HHS Public Access

Author manuscript

Curr Biol. Author manuscript; available in PMC 2022 December 06.

Published in final edited form as:

Curr Biol. 2021 December 06; 31(23): 5176–5191.e5. doi:10.1016/j.cub.2021.09.037.

Relative salience signaling within a thalamo-orbitofrontal circuit governs learning rate

Vijay Mohan K Namboodiri[1,2], **Taylor Hobbs**[1], **Ivan Trujillo Pisanty**[1], **Rhiana C Simon**[3], **Madelyn M Gray**[3], **Garret D Stuber**[1,3,4]

[1]The Center for the Neurobiology of Addiction, Pain, and Emotion, Department of Anesthesiology and Pain Medicine, Department of Pharmacology, University of Washington, Seattle, WA 98195, USA

[2]Current Address: Alcohol and Addiction Research Group, Department of Neurology, Neuroscience Graduate Program, University of California at San Francisco, San Francisco, CA, 94158, USA

[3]Graduate Program in Neuroscience, University of Washington, 98195

SUMMARY

Learning to predict rewards is essential for the sustained fitness of animals. Contemporary views suggest that such learning is driven by a reward prediction error (RPE) — the difference between received and predicted rewards. The magnitude of learning induced by an RPE is proportional to the product of the RPE and a learning rate. Here we demonstrate using two-photon calcium imaging and optogenetics in mice that certain functionally distinct subpopulations of ventral/medial orbitofrontal cortex (vmOFC) neurons signal learning rate control. Consistent with learning rate control, trial-by-trial fluctuations in vmOFC activity positively correlates with behavioral updating when RPE is positive, and negatively correlates with behavioral updating when RPE is negative. Learning rate is affected by many variables including the salience of a reward. We found that the average reward response of these neurons signals the relative salience of a reward, as it decreases after reward prediction learning or the introduction of another highly salient aversive stimulus. The relative salience signaling in vmOFC is sculpted by medial thalamic inputs. These results support emerging theoretical views that the prefrontal cortex encodes and controls learning parameters.

eTOC

[4]Corresponding Author and Lead Contact; gstuber@uw.edu.

AUTHOR CONTRIBUTIONS: V.M.K.N., T.G.H., I.T.P., R.S., and M.M.G. performed experiments. V.M.K.N. and M.M.G. performed analyses. V.M.K.N., and G.D.S. designed experiments and wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS : None.

Here, by developing a theoretical approach to test neuronal control of learning rate (a key reinforcement learning variable), Namboodiri et al. show that outcome responses in a thalamo-orbitofrontal circuit guide learning rate by signaling the relative salience of outcomes.

INTRODUCTION

A simple, yet powerful model for learning that a cue predicts an upcoming reward is to update one's predictions of future reward via a reward prediction error signal (RPE)—the difference between a received and predicted reward [1,2]. This RPE model largely explains the response dynamics of midbrain dopaminergic neurons [2–4]—the main identified neural system responsible for reward prediction learning [5–7]. Even though the bulk of experimental work into the neuronal mechanisms of reinforcement learning focuses on the mesolimbic dopamine circuitry, a rich computational literature argues that the dopamine RPE circuitry cannot operate in isolation and that the brain likely contains complementary systems for learning [8–10]. For instance, reinforcement learning algorithms contain parameters such as learning rate that also need to be learned. Since the net magnitude of learning due to an RPE is the product of the RPE and a learning rate for the reward, optimally tuning the learning rate (amount by which RPE updates reward prediction) can be highly beneficial to adapt learning to one's environment [8–12].

Prior computational models have proposed many possibilities for how animals set learning rates. One of the earliest models to suggest that the learning rate is itself dynamic was the Pearce-Hall associative learning model, which proposed that learning rate is proportional to the absolute magnitude of the previous RPE, a variable often referred to as salience or associability [13]. Other models propose that learning rates depend on the expected and unexpected uncertainty in the environment [10–12,14–16]. Such learning of learning rate is a special case of the learning of parameters for reinforcement learning—previously referred to as "meta learning" or "meta-reinforcement learning" [8]. Adaptive learning rates are also a consequence of more recently proposed second-order learning mechanisms operating on top of an initial model-free reinforcement learning mechanism (also referred to as "meta-reinforcement learning")[9]. Regardless of the details, a common consequence of all the above learning algorithms is that the learning adapts to the properties and statistics of the environment.

Recent research has focused on employing these models to test whether activity in different brain regions correlates with learning rate. A key challenge in this endeavor is that learning rate is a latent cause of behavior that can be hard to measure directly. The most common approach to solve this problem has been to fit behavior using reinforcement learning models under conditions that are theoretically hypothesized to vary learning rate (e.g., systematic variation of uncertainty) [12,17–20]. Using such approaches, multiple neural signals have been found to correlate with learning rate, including in the anterior cingulate cortex (ACC) [10,12], dorsal raphe serotonergic system [17,21], dorsomedial prefrontal cortex [18,20,22] and anterior insula [20]

Here, we develop an alternative approach to test learning rate control and test the hypothesis that the orbitofrontal cortex (OFC) controls the learning rate of a reward. We tested

this hypothesis for a few reasons. First, prior research has demonstrated causal roles for the orbitofrontal cortex (OFC) in reward learning and adaptation [23–27]. Second, OFC receives inputs from all the regions mentioned above [28–30]. Third, OFC activity has been hypothesized to convey expected uncertainty, a key variable governing learning rate [10]. Fourth, we previously observed that the suppression of reward responses in vmOFC was sufficient to reduce the rate of behavioral learning [26]. Fifth, OFC activity correlates with the attribution of stimulus salience, a key variable thought to modulate learning rate [31]. Thus, we tested whether vmOFC neurons act as a controller of learning rate by signaling the relative salience of rewards, and whether such signaling is mediated, at least in part, by one of its major inputs from medial thalamus.

RESULTS

Theoretical prediction for learning rate control

We first developed the following theoretical approach to identify a learning rate control signal during cue-reward learning. A model-free learning agent that learns the value of the cue will update cue value whenever there is an RPE at the time of the outcome. On trials in which there is a positive RPE, there will be a positive change (i.e., increase) in cue value (Figure 1A). The magnitude of this change will depend on both the RPE and the learning rate. Importantly, while RPE can be positive or negative on a trial, the learning rate is a positive number. Hence, if the learning rate increases for a given positive RPE, there will be a proportionally larger increase in cue value on the next trial. Similarly, on trials in which there is a negative RPE, there will be a negative change (i.e., decrease) in cue value (Figure 1A). If the learning rate increases for a given negative RPE, there will be a proportionally larger decrease in cue value on the next trial.

If the learning rate control system has some intrinsic noise, the trial-by-trial variability in learning rate will cause corresponding trial-by-trial variability in cue value update. If we plot the relationship between learning rate on a trial and the magnitude of cue value update on the next trial, the slope of this relationship should be the RPE (since $\Delta\text{cue value}_{n+1} = \text{RPE}_n \times \text{learning rate}_n$). Thus, this slope should be positive on positive RPE trials and negative on negative RPE trials (Figure 1B, C). In contrast with the above predictions, as $\Delta\text{cue value}_{n+1} = \text{RPE}_n \times \text{learning rate}_n$, the slope of the relationship between RPE on a trial and the subsequent cue value update is the learning rate. Hence, this slope should always be positive regardless of the sign of RPE as the learning rate is a positive quantity. Similarly, because RPE is the incentive value of outcome minus cue value, the slope of the relationship between the incentive value of outcome on a trial and the cue value update on the next trial is also the learning rate (i.e., $\Delta\text{cue value}_{n+1} = (\text{outcome}_n - \text{cue value}_n) \times \text{learning rate}_n$), and hence always positive. In other words, unlike learning rate, an increase in RPE or outcome value on a trial should always cause an increase in cue value on the next trial. To sum up, the above test for positive correlation with cue value update after positive RPE trials and negative correlation with cue value update after negative RPE trials is a discriminative test for learning rate control.

To test whether these theoretical conclusions are valid during reinforcement learning under noise, we first simulated a temporal difference learning agent that sets learning rate to

a constant plus some noise (Figure 1D) (see Methods). To equate the magnitudes of the positive and negative RPEs, we simulated conditions in which the cue predicted reward at 50% probability. Here, rewarded trials result in a positive RPE and unrewarded trials result in a negative RPE. Under these conditions with a noisy learning rate controller, we indeed observed a positive correlation between the learning rate control signal and the subsequent cue value update on positive RPE trials and a negative correlation on negative RPE trials (Figure 1E, F). We next tested if this prediction for a learning rate controller remains true for various conditions, including a non-linear mapping between cue value and reward-seeking behavior, a learning rate signal that is a function of volatility in the environment plus some noise, and when the reward probability is 80% or 20%. In all cases, we found that the sign of the slope between a learning rate control signal on a trial and the subsequent cue value update is the sign of the RPE on the trial (Figure S1). We also verified that this is true immediately after the reward probability is reduced from 100% to 50%, i.e., as the new cue value is being learned, or long after the new cue value has been stably learned (Figure S1). Thus, this test for learning rate control remains true under a variety of assumptions.

To sum up, we used this discriminative test to identify whether vmOFC activity abides by learning rate control. The above predictions for learning rate control are also true for a signal that controls the absolute magnitude of the RPE. Indeed, absolute magnitude of the RPE has itself been hypothesized to be a driver of learning rate [13]. Hence, we will consider a signaling of the absolute magnitude of the RPE as another model of learning rate control for now and will consider this possibility later in the manuscript.

Trial-by-trial and average activity in vmOFC outcome responses reflects learning rate control

To investigate the above test for learning rate control, we behaviorally measured cue-reward learning on a discriminative Pavlovian trace conditioning task in head-fixed mice [26,32] (Figure 2A). In this task, an auditory cue paired with a delayed reward (labeled CS+ for conditioned stimulus predictive of reward) and another auditory cue paired with no reward (labeled CS−) were randomly interleaved across trials (Figure 2A). Once learned, mice show anticipatory licking following CS+, but not CS− (Figure 2B). Thus, the anticipatory licking is a behavioral proxy for reward prediction/cue value.

We next validated that updates in anticipatory licking are indeed a proxy for cue value updates. To this end, we tested whether anticipatory licking updates positively after positive RPE and negatively after negative RPEs, consistent with cue value updating. To reliably induce positive and negative RPEs, we reduced the probability of reward following CS+ to 50% (Figure 2C). In this session, reward receipt following CS+ induces a positive RPE (as received reward is larger than the predicted reward) and reward omission following CS+ induces a negative RPE (as received reward is less than the predicted reward). We measured the update in anticipatory licking due to positive and negative RPEs. We found that if a given CS+ trial is rewarded (i.e., positive RPE), animals increased their anticipatory licking on the next CS+ trial (may occur after interleaved CS− trials) (Figure 2D). Similarly, if a given CS+ trial is unrewarded (i.e., negative RPE), animals reduced their anticipatory licking on

the next CS+ trial (Figure 2D). Thus, consistent with cue value update, anticipatory licking increases after a positive RPE and decreases after a negative RPE.

However, it may be possible that such updating is instead driven by learning independent changes such as the motivation to lick. If so, after a rewarded CS+ trial, an increased motivation to lick should result in an increase in anticipatory licking even on the subsequent CS− trial. To test this possibility, we identified CS− trials that immediately followed CS+ trials and assayed whether anticipatory licking on these CS− trials depended on the outcome of the previous CS+ trial. We found that the update in CS− anticipatory licking did not depend on whether the previous CS+ trial was rewarded (Figure 2E). One concern with this control analysis could be that motivation may be present only when there is a reward expectation. Since animals do not expect reward after CS−, a motivation confound may not be testable on these trial types. However, we found this assumption to not be true. This is because licking during the pre-cue baseline, a period with no reward expectation, nevertheless showed clear effects of motivation depending on the previous trial's outcome (Figure S2). However, once this baseline shift in licking due to motivation was subtracted, updates in cue-induced anticipatory licking was quantitatively consistent with RPE-induced updates in cue value (Figure S2). Thus, an update in anticipatory licking across CS+ trials provides a behavioral measure of cue value update.

We next tested whether the outcome responses of vmOFC neurons is consistent with learning rate control. If so, based on the results in Figure 1D-F, we would predict that vmOFC activity on a CS+ trial would be positively correlated with the subsequent CS+ anticipatory licking update after positive RPE, and negatively correlated with the subsequent CS+ anticipatory licking update after negative RPE. To measure activity in a large number of individual vmOFC neurons during reward prediction learning, we used two-photon microendoscopic calcium imaging [26] (Figure 2F). We found that the reward responses of vmOFC neurons (measured within 3 s after reward delivery) showed positive correlations with the subsequent update in CS+ anticipatory licking on positive RPE trials (Figure 2G-I; see Table S1 for a compilation of all statistical results including all details and sample sizes for all figures). We further found that the reward omission responses of vmOFC neurons (measured within 3 s after reward omission) showed negative correlations with the subsequent update in CS+ anticipatory licking on negative RPE trials (Figure 2G-I). Thus, these results are consistent with learning rate control and rule out the alternative possibilities of vmOFC responses controlling RPE or reward value. This is because these models predict a positive correlation of neural responses with cue value update on negative RPE trials. These results further rule out vmOFC reward responses controlling learning independent factors such as motivation. Mathematically, learning independent factors can be treated as a baseline shift independent of RPE in the equation shown in Figure 1A (i.e., ΔCue Value = RPE*learning rate + learning independent factors). Thus, the slope between such learning independent factors and cue value update should be independent of the RPE. In other words, an increase in learning independent factors such as motivation should produce more anticipatory licking on the next trial even if the current trial has a negative RPE. Thus, the above results are also inconsistent with an encoding of motivation. Finally, we previously demonstrated that vmOFC outcome responses causally control behavioral updating [26]. Specifically, we observed that inhibition of vmOFC outcome responses impairs

CS+ anticipatory licking update following RPE (replotted in Figure S2). Together, these results strongly support the conclusion that vmOFC outcome responses control learning rate.

We next tested whether the average activity of vmOFC neurons abides by additional predictions of learning rate signaling. For convergence of reward prediction learning in a stationary environment, theoretical reinforcement learning models require the average learning rate to reduce over time [8,11,33]. Thus, a neural signal for learning rate should reduce over the course of initial learning when reward probability following CS+ was stably maintained at 100%. Consistent with this, we found that the overall response of vmOFC neurons to reward receipt reduced after learning (Figure 3A-C).

As there is considerable heterogeneity in vmOFC neural responses, we next identified neuronal subpopulations with similar task-induced activity patterns using a clustering approach [26]. This approach allowed us to identify neurons that are active at similar times in the task, i.e. neuronal ensembles in vmOFC. Using this approach, we previously showed that vmOFC neurons cluster into nine different neuronal subpopulations/ensembles based on their time-locked activity profiles to the CS+ and CS− late in learning (Figure 3D) [26]. Each neuron was assigned a cluster identity based on its activity late in learning when the reward probability was 100%. Due to our ability to longitudinally track the same neurons [26], we were able to evaluate the response of these neuronal clusters under many other task conditions discussed later. Among all these clusters of neurons, we found positive reward responses early in the learning of the CS+-reward association (Figure 3D, E). Once mice learned to predict the upcoming reward upon CS+ exposure, there was a reduction in the reward responses of many clusters (Figure 3D, E). Some clusters (most prominently clusters 1 and 3) even displayed a negative reward response late in learning (Figure 3D). These data show that the suppression of reward responses after reward prediction is specific to some neuronal subpopulations within vmOFC.

We next tested one more prediction for learning rate signaling. Learning rate should be higher in the presence of unexpected environmental variability [8,11,33]. Consistent with this, compared to the session late in learning with 100% reward probability, the responses on rewarded trials was higher after switching to either the 50% reward probability session or another session with unpredictable rewards in the intertrial interval, both sessions with unexpected environmental variability (Figure S3). Thus, the reward responses of some vmOFC neuronal subpopulations, including those projecting to ventral tegmental area (VTA)—a midbrain regulator of learning—abide by this additional prediction for learning rate signaling.

Reward responses of specific vmOFC neuronal subpopulations signal the relative salience of a reward

As mentioned in the introduction, learning rate can be modulated by several variables. For instance, the Pearce-Hall model proposes that learning rate is modulated by the absolute magnitude of RPE [13]. Though this variable is thought to modulate the salience of cues predicting rewards, it is calculated at the time of the outcome. For simplicity, we will refer to this salience calculated at outcome time as outcome salience. Compared to an unpredicted reward, a predicted reward in a stationary environment should have low salience

and learning rate, as the absolute magnitude of the RPE at its occurrence is low. A signaling of salience is consistent with many of the above findings (though see **Supplementary Note 2** for a detailed treatment). However, learning rate may also be modulated by the relative salience of a stimulus in relation to all other stimuli in the environment [34,35]. The relative salience of a stimulus in comparison to other stimuli has been shown to drive attentional capture by that stimulus [36,37]. Unlike the absolute magnitude of the RPE, the relative salience of a stimulus depends (by definition) on the other stimuli that are also present in the environment. Therefore, the relative salience of a stimulus such as an unpredicted reward should reduce when a much more salient stimulus is introduced in the rewarding context. We next tested if the vmOFC reward responses are consistent with a signaling of the absolute magnitude of RPE or the relative salience of a reward.

To this end, we investigated whether vmOFC reward responses are suppressed in a context independent of reward prediction due to the presence of another salient stimulus. Animals typically have higher learning rates for punishments than for rewards, and for salient stimuli compared to relatively less salient stimuli [38–42]. For instance, prediction of highly salient aversive stimuli such as foot shocks or quinine (a bitter tastant) often occurs in single trials [43,44]. We thus hypothesized that delivering rewards in a context that also includes the delivery of salient aversive stimuli would result in a suppression of vmOFC reward responses due to a relative reduction in the salience of sucrose, *independent* of the suppression due to reward prediction. To minimize sensory confounds, we used an aversive stimulus delivered using the same sensory modality as the reward (i.e., taste). We thus intermittently and randomly (i.e. unpredictable) delivered drops of either sucrose or high concentration quinine (1.5–2.5 mM) in a 3:1 ratio to headfixed mice (Figure 4A) (sucrose-quinine experiment). Since the liquid deliveries were unpredictable and mostly sucrose, the animals consistently sampled the liquid to ascertain whether a given drop was rewarding or aversive. On a given trial, mice quickly suppressed their licking if the liquid was quinine, demonstrating aversion (Figure 4B, C).

In this task, quinine may also have been more salient than sucrose for two additional reasons. One, the mice in our task had considerable experience receiving sucrose under the two-photon microscope. However, quinine was a novel stimulus in the sucrose-quinine session. Novelty typically increases salience. Two, the presence of quinine reduces the context-reward association such that a high frequency of quinine could make licking in the context aversive, thereby making an estimation of quinine frequency more salient for deciding whether to lick. Regardless of the exact mechanism, this experiment thus allowed us to test whether vmOFC sucrose responses reduce in a manner consistent with the signaling of a reduction in the relative salience of sucrose.

Consistent with a signaling of relative salience, we found that unpredicted sucrose responses in vmOFC neurons were suppressed when delivered in a context containing quinine (Figure 4D, E). In comparison, reward responses early in learning (Figure 4D, E, Figure 3B-D, Figure S4), and unpredicted reward responses after learning in the absence of quinine (Figure 4F), were positive across all clusters. To further demonstrate the effect of quinine presentation on the response to sucrose, we compared the sucrose trials prior to the first experience of quinine against the remaining sucrose trials in the first sucrose-quinine session

(Figure 4G). We found that the sucrose responses reduced considerably within the same session after the first experience of quinine. Overall, these results support relative salience signaling. We will note however that since this task does not differentiate motivational versus learning effects on licking, we cannot test whether learning rates are larger for quinine compared to sucrose in this task. Lastly, our use of a highly aversive concentration of quinine allowed us to demonstrate that an encoding of relative value or a lingering taste of quinine affecting sucrose encoding are not consistent with these observations (see **Supplementary Note 2** in [45]).

One potential concern with the above results could be whether the reward response adaptation due to reward prediction or quinine are correlated within the *same* neurons. To test this, we evaluated the activity of only those neurons that were longitudinally tracked across these sessions (Figure 5). We found that the reward responses were correlated across all three conditions in these neurons (Figure 5, correlations quantified in Table S1). Thus, the reward response adaptation due to reward prediction or the presence of quinine are correlated across the vmOFC neuronal population. This is also true for reward responses in the session with 50% reward probability (Figure S5). We further ruled out the possibility that vmOFC reward responses simply reflect an efference copy of an arousal signal that may vary across these conditions (Figure S5). To this end, we tested whether changes in pupil diameter in darkness, a measure of arousal, show similar changes as those seen in vmOFC subpopulations. We found that the mean pupil diameter responses to sucrose early in learning and after 50% reward probability reduction do not correlate with the corresponding vmOFC reward responses (Figure S5). Nevertheless, consistent with a reduction in the relative salience of sucrose in a context containing quinine, we observed a highly dampened pupil dilation for sucrose in this context (Figure S5). These observations show that a simple efference copy of an arousal signal is not sufficient to explain vmOFC reward responses. Thus, consistent with a system that signals the relative salience of a reward, vmOFC reward responses reduce in two independent settings due to reward prediction or the reduction of relative salience of unpredicted rewards. Cumulatively, the reward responses observed in vmOFC rule out typically assumed reinforcement learning variables such as RPE, absolute magnitude of RPE, reward value, relative reward value and expected future value (treated in detail in **Supplementary Note 2** of [45]).

Medial thalamic inputs to vmOFC control relative salience signaling in vmOFC

We next assessed the neuronal circuit mechanism for the relative salience signaling in vmOFC neurons. We hypothesized that reward responses in vmOFC may at least be partially controlled by inputs from medial thalamus (mThal). This is because a wide array of reward responsive regions such as basolateral amygdala, other prefrontal cortical regions, and pallidal regions project to the medial thalamus and can indirectly control vmOFC reward responses through mThal [46,47]. Further, disconnection studies have shown that interactions between mThal and OFC are necessary for reward related decision-making [48]. Despite this, whether reward responses in mThal→vmOFC input exhibit relative salience signaling and whether this input causally affects relative salience signaling in vmOFC is unknown. We first identified the anatomical locations of thalamic cell bodies projecting to vmOFC using viral [49] and non-viral [32] retrograde tracing approaches (Figure 6A). The

predominant thalamic structures projecting to vmOFC are the anteromedial and mediodorsal thalamic nuclei (Figure 6B, Figure S6). We then investigated the reward response plasticity of input from these regions to vmOFC. We compared unpredicted reward responses of mThal→vmOFC axons in sessions without and with quinine (Figure 6C-E). We found largely positive responses in these axons in response to unpredicted sucrose rewards in the absence of quinine (Figure 6F, Figure S6). These reward responses were suppressed in a session containing quinine (Figure 6F), showing qualitative correspondence with the reward response adaptation observed in vmOFC neuronal responses.

These results suggest that mThal input might contribute to the relative salience signaling observed in vmOFC neurons. To test the causal influence of mThal input on vmOFC reward responses, we optogenetically inhibited mThal→vmOFC input after reward delivery while imaging from vmOFC neurons (Figure 7A). To remove light artifacts, we discarded the imaging frames during optogenetic inhibition, and evaluated reward responses right after the termination of inhibition. Since GCaMP6s responses are slow with a decay time of roughly two seconds [50], a change in activity during the inhibition will be apparent even after the inhibition for up to two seconds. We found that individual neurons showed both positive and negative modulation of activity due to mThal inhibition (Figure 7B). To test whether mThal input affects vmOFC reward response adaptation, we inhibited mThal→vmOFC axons in a session containing unpredicted deliveries of sucrose and quinine. We found that the reduction in sucrose responses due to the presence of quinine was significantly dampened upon mThal inhibition, in all clusters except clusters 4 and 5 (Figure 7D). There was also a non-selective change in quinine responses in some clusters (Figure S7). Two potential confounds for this experiment are that vmOFC responses may reflect the presence of light due to the LED, or that mThal→vmOFC axons may show rebound excitation after the one second inhibition. We ruled these confounds out because we observed no effect on vmOFC neurons during spontaneous inhibition of mThal→vmOFC axons in the absence of rewards, or in virus control animals without opsin expression (i.e. with LED but no inhibition) (Figure S7). Therefore, these results strongly support a causal role for mThal→vmOFC axons in controlling the relative salience signaling in vmOFC.

Considering the cluster-wise variability in the effect of mThal→vmOFC inhibition on vmOFC reward responses, we next tested whether this variability is related to the variability in response adaptation across clusters. If mThal→vmOFC is responsible for controlling relative salience signaling and not just for controlling the positive reward responses (see Discussion below), the cluster-wise variability in relative salience signaling should predict the cluster-wise variability in the effect of mThal→vmOFC inhibition on vmOFC reward responses. We indeed found that the suppression of reward response due to the presence of quinine (Figure 4) predicts the average effect of mThal inhibition on a given vmOFC neuronal cluster/subpopulation (Figure 7E). Overall, these results demonstrate that the relative salience signaling in vmOFC neurons depends on mThal inputs.

DISCUSSION

Recent studies have begun investigating OFC activity using two-photon calcium imaging [26,51–53]. Using this approach to longitudinally track neuronal activity across tasks, we

identified neuronal subpopulations in vmOFC that signal learning rate. We propose that the most parsimonious explanation of our major findings is a learning rate control signal (see **Supplementary Note 2** in [45]). Consistent with this signal acting to control learning rate, there is significant correlation between vmOFC outcome response on a CS+ trial and the behavioral updating on the subsequent CS+ trial. Crucially, the sign of this correlation is the sign of the RPE on a trial—a strong prediction for learning rate control. Along with our previous findings that vmOFC reward responses causally controls behavioral learning based on recent reward history [26], the current results provide strong evidence that vmOFC neurons act to control learning rate. Interestingly, the learning rate signaling in vmOFC is consistent with a general signaling of the relative salience of a reward across different contexts. We then show that medial thalamic inputs to vmOFC exhibit qualitatively similar response adaptation as vmOFC neurons, and causally control the relative salience signaling of specific vmOFC subpopulations. Overall, these results bolster the emerging theoretical view that the prefrontal cortex acts as a meta learning system that can adaptively control learning rate [9], in addition to representing cognitive parameters for learning such as uncertainty [10], confidence [54–56], surprise [57], value [58,59] and volatility [12].

A key challenge in identifying neuronal encoding of learning rate control is that learning rate is a latent cause of behavior. Our approach for the identification of learning rate control is different from that used commonly in the field. Prior studies first obtain an independent measure of learning rate by fitting a reinforcement learning model to the behavioral changes due to expected and unexpected environmental variability [12,17–20]. Since behavioral learning depends on both learning rate and RPE, such an approach requires the careful dissection of RPE magnitude and learning rate on a trial-by-trial basis. Once a trial-by-trial estimate of learning rate is thus obtained, these studies look for correlation between trial-by-trial changes in neural activity with this independently measured estimate of learning rate. Instead, we develop a simpler, but powerful, discriminative test for learning rate control. The trial-by-trial activity of neurons controlling learning rate should correlate with the subsequent trial's cue value update (measured by reward seeking), with a positive correlation after positive RPE trials and a negative correlation after negative RPE trials. This prediction results from the fact that learning rate acts as a positive multiplicative gain on the influence of RPE on cue value update. Importantly, this prediction is true under a variety of additional assumptions including a dependence of learning rate on volatility, a non-linear mapping between cue value and cue-induced reward seeking, or the probability of the reward following the cue. Thus, we believe that this approach can be used as a general test for identifying neuronal activity consistent with learning rate control.

In typical reinforcement learning algorithms, learning is "complete", i.e. reward is predicted, when RPE at reward receipt becomes zero. Nevertheless, when rewards are delayed from their predictors, the resultant temporal uncertainty in the subjective estimation of the delay causes RPE to be significantly positive even after learning. Indeed, under delays less than even two seconds, midbrain dopaminergic neurons exhibit significant positive responses to a predicted reward even after extensive training [6,60–63]. Thus, it is not possible to evaluate whether a received reward was fully predicted using only the RPE signal on a trial. The reward response adaptation observed in some vmOFC subpopulations may provide an

explicit signal that a delayed reward is as predicted as the reward response reverses in sign once a delayed reward is predicted. Hence, the signal from vmOFC can counteract the positive RPE signal from midbrain dopaminergic neurons and signal that a delayed reward is predicted despite the uncertainty in subjective timing. A similar "brake" on learning after the prediction of a delayed reward has been assumed in computational models of the learning of reward timing [64,65]. In this sense, the vmOFC reward response adaptation takes into account whether the temporal uncertainty in a delayed reward is *expected* uncertainty due to the uncertainty in estimation of the fixed delay, a computation that has been behaviorally shown to exist in rodents [66]. This is also consistent with a prior study on the ventral OFC [67].

We have shown that the general role of OFC in reward learning results at least in part from the control of behavioral learning rate. Consistent with a control of behavioral learning rate by OFC, we previously found that inhibition of vmOFC→VTA neurons during the reward, but not the cue, suppresses behavioral learning based on recent reward history [26]. A previous study also found that inhibition of lateral OFC (likely also containing some ventral OFC) during the reward, but not cue, period of an instrumental task, affected behavioral adaptation dependent on reward history [23]. Further, lesions of medial OFC affect learning and representation of outcomes associated with an action [68]. This finding is potentially consistent with the control of learning rate, as lesion of medial OFC may cause "over-learning" of an action, thereby making it less sensitive to the expected outcome.

Careful future experiments are needed to examine the generality of these findings to different behavioral tasks and states. For instance, it has been argued that OFC function might be fundamentally different in Pavlovian versus instrumental tasks due to the differences in the mental representation of these tasks [68]. Thus, it remains to be seen whether our findings would generalize to instrumental learning. Further, adaptive control of learning rate is typically studied in the context of dynamic uncertain environments. While some aspects of our task are dynamic (e.g. change in contingency while recording from the same neurons), it is necessary to test the generality of these findings in relation to the control of learning rate by expected and unexpected uncertainty [10–12,17]. It will also be interesting to test whether OFC outcome representation in bandit tasks relate to learning rate [69]. One interesting aspect of our task is that some clusters encode the outcome of a trial well past the reward was delivered (e.g., cluster 4, 6 in Figure 3D). Such long-lasting outcome encoding has previously been observed across different task conditions and species in the OFC [54,69,70]. A limitation of our study is that it was conducted under moderate water deprivation and with a specific type of reward (sucrose) and aversive stimulus (quinine). It is likely that varying the levels of water deprivation will change the behavioral salience of these liquids. Thus, careful studies are required to ascertain the generality of these findings across behavioral states and different types of rewards/punishments. Relatedly, an aberrant learning rate for drugs of abuse in OFC (e.g. higher learning rate for positive RPEs compared to negative RPEs) might partially explain the role of OFC in drug addiction [71,72]. If our findings generalize to these different settings, that would suggest that OFC acts more generally as a system that prioritizes currently available rewards or punishments for learning, based on the current behavioral state.

Future studies are also required to test whether relative salience signaling occurs in other brain regions. The anterior cingulate cortex has often been suggested to reflect variables required for adaptive learning rate control [10,12,57]. Correlates of Pearce-Hall like salience have also been found in amygdala [73]. Nevertheless, it is unclear if these regions exhibit a strong reduction in responses to delayed, but predicted rewards and rewards presented in a context containing more salient unconditioned stimuli. Interestingly, the ventromedial prefrontal cortex, an area adjacent to orbitofrontal cortex, primarily shows lower responses to aversive outcomes than rewarding outcomes at the single neuron level [74]. This appears different from the results here and suggests that there may be considerable differences between the encoding of nearby prefrontal cortical regions. An interesting distinction from the function of other medial prefrontal cortical regions is that vmOFC activity does not control the expression of learned licking behavior [26], but regions such as the prelimbic cortex do [32,75].

Another major unresolved question is how OFC controls behavioral learning rate by interacting with the midbrain dopaminergic neurons signaling RPE [2,7]. OFC has been shown to project heavily to the striatum [72,76,77]. This projection may be particularly important for an interaction with the dopaminergic system. Though we have investigated vmOFC→VTA neurons in this and our previous study [26], this projection passes through the striatum, and may send collaterals to the striatum. Thus, we cannot rule out the possibility that collaterals in the striatum mediate some of the effects of vmOFC→VTA inhibition in our previous study [26]. One especially promising circuit mechanism may be via the low threshold spiking interneurons in dorsal striatum, which show similar reduction in reward responses during instrumental learning [78]. If true, this would raise an interesting possibility that the role of this output and OFC in the control of habitual behavior or behavior insensitive to negative outcomes [72,76,77,79] may be due to its role in controlling learning rate (e.g. learning rate to negative outcomes set too low). OFC could also control behavioral learning rate through its interactions with anterior cingulate cortex [10].

Though the circuit mapping experiments performed here cannot ascertain whether the role of mThal input to vmOFC is unique among other inputs, three features of the observed results are worth highlighting. One, inhibition of mThal input to vmOFC changes activity in vmOFC clusters proportional to the amount of adaptation in reward responses (Figure 7E). In fact, we showed that 50% of the cluster-wise variance in the effect of mThal inhibition is explained by the variance in adaptation between clusters. Such a robust causal relationship between mThal input and vmOFC output suggests that mThal might indeed be one of main inputs contributing to the reward response adaptation observed in vmOFC. Two, a concern could be that that any strong input with reward activity would control reward response adaptation in vmOFC. An implicit assumption of this concern is that the observed effect is simply a reflection of the strength of reward encoding. The cluster with the strongest reward response in vmOFC is cluster 5 (Figure 3, 4). If a strong input to vmOFC contributes to the strongest output (i.e., cluster 5), we would expect a significant change in the activity of cluster 5 due to input inhibition. Yet, this cluster had one of the lowest effects of mThal→vmOFC inhibition (Figure 7). Thus, the effect of inhibition is not simply related to the strength of vmOFC activity. In this light, it is unlikely that any strong input would reproduce this effect. Lastly, thalamic outputs are excitatory [80] and exhibit

positive reward responses (Figure 6F). Yet, inhibition of mThal→vmOFC input increases neuronal response to reward in vmOFC excitatory neurons. These results imply a key role for vmOFC inhibitory interneurons in shaping the responses of vmOFC output neurons. This means that while mThal inputs are integral for the reward response adaptation in vmOFC output neurons, the computation happens within the vmOFC local circuit. Despite these noteworthy features of the mThal input to vmOFC, future work needs to address whether other inputs contribute to learning rate control by vmOFC. For instance, dorsal raphe serotonergic neurons have been shown to control learning rates [17] and may potentially do so through their interactions with OFC.

A potential confound for the response change in vmOFC due to inhibition of mThal→vmOFC input is that the observed vmOFC response changes may be due to a change in behavior resulting from the input inhibition. However, this is unlikely as the inhibition we performed is unilateral and restricted to the field of view under the lens. Indeed, based on tissue scattering and the LED power, it is likely that the inhibition of thalamic axons occurs only within 200 μm or less of tissue under the lens. In the behavior we measured (licking), we observed no effect of inhibition of thalamic axons (Figure S7C). Considering that behavioral effects in prefrontal circuits are typically only apparent in strong bilateral inhibition/lesion, it is highly unlikely that the small-scale inhibition in the local circuit under the lens produced any unmeasured behavioral effects that indirectly modulated vmOFC neuronal responses.

In conclusion, we have shown that vmOFC reward responses signal the learning rate for rewards. Whether these results generalize to a broad role for OFC in prioritizing environmental stimuli for learning remains to be tested. Nevertheless, the identification of mThal→vmOFC circuit as one involved in the control of learning rate opens the possibility to study the neural circuit control of the learning of parameters for learning, i.e., meta learning.

STAR Methods

RESOURCE AVAILABILITY

Lead Contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Garret D Stuber (gstuber@uw.edu)

Materials Availability: This study did not generate new unique reagents.

Data and Code Availability:

- Microscopy and behavioral data reported in this paper for clustering is available on the Stuber lab Github page, along with the original code for clustering and the learning rate simulations (https://github.com/stuberlab).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects and Surgery: All experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of North Carolina and University of Washington and accorded with the Guide for the Care and Use of Laboratory Animals (National Institutes of Health). Adult male and female wild type C57BL/6J mice (Jackson Laboratories, 6–8 weeks, 20–30 g) were group housed with littermates and acclimatized to the animal housing facility until surgery. Survival surgeries were stereotaxically performed while maintaining sterility, as described previously [26,81]. Induction of anesthesia was carried out by using 5% isoflurane mixed with pure oxygen (1 L/min) for roughly thirty seconds to a minute, after which anesthesia was maintained using 0.6–1.5% isoflurane. The surgeon monitored respiratory rate intermittently to ensure appropriate depth of anesthesia. The animals were placed on a heating pad for thermal regulation. Data from animals used in Figure 2–5, Figure S2, Figure S3 and Figure S4A, B were collected at UNC. For these animals, pre-operative buprenorphine (0.1 mg/kg in saline, Buprenex) treatment was given for analgesia. Eyes were kept moist using an eye ointment (Akorn). 2% lidocaine gel (topical) or 1mg/kg lidocaine solution was applied or injected onto the scalp prior to incision. Details of viral injection, lens and optic fiber implantation are provided below. A custom-made stainless-steel ring (5 mm ID, 11 mm OD, 2–3 mm height) was implanted on the skull for headfixation and stabilized with skullscrews as well as dental cement. Animals either received acetaminophen (Tylenol, 1 mg/mL in water) in their drinking water for 3 days, or 5 mg/kg carpofen 30 min prior to termination of surgery for post-operative analgesia. Animals were given at least 21 days (and often, many more) with ad libitum access to food and water to recover from surgery. Following recovery, animals used for behavioral studies were water deprived to reach 8590% of their pre-deprivation weight and maintained in a state of water deprivation for the duration of the behavioral experiments. Animals were weighed and handled daily to monitor their health. The amount of water given daily was between 0.6–1.2 mL and was varied for each animal based on the daily weight. A total of 38 mice were included in this study: 5 OFC-CaMKII imaging (0 female, Figure 2–4), 7 OFC-VTA imaging (0 female, Figure S3, 4), 5 mThal→vmOFC axon imaging (3 female, Figure 6, Figure S6), 7 mThal→vmOFC inhibition or control inhibition during vmOFC imaging (3 female, Figure 7, Figure S7), 10 anatomical tracing (8 female, Figure 6B), and 4 pupil diameter tracking (0 female, Figure S5).

METHOD DETAILS

Head-fixed behavior: Trace conditioning was done exactly as before [26]. A brief outline of these methods is summarized here. Water deprived mice were first trained to lick for random unpredictable sucrose (10–12.5%, ~2.5 μL) deliveries in a conditioning chamber. Mice received one of two possible auditory tones (3 kHz pulsing tone or 12 kHz constant tone, 75–80 dB) that lasted for 2 seconds. A second after the cues turned off, the mice received a sucrose reward following one of the tones (designated CS+), whereas the other tone resulted in no reward (designated CS−). The identity of the tones was counterbalanced across mice in all experiments. The cues were presented in a pseudorandom order and in equal proportion until a total of 100 cue presentations (trials) were completed. The intertrial interval between two consecutive presentations of the cues was drawn from a

truncated exponential distribution with mean of 30 s and a maximum of 90 s, with an additional 6 s constant delay. Early in learning (Figure 2) was defined as the first session of conditioning. Late in learning (Figure 2) was defined as the day that the area under a receiver operating characteristic curve (auROC) of lick rates to CS+ versus CS− remained high and stable (auROC larger than 0.7 on at least 2 consecutive sessions or larger than 0.85). Two contingency degradation experiments were performed as described previously [26], with reward probability reduced 50% in one (Figure 2,3) and background unpredictable rewards introduced in the intertrial interval in the other, with reward probability set to 100% (Figure S3, Figure 4). Exact parameters for the rates of unpredicted rewards were described previously [26]. For the sucrose and quinine experiment (Figure 4), drops of sucrose (10–12.5%, ~2.5 μL) or quinine hydrochloride dihydrate (1.5–2.5 mM, ~2.5 μL) were randomly delivered at a 3:1 ratio (60 sucrose drops and 20 quinine drops). The interdrop interval was a minimum of 13–18 s and a maximum of 2328 s. These were chosen to maintain a sufficient interval between consecutive drops so as to prevent any bleed-through of GCaMP fluorescence. Even though the hazard rate is not flat for these intervals, the animals did not show any behavioral evidence of temporal expectation of the delivery times.

2-photon microscopy: The methods were similar to those published previously [26]. We used a calcium indicator (GCaMP6s) to image calcium changes using 2-photon microscopy. The injection coordinates and volumes for virus as well as the coordinates for implanting a gradient refractive index (GRIN) lens were as published previously [26]. For mThal axon imaging (Figure 6) or inhibition (Figure 7), we injected 400500 nL of AAVDJ-CaMKIIα-GCaMP6s (at an effective titer of ~$1–2\times10^{12}$ infectious units per mL, UNC Vector Core) or AAV5-CaMKII-eNpHR3.0-mCherry (~4×10^{12} infectious units per mL, UNC Vector Core) unilaterally in mThal (−1.3 AP, 0.5 ML, −3.5 DV from bregma). We used a resonant scanner (30 Hz frame rate acquisition, Olympus Fluoview FVMPE-RS) and performed an online averaging of 6 times to get an effective frame rate of 5 Hz, to minimize the size of recorded files as we had negligible motion artifacts. A GaAsP-PMT with adjustable voltage, gain and offset was used, along with a green filter cube. We used a long working distance 20x air objective, specifically optimized for infrared wavelengths (Olympus, LCPLN20XIR, 0.45 NA, 8.3 mm WD). We imaged either at 955 or 920 nm using a Ti-sapphire laser (SpectraPhysics, ~100 fs pulse width) with automated alignment. The animals were placed on a 3-axis rotating stage (Thorlabs, TTR001/M) to precisely align the surface of the GRIN lens to be perpendicular to the light path, such that the entire circumference of the lens is crisply in focus (within 1–2 μm). The imaging acquisition was triggered by a custom Arduino code right before the start of a behavioral session, and a TTL output of every frame was sent as an input to the Arduino. The imaging acquisition was triggered off at the end of the behavioral session (~ one hour). The data in Figure 2, 3 (OFC-CaMKII) and Figure S3 (OFC→VTA) were re-analyzed based on data from a previous publication [26]. In every mouse, one z-plane was imaged throughout acquisition so that the same cells could be tracked through learning [26]. After mice were trained, other z-planes were also imaged (one per session) to get a measure of the total functional heterogeneity in the network. A total of 2–6 z-planes per mouse were imaged in the OFC-CaMKII group, whereas 1–3 z-planes were imaged in the OFC-VTA group. Thus, responses early in learning were from the plane tracked throughout learning, whereas the responses late in learning, including the

contingency degradation sessions, were from all imaged planes. The sucrose quinine session was run for most imaging planes at the end of the conditioning experiments [26]. The data in Figure 4 were collected from the same animals as in Figure 2, 3, but have never been published previously.

Optogenetics during imaging: These animals received an injection of AAVDJ-CaMKIIα-GCaMP6s in vmOFC and AAV5-CaMKIIαeNpHR3.0-mCherry (experimental) or AAV5-CaMKIIα-mCherry (control) in mThal. All animals showed significant expression of the opsin in mThal. Light was delivered for optogenetic inhibition to the full field of imaging by an LED kit with a peak wavelength of 615 nm (FV30SP-LED615, Olympus) [82]. The frames containing light artifacts due to LED illumination were masked out for analysis (shown as white bars in Figure 7B, C). The same preprocessing pipeline as before (including motion correction, signal extraction and neuropil correction) was employed on this masked data. These animals were first trained to lick in response to random sucrose deliveries. We then performed optogenetic inhibition during sucrose consumption for multiple imaging planes. A random half of the trials received inhibition. The effect of the LED was calculated by comparing sucrose fluorescence on the trials with and without LED. The animals were subsequently trained on the trace conditioning paradigm (no imaging). Once anticipatory licking was high and stable, we imaged the same neurons that were imaged earlier to obtain PSTHs around CS+ and CS−. These PSTHs were used to classify neurons into the clusters identified using the much larger population of neurons in Figure 3. The classification was done using a linear support vector classifier (Scikitlearn), as was used previously for classifying OFC→VTA neurons [26]. We then performed inhibition of mThal axons while imaging from vmOFC neurons during the sucrose and quinine session (Figure 7C, D). Here, to obtain a sufficient number of trials with and without inhibition, we first performed recordings without inhibition (80 trials), followed by with inhibition (80 trials). The effect of LED was calculated by comparing sucrose or quinine fluorescence with and without LED.

Retrograde tracing, histology and microscopy: 400 nL of rAAV2retro-hSyn-eYFP (~2×10^{12} infectious units/mL) or CTB-488 were stereotaxically injected at roughly 2.5 AP, 0.5 ML and 2.3–2.5 DV from bregma using the surgical methods described above. 3–5 weeks after surgery, animals were euthanized with an overdose of pentobarbital (~390 mg/kg, Somnasol, Covetrus EU-HS-045–100-0), and transcardially perfused with 4% paraformaldehyde (PFA, Sigma-Aldrich, #158127). Perfused brains were incubated in 4% PFA overnight and moved to a 30% sucrose solution (Sigma-Aldrich, #S0389) for ~2 days prior to cryosectioning. 40 μm thick sections were used in tracing experiments. For retroAAV2 thalamic labeling (Figure 6B, Figure S6), eYFP signal was enhanced and stabilized using a chicken anti-GFP antibody (Aves Lab, #GFP-1020, 1:500 dilution), paired with a donkey anti-chicken secondary (Jackson Immunoresearch, #703–545-155, 1:1000 dilution). GFP and eYFP have highly similar protein sequences, which allows the use of a GFP antibody for immunostaining. Brain sections were imaged using a 20x air objective on a confocal microscope (Olympus Fluoview FV3000). Resulting image tiles were stitched, and Z-stacks were taken at ~1 μm intervals and averaged across slices yielding a maximum intensity projection image. Brain atlas outlines (https://mouse.brainmap.org/static/atlas) were overlaid onto each image to allow assignment of thalamic subregions, in which labeled

cells were counted using ImageJ (https://imagej.net/Fiji). Percentage of total thalamic cells labeled (Figure 6B) was quantified as the number of cells per region divided by the sum total of all counted eYFP+ or CTB+ cells. The intermediodorsal nucleus was counted as part of the mediodorsal region.

Pupil measurements: Pupil area measurements were performed on an independent cohort of animals that went through the complete head-fixed behavior paradigm. Analysis was performed for the unpredicted sucrose without quinine condition, the early in learning condition, the 50% reward probability and the unpredicted sucrose and quinine delivery conditions. Pupil recordings were performed using a monochrome USB 2.0 CMOS Camera (ThorLabs DCC1545M) at 5 Hz. A triggered red LED flash in the inter-trial interval was used to align behavior and camera recordings. This flash occurred outside of the analyzed window for all recordings. To align the behavior and pupil recordings, LED flashes were detected using ImageJ and MATLAB to identify large fluctuations in pixel intensity that corresponded to LED onset from an ROI that contained the LED. The LED onset timestamps for the video recordings were aligned to the LED trigger timestamp in the behavior recordings on a trial-by-trial basis to account for dropped frames across the session, though very few frames were dropped overall. After the video and behavior recordings were aligned, we extracted data from the interval spanning 3 seconds before cue presentation to 3 seconds after reward delivery for the early in learning and 50% reward probability conditions, and 3 seconds before and after the first lick following uncued fluid delivery for the unpredicted sucrose without quinine and unpredicted sucrose with quinine sessions. We preprocessed the data with two runs of the CLAHE ImageJ plug-in [83] with default parameters to enhance local contrast for pupil discrimination. After preprocessing, we performed an average intensity grouped z-projection of 5 frame bins to reduce the data to 1 Hz for manual annotation of the pupil. We then drew an ellipse bounding the pupil for each resulting frame and extracted the area of the ellipse for each frame. For sessions with multiple trial types, we did not attach trial identity data to each trial until all pupil measurements were complete.

Temporal Difference (TD) Learning simulations: Here, we describe the methods used for the simulations used in Figure 1D-F and Figure S1. For these simulations, we assumed the same task structure as in our behavioral task. Thus, there were both CS+ and CS− cues predictive of reward and reward omission respectively. The cue-outcome delays were 3 s and the intertrial interval (ITI) was exponentially distributed with a mean of 30 s and a maximum of 90 s. The outcome periods were 3 s each for either cue. To model the structure of this task, we used a complete serial compound state space with a single state lasting 3 s. Thus, there was a single state for each cue and a single state for the outcome. Importantly, the ITI was similarly split to 30 states (the maximum possible ITI was divided into 3 s states starting after the previous outcome state). We have previously noted that this commonly used state space makes problematic assumptions about real animals [84], but nevertheless use it here for its simplicity in illustrating our main claim about learning rate. The task was simulated based on the temporal parameters listed above and the timeline for the task was converted to a timeline of states, with each time step lasting 3 s. The temporal difference value function

and update were then applied on this state space. Specifically, the value function was written as

$$V(s_t) = \langle r_t + \gamma V(s_t + 1) \rangle$$

where V represents state value, s represents state, the subscript represents time step, γ represents the discount factor (assumed to be 0.99), and the angular bracket represents the expected value.

During learning, state values were learned by the TD learning rule

$$V(s_t) \leftarrow V(s_t) + \alpha_t \delta_t$$

$$\delta_t = r_t + \gamma V(s_t + 1) - V(s_t)$$

where α represents learning rate and δ represents RPE.

For the learning rate controller used in Figure 1, we $\alpha_t = 0.1 + 0.02\mathcal{N}(0, 1)$ assumed that where the second term on the right-hand side is a normally distributed noise term. A total of 10,000 randomly interleaved CS+ and CS− trials (with equal frequency) were simulated at 100% reward probability, followed by the same number of trials at 50% reward probability. The plots in Figure 1E, F were made based on the learning rates at outcome times and the change in state value of the CS+ on the next CS+ presentation in the 50% probability condition.

In Figure S1, we assumed that the learning rate is itself learned based on the volatility in the environment. We defined volatility (υ) as

$$\upsilon = \frac{\kappa |z_{RPE}|}{1 + \kappa |z_{RPE}|}$$

0where κ is a constant set to 5 and $|Z_{RPE}|$ refers to the absolute magnitude of the z-score of the current RPE against a rolling window (5,000 time steps) of previous RPEs. This form was used to keep the volatility between 0 and 1, and to ensure that whenever the current RPE deviates significantly from the previous RPEs in either direction, the environment is deemed to be volatile, i.e., there is high unexpected uncertainty about the environment. The learning rate was then updated based on this volatility using the following relationship.

$$\alpha_t \leftarrow \alpha_t + \eta(\upsilon_t - \alpha_t) + 0.01\mathcal{N}(0, 1)$$

where η is the learning rate for the learning rate and was set to 1. This learning rate ensured that whenever the volatility of the environment was deemed to be high, the learning rate was high, but at other times, the learning rate was low. To ensure that learning rate is always positive, we set its value to 0 whenever it became negative based on the above learning rule.

For the non-linear mapping between cue value and reward-seeking behavior assumed in Figure S1, we assumed the following sigmoidal relationship.

$$L = \frac{L_{\max}}{1 + e^{-(V - V_{0.5})/\tau}}$$

where L is the lick rate on a trial and L_{max} is the maximum possible lick rate on a trial This is a sigmoid with half-point of $V_{0.5}$ and rate of slope around this half-point controlled by τ. These parameters were set to $L_{max} = 10$, $V_{0.5} = 3$ and $\tau = 0.4$.

For the linear mapping between cue value and reward-seeking behavior, we assumed $L = V$.

QUANTIFICATION AND STATISTICAL ANALYSIS

Imaging data analysis: Preprocessing (motion correction, manual ROI detection, signal extraction, neuropil correction) using SIMA [85] for vmOFC cell body imaging was as described previously [26]. For axon imaging, we used the same approach as for cell bodies for motion correction. For detection of axonal ROIs, we employed a manual hand-drawing method but with different criteria: 1) we drew ROIs only around parts of an axon that showed no resolvable overlap with other fluorescent regions, thereby making the ROIs small, 2) we drew ROIs along a single axon (at least those that can reliably be tracked along the imaging focal plane) only once (more on this below), 3) we drew ROIs around regions of an axon that are definitely within the imaging plane (these are often bright and sharply in focus), 4) we drew ROIs that were at least 10 pixels or so, in order to minimize noise. An example plane with manually annotated ROIs is shown in Figure S6. Despite these precautions, it is impossible to know whether different ROIs are from the same underlying axon, as axons move in and out of the imaging plane and are highly branched. Thus, we do not make any claims about the individual ROIs shown in Figure 6F representing individual axons (though the example traces in Figure 6E, representing the axonal ROIs with the largest skew in activity, show distinct patterns). A common approach to identify ROIs that are putatively from the same axon is to remove segments that show high correlations. However, this threshold depends crucially on the signal to noise ratio of the recording. If the signal to noise ratio is low, segments of the same axon will show low correlations due to the noise dominating the fluorescence. Thus, the decision to set the threshold often becomes subjective, especially considering variability in signal to noise ratio between animals resulting from variability in GCaMP expression. We avoid this problem by not claiming that different ROIs are necessarily from different axons. Instead, we quantify both a population mean of all ROIs and separately perform a clustering analysis of ROIs based on their response profiles to identify heterogeneous response profiles (more on this below). We did not perform any neuropil correction for axon recordings as the neuropil is primarily due to signals of interest (i.e. from axons). Due to this reason, our earlier approach of ensuring no overlap with other axons is important to get isolation of signals.

The clustering analysis for vmOFC neurons was performed as previously published [26]. Importantly, all clustering was performed based on the peristimulus time histograms (PSTHs) around CS+ and CS− in the session after learning with 100% reward probability.

So, any neuron that has an assigned cluster identity was recorded under the 100% contingency late in learning. Once a cluster ID was assigned to a neuron, the same cluster ID was used for that same neuron in all other sessions. This was possible because 2-photon imaging allowed longitudinal tracking of the exact same set of neurons across many days and tasks [26].

The clustering of axon ROIs (Figure S6) used the same approach but used PSTHs around sucrose alone (Figure S7G), or around both sucrose and quinine (Figure S6H). As we did not longitudinally track axons across these sessions, we performed clustering separately for these two sessions. Nevertheless, qualitative correspondence between the sucrose responses of some clusters can be seen by comparing Figure S6G and H. The benefit of performing cluster analysis for axons is that different identified clusters are almost certain to be from different underlying axons, as the clustering is done on average responses across trials, thereby reducing noise. Thus, we do not assess axon-by-axon heterogeneity as we cannot reliably identify individual axons. We can nevertheless assess heterogeneity of information encoding in the axonal population by interpreting cluster-wise differences in response profiles. This is philosophically similar to our approach with clustering of vmOFC neurons. We only interpret average results across all neurons within a cluster, thereby treating a cluster as a unit of information representation. The one big difference between the identification of cell body and axon clusters is that unlike in the cell body case, we cannot assess prevalence of each cluster among the axonal population. This is because the ROIs making up any cluster might potentially be overlapping and represent the same axon.

In Figure 3, we quantified reward responses as the coefficients of a General Linear Model (GLM) fit to reward delivery [26]. Importantly, this GLM approach was applied to the deconvolved calcium fluorescence to remove fluorescence changes purely due to the dynamics of GCaMP6s. We employed a GLM approach primarily to separate lick related responses and reward prediction or receipt responses, as both licking and rewards generally produced positive responses [26]. We did not employ a similar GLM approach for analyzing the sucrose and quinine responses in Figure 4 as these responses were evidently dissociated from licking responses. This is because licking was high for sucrose alone, high for sucrose in the sucrose-quinine session and low for quinine in the sucrose-quinine session; yet, the responses were generally high, low and high respectively, for these conditions.

To obtain the average population slope in Figure 2I, we first fitted a best-fit linear regression to the trial-by-trial variability in OFC reward response (separately in rewarded and unrewarded trials) with the licking update on the next trial for each neuron. We then averaged these slopes across all recorded neurons to obtain the average population slope separately for the positive RPE case (i.e. rewarded trials in the 50% reward probability session) and the negative RPE case (i.e. unrewarded trials in the 50% reward probability session). To obtain the visualization plot shown in Figure 2H, we first z-scored both the trial-by-trial variability of a neuron and the licking update (i.e. both axes) for each neuron and pooled these data for all neurons. We then binned the data along the activity z-score and calculated the corresponding mean lick update for that bin (using numpy.digitize with right=False, i.e., a datapoint is counted in a given bin if it is greater than or equal to the lower edge of the bin and less than the higher edge). The plot shows the lower edges of

the bins, with $z < -3$ not shown due to low sampling. The z-scoring was performed prior to pooling to avoid confounding within-neuron and between-neuron variability. The actual quantification was done on the raw slopes instead of the z-scored data to avoid equating neurons with considerable difference in the trial-by-trial variability of reward responses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank P. Phillips, J. Berke, D. Ottenheimer, A. Mohebi, K. Ishii, R. Gowrishankar, S. Piantadosi, J. Rodriguez-Romaguera and Z.C. Zhou for comments on the manuscript, and S. Mihalas and all Stuber lab members for helpful discussions. We thank Karl Deisseroth (Stanford University) and the GENIE project at Janelia Research Campus for viral constructs.

Funding: This study was funded by grants from the National Institute of Mental Health (R00MH118422, V.M.K.N.; F31 MH117931, R.S.), National Institute of Drug Abuse (R37-DA032750 & R01-DA038168, G.D.S., and P30-DA048736) and Brain and Behavior Research Foundation (NARSAD Young Investigator Award, V.M.K.N.).

REFERENCES

1. Rescorla RA, and Wagner AR (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory 2, 64–99.
2. Schultz W, Dayan P, and Montague PR (1997). A Neural Substrate of Prediction and Reward. Science 275, 1593–1599. [PubMed: 9054347]
3. Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, and Uchida N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. Nature 525, 243–246. [PubMed: 26322583]
4. Mohebi A, Pettibone JR, Hamid AA, Wong J-MT, Vinson LT, Patriarchi T, Tian L, Kennedy RT, and Berke JD (2019). Dissociable dopamine dynamics for learning and motivation. Nature 570, 65–70. [PubMed: 31118513]
5. Chang CY, Esber GR, Marrero-Garcia Y, Yau H-J, Bonci A, and Schoenbaum G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci 19, 111–116. [PubMed: 26642092]
6. Le K., Claa LD., Hachisuka A., Bakhurin KI., Nguyen J., Trott JM., Gill JL., and Masmanidis SC. (2020). Temporally restricted dopaminergic control of reward-conditioned movements. Nat. Neurosci 23, 209–216. [PubMed: 31932769]
7. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, and Janak PH (2013). A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16, 966–973. [PubMed: 23708143]
8. Schweighofer N, and Doya K. (2003). Meta-learning in reinforcement learning. Neural Netw 16, 5–9. [PubMed: 12576101]
9. Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, Hassabis D, and Botvinick M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci 21, 860–868. [PubMed: 29760527]
10. Soltani A, and Izquierdo A. (2019). Adaptive learning under expected and unexpected uncertainty. Nat Rev Neurosci 20, 635–644. [PubMed: 31147631]
11. Iigaya K. (2016). Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. eLife 5, e18073.
12. Behrens TEJ, Woolrich MW, Walton ME, and Rushworth MFS (2007). Learning the value of information in an uncertain world. Nat Neurosci 10, 1214–1221. [PubMed: 17676057]

13. Pearce JM, and Hall G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87, 532–552. [PubMed: 7443916]
14. Courville AC, Daw ND, and Touretzky DS (2006). Bayesian theories of conditioning in a changing world. Trends in Cognitive Sciences 10, 294–300. [PubMed: 16793323]
15. Preuschoff K, and Bossaerts P. (2007). Adding prediction risk to the theory of reward learning. Ann. N. Y. Acad. Sci 1104, 135–146. [PubMed: 17344526]
16. Monosov IE (2020). How Outcome Uncertainty Mediates Attention, Learning, and DecisionMaking. Trends in Neurosciences 43, 795–809. [PubMed: 32736849]
17. Grossman CD, Bari BA, and Cohen JY (2020). Serotonin neurons modulate learning rate through uncertainty. bioRxiv, 2020.10.24.353508.
18. Hernaus D, Xu Z, Brown EC, Ruiz R, Frank MJ, Gold JM, and Waltz JA (2018). Motivational deficits in schizophrenia relate to abnormalities in cortical learning rate signals. Cogn Affect Behav Neurosci 18, 1338–1351. [PubMed: 30276616]
19. Jepma M, Murphy PR, Nassar MR, Rangel-Gomez M, Meeter M, and Nieuwenhuis S. (2016). Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. PLOS Computational Biology 12, e1005171.
20. McGuire JT, Nassar MR, Gold JI, and Kable JW (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. Neuron 84, 870–881. [PubMed: 25459409]
21. Iigaya K, Fonseca MS, Murakami M, Mainen ZF, and Dayan P. (2018). An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. Nat Commun 9, 2477. [PubMed: 29946069]
22. Wu X, Wang T, Liu C, Wu T, Jiang J, Zhou D, and Zhou J. (2017). Functions of Learning Rate in Adaptive Reward Learning. Front. Hum. Neurosci 11.
23. Constantinople CM, Piet AT, Bibawi P, Akrami A, Kopec C, and Brody CD (2019). Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. Elife 8.
24. Jones JL., Esber GR., McDannald MA., Gruber AJ., Hernandez A., Mirenzi A., and Schoenbaum G. (2012). Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values. Science 338, 953–956. [PubMed: 23162000]
25. Miller KJ, Botvinick MM, and Brody CD (2018). Value Representations in Orbitofrontal Cortex Drive Learning, but not Choice. bioRxiv, 245720.
26. Namboodiri VMK, Otis JM, Heeswijk K. van Voets ES, Alghorazi RA, Rodriguez-Romaguera J, Mihalas S, and Stuber GD (2019). Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. Nat. Neurosci 22, 1110. [PubMed: 31160741]
27. Wilson RC, Takahashi YK, Schoenbaum G, and Niv Y. (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–279. [PubMed: 24462094]
28. Barreiros IV, Panayi MC, and Walton ME (2021). Organization of Afferents along the Anterior–posterior and Medial–lateral Axes of the Rat Orbitofrontal Cortex. Neuroscience 460, 53–68. [PubMed: 33609638]
29. Morecraft RJ, Geula C, and Mesulam M-M (1992). Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. Journal of Comparative Neurology 323, 341–358.
30. Ren J, Friedmann D, Xiong J, Liu CD, Ferguson BR, Weerakkody T, DeLoach KE, Ran C, Pun A, Sun Y, et al. (2018). Anatomically Defined and Functionally Distinct Dorsal Raphe Serotonin Sub-systems. Cell 175, 472–487.e20.
31. Boehme R, Deserno L, Gleich T, Katthagen T, Pankow A, Behr J, Buchert R, Roiser JP, Heinz A, and Schlagenhauf F. (2015). Aberrant Salience Is Related to Reduced Reinforcement Learning Signals and Elevated Dopamine Synthesis Capacity in Healthy Adults. J. Neurosci 35, 10103–10111.
32. Otis JM, Namboodiri VMK, Matan AM, Voets ES, Mohorn EP, Kosyk O, McHenry JA, Robinson JE, Resendez SL, Rossi MA, et al. (2017). Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. Nature 543, 103–107. [PubMed: 28225752]
33. Sutton RS, and Barto AG (1998). Introduction to Reinforcement Learning 1st ed. (MIT Press).
34. Bower GH, and Trabasso T. (1964). Concept Identification. In Studies in Mathematical Psychology, Atkinson RC, ed. (Stanford University Press).

35. Downing BD (1968). Salience and learning rate in concept identification. Psychon Sci 10, 73–74.
36. Siebold A, and Donk M. (2014). On the Importance of Relative Salience: Comparing Overt Selection Behavior of Single versus Simultaneously Presented Stimuli. PLOS ONE 9, e99707.
37. Zehetleitner M, Koch AI, Goschy H, and Müller HJ (2013). Salience-Based Selection: Attentional Capture by Distractors Less Salient Than the Target. PLoS One 8, e52595.
38. Frank MJ., Doll BB., Oas-Terpstra J., and Moreno F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12, 1062–1068. [PubMed: 19620978]
39. Galea JM, Mallia E, Rothwell J, and Diedrichsen J. (2015). The dissociable effects of punishment and reward on motor learning. Nat. Neurosci 18, 597–602. [PubMed: 25706473]
40. Gershman SJ (2015). Do learning rates adapt to the distribution of rewards? Psychon Bull Rev 22, 1320–1327. [PubMed: 25582684]
41. Kojima S, Yamanaka M, Fujito Y, and Ito E. (1996). Differential Neuroethological Effects of Aversive and Appetitive Reinforcing Stimuli on Associative Learning in Lymnaea stagnalis. jzoo 13, 803–812.
42. Mackintosh NJ (1976). Overshadowing and stimulus intensity. Animal Learning & Behavior 4, 186–192. [PubMed: 964444]
43. Slotnick B, and Coppola DM (2015). Odor-Cued Taste Avoidance: A Simple and Robust Test of Mouse Olfaction. Chem Senses 40, 269–278. [PubMed: 25787943]
44. Ader R, Weijnen JAWM, and Moleman P. (1972). Retention of a passive avoidance response as a function of the intensity and duration of electric shock. Psychon Sci 26, 125–128.
45. Namboodiri VMK, Hobbs T, Pisanty IT, Simon RC, Gray MM, and Stuber GD (2021). Relative salience signaling within a thalamo-orbitofrontal circuit governs learning rate. bioRxiv, 2020.04.28.066878.
46. Mitchell AS, and Chakraborty S. (2013). What does the mediodorsal thalamus do? Front Syst Neurosci 7, 37. [PubMed: 23950738]
47. Jankowski MM, Ronnqvist KC, Tsanov M, Vann SD, Wright NF, Erichsen JT, Aggleton JP, and O'Mara SM (2013). The anterior thalamus provides a subcortical circuit supporting memory and spatial navigation. Front Syst Neurosci 7.
48. Izquierdo A, and Murray EA (2010). Functional interaction of medial MD thalamus but not nucleus accumbens with amygdala and orbital prefrontal cortex is essential for adaptive response selection after reinforcer devaluation. J Neurosci 30, 661–669. [PubMed: 20071531]
49. Tervo DGR, Hwang B-Y, Viswanathan S, Gaj T, Lavzin M, Ritola KD, Lindo S, Michael S, Kuleshova E, Ojala D, et al. (2016). A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. Neuron 92, 372–382. [PubMed: 27720486]
50. Chen T-W, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499, 295–300. [PubMed: 23868258]
51. Banerjee A, Parente G, Teutsch J, Lewis C, Voigt FF, and Helmchen F. (2020). Valueguided remapping of sensory cortex by lateral orbitofrontal cortex. Nature 585, 245–250. [PubMed: 32884146]
52. Jennings JH., Kim CK., Marshel JH., Raffiee M., Ye L., Quirin S., Pak S., Ramakrishnan C., and Deisseroth K. (2019). Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. Nature 565, 645–649. [PubMed: 30651638]
53. Wang PY, Boboila C, Chin M, Higashi-Howard A, Shamash P, Wu Z, Stein NP, Abbott LF, and Axel R. (2020). Transient and Persistent Representations of Odor Value in Prefrontal Cortex. Neuron 108, 209–224.e6.
54. Hirokawa J, Vaughan A, Masset P, Ott T, and Kepecs A. (2019). Frontal cortex neuron types categorically encode single decision variables. Nature 576, 446–451. [PubMed: 31801999]
55. Kepecs A, Uchida N, Zariwala HA, and Mainen ZF (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature 455, 227–231. [PubMed: 18690210]
56. Masset P, Ott T, Lak A, Hirokawa J, and Kepecs A. (2020). Behavior- and Modality-General Representation of Confidence in Orbitofrontal Cortex. Cell 182, 112–126.e18.

57. Hayden BY, Heilbronner SR, Pearson JM, and Platt ML (2011). Surprise Signals in Anterior Cingulate Cortex: Neuronal Encoding of Unsigned Reward Prediction Errors Driving Adjustment in Behavior. J Neurosci 31, 4178–4187. [PubMed: 21411658]
58. Ballesta S, Shi W, Conen KE, and Padoa-Schioppa C. (2020). Values encoded in orbitofrontal cortex are causally related to economic choices. Nature 588, 450–453. [PubMed: 33139951]
59. Kuwabara M, Kang N, Holy TE, and Padoa-Schioppa C. (2020). Neural mechanisms of economic choices in mice. eLife 9, e49669.
60. Coddington LT, and Dudman JT (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat. Neurosci 21, 1563–1573. [PubMed: 30323275]
61. Cohen JY, Haesler S, Vong L, Lowell BB, and Uchida N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482, 85–88. [PubMed: 22258508]
62. Engelhard B, Finkelstein J, Cox J, Fleming W, Jang HJ, Ornelas S, Koay SA, Thiberge SY, Daw ND, Tank DW, et al. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature 570, 509–513. [PubMed: 31142844]
63. Kobayashi S, and Schultz W. (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci 28, 7837–7846. [PubMed: 18667616]
64. Gavornik JP, Shuler MGH, Loewenstein Y, Bear MF, and Shouval HZ (2009). Learning reward timing in cortex through reward dependent expression of synaptic plasticity. PNAS 106, 6826–6831. [PubMed: 19346478]
65. Namboodiri VMK, Huertas MA, Monk KJ, Shouval HZ, and Hussain Shuler MG (2015). Visually cued action timing in the primary visual cortex. Neuron 86, 319–330. [PubMed: 25819611]
66. Kheifets A., Freestone D., and Gallistel CR. (2017). THEORETICAL IMPLICATIONS OF QUANTITATIVE PROPERTIES OF INTERVAL TIMING AND PROBABILITY ESTIMATION IN MOUSE AND RAT. J Exp Anal Behav 108, 39–72. [PubMed: 28653484]
67. Stolyarova A, and Izquierdo A. (2017). Complementary contributions of basolateral amygdala and orbitofrontal cortex to value learning under uncertainty. Elife 6.
68. Bradfield LA, and Hart G. (2020). Rodent medial and lateral orbitofrontal cortices represent unique components of cognitive maps of task space. Neurosci Biobehav Rev 108, 287–294. [PubMed: 31743727]
69. Costa VD, and Averbeck BB (2020). Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore–Exploit Tradeoffs. J. Neurosci 40, 2553–2561. [PubMed: 32060169]
70. Simmons JM, and Richmond BJ (2008). Dynamic changes in representations of preceding and upcoming reward in monkey orbitofrontal cortex. Cereb Cortex 18, 93–103. [PubMed: 17434918]
71. Everitt BJ, Hutcheson DM, Ersche KD, Pelloux Y, Dalley JW, and Robbins TW (2007). The orbital prefrontal cortex and drug addiction in laboratory animals and humans. Ann. N. Y. Acad. Sci 1121, 576–597. [PubMed: 17846151]
72. Pascoli V, Hiver A, Van Zessen R, Loureiro M, Achargui R, Harada M, Flakowski J, and Lüscher C. (2018). Stochastic synaptic plasticity underlying compulsion in a model of addiction. Nature 564, 366–371. [PubMed: 30568192]
73. Roesch MR, Esber GR, Li J, Daw ND, and Schoenbaum G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. Eur. J. Neurosci 35, 1190–1200. [PubMed: 22487047]
74. Monosov IE, and Hikosaka O. (2012). Regionally Distinct Processing of Rewards and Punishments by the Primate Ventromedial Prefrontal Cortex. J. Neurosci 32, 10318–10330.
75. Parent MA, Amarante LM, Liu B, Weikum D, and Laubach M. (2015). The medial prefrontal cortex is crucial for the maintenance of persistent licking and the expression of incentive contrast. Front Integr Neurosci 9, 23. [PubMed: 25870544]
76. Gremel CM, Chancey JH, Atwood BK, Luo G, Neve R, Ramakrishnan C, Deisseroth K, Lovinger DM, and Costa RM (2016). Endocannabinoid Modulation of Orbitostriatal Circuits Gates Habit Formation. Neuron 90, 1312–1324. [PubMed: 27238866]
77. Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, Lee D, and Taylor JR (2019). Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. Neuron 0.

78. Holly EN, Davatolhagh MF, Choi K, Alabi OO, Vargas Cifuentes L, and Fuccillo MV (2019). Striatal Low-Threshold Spiking Interneurons Regulate Goal-Directed Learning. Neuron 103, 92–101.e6.

79. Morisot N, Phamluong K, Ehinger Y, Berger AL, Moffat JJ, and Ron D. (2019). mTORC1 in the orbitofrontal cortex promotes habitual alcohol seeking. Elife 8.

80. Halassa MM, and Sherman SM (2019). Thalamocortical Circuit Motifs: A General Framework. Neuron 103, 762–770. [PubMed: 31487527]

81. Resendez SL., Jennings JH., Ung RL., Namboodiri VMK., Zhou ZC., Otis JM., Nomura H., McHenry JA., Kosyk O., and Stuber GD. (2016). Visualization of cortical, subcortical and deep brain neural circuit dynamics during naturalistic mammalian behavior with head-mounted microscopes and chronically implanted lenses. Nat Protoc 11, 566–597. [PubMed: 26914316]

82. Otis JM, Zhu M, Namboodiri VMK, Cook CA, Kosyk O, Matan AM, Ying R, Hashikawa Y, Hashikawa K, Trujillo-Pisanty I, et al. (2019). Paraventricular Thalamus Projection Neurons Integrate Cortical and Hypothalamic Signals for Cue-Reward Processing. Neuron 103, 423–431.e4.

83. Zuiderveld K. (1994). Contrast limited adaptive histogram equalization. In Graphics gems IV (Academic Press Professional, Inc.), pp. 474–485.

84. Namboodiri VMK (2021). What is the state space of the world for real animals? bioRxiv, 2021.02.07.430001.

85. Kaifosh P, Zaremba JD, Danielson NB, and Losonczy A. (2014). SIMA: Python software for analysis of dynamic fluorescence imaging data. Front Neuroinform 8, 80. [PubMed: 25295002]

HIGHLIGHTS

- Novel approach to test if neural activity abides by learning rate control.
- Outcome response in orbitofrontal cortex is consistent with learning rate control.
- OFC outcome response signals relative salience of an outcome.
- Medial thalamic input to OFC causally mediates encoding in OFC.

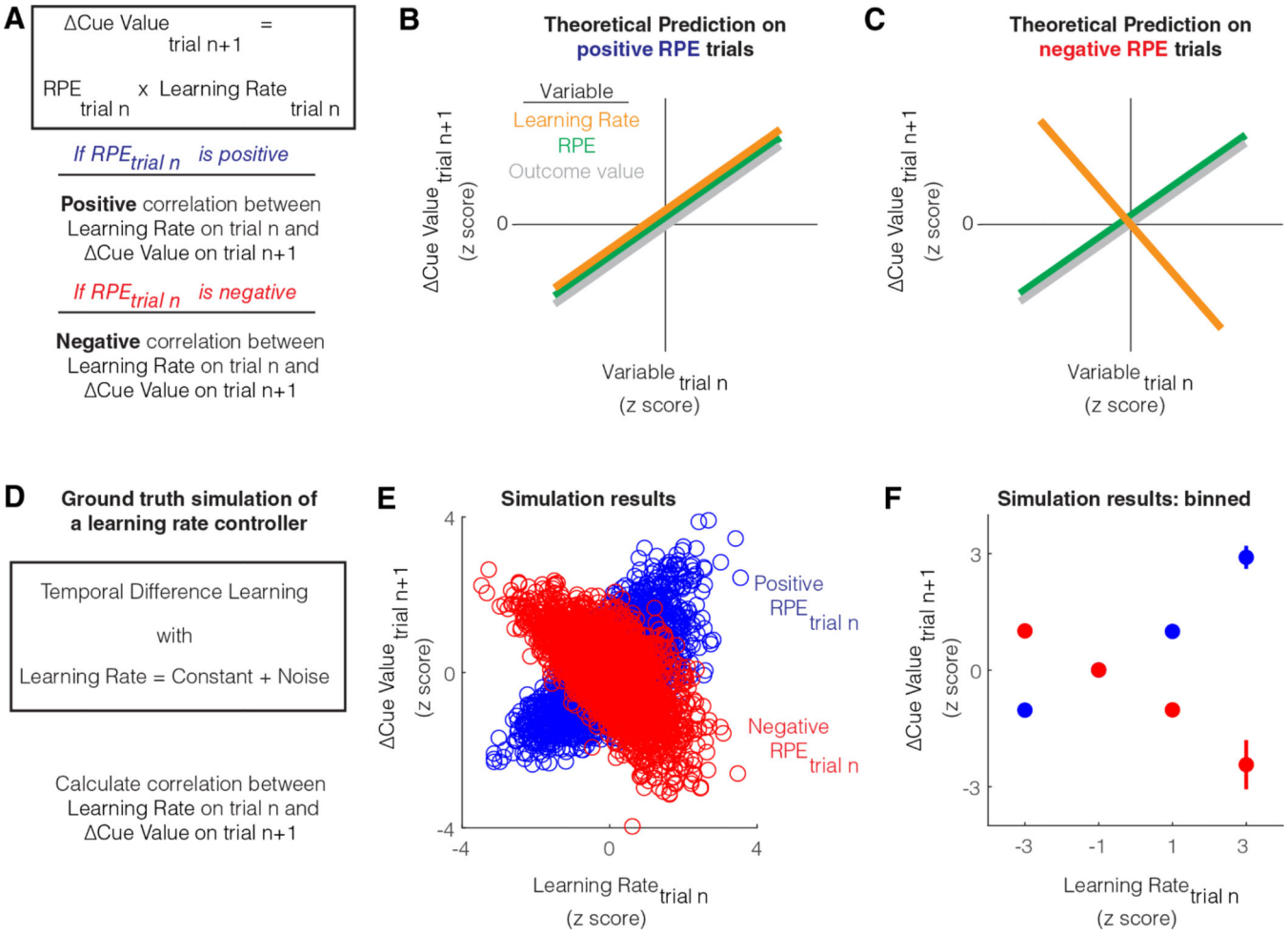

Figure 1: Theoretical test for learning rate control

A. Schematic showing that during cue-reward learning, the update in cue value on a trial from the previous trial is the RPE on the previous trial multiplied by the learning rate on the previous trial. This relationship implies that the sign of the dependence between cue value update and learning rate on a trial is the same as the sign of the RPE on that trial.

B. Theoretical prediction for the dependence between trial-by-trial update in cue value on consecutive trials, and trial-by-trial fluctuations in different variables, for positive RPE trials. When RPE is positive, an increase in learning rate would cause a proportional increase in cue value. When RPE is positive, an increase in either outcome value or RPE would also cause an increase in cue value.

C. Theoretical prediction for the dependence between trial-by-trial update in cue value on consecutive trials, and trial-by-trial fluctuations in different variables, for negative RPE trials. When RPE is negative, an increase in learning rate would cause a *decrease* in cue value. This is because the learning rate modulates the magnitude of the cue value update in the direction of the RPE. On the other hand, when RPE is negative, an increase in either outcome value or RPE would still cause an increase in reward prediction. This is because the slope for the relationship between RPE or outcome value and cue value update, is the learning rate, a positive quantity.

D. Simulation of a temporal difference learning agent that sets learning rate to a constant with some added noise. In this learning agent, we calculate the correlation between learning rate on a given trial and the subsequent update of cue value.

E. Results of simulations on positive and negative RPE trials during cue-reward learning when reward probability was set to 50%. The schematized predictions from **B** and **C** are verified in these simulations. Both learning rate and cue value update are z-scored similar to that shown in **B** and **C**, since the primary outcome of interest is the sign of the correlation. Please note that low learning rates produce minimal cue value updates for both positive and negative RPEs, even though the z-scores are divergent.

F. Summary of results from **E** calculated by binning z-scores of learning rate. Simulations for more various other scenarios are plotted in Figure S1.

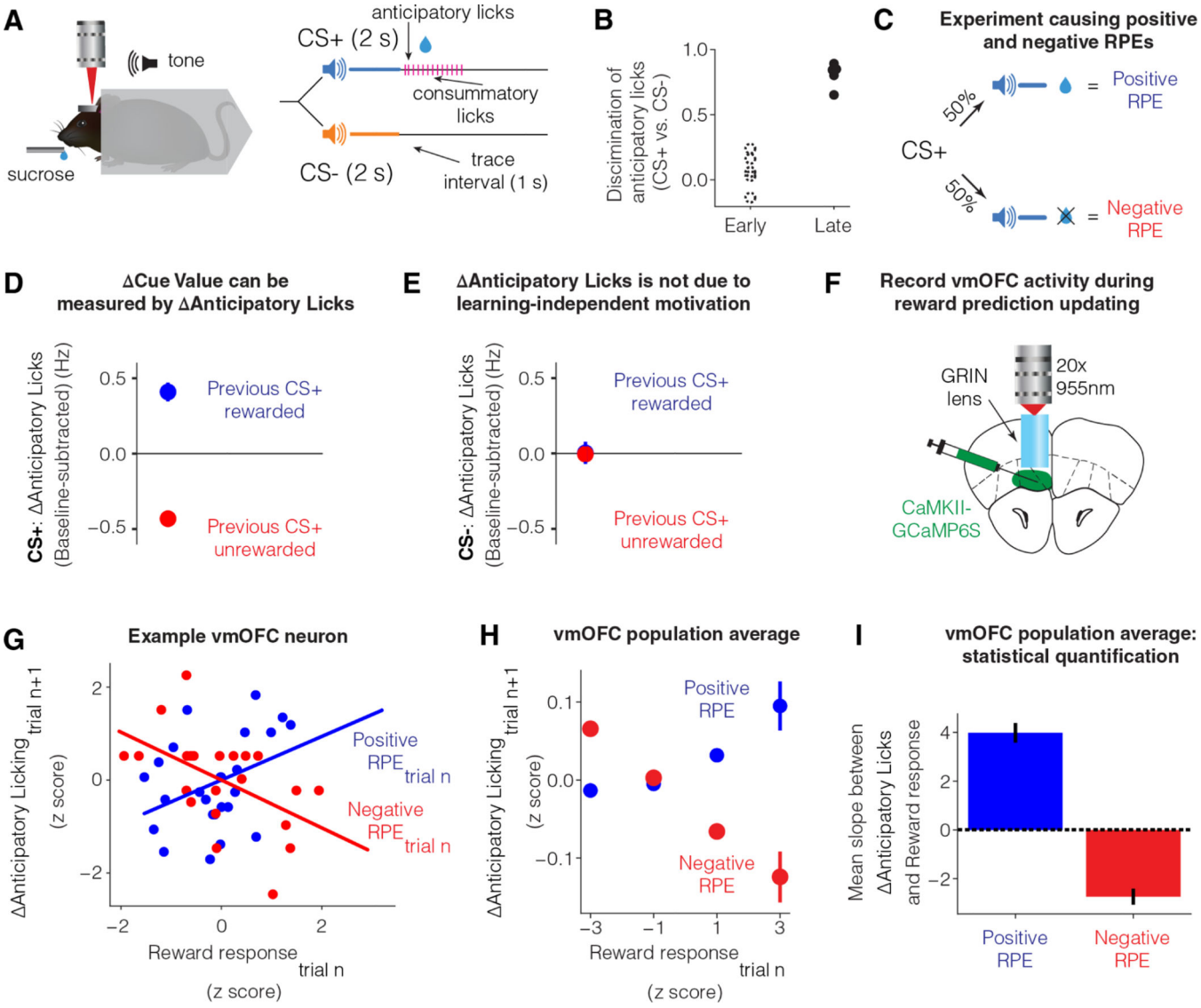

Figure 2: Trial-by-trial fluctuations in vmOFC reward responses reflect learning rate control

A. Differential trace conditioning task in headfixed mice [26,32].

B. Behavior early and late in learning. Early session was defined as the first day of learning and late session as the day when anticipatory licking in response to CS+ was high and stable (Materials and Methods) [26]. Cue discrimination was measured as two times the area under a receiver operator characteristic curve between lick counts after CS+ and lick counts after CS−, minus one (Materials and Methods) [26].

C. Schematic showing that in a session with reward probability of 50%, RPE will be positive on rewarded trials and negative on unrewarded trials.

D. The change in anticipatory licking on consecutive CS+ trials (potentially with interceding CS− trials) is reliably positive after rewarded CS+ trials (positive RPE) and negative after unrewarded CS+ trials (negative RPE) (n=34 sessions from n=12 imaging mice). Thus, update in anticipatory licks can be used to estimate update in cue value. See Table S1 for all

statistical results in the manuscript for all figures, including all statistical details and sample sizes.

E. A potential concern is that receiving reward on a CS+ trial might increase general motivation to lick on the next trial, independent of cue value learning. If true, CS− trials immediately following a rewarded CS+ trial should show higher licking compared to CS− trials immediately following an unrewarded CS+ trial. This panel plots the update in anticipatory licking on CS− trials based on whether the immediately preceding CS+ trial was rewarded or unrewarded. The lack of licking update shows that the effect in **D** is not a learning independent motivation signal. See Figure S2 for evidence of motivation signals during the pre-cue baseline period.

F. Schematic showing the recording of vmOFC activity using two-photon microendoscopic calcium imaging.

G. Data from an example neuron showing the dependence between trial-by-trial update in anticipatory licking on CS+ trials, and trial-by-trial fluctuations in response on rewarded trials (positive RPE) and unrewarded trials (negative RPE). The lines show the best fit regression in each condition. The observed relationship is as expected if vmOFC controls learning rate on a trial-by-trial basis.

H. Z-scored, pooled, and binned data across all vmOFC neurons to visualize the dependence between trial-by-trial response fluctuations and licking update for the population of vmOFC neurons on positive and negative RPE trials. Each neuron's data were z-scored separately for each axis, all z-scored data were then pooled, and binned into the four bins shown in the plot. Error bars are standard error of the mean. These data are shown purely for an intuitive visualization of the average relationship between these variables in the vmOFC population.

I. Statistical quantification of the average slope between reward response on a trial and licking update on the next trial across all neurons on both positive and negative RPE trials. No z-scoring was performed here to avoid assigning an equal weight to neurons with low or high trial-by-trial variability in responses.

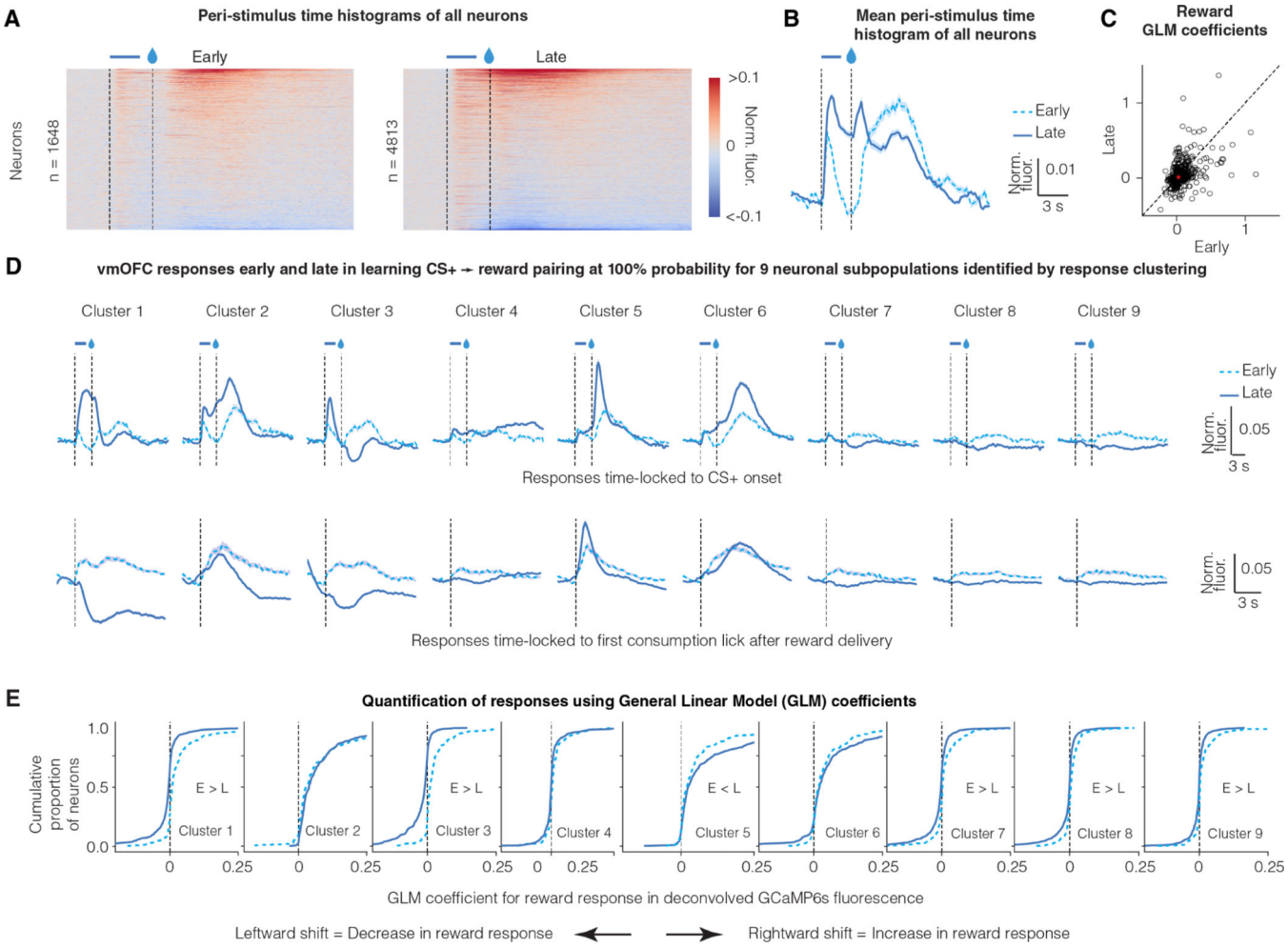

Figure 3: Reward responses of some vmOFC neuronal subpopulations reduce after reward prediction learning

A. Peri-Stimulus Time Histograms (PSTHs) of GCaMP6s fluorescence early and late in learning for all vmOFC neurons expressing CaMKIIα (excitatory neurons). All recorded neurons from both timepoints are shown.

B. Average PSTH across all recorded neurons early and late in learning

C. General Linear Model (GLM) coefficients for reward response early and late in learning from those neurons that were longitudinally tracked between these sessions (Materials and Methods) [26] (n = 1,590 tracked neurons). We performed the GLM analyses on deconvolved fluorescence traces to remove lick response confounds and the slow decay of GCaMP6s dynamics [26]. The dashed line is the identity line where responses early and late are equal. The average response early and late is indicated by the red asterisk. On an average, the neuronal response to reward reduces significantly after reward prediction learning.

D. PSTHs of GCaMP6s fluorescence early and late in learning for nine subpopulations identified using a clustering of late responses [26]. Each line corresponds to the average of CS+ PSTH across all neurons within a cluster (see text for rationale). The identification of neuronal subpopulations within CaMKIIα expressing vmOFC neurons using clustering algorithms, and the PSTHs late in learning, were published previously [26]. The top row

shows responses time locked to cue onset and the bottom row shows responses time locked to reward consumption. Clusters 1 and 3 reverse the sign of their reward responses from positive (i.e. greater than baseline) early in learning to negative (i.e. less than baseline) late in learning. Error shadings correspond to confidence intervals.

E. Cumulative distribution function for General Linear Model (GLM) coefficients for reward response early and late in learning (Materials and Methods) [26]. The y-axis effectively percentiles the responses shown on the x-axis. So, a reduction in reward response late in learning causes a leftward shift in the curves. These results demonstrate positive reward responses in all clusters early in learning, but negative reward responses in some clusters (1, 3, 7, 8 and 9) late in learning, showing a flip in sign.

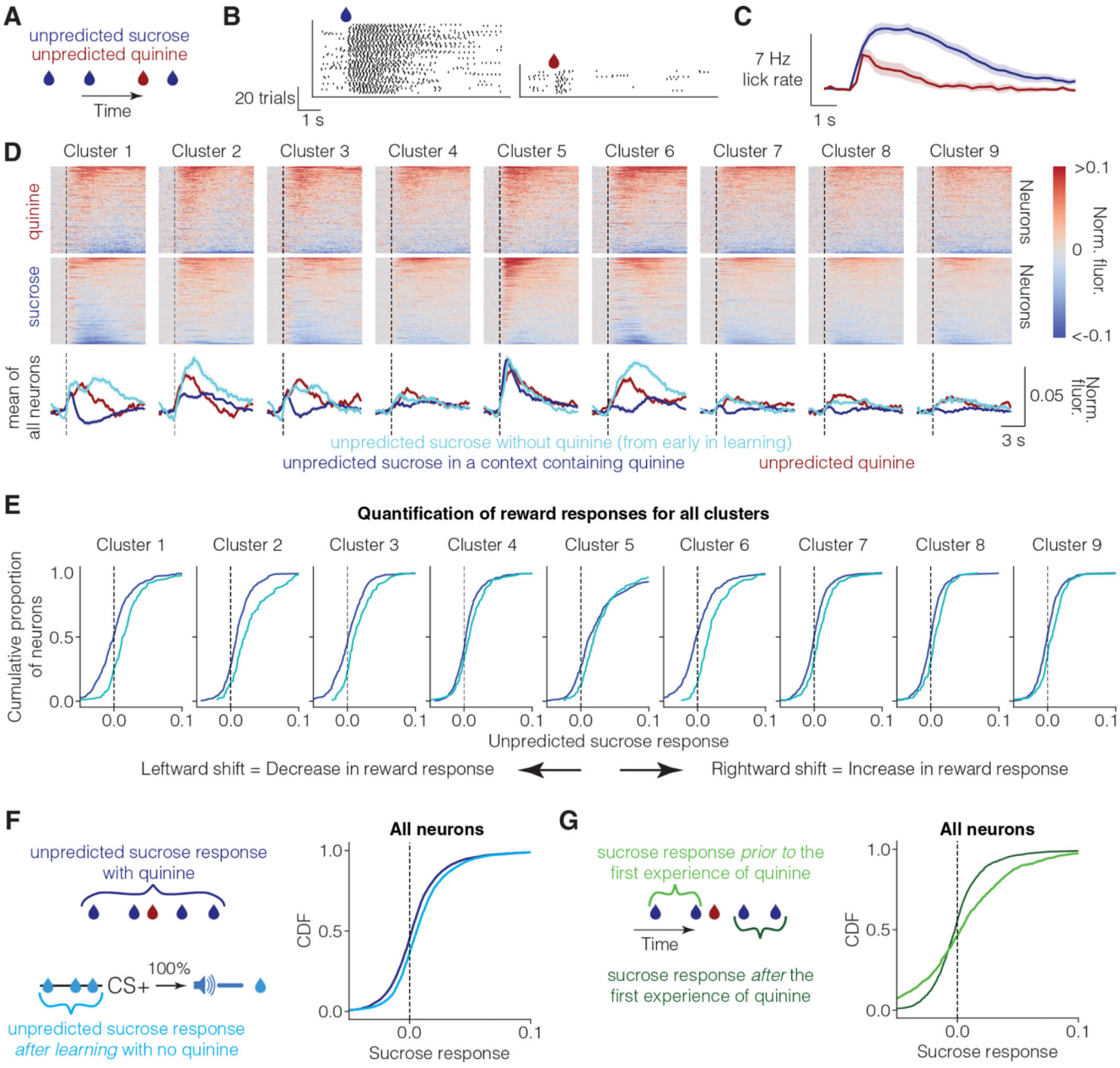

Figure 4: Unpredicted reward responses of vmOFC subpopulations reduce in a context containing a highly salient aversive stimulus.

A. Schematic of unpredicted sucrose and quinine delivery. Unpredicted sucrose (10%) and quinine (1.5–2.5 mM) are delivered pseudorandomly in a 3:1 ratio (Methods). In this experiment, licks are necessary to sample the liquid.

B. Raster plot of licking (black ticks) from an example behavioral session from one animal. Animals lick at high rates after sucrose delivery, but immediately stop licking after sampling quinine deliveries.

C. Average lick rate across all animals and sessions (n=26 sessions from n=11 imaging mice). The histograms are time-locked to liquid delivery.

D. The average PSTH for sucrose and quinine responses for all OFC-CaMKII neurons (n=3,716 neurons from 5 mice), aligned to the first lick after liquid delivery (i.e. initiation of consumption). The line graphs at the bottom show the average across all neurons within a cluster. The cyan line shows the average response of each cluster to reward early in learning aligned to first lick after liquid delivery (same as in Figure 3A). Responses of OFC→VTA neurons are shown in Figure S4.

E. Cumulative distribution function for fluorescence responses to unpredicted sucrose in the sucrose and quinine session, and the sucrose responses early in learning. In this case, since the response is evidently dissociated from licking (see text), we did not perform a GLM analysis.

F. Cumulative distribution function for reward response for all vmOFC neurons for the sucrose quinine session and another control session in which unpredicted rewards were delivered during conditioning after learning ("Background" session in [26]). The positive responses to unpredicted sucrose without quinine after learning further supports the positive unpredicted sucrose responses observed early in learning (shown in **D, E**). Positive unpredicted sucrose responses in the absence of quinine are also replicated in two separate cohorts in Figure S4.

G. Cumulative distribution function for fluorescence responses to unpredicted sucrose in the sucrose and quinine session *prior to the first ever experience* of unpredicted quinine, and responses to unpredicted sucrose after the first experience of unpredicted quinine within the same session (i.e. same neurons). The average response of vmOFC neurons reduces after the first experience of quinine (i.e. leftward shift in curves). Note that these data only include the neurons that were recorded from the first sucrose and quinine session to only include trials before the first ever experience of quinine by the animals. Note also that the number of trials prior to the first experience of quinine is quite low (n = 2.6 trials on average), adding to the variability of responses in this condition.

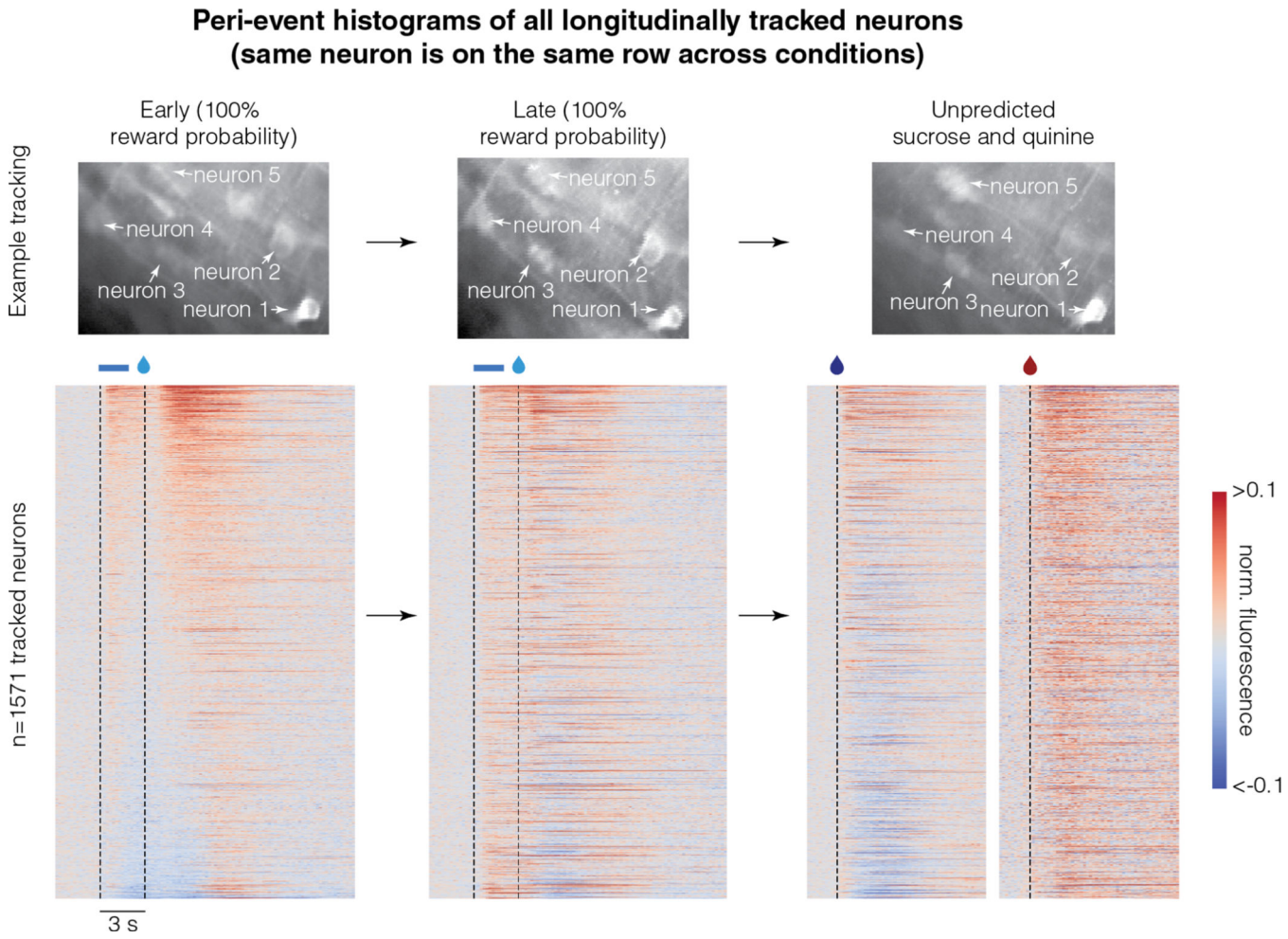

Figure 5. Reward responses of the longitudinally tracked neurons are correlated across three conditions:

Top row shows example longitudinally tracked neurons. Here, the intensity of a pixel corresponds to activity and hence, different brightness across sessions corresponds to different activity levels. The bottom row shows the peri-event histograms of all longitudinally tracked neurons. Each row in this heatmap across the different conditions corresponds to the same neuron. Neurons are sorted by their average activity early in learning. These data show that response to reward is correlated across all conditions (quantified in Table S1). For instance, the neurons that show the lowest amount of activity early in learning tend to be the neurons that show inhibitory reward responses late in learning or in the sucrose and quinine experiment (correlations quantified in Table S1). Please note that the responses of some neurons are saturated in the color map to ensure that the response patterns of most neurons are visible.

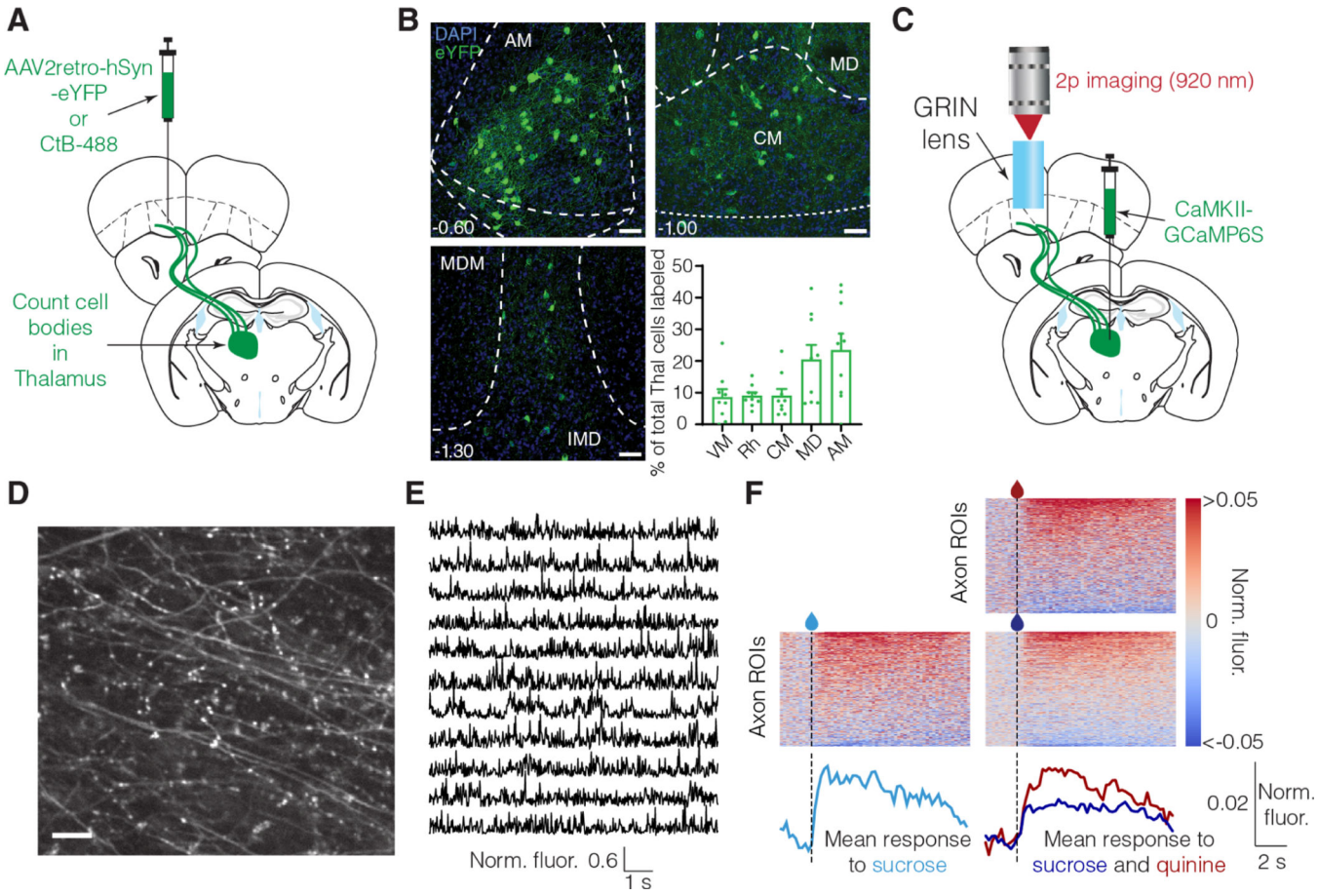

Figure 6: Medial thalamus (mThal) conveys reward responses to vmOFC and shows qualitatively similar reward response adaptation as vmOFC neurons.

A. Surgery schematic for retrograde anatomical tracing, showing injections of either retrogradely traveling virus (AAV2retro) or Cholera Toxin-B (CTB).

B. mThal cell bodies projecting to OFC counted using CTB and AAV2retro labeling. Representative images show AAV2retro expression (see Figure S6 for CTB expression). Top 5 thalamic regions are shown. See Figure S6 for counts in all thalamic nuclei, split by AAV2retro and CTB injections. AM: Anteromedial, MD: Mediodorsal, CM: Centromedial, Rh: Rhomboid, VM: Ventromedial. Scale bar = 50 μm.

C. Surgery schematic for mThal axon imaging in vmOFC.

D. Example zoomed-in mThal axon standard deviation projection image showing individual axons in vmOFC. The scale bar corresponds to 10 μm.

E. Example mThal GCaMP traces from individual axonal regions of interest (ROIs) (Methods).

F. Heat maps show trial-averaged responses from individual axon ROIs (*do not necessarily correspond to distinct axons,* see Figure S6 and Methods for details and interpretation) to unpredicted sucrose alone (left) or unpredicted sucrose and quinine (right, similar experiment as Figure 4A). The bottom traces show the average responses across all segmented mThal axon ROIs aligned to the first lick after liquid delivery (dashed line). The same animals (n=3) were used under all conditions, to be directly comparable; see Figure S6

for data from two more animals in the sucrose only condition. Mean sucrose response across the population is lower in the session with quinine compared to the session without quinine. The statistical test was applied using the mean fluorescence across all ROIs per animal as an independent measure. This adaptation is qualitatively similar to that seen in some vmOFC clusters (Figure 4).

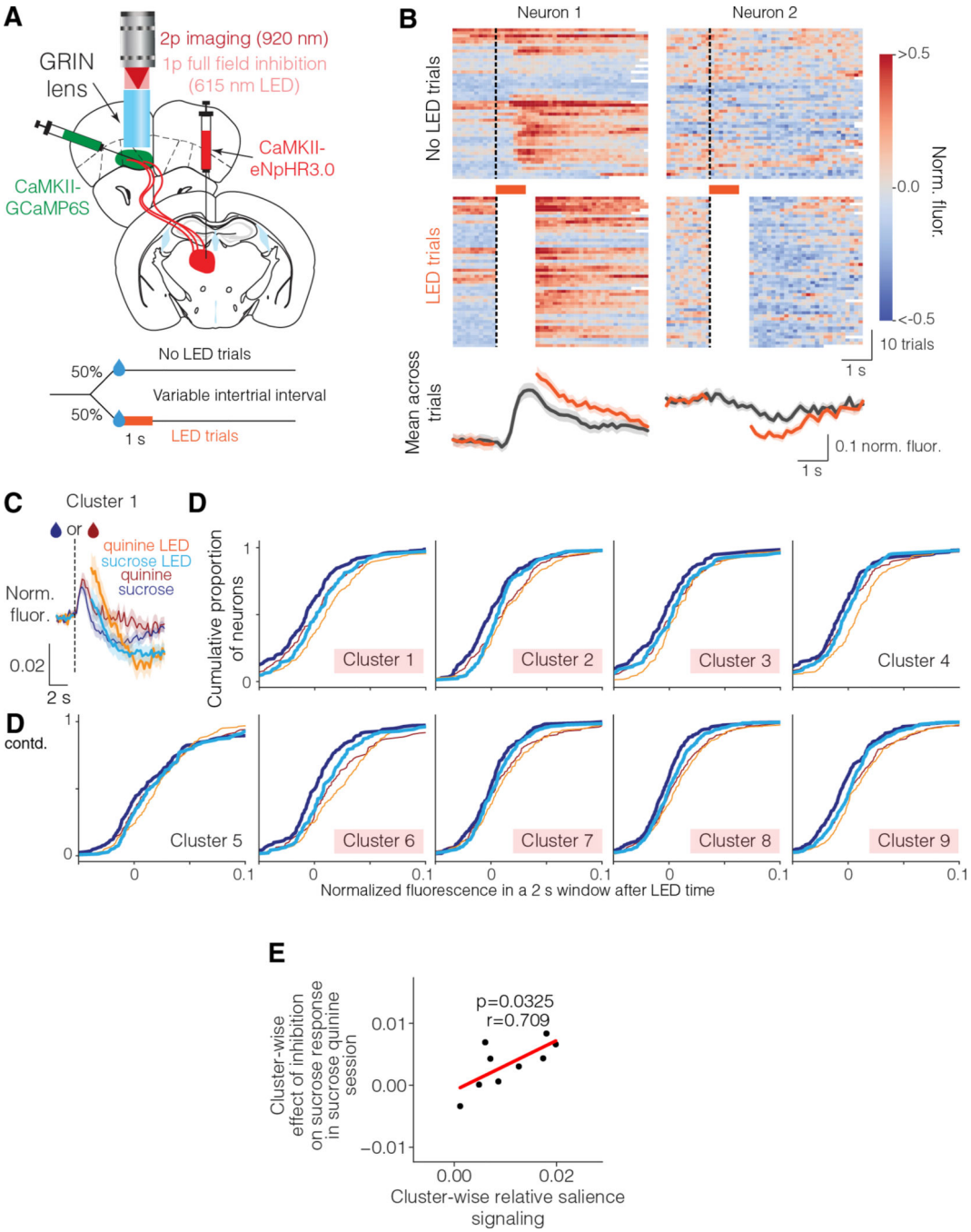

Figure 7: Medial thalamic input to vmOFC guides vmOFC reward response adaptation.

A. Schematic of mThal inhibition while imaging vmOFC CaMKIIα expressing neurons. Top shows surgery schematic and bottom shows experiment schematic.

B. Example neurons showing effect of mThal inhibition on unpredicted sucrose responses. Left neuron shows an increase in activity due to mThal inhibition, whereas the right neuron shows a decrease in activity. Frames around LED illumination were masked out (shown as white) to prevent light artifacts in imaging.

C. PSTH of cluster 1 showing effect of mThal inhibition on sucrose and quinine responses.

D. Empirical cumulative distribution functions showing average fluorescence with and without LED for individual neurons within each cluster for sucrose and quinine responses (n=1,645 neurons in total from 3 mice). This shows the full distribution of the population responses, with a rightward shift signifying an increase in activity. Here, the red shadings correspond to clusters showing significant mean effect on their sucrose responses.
E. Cluster-wise relationship between relative salience signaling (measured by suppression in reward response due to the presence of quinine, Figure 4) and effect of mThal inhibition on sucrose response (i.e., change in sucrose response due to LED as shown in **D** minus the change in spontaneous response due to LED as shown in Figure S7). There is a strong positive correlation (~50% explained variance).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
chicken anti-GFP antibody (1:500 dilution)	Aves Lab	#GFP-1020
donkey anti-chicken secondary antibody (1:1000 dilution)	Jackson Immunoresearch	#703-545-155
Bacterial and virus strains		
AAVDJ-CaMKIIα-GCaMP6s (effective titer of ~1–2×10^{12} infectious units per mL)	UNC Vector Core	NA
AAV5-CaMKII-eNpHR3.0-mCherry (~4×10^{12} infectious units per mL)	UNC Vector Core	NA
AAV5-CaMKIIα-mCherry (~4×10^{12} infectious units per mL)	UNC Vector Core	NA
rAAV2retro-hSyn-eYFP (~2×10^{12} infectious units/mL)	UNC Vector Core	NA
Biological samples		
Chemicals, peptides, and recombinant proteins		
Cholera Toxin subunit B conjugated with AlexaFluor 488 (CTB-488)	Molecular Probes	NA
Critical commercial assays		
Deposited data		
Imaging and behavioral data for clustering	This paper	10.5281/zenodo.5507624
Experimental models: Cell lines		
C57BL mice	Jackson Laboratory	NA
Experimental models: Organisms/strains		
Oligonucleotides		
Recombinant DNA		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Calcium imaging data analysis scripts (Python)	This paper	10.5281/zenodo.5507624
ImageJ	NIH	https://imagej.nih.gov/ij
Python 2.7	Anaconda	https://www.anaconda.com/
SIMA v1.3	[85]	https://github.com/losonczylab/sima
Temporal difference learning simulations	This paper	10.5281/zenodo.5507624
Other		