neurosci.* **21**, 1421–1430 (2018).

17. Dodson, P. D. *et al.* Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E2180–8 (2016).

18. de Jong, J. W. *et al.* A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. *Neuron* **101**, 133–151.e7 (2019).

19. Hamid, A. A., Frank, M. J. & Moore, C. I. Dopamine waves as a mechanism for spatiotemporal credit assignment. *BioRxiv* (2019).

20. Bova, A., Gaidica, M., Hurst, A., Iwai, Y. & Leventhal, D. K. Precisely-timed dopamine signals establish distinct kinematic representations of skilled movements. doi:10.1101/2020.07.29.227298.

21. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).

22. Hughes, R. N. *et al.* Ventral Tegmental Dopamine Neurons Control the Impulse Vector during Motivated Behavior. *Curr. Biol.* **30**, 2681–2694.e5 (2020).

23. Kremer, Y., Flakowski, J., Rohner, C. & Lüscher, C. Context-dependent multiplexing by individual VTA dopamine neurons. *Journal of Neuroscience* **40**, 7489–7509 (2020).

24. Coddington, L. T. & Dudman, J. T. Learning from Action: Reconsidering Movement

Signaling in Midbrain Dopamine Neuron Activity. *Neuron* **104**, 63–77 (2019).

25. Hollon, N. G. *et al.* Nigrostriatal Dopamine Signals Sequence-Specific Action-Outcome Prediction Errors. *Cold Spring Harbor Laboratory* 2021.01.25.428032 (2021) doi:10.1101/2021.01.25.428032.
26. Syed, E. C. J. *et al.* Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nat. Neurosci.* **19**, 34–36 (2016).
27. Jin, X. & Costa, R. M. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457–462 (2010).
28. Collins, A. G. E. & Cockburn, J. Beyond dichotomies in reinforcement learning. *Nat. Rev. Neurosci.* **21**, 576–586 (2020).
29. Sauce, B. & Matzel, L. D. The causes of variation in learning and behavior: why individual differences matter. *Front. Psychol.* **4**, 395 (2013).
30. Kober, J., Bagnell, J. A. & Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of* (2013).
31. Tsai, H.-C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080–1084 (2009).
32. Rossi, M. A., Sukharnikova, T., Hayrapetyan, V. Y., Yang, L. & Yin, H. H. Operant self-stimulation of dopamine neurons in the substantia nigra. *PLoS One* **8**, e65799 (2013).
33. 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).
34. Ilango, A., Kesner, A. J., Broker, C. J., Wang, D. V. & Ikemoto, S. Phasic excitation of ventral tegmental dopamine neurons potentiates the initiation of conditioned approach behavior: parametric and reinforcement-schedule analyses. *Front. Behav. Neurosci.* **8**, 155 (2014).
35. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**, 229–256 (1992).

36. Silver, D. A. RL Course by David Silver - Lecture 7: Policy Gradient Methods. *YouTube* https://www.youtube.com/watch?v=KHZVXao4qXs (2015).

37. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533**, 402–406 (2016).

38. Gadagkar, V. *et al.* Dopamine neurons encode performance error in singing birds. *Science* **354**, 1278–1282 (2016).

39. Rueda-Orozco, P. E. & Robbe, D. The striatum multiplexes contextual and kinematic information to constrain motor habits execution. *Nat. Neurosci.* **18**, 453–460 (2015).

40. Tan, H. *et al.* Human subthalamic nucleus in movement error detection and its evaluation during visuomotor adaptation. *J. Neurosci.* **34**, 16744–16754 (2014).

41. Turner, R. S. & Desmurget, M. Basal ganglia contributions to motor control: a vigorous tutor. *Curr. Opin. Neurobiol.* **20**, 704–716 (2010).

42. Smith, M. A., Brandt, J. & Shadmehr, R. Motor disorder in Huntington's disease begins as a dysfunction in error feedback control. *Nature* **403**, 544–549 (2000).

43. Báez-Mendoza, R. & Schultz, W. Performance error-related activity in monkey striatum during social interactions. *Sci. Rep.* **6**, 37199 (2016).

44. Chen, R. *et al.* Songbird Ventral Pallidum Sends Diverse Performance Error Signals to Dopaminergic Midbrain. *Neuron* **103**, 266–276.e4 (2019).

45. Seiler, J. L., Cosme, C. V., Sherathiya, V. N., Bianco, J. M. & Lerner, T. N. Dopamine Signaling in the Dorsomedial Striatum Promotes Compulsive Behavior. doi:10.1101/2020.03.30.016238.

46. Coddington, L. T. & Dudman, J. T. In Vivo Optogenetics with Stimulus Calibration. *Methods Mol. Biol.* **2188**, 273–283 (2021).

47. Willuhn, I., Burgeno, L. M., Everitt, B. J. & Phillips, P. E. M. Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20703–20708 (2012).

48. Brown, H. D., McCutcheon, J. E., Cone, J. J., Ragozzino, M. E. & Roitman, M. F. Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. *Eur. J. Neurosci.* **34**, 1997–2006 (2011).
49. Miconi, T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife* **6**, (2017).
50. Fiete, I. R., Fee, M. S. & Seung, H. S. Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *J. Neurophysiol.* **98**, 2038–2057 (2007).
51. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
52. Fee, M. S. The role of efference copy in striatal learning. *Curr. Opin. Neurobiol.* **25**, 194–200 (2014).
53. Dudman, J. T. & Gerfen, C. R. The basal ganglia. *The Rat Nervous System (Fourth Edition)* (2015).
54. Haber, S. N. The place of dopamine in the cortico-basal ganglia circuit. *Neuroscience* **282**, 248–257 (2014).
55. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine neurons learn relative chosen value from probabilistic rewards. *Elife* **5**, (2016).
56. Rutledge, R. B. *et al.* Dopaminergic drugs modulate learning rates and perseveration in Parkinson’s patients in a dynamic foraging task. *J. Neurosci.* **29**, 15104–15114 (2009).
57. Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *Elife* **5**, (2016).
58. 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).
59. Berridge, K. C., Robinson, T. E. & Aldridge, J. W. Dissecting components of reward: ‘liking’, ‘wanting’, and learning. *Curr. Opin. Pharmacol.* **9**, 65–73 (2009).
60. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. Neurosci.* **22**,

1761–1770 (2019).

61. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).

62. Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).

63. Sharpe, M. J. *et al.* Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).

64. Momennejad, I. *et al.* The successor representation in human reinforcement learning. *Nat Hum Behav* **1**, 680–692 (2017).

65. Bennett, D., Niv, Y. & Langdon, A. Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior. (2021) doi:10.31234/osf.io/ew58m.

66. Hadjiosif, A. M., Krakauer, J. W. & Haith, A. M. Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning. *J. Neurosci.* (2021) doi:10.1523/JNEUROSCI.2125-20.2021.

67. Gallistel, C. R. *The Organization of Action: A New Synthesis*. (Psychology Press, 2013).

68. Elber-Dorozko, L. & Loewenstein, Y. Striatal action-value neurons reconsidered. *Elife* **7**, (2018).

69. Stelly, C. E., Tritley, S. C., Rafati, Y. & Wanat, M. J. Acute Stress Enhances Associative Learning via Dopamine Signaling in the Ventral Lateral Striatum. *J. Neurosci.* **40**, 4391–4400 (2020).

70. Amo, R., Yamanaka, A., Tanaka, K. F. & Uchida, N. A gradual backward shift of dopamine responses during associative learning. *bioRxiv* (2020).

71. Mohebi, A. *et al.* Dissociable dopamine dynamics for learning and motivation. *Nature* **570**, 65–70 (2019).

72. Saunders, B. T., Richard, J. M., Margolis, E. B. & Janak, P. H. Instantiation of incentive value and movement invigoration by distinct midbrain dopamine circuits.

doi:10.1101/186502.

73. Flagel, S. B. *et al.* A selective role for dopamine in stimulus-reward learning. *Nature* **469**, 53–57 (2011).

74. Arbuthnott, G. W. & Wickens, J. Space, time and dopamine. *Trends Neurosci.* **30**, 62–69 (2007).

75. Lammel, S., Ion, D. I., Roeper, J. & Malenka, R. C. Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron* **70**, 855–862 (2011).

76. Lerner, T. N. *et al.* Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell* **162**, 635–647 (2015).

77. Stelly, C. E., Girven, K. S., Lefner, M. J., Fonzi, K. M. & Wanat, M. J. Dopamine release and its control over early Pavlovian learning differs between the NAc core and medial NAc shell. *Neuropsychopharmacology* (2021) doi:10.1038/s41386-020-00941-z.

78. Hamid, A. A., Frank, M. J. & Moore, C. I. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* (2021) doi:10.1016/j.cell.2021.03.046.

79. Redish, A. D. Addiction as a computational process gone awry. *Science* **306**, 1944–1947 (2004).

80. Vandaele, Y. & Ahmed, S. H. Habit, choice, and addiction. *Neuropsychopharmacology* **46**, 689–698 (2021).

81. Pascoli, V. *et al.* Stochastic synaptic plasticity underlying compulsion in a model of addiction. *Nature* **564**, 366–371 (2018).

82. Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O. & Clune, J. First return, then explore. *Nature* **590**, 580–586 (2021).

83. Park, J., Coddington, L. T. & Dudman, J. T. Basal Ganglia Circuits for Action Specification. *Annu. Rev. Neurosci.* **43**, (2020).

84. A Michaels, J. & Scherberger, H. HebbRNN: A reward-modulated Hebbian learning rule for recurrent neural networks. *J. Open Source Softw.* **1**, 60 (2016).

85. Driscoll, L. N., Golub, M. D. & Sussillo, D. Computation through Cortical Dynamics. *Neuron* vol. 98 873–875 (2018).

86. Vogels, T. P., Rajan, K. & Abbott, L. F. Neural network dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).

87. Shindou, T., Shindou, M., Watanabe, S. & Wickens, J. A silent eligibility trace enables dopamine‐dependent synaptic plasticity for reinforcement learning in the mouse striatum. *Eur. J. Neurosci.* **49**, 726–736 (2019).

88. Noorani, I. & Carpenter, R. H. S. The LATER model of reaction time and decision. *Neurosci. Biobehav. Rev.* **64**, 229–251 (2016).

89. Paxinos, G. & Franklin, K. B. J. *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates*. (Academic Press, 2019).