Nature Communications

https://doi.org/10.1038/s41467-025-68215-0

Article in Press

Time-dependent deployment of medial prefrontal cortical representations in male mice

Received: 17 January 2024

Accepted: 18 December 2025

[redacted]

Cite this article as: Lopez-Yepez, J.S., Barta, A., Martin, J. *et al.* Time-dependent deployment of medial prefrontal cortical representations in male mice. *Nat Commun* (2026). https://doi.org/10.1038/s41467-025-68215-0

Junior Samuel Lopez-Yepez, Anna Barta, Juliane Martin, Maria Moltesen, Tsz-Fung Woo, Oliver Hulme, Ebru Demir & Duda Kvitsiani

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Time-dependent deployment of medial prefrontal cortical representations in male mice

Junior Samuel Lopez-Yepez[1,2#], Anna Barta[2#], Juliane Martin[2], Maria Moltesen[2], Tsz-Fung Woo[2], Oliver Hulme[3,4], Ebru Demir[5,6], Duda Kvitsiani[2,5,6*]

1. Aarhus University, Department of Chemistry, Langelandsgade 140, Building 1513, 431, 8000 Aarhus C, Denmark

2. Aarhus University, Danish Research Institute of Translational Neuroscience (DANDRITE), Ole Worms Allé 6, Building 1182, 8000 Aarhus C, Denmark

3. Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre,

Hvidovre, Denmark

4. London Mathematical Laboratory, London, United Kingdom

5. Department of Biomedical Sciences, Southern Illinois University School of Medicine, Carbondale, IL, USA

6. School of Electrical, Computer, and Biomedical Engineering, Southern Illinois University, Carbondale, IL, USA

[*]Corresponding author: dkvitsiani38@siumed.edu

#These authors contributed equally: Junior Samuel Lopez-Yepez and Anna Barta

Abstract

In reward foraging tasks, prefrontal neurons track reward history, yet animals also show persistent choice-history biases. How these histories are represented in prefrontal circuits and guide animals' decisions remains unknown. We asked whether past rewards and choices are incorporated by leaky integration or carried as discrete, history-specific codes, and how these codes are recruited under different task demands. We recorded medial prefrontal cortex (mPFC) activity while mice performed probabilistic reward foraging task and fit a reinforcement-learning model whose decision variable, combining reward and choice histories, captured behavior. Neurons represented history-specific rewards and choices while integrating them consistent with their behavioral impact. We then altered reward contingencies and inter-choice intervals and transiently inactivated mPFC. Neural representations adapted to changing task demands, yet the behavioral impact of inactivation was sensitive to inter-choice interval and reward contingencies. We conclude that mPFC hosts redundant computations whose influence is gated by timing and task structure.

Introduction

Reinforcement learning (RL) models achieved tremendous success in explaining animal behavior and helped us interpret the activity of neurons in diverse brain areas[1–3]. This success is particularly prominent in reward foraging tasks where animals or artificial agents have to come up with the policy to maximize reward harvesting efficiency amid changing environmental conditions[4]. Some of the well-established reward foraging tasks that mimic naturalistic habitats assume nonstationary reward statistics [5–8]. For example, in two alternative choice tasks with variable interval reinforcement schedule (VI) reward statistics on the unchosen options do not remain fixed, instead they grow, simulating the accumulation of food on unvisited patches[9]. In such tasks, the decision-making process is influenced not only by past rewards but also by past choices, as the probability of rewards on unchosen options increases, incentivizing animals to alternate between options, even if they have previously received rewards [5–8].

RL models developed to solve this type of foraging tasks incorporate both reward history and choice history components into the decision process[6,7,10]. These models have played a pivotal role in explaining animal behavior and shed a new light on the representation of action values, typically computed as the weighted sum of past rewards. However, the significance of choice history representations and how they interface with reward history representations in these tasks remains unclear. The nature of representations within the prefrontal areas warrants further investigation, as cortical representations are distributed widely[11–13], all the while maintaining functional specificity[14,15]. Additionally, due to the nature of the multimodal input prefrontal areas receive[16] representations are highly diverse[17] and mixed[18,19] , although others argued for categorical representations[20]. Consequently, when animals engage in reward foraging tasks, neurons in these brain regions are often found to represent a multitude of behavioral variables. This is especially true at the population level, since in a high-dimensional space various behavioral variables can be decoded simultaneously[21]. The theoretical and experimental work further suggests that prefrontal areas constitute self-sufficient learning systems containing all the necessary representations to master the tasks at hand[22,23]. Therefore, a fundamental question that has received little attention in previous studies, with few exceptions[15] is: which representations are functionalized by the animals and influence their decision-making processes ?

A survey of the function of the medial prefrontal cortex (mPFC) reveals two prominent trends. One line of research, initiated several decades ago, links mPFC to working memory[24], while another set of studies, as mentioned above, grounded in the field of RL, contends that mPFC guides animals' decisions by representing decision variable (DV) of the RL model[3,7,25,26]. Yet, it remains unclear whether the DV representations and their behavioral deployment are independent of the mPFC's memory function or do they work in tandem. In other words, does the mPFC represent and use DV, regardless of the temporal intervals between task relevant events or the representations of DV and its behavioral deployment depend on the memory load?

To address those issues, we set up a probabilistic reward foraging task for mice involving two alternative ports which required animals to maintain a memory of their past rewards and past choices to make decisions [5–7]. The probability of reward at each port was a function of "set reward probabilities" that determined the chance of a reward arriving at each port, while also being a function of past choices. Using a previously developed reinforcement learning model, the double trace (DT) model, which integrates choice and reward history into the DV[27], we investigated how the integration of reward history and choice history representations aligned with the animals' behavior.

To investigate whether choice history representations conformed to the task structure and were not purely motoric in nature, we utilized two tasks: one with reward probabilities contingent on previous choices as described above (VI task) and another, a Variable Ratio Reinforcement task (VR), in which this dependency was removed.

While we identified all components of the RL model in the mPFC, it remained unclear whether these representations were actively employed by the animals. To address this, we temporarily inactivated the mPFC and examined how animals integrated reward and choice history. For this, we manipulated reward contingencies through VI and VR tasks and varied the temporal intervals between events to investigate whether the computation of the DV and its deployment by the mPFC were influenced by the memory load and task structure.

Results

Past rewards and choices maximize reward-harvest efficiency.

We set up a discrete version of the variable interval (VI) schedule of reinforcement task for two options [5,7](Fig.1A, see Materials and Methods). In this task, the mice initiated a trial by poking their noses into the center port. After a brief period of waiting (uniformly distributed, from 0.2 to 0.5 s), the mice had to choose between the left and right ports to collect water rewards (2 µl). The inter-trial interval with the mean of 1.1557 ± 0.0327 (standard error of the mean, s.e.m.) seconds was determined solely by animals' decision to initiate the next center port poke after completing a previous trial. In every trial, rewards on two ports (left and right) were assigned according to independent set reward probabilities. Once a reward was assigned to a given port, it remained available until the animal chose that port. The probability of obtaining a reward from a given port in each trial changed as a function of past choices (see equations (eqs.) 1 and 2, Materials and Methods). The set reward probabilities changed for every block of 35 to 200 trials. Such randomized block lengths helped us to minimize the expectation from animals for block transitions. The pair of set reward probabilities for left and right ports were the following tuples [0.1:0.4], [0.4:0.1], [0.25:0.25], [0.6:0.1] and [0.1:0.6], respectively (Materials and Methods). We collected a total of 82 behavioral sessions from 5 mice comprising 71,737 trials. On average each session included 875 ± 37 trials with the number of trials per session ranging from minimum of 129 to a maximum of 1814.

In the VI task, choice ratios closely matched reward ratios on a local scale (Fig.1B) and on a global scale [5,7] (Fig.1C), showing slightly undermatching[28]. To investigate what strategies animals used to maximize reward harvesting efficiency we compared reward rates (fraction of rewards per trial) to those of several artificial agents. In the VI task, it becomes advantageous to switch ports even after experiencing the rewards, because reward probabilities increment on unchosen port. For instance, simply alternating choices yields higher reward rates when compared to consistently selecting the port with higher set reward probabilities (Fig.1D). Hence, it is reasonable to assume that animals may use some form of alternation strategy as it was documented in previous studies[5,7,29,30]. Remarkably, mice in the VI task surpassed the performance of the alternation strategy (both for the example animals and all animals combined, Mann–Whitney U-test $p < 0.001$), prompting us to investigate the precise strategy employed by the animals (Fig.1D). As shown before[5–7,11,27] logistic regression of current choices on past rewards and past choices revealed that past rewards increased the probability of the port being chosen again, however the further back in time the reward was observed, the less effect it had on the port being chosen. Past choices had non-monotonic effects on current choices, meaning immediate past choices had the negative effect and choices further back in history had positive effects on current choices (Fig.1E and SFig.1A). Non monotonic choice history effects could be partially explained as a result of animals alternating when experiencing no rewards on immediate past trials and persevering on choices that happened two or more trials back in history (SFig.1B). To test if animals used alternation as a low-level default strategy or adjusted the alternation as a function of difference in set reward probabilities, we measured rate of alternation as a function of difference in set reward probabilities. We saw modest, but significant decrease in alternation rate as difference in set reward probabilities increased (Fig.1F).

Response time (RT) defined as the time elapsed from leaving the center port to poking left or right port was used as an alternative measure of performance. We found RTs were largely consistent with the effect of past rewards and past choices on current choices. RTs for the current choice were reduced when animals recently experienced a reward for that option (SFig.1C, regression coefficients, were $-1.7537*10^{-3} \pm 1.05 *10^{-3}$ for rewards one trial back on the same side and $-0.443*10^{-3} \pm 0.714*10^{-3}$ for rewards one trial back on the opposite side), and RTs increased when the animal chose the same option. Consequently, RTs decreased when the animal chose the alternative option in the previous trial (SFig.1C, regression coefficients were $4.85 *10^{-3} \pm 2.2*10^{-3}$ for combined choice effects one trial back on the same side).

Supported by previous work[27] and based on our findings, we conclude that the incorporation of reward and choice history into the decision-making process contributes to animals' ability to maximize their reward harvesting efficiency.

mPFC integrates past rewards and choices consistent with behavior

If choice and reward history effects are impacting animals' decisions, how are they represented on a single neuron level? More specifically, we asked whether individual neurons combine reward and choice histories in a way that matches their observed impact on the animal's current choices.

We recorded the activity of neurons (709 neurons from 5 mice) in the mPFC[31] (SFig.2A), using tetrodes. We used the manual spike sorting software MClust (AD Redish) to isolate single neurons (median LRatio = 0.0287, 95% confidence interval < 0.2325, median isolation distance = 20.8075, 95% confidence interval >10.4417, violation of refractory period <0.2763% of total spikes[32]). The recorded neurons showed the characteristic distribution of firing rates vs. spike waveforms (SFig.2B)[33].

Next, we sought to understand how behavior of animals could be explained by neural activity. As behavioral analysis showed, regression coefficients on average for immediate past trials for rewards and choices showed opposite signs (Fig.1E). Whereas immediate past rewards promote choice for the same port, immediate past choices promote choices for the alternative port. Hence, should individual neurons integrate the immediate past reward and choice in accordance with their impact on current choices, we would expect neurons to discriminate between these events in a contrasting manner. Individual example neuron (Fig.2A) showed opposite modulation of firing rate for immediate past reward vs. no reward for the right port and for immediate past choices for the right vs. left port. To determine how neurons discriminated between recent rewards and choices, we measured their selectivity using a method based on the difference in firing rate distributions. This was done using the area under the receiver operating curve (AUC) as the key metric, with the AUC scores normalized between -1 and 1 (see Materials and Methods for details). For choice selectivity positive AUC scores indicated preference for the right port and negative scores preference for the left port. We calculated the selectivity scores for the immediate past rewards from both the left and right separately. Then, to align these scores with the choice selectivity score, we reversed the sign of the AUC scores for the left port and combined them with the selectivity scores for the right port. This approach allowed us to measure the selectivity for rewards at the chosen port (further details in Materials and Methods). We observed subtle yet statistically significant negative correlation between selectivity of neurons for immediate past rewards and immediate past choices. (Fig. 2B).

The selectivity analysis was limited to one trial back history, while past rewards and choices beyond one trial back in history had effects on current choices. This was particularly clear when analysing choice history effects as they have non monotonic effects on current choices. Namely, negative contribution of immediate past choices on current choices may be overridden by positive contribution of past choices further back in history. To consider entire reward and choice history effects we approximated these effects using three exponentials. One for reward history and two for choice history to capture its non-monotonic shape. We used a reinforcement learning model - the double trace (DT) model - that incorporates reward and choice histories (Fig.2C, Materials and Methods section DT model for detailed description) into the decision rule[27]. In the DT model past reward contributions are captured by the *Q* values, that are updated for left and right ports separately as a function of reward outcome from those ports. Non-monotonic choice history effects (Fig.1E) are captured by the weighted sum of fast *F* and slow *S* choice traces for each port that approximate immediate past and distant past choice history effects on current choices, respectively. The integration of weighted sum of choice traces (we also refer to it as choice history effects) and *Q* values (or reward history effects) for the left and right ports define the decision variable (*h*), which via softmax selection rule is converted to choice probability P(c).

Prior work found that the DT model was superior in performance compared to other RL models of the same class (Q-learning models) in terms of its predictive adequacy of the animals’ choices and in terms of its normative performance (SFig.2C and D, see the Methods section on the model selection)[27]. Here we further explored if animals deployed the DT model to maximize reward harvesting efficiency and if neurons represented current choices consistent with the effects of past choices and past rewards.

First, we showed that reward harvesting efficiency measured as regret, difference in reward harvesting efficiency between model and optimal agent (Optimal_Baiting model) that has access to the reward probabilities on a trial by trial basis negatively correlated with the DT model's ability to predict animals choices (Fig.2D). Furthermore, we showed that as animals progressed over the sessions the DT models predictive accuracy improved (SFig.2E) suggesting that animals were learning to deploy the DT model over time.

Second, we noted that in majority of the sessions reward history effects captured by the *Q* value and choice history effects captured by weighted sum of *F* and *S* components of the DT model had predominantly opposing effects on current choices (Fig.2E and SFig.2F). If individual neurons are encoding *Q* value and choice history effects, it would follow that neurons should inherit the structure of these correlations. To test this, we analysed trial by trial correlation of firing rate of each neuron to *Q* value and choice history effects. We observed that on average the strongest correlations with DT model variables were at the decision epoch (SFig.2G) with the fast choice component *F* having the best correlation with individual firing rates of neurons (SFig.2H). Next, we observed that neurons encoded *Q* value and choice history effects consistent with their impact on current choices (Fig.2F and SFig.2I). We tested whether neurons encoded *Q*-value and choice history effects in opposing directions using a linear mixed effect model (LMM). The dependent variable was each neuron's correlation between firing rate and *Q*-value; the fixed effect was the neuron's correlation with choice history effect and animal ID was included as a random intercept. P-values ($p < 0.001$) were obtained via a within-animal permutation test that shuffled the correlations of choice-history effects. While majority of neurons represented *Q* value or fast and slow component of the choice trace, some also encoded mixture of the model variables (Fig.2G) as expected from prefrontal cortical areas that have highly mixed representations.

Decoding accuracy of the mPFC neurons of choices predicted by the DT model showed positive correlations with the DT model's performance (fit of the DT model, SFig.2J), while there was no correlation with the reward harvesting efficiency (SFig.2K).

Neurons show discrete coding of past rewards and choices

In both experimental and theoretical studies representations in mPFC are not limited to variables given by the RL model but encompass also other variables that can be used to derive the RL model variables. For example, individual neurons may preferentially represent rewards for particular trials back in history with more neurons representing more recent rewards. Such diminishing responses to past rewards on a population level can be used to derive the *Q* values[25,34].

We found that in many sessions reward and particularly choice history could be decoded comparable to the choices predicted by the DT model (Fig.3A). To further understand the nature of these representations at the individual neuron level we looked at how neurons discriminated reward and choices for particular trials back in history. We looked at the responses of neurons conditioned on reward outcomes for specific trials back in history. If mPFC individual neurons represent *Q* values, we could expect them to operate as leaky integrators [35] following standard RL value update form (eq. 14 also see Fig.2C) and distinguish immediate past rewards from no rewards with highest accuracy. Alternatively, if mPFC neurons discriminate reward and choices preferentially from various past trials, their representations may be constrained to specific trials or temporal intervals in the history.

We observed that some neurons (13% of analysed neurons out of total 709 analysed neurons) discriminated rewarded from non-rewarded trials (selectivity AUC scores, with $p < 0.05$ in permutation test) that happened two trials back in history (Materials and Methods). The example neuron (Fig.3B) that showed higher selectivity for right rewards 2 trials back (AUC score = 0.34, $p < 0.001$) than 1 or 3 trials back (AUC score = 0.56, p value = 0.01, for 1 trial back and AUC score = 0.45, $p = 0.02$ for three trials back right rewards) discriminated rewarded trials from non-rewarded trials, and this discrimination was strongest when rewards happened two trials back in history. This was also seen on a population level when we selected cells tuned to rewards for specific trials back in history (SFig.3A).

To account for representations of rewards and choices not limited to two trials back in history, we conducted a comprehensive analysis of single neuron responses in the mPFC. Using regularized regression analysis, we examined the firing rate of each neuron during the decision epoch (Fig.3C) and regressed it against rewards and choices from up to 10 trials back, as well as the current choices (Materials and Methods, see section on linear regression analysis for neurons).

45% of the recorded cells (320 cells) showed a significant regression coefficient for at least one of the events analysed (SFig.3B). The number of neurons with significant regression coefficients for both rewards and choices showed surprising preference to events two or more trials back in history (Fig.3C). The preferential encoding of rewards and choices for the past trials was not due to exclusion of the regressors that did not pass the statistical test or regularization method (SFig.3C). We further tested the reliability of the regressors by splitting into half each session and estimating the preferred regressor (lag) for each neuron. Scatter plot showed significant number of neurons ($p < 0.001$) having the same preferred lags for two halves of the session (Fig.3D). P values were derived by permutation test, shuffling the lags for one half of the session 1000 times and counting how many times the neurons had the same preferred lag for two halves. Regressors in addition showed monotonically decaying trends towards past trials (Fig.3E) consistent with the idea that at the population level mPFC represented the DVs.

We further tested whether neurons were preferentially responsive to rewards and choices occurring at specific time intervals in history as opposed to occurring at specific trials in history. We used time windows of 0.2 s. (total duration of 8 s) to chunk the history into time intervals and performed the same regression analysis described above. We observed that the maximum regression coefficients were specific to time intervals in the past (SFig.3D). As in the previous analysis, most neurons showed the most robust modulation to immediate past events (SFig.3E).

To test if history-specific representations arise as a function of recurrency within the cortical circuits we used recurrent neural network (RNN) model (layrecnet function from Matlab) that was trained to output the DV given the choice and reward input. More specifically the RNN with one hidden layer received the same reward and choice data as animals on its two input neurons and was trained to produce the same Q value and choice history effects for left and right ports on its output (Fig.3F). We observed the history-specific tuning of neurons to rewards and choices (Fig.3G). The number of neurons showed monotonic decay as function of the history (Fig.3H). We also note that choice history effects with its non-monotonic shape can be recovered in both biological neurons and in RNN (SFig.3F). Overall, the reward and choice history-specific neurons and their distribution resembled the representations from biological neurons (SFig.3G-I)

To take advantage of the history-specific responses in neurons [36] we built a support vector machine (SVM) classifier to see how well population activity (all neurons recorded in a session) could linearly separate current choices, past choices, and past rewards up to two trials back (total number of states 16). We built multiple weak pairwise linear classifiers to discriminate among different states (Materials and Methods) [19].The neural population could separate above chance level past immediate choices, current choices, and past rewards up to two trials back (SFig. 4).

We independently validated trial and time history representations on a different cohort of mice (n = 6). We used the Kilosort2 [37] spike sorting algorithm and post-processing steps to identify single units (Materials and Methods). 99% of isolated units with <0.63% of spikes violated the refractory period. We recorded 3805 neurons in the VI task in the mouse mPFC (SFig.5A), finding 1458 neurons (with the false discovery rate p <0.00001) in 214 sessions with a total of 129,072 trials that showed significant regression coefficients to any of the behavioral events (past rewards and past choices up to 10 trials back). Our analysis confirmed the previous results (Fig.3B-D) that neurons showed preferential activity for trial- and time-specific events (SFig.5B–E for trial history-specific regression analysis and SFig.5F–G for time history-specific regression analysis). Choice history regression coefficients also showed non-monotonic decay towards past trials (SFig.5I).

We conclude that neurons in mPFC besides representing DV also show preference to encode history specific past rewards and choices.

Neural responses in mPFC capture change in task demand

In prior studies, it was demonstrated that medial prefrontal cortex controls flexibility of animals to update the task contingencies as opposed to executing already well learned rules[38]. Since our mice were run on VI task for extended sessions one could speculate that the task became less mPFC and more basal ganglia dependent[39]. To test if mPFC neurons could track and update the changing task demands we subjected animals to VI and VR tasks. We first ran animals on VI tasks for few sessions (6.8 ± 3.4351) and then switched to run on VR task. The total number of sessions per animals was 26 ± 3.

Because in VI task reward probabilities are conditioned on the past choices, animals often develop alternation choice bias, reflecting baiting rates [5–7]. Unlike the VI task, in the VR task, reward probabilities are determined only by set reward probabilities. In VR tasks, animals tend to choose the higher reward probability option independent of the reward outcome, resulting in perseverance choice bias [8]. If neurons in the mPFC represent change in task demand, they should represent change of past choice effects on current choices.

We collected 67 sessions from 5 mice (total number of trials = 126083), with 865 ± 43.4 trials per session in the VI task, and 62 sessions with a 794 ± 39.8 trials per session for the VR task. First, we tested whether mice could estimate the preference for options consistent with VI and VR task demands. We verified that reward rates on alternation (left-right or right-left choices) minus reward rates on perseverance (right-right or left-left choices) for the example animal (Fig.4A, dark green line), were significantly different (Mann-Whitney U-test $p < 0.001$) in VI task (0.058 ± 0.001) and in the VR tasks (0.081 ± 0.001). The same reward rates for all animals and all sessions combined in the VI task were higher than for those in the VR task (Fig.4B, left panel). We observed a higher alternation rate in the VI task compared to the VR task at the individual animal level (Fig.4A light green line, 0.61 ± 0.004 in VI task and 0.32 ± 0.004 in VR tasks were significantly different (Mann-Whitney U-test p <0.001). The same was true for all animals and sessions combined (Fig. 4B, right panel).

The behavioral adjustment from the VI to the VR task was seen on learning rates (SFig.6A upper panel) and on choice history effects. Immediate choice history effects had a sign reversal from negative in the VI task to positive values in the VR task (Fig. 4C, rightmost panel). Past rewards, unlike immediate past choices, had mixed effects on the current choices in VI or VR tasks. Left immediate rewards were slightly, but significantly higher in the VI task than in the VR task and right rewards two or more trials back in history had stronger effects on the current choices. We also noted that correlations between Q values and choice history effects shifted towards more negative scores (Fig. 4D) from VR (correlation of Q and ϕ*F + θ*S for left option 0.415 ± 0.033 and for right option 0.414 ± 0.036) to VI task (for left option 0.099 ± 0.043 and right option 0.085 ± 0.047, permutation test p <0.001) also seen on the correlation of Q values with the individual choice history effects (SFig.6A lower panel). This result again reaffirmed that Q values and choice history effects antagonise each other in the VI task but have the same effects on choices in the VR task.

Next, we recorded from mPFC (SFig.6B) 1642 neurons in the VI and 1465 neurons in the VR tasks from the same animals (n = 5 animals, n = 67 in VI and n = 62 sessions in VR tasks) using the Kilosort2-based sorting algorithm as described previously (violation of refractory period in 99% of single units was <0.38%). Neurons for VI and VR tasks were recorded from different sessions. Using the same analysis as before (Fig.2B) the reward selectivity for chosen port and choice selectivity showed slightly negative and positive correlations in VI and VR tasks respectively (Fig.4E), aligning well with their effects on current choices. To test whether task context modulates the coupling between choice- and reward-history selectivity, we fit a LMM: AUC_choice ~ AUC_reward * Task_type + side + (1 | AnimalID). Here, AUC_choice is the AUC for immediate-past choices, AUC_reward is the AUC for immediate-past rewards, Task_type indicates VI vs. VR, and side codes left vs. right ports. The interaction (AUC_reward * Task_type) tests whether the reward–choice relationship depends on task type. Significance ($p < 0.05$) was assessed via 1,000 label permutations of Task_type within each animal to generate the null distribution. We extended the analysis to the full history of rewards and choices using a DT model. For each neuron, we correlated firing rate with DT-derived variables: *Q*-value and choice-history effects (the weighted sum of fast and slow choice traces). These correlations were task dependent (Fig. 4F; SFig. 6C). To test this formally, we fit an LMM: Coeff_choice ~ Coeff_value * Task_type + (1 | AnimalID), where Coeff_choice is the neuron's correlation with choice-history effects and Coeff_value is its correlation with *Q*-value. Significance ($p < 0.001$) was assessed by shuffling Task_type within each animal. Together, these results indicate that the coupling between value and choice-history encoding varies with task demands. Next, we asked if performance of the animal on VI and VR tasks were adaptive, in a sense that the change of behavioral strategy led to increase in reward harvesting efficiency. To assess this, we tested the performance of the DT model in a simulated environment. We used the same reward probabilities and block lengths that the animal experienced in the task from which its parameters were derived for intra-task testing. For cross-task testing, we applied these parameters to an alternative task. For example, in intra-task testing DT model parameters that were derived from a given session of VI (or VR) task were tested on simulated environment with the same reward statistics from VI (or VR) task of the remaining sessions. In cross-task testing VI (or VR) model parameters derived from a given session were tested on simulated sessions with the VR (or VI) task (Fig.4G). As a metric of performance, we used reward rates (rewards per trial) collected by the model. Our analysis showed that intra-task testing was achieving more reward rates then cross-task testing (Fig.4H). Next, we speculated that if neurons in the mPFC track the DT model decision variable to maximize the reward harvesting efficiency, then population decoding accuracy should correlate with the task performance. The task performance was (computed as a difference between intra-task and cross-task reward rates) showed positive correlations with the decoding accuracy (Fig.4I left panel). Elastic-net regression revealed that among other variables such as intertrial interval (ITI), DT model fit and response time (RT), task performance was better at explaining the decoding accuracy (Fig.4I, right panel). Furthermore, DT model fit and population decoding accuracy showed positive, but not significant correlations suggesting that the observed correlations between cross task performance and population decoding accuracy was not due to the correlations between DT model fit and population decoding accuracy (SFig.6D).

This result aligns well with the published work[40] on the involvement of prefrontal areas in adapting to task demands and shows that individual neurons integrate reward and choice histories in agreement with task demands.

Behavioral impact of mPFC rises with longer inter trial intervals

If mPFC neurons deploy representations of behavioral variables (reward and choice history, *Q* values, choice history effects, etc) the inactivation of the mPFC should hinder animals' performance. We tested the role of mPFC in VI and VR tasks. We trained 38 mice on VR (6.3 ± 0.15, sessions per animal) and VI tasks (9.4 ± 0.14, sessions per animal), with a total of 545 sessions and 746 ± 13 trials per session. Next, we injected the AAV virus AAV- CaMKIIa -hM$_4$D-mCherry (hM$_4$D receptor belongs to class of Gi-DREADDs)[41] or AAV- CaMKIIa -GFP (referred to GFP) expressing hM$_4$D or GFP protein respectively in the mPFC (Fig.5A) and performed acute inactivation by alternating injections of clozapine N-oxide (CNO) and saline on consecutive days (Fig.5B).

To test the effect of the inactivation on the activity of neurons, we recorded neurons in a separate animal unilaterally injected with the AAV- CaMKIIa -hM$_4$D-mCherry virus in mPFC. As a control for the inactivation, we also recorded neurons in parietal cortex not injected with the virus. The inactivation reduced the firing rate of neurons that was specific to the injection site (SFig.7A-C).

Significant differences were found in the reward harvesting efficiency in VR tasks between experimental (hM$_4$D expressing animals treated with saline and CNO), but not in the control (GFP animals) groups (SFig.7D, left panel). No significant drop was found in the reward harvesting efficiency in the hM$_4$D expressing animals treated with saline and CNO in the VI task (SFig.7D, right panel).

Next, we examined the effect of the inactivation on the reward and choice history effects. In animals that expressed hM$_4$D treated with CNO, we observed change in both choice and reward history effects in the VR task (Fig.5C, left panel). Only reward history effects were affected in the VI task (Fig.5C, right panel). We could not see any change in the bias in VR (0.4009 ± 0.0321 for Saline and 0.4382 ± 0.0375 CNO groups in hM$_4$D animals, $p = 0.54$ LMM with animal ID as random effects and drug manipulation as fixed effects, residuals did not show deviation from normal distribution.) task and no change in VI task (0.3021 ± 0.0201 Saline treatment group and 0.3354 ± 0.0230 CNO treatment group of hMD$_4$D animals. $p = 0.18$ using LMM with animal ID as random and drug manipulation as fixed effects, residuals did not show deviation from normal distribution). We did not see any significant change in bias in either VI or VR task in GFP animals. The exclusive effect of the inactivation in VI task on reward history effects is difficult to explain if we assume that mPFC implements the DV according to DT model, since inactivation should affect both reward and choice history effects. Here we provide an alternative interpretation of these results. Animals spend more time in alternating between ports in VR task and more time in perseverance in VI task. This is because, in the VI task, animals receive higher reward rates when they alternate (as shown in Fig. 4B) compared to the VR task. Similarly, in the VR task, perseverance is more beneficial since reward probabilities remain fixed. Therefore, even in cases where there's no immediate past reward, it still makes sense for the animal to persist in choosing the same port with the higher reward probability. We checked the time interval between consecutive choices for perseverance and alternation in VI and VR tasks in CNO-treated and saline-treated animals. In the VI task, mice spent more time between consecutive choices (we defined this as inter-choice interval (ICI)) on perseverance than on alternation, and the opposite was true for the VR task (Fig.5D, SFig.7E). A three-way ANOVA (Supplementary Table 1) revealed a significant main effect of the behavioral task ($F_{(1,538)}= 14.77$, $p <0.001$ for alternation and $F_{(1,538)}= 5.62$, $p <0.05$ for perseverance) on ICI, while no other grouping variables (drug, virus) or interaction terms showed significant effects. We also tested the effect of the task type (VI or VR) on ICI for alternation and perseverance separately using LMM. We used ICI as dependent variable, task type as fixed effect and animalID as random effects. P values ($p < 0.01$ for alternation and $p < 0.001$

for perseverance) were derived from the shuffled distribution by permuting the task type within each animal. Thus, we speculated that choice, and reward history effects could be explained by variance in the time interval between the choices.

To directly test if mPFC is specifically deployed for time-separated events, we increased the time interval between choices (Fig.5E) in both VR and VI tasks. The waiting interval time for the centre port was between 2.0–2.5 s. We collected 143 sessions with 572 ± 13 trials per session from 14 mice (11.1 ± 0.1 session per animal) in VI task and 36 sessions, with 355.7 ±16.7 trials per session from 6 mice (5.7 ± 1.2 sessions per animal) in VR task. In VR task with long delay animals were run on 0.3 vs. 0.3 and 0.1 vs. 0.6 probability with the block length of 15-30 trials. The effect of mPFC inactivation was clear on both reward and choice history effects in VI task and in VR task (Fig.5F). We did not see any change in the bias in VR (0.4326 ± 0.0482 for Saline and 0.4236 ± 0.1034 CNO groups in hM_4D animals, $p = 0.48$ using linear mixed model with animal ID as random effects and drug manipulation as fixed effects, residuals showing no deviation from normal distribution) and VI tasks with long delay (0.3303 ± 0.0313 for Saline and 0.3316 ± 0.0430 CNO groups in hM_4D animals, $p = 0.1$ using linear mixed model with animal ID as random effects and drug manipulation as fixed effects, residuals showing no deviation from normal distribution). Furthermore, effect size of the inactivation was significantly higher in long delay versions of the VI task (Fig.5G) but not in VR task. We tested the significance of the effect size separately for VR and VI tasks using LMM. Here effect size for all regressors were tested if they had a linear dependence with the delay condition as fixed effects and animalID as random effects (Effect_size ~ delay + (1|AnimalID). P values were derived from permutation test by shuffling the drug condition (CNO vs. Saline) within each animal 1000 times. We found significant effects $p < 0.01$ in VI task only for the hM4D(Gi) expressing animals but not for the GFP animals in VI and VR tasks and hM4D(Gi) animals in VR task. In an alternative way to test the effect of temporal gap between trials on the reward and choice history effects we again split the ICI interval into perseverance and alternation trials, focusing on the regressors for immediate past rewards and choices and combining VI sessions with short and long delay.

The same was done for the VR task. Using LMM, we evaluated how regression coefficients depended on the fixed effects of the drug, ICI interval (ICI_X), and their interaction, with animalIDs as random effects. The model specification was: Regressor ~ drug + ICI_X*drug + (1|AnimalID), where ICI_X represents the inter-choice interval for alternation (A) or perseverance (P). Significant interaction effects ($p < 0.01$) were observed only for hM4D(Gi) animals for rewards and choices one trial back in VI task and only for choices in VR task (Supplementary Table 2 and 3). To directly compare the effect of inactivation between VI and VR tasks as a function of delay we used the LMM that tested the interaction effect of delay with the task type (VI or VR). We used the following model Effect_size ~ delay*Task + (1|AnimalID) and run it for GFP and hM4D(Gi) animals separately. P values were derived from the shuffled distribution by permuting the drug labels (CNO vs. Saline) within the same animal. We show that delay had stronger effects in VI task compared to VR task ($p < 0.001$) only in hM4D(Gi) expressing animals but not in GFP animals.

We observed more pronounced reduction in reward rates in the long-delay version of the VI task compared to the VR task for the experimental group administered with CNO (SFig.7F).

Overall, we observed stronger effects of mPFC inactivation in VI task compared to VR task and this effect showed dependence on temporal separation of the task relevant events.

Delays reshape reward and choice history representations in mPFC

If above conclusion is true, then one should also see the change in neural representations when we manipulate the delay between temporal intervals between task relevant events. This notion aligns with a substantial body of literature supporting the prefrontal areas' involvement in tasks demanding the retention of event memories in short-term working memory[42–46]

We examined this question using a version of the VI task that imposed long temporal delays between choices. We kept the set reward probabilities and block length the same, but we introduced two block types with variable delays. Mice had to wait for 0.2–0.5 and 2–2.5 s in the short and long delay block types, respectively (Fig. 6A). We collected 84 sessions from 4 mice with a total of 42,654 trials (504 ± 47 trials per session). We recorded 2006 single units (99% of neurons showed a violation of refractory period, with less than <0.5% of total spikes) in mPFC from these animals (SFig.8A).

On a behavioral level we observed that mice alternated their choices more in short delay blocks compared to long delay blocks (first regression coefficient showed -2.4609 ± 0.03 in short delay blocks and -0.6483 ± 0.02 in long delay blocks. Fig.6B). The reward history effects showed no statistical difference (first regression coefficient for short and long delay blocks, p = 0.65 for right rewards and p = 0.54 for left rewards, Fig.6B). Consistent with this, the DT model's *Q* values and choice history effects showed more negative correlations in short delay than in long delay blocks (Fig.6C and SFig.8B) confirmed by the LMM that tested the delay condition on the correlation between *Q* values and choice history effects. Namely we tested the following model Corr ~ delay + side + (1|animalID). Here Corr stands for the session-by-session correlation of Q values and choice history effects, delay is the condition (short or long) and side separates left and right sides, both as fixed effects. animalID is used as a random effect from each animal. p values (p < 0.01) were derived from the shuffled distribution by permuting the delay condition within each animal. DT model parameters did not show significant difference between short and long delay conditions (SFig.8C). To test in which environment was the animal more efficient in reward harvesting efficiency, we measured regret in short vs. long delay blocks. Surprisingly, in the long delay blocks regret was lower, suggesting that animals were using more optimal strategies (Fig.6D).

On the single neuron level choice history representations were more dominant in short than in long delay blocks, mirroring the behavioral effects of past choices. Meanwhile, reward history representations were more dominant in long delay blocks (Fig.6E). However, while the same neurons maintained choice history representations a distinct set of neurons was responsible for reward history representations in short vs. long delay blocks (Fig.6F).

To test if mPFC representations tracked the animals performance, we decoded the choices given by the DT model and correlated the decoding accuracy of the recorded population with the accuracy of the DT model's prediction. We found that the correlation between population decoding accuracy and DT model's performance was stronger in long than in short delay blocks (SFig.8D). This is despite that DT model's fit was higher in short (0.7382 ± 0.0265) vs. long (0.6440 ± 0.0239) delay task (Mann-Whitney U test $p < 0.001$) and population decoding accuracy was not different in short (0.1512 ± 0.0177) vs. long delay blocks (0.1014 ± 0.0206). Furthermore, regret was also negatively correlated with the population decoding accuracy only in long delay blocks (SFig.8E). The absence of the correlation in the short delay blocks in comparison to the VI task without delay manipulation (SFig.2J and K) could be due to the small number of trials in delay manipulation (237.77±22.00 in short blocks and 219.73±22.00 in long blocks) compared to the one without delay manipulation (875 ± 37).

Overall, our data show that reward and choice history representations are dynamically adjusted as temporal separation of task relevant events increases.

Discussion

Reward foraging animals in nature are faced with non-stationary environment[47]. Typically rewards in unvisited food patches grow. Thus, animals need to keep track of not only how rich the current patch is but also how long they have stayed in the current patch. The VI task approximates such scenarios [5,48] and allowed us to explore how reward history and choice history were represented and used by prefrontal circuits. While choice history effects have been well documented in various animals in VI tasks[5–7,11,30] their representations and especially its integration with reward history has not been explored. The DT model, designed to integrate the effects of choice and reward history into the decision-making process, demonstrated superior predictive accuracy for the animals' choices compared to existing RL models of the same category [27]. We showed that neurons incorporate choice and reward history effects consistent with animals' decisions. The representations in mPFC as expected were highly mixed and comprised of different DT model variables like Q values, choice history effects and the combination of Q values. Furthermore, neural representations preferentially encoded reward and choice events from specific trials in history. As the temporal distance from the current trial increased, fewer neurons encoded events. This phenomenon could enable downstream targets to compute DT model variables at the population level.

Such history-specific tuning of neural responses was seen also in recurrent neural network model. These findings indicate that the recurrency in the network may have generated trial-history specific representations. Manipulating the task structure that imposed opposing choice history effects in VI and VR tasks was followed by concomitant changes in the behavior and representations, suggesting that mPFC was playing an adaptive function to adjust animals' behavior to the changing task demands. However, the inactivation of the mPFC revealed the task and temporal context specific effects. Namely, we observed that delay manipulation affected VI task stronger than the VR task. Different from the work of Bari et al.[7] inactivation of mPFC produced no significant change in bias. [7]This discrepancy may arise from variations in task specifics, such as head-fixed versus freely behaving conditions, differing lengths of inter-trial intervals (ITI), and the pretraining methods employed.

Finally, we note that it is difficult to understand the behavioral function of the mPFC based solely on the neural activity and its representations. The neural representations with the delay manipulation in VI task and inactivation effects in short and long delay version of the VI task are consistent with each other. Namely that as the population decoding accuracy and DT model's fit show positive correlations in long delay version of the VI task, inactivation effects also show stronger effects in long delay version of the VI task. However, the mere presence of these correlations does not guarantee that decision variables given by the DT model are used by the animals. In the short version of the VI task, we did observe also positive correlations between DT model's fit and population decoding accuracy (SFig.2J), however mPFC inactivation had modest behavioral effects, changing only the reward history effects (Fig.5C). This was not clear from mere neural representations, suggesting that it is difficult to infer from the pure representations which decision variables are used by the animals.

mPFC has been implicated in a wide range of tasks with a diversity of functions. In early studies, mPFC was considered to maintain a working memory in spatial navigation tasks[42,43] and later this function was expanded to non-spatial working memory tasks[24]. In parallel to working memory studies, research using decision making paradigms have shown that mPFC participates in conflict resolution[49] , learning the action sequences[50], effort-based decisions[51], learning the action values[52], learning the structure of the task[26], just to name a few. While these studies are shedding light on the diverse role mPFC plays in behavior they all view mPFC as the learning system that updates the action values based on the environmental state the animal is in. This is consistent with the wider theoretical framework that suggested mPFC comprises the meta reinforcement learning system that contains all the sufficient ingredients to implement the reinforcement learning model[22]. Our work is consistent with that idea, with one caveat. The implementation of the RL model is sensitive to the temporal separation of behaviorally relevant events. Which brings the original proposed function of mPFC of working memory into the RL framework. Indeed, the tasks that argue for mPFC involvement in value computations impose long inter-trial intervals that naturally make the task mPFC dependent[7,25].

What makes mPFC uniquely positioned to serve this function? The basal ganglia circuits are also well positioned to learn the action values[53]. With the help of dopaminergic system broadcasting reward prediction errors[54] striatal neurons have been shown to implement the RL algorithm by learning the action values[6]. We conjecture that the basal ganglia circuit can function when outcomes and actions are happening in close temporal proximity. So that proper credit is assigned to the right actions. However, when actions and outcomes are separated with long temporal delays the basal-ganglia circuit (or other subcortical circuits) alone cannot assign the credit to the right actions. Instead, cortical circuits that possess highly recurrent dynamics are in a good position to keep track of the past actions[55] and with the help of the reward prediction error are able to update the value of past actions. The mPFC's history-specific representations of past choices [56] indicate that credit can be assigned to those previous choices which contributed to obtaining rewards. This leads to the next question, what makes mPFC engage in task with long temporal delays?

Ventral tegmental area (VTA) dopaminergic neuron responses are positively correlated with the temporal gaps between unexpected reward outcomes[57]. Based on this we hypothesize that cortico-striatal connections are gated during long temporal delays because during long intervals dopaminergic neurons respond more vigorously to the unexpected reward. This idea is supported by anatomical connections between mPFC and dorsomedial striatum medium spiny neurons that express dopamine receptors[58] and receive strong dopaminergic input from VTA[59]. If representations from mPFC propagate to the striatal circuits, it is not surprising to find the representation of reward and choice history in striatal neurons that receive direct connections from mPFC[60].

We find little evidence that mPFC maintains event memories via persistent activity. Instead, we identify neurons whose selectivity peaks for rewards and choices occurring two or more trials back, a pattern that is difficult to reconcile with simple leaky integration models. Prior reports in dorsal anterior cingulate cortex of rhesus monkeys [56] and rodent cortex[11,25,34] documented reward-history signals one or two trials back but did not address whether these reflect leaky integration or history-selective coding, leaving the issue open. Our results address this gap by showing individual neurons with maximal tuning to two or more trials back in history, rather than the monotonic, decay predicted by leaky integration[35].

Temporal filtering of events suggests that the memory representations can depend on the current context. Indeed, in mouse auditory cortex current sensory input determines the memory representation of the past sensory input[36] and such postdictive effects have been described in human electroencephalogram recordings[61]. While such effect has not been reported in mouse mPFC according to the DT model (which is true also for many of the other Q-RL models) the unchosen option values decay while chosen option values are updated. This suggests that the memory of rewards is conditioned on the current choice which can be considered as the postdictive effects. The more direct evidence of the postdictive effects showed up when we examined the choice history representations. Choice history representations in VI and VR tasks showed clear adaptation to the changing task structure, although this was not performed on the same neurons and inferences were made from different sessions.

The choice history effects, and its neural representations received little attention in previous studies. Part of the reason is that it is not clear, at least in the VI task, what generates non monotonic function of choice history effects. We previously speculated that this functional form may arise from animals assuming uniform distribution of set reward probabilities[27]. However, we also showed that choice history effects change as set reward probabilities change[27] so understanding what drives choice history effects is fundamental to derive its generative mechanism. In either case choice history effects suggest the parallel computations taking place in mPFC, one that computes values based on the reward outcome and the other that forms a habit-like policy and ignores the trial-by-trial outcome. What was surprising that both of these representations may be harboured in the same neurons as mixed selectivity in our dataset was widespread. This suggests that individual neurons in mPFC implement two complementary strategies: one that is flexible and adaptive on a trial-by-trial basis and other that is more rigid but still quite efficient. Efficiency of a choice-based strategy was close to what animals achieved when using both reward history and choice history effects. The balance between these two strategies must depend on the environmental statistics. Although we show that choice-based strategy can be updated as animals change their behavior from VI to VR task, parametrically manipulating the choice-based strategy is needed to understand the interplay between these two systems.

One conceptual framework that can normatively explain the tradeoff between adaptive and more rigid strategies is policy compression[62] which posits that capacity-limited agents trade off expected reward against policy complexity, measured by the mutual information between states and actions, I(S;A). Low-complexity (compressed) policies are relatively state-agnostic and manifest as choice bias (strong choice-history influences), whereas higher-complexity policies are more state-dependent and emphasize reward-history information. Formally, the optimal policy maximizes expected reward subject to a bound on I(S;A); equivalently, behavior can be viewed as minimizing policy complexity for a desired reward rate (aspiration level). Under this view, when attainable reward rates are depressed (e.g., VI with added delays), maintaining a given aspiration level requires allocating more policy complexity—i.e., using more state information. This predicts stronger reward-history representations and weaker choice-history signals; when rewards are richer or states are less informative, policies compress and the opposite pattern emerges. This reward–complexity trade-off provides a normative explanation of increase in reward history representations and decrease in choice history representations in long delay VI task compared to the short one.

Our work, while shedding light on the role of the mPFC in decision-making, also had several shortcomings. 1. We could not answer which part of mPFC is responsible for the observed behavioral effects. While there are divisions in mPFC with different functions[63], our analysis of recorded data could not reveal region-specific representations. This may be due to the fact that we never recorded the different regions of mPFC within the same sessions and will require future work to record concurrently different parts of mPFC. 2. We did not explore how neural representations change as animals adapt their strategies when transitioning from VI to VR tasks. This was primarily because it was challenging to train animals to switch between the two tasks within a single session. The difficulty likely stems from the low reward probabilities used in these tasks (commonly 0.1 vs. 0.4 and 0.25 vs. 0.25), which required animals to perform a large number of trials to integrate outcome histories and discern whether they were in a VI or VR task. Future studies using greater differences in reward probabilities may help address this limitation. 3. We inactivated the mPFC using CNO and saline injections on an alternating schedule, which could have potentially biased animal behavior by causing them to anticipate drug versus vehicle injections. However, we believe this is unlikely because control animals expressing GFP showed no behavioral differences between CNO and saline injections.

Methods

Experiments on mice were conducted in compliance with institutional and national guidelines for the ethical treatment of animals. Approval was obtained from the Danish Animal Experiments Inspectorate under the Ministry of Justice (Permit 2017-15-0201-01357) and adhered to the U.S. National Institutes of Health Guide for the Care and Use of Laboratory Animals. Additionally, all procedures were approved by the Southern Illinois University Institutional Animal Care and Use Committee (IACUC protocol 23-005) on 20th May 2024.

Animals and surgery

The animals were (until microdrive implantation) group-housed with a maximum of 4 animals per cage under a regular 12 hours light/dark cycle.

We used a total of 65 male mice. These mice were 4-6 weeks old (C57BL\J6 background) at the time of viral injections or microdrive implantation. 21 animals were implanted with microdrive, all on their right hemisphere. The remaining 44 mice were used for $hM_4D(Gi)$ mediated inactivation experiments. Out of 21 animals used for recording single neurons, 11 mice were implanted for regular VI task, of which 5 mice were used to identify single units with spike sorting algorithm MClust and 6 mice were used to identify single units with modified version of the Kilosort2 (see details below). The remaining 4 mice were run with the delayed version of the VI task and 5 were used for running concurrently on VI and VR tasks. One mouse was used for concurrent recording and inactivation experiments. All these 10 animals were used to identify single neurons using modified version of Kilosort2 algorithm.

Before the microdrive implantation, animals were anesthetized by intraperitoneally injected Ketaminol (10 mg/ml)/ Xylazine (1.6 mg/ml) mixture. We injected 0.1mg per gram of body weight Ketaminol and 0.01mg of body weight Xylazine. Animals received a supplemental dose of anesthetics in 30-90 min. intervals to maintain the depth of anesthesia. After we confirmed the absence of pain reflexes, we shaved the head of the animal, disinfected with 70% ethanol and subcutaneously injected a mixture of lidocaine (10mg/ml) and norepinephrine(5µgr/ml) with a total volume ~10-20 µl. around the head area. Subsequently, the animal was head-fixed into a stereotaxic frame (David Kopf Instruments, USA). Following this, the head skin was either completely removed (for microdrive implantation) or cut in the middle (for viral injections), and the skull was exposed and dried thoroughly with hemostatic sponges (Ferrosan Medical Devices, Denmark). Taking Bregma as the reference point, the implantation site was marked at 1.6mm. anterior and 0.3-0.5mm lateral to Bregma. The skull surrounding the implantation site was covered with C\B-Metabond (Parkell, USA) to enhance the adhesion of the implant to the skull. The brain surface was exposed by drilling the cranial window. The microdrive that housed 8 tetrodes composed of nichrome wires (PX000004, Sandvik) was positioned above the brain surface according to the coordinates and lowered slowly until the guiding tube surrounding the tetrodes touched the skull. During the lowering process, the position of the slightly exposed tetrodes (300-500 micron extended from the guide tube) was monitored to measure penetration of the brain surface. One 0.25-mm diameter stainless steel wire (Alpha Wire Company, USA) was stripped at the end and inserted ~0.5mm below the dura to serve as ground and reference. We secured this wire to the skull by a thin layer of ultraviolet light curable dental cement (Vitrebond Plus, 3M Company, USA). The other end connected to the electrode interface board (EIB-36-PTB, Neuralynx,Inc) of the microdrive that also housed the tetrodes. Finally, the tetrodes and brain were protected by applying a drop of ocular lubricant (Dechra, USA) followed by the thick layer of the dental cement Paladur (Kulzer, Germany). Tetrodes were further lowered by 320µm using the screw on the microdrive. Ketoprofen, (5 mg/kg) was administered postoperatively.

Behavioral training

The water deprivation of the mice was initiated 72 hours before the training began. The behavioral sessions were performed in a custom-built behavior box equipped with three nose pokes (ports). Each port contained a stainless-steel hypodermic tube (15GA) tube connected to solenoid valve (LHDA1233115H, Lee Co. USA) to deliver water reward, an infrared light emitting diode (480-1969-ND, Digikey, USA) and a phototransistor (480-1958-ND, Digikey, USA) which allowed precise detection of the exact entry and exit times of the animals. The nose ports were also equipped with white light emitting diodes (VAOL-3LWY4-ND, Digikey, USA) to signal active ports during training sessions (see below). The LEDs, phototransistors and valves were connected via custom circuit to Bpod (Sanworks, USA). The behavioral protocols were written in MATLAB (Mathworks,Inc. USA) and were controlled by Bpod software.

Animals were trained to poke their noses to the center port, followed by poking into either of the side ports according to the following schedule. At first, rewards were automatically delivered to both side ports after center port entrance. As soon as the animals completed 20-30 such trials, we conditioned the reward delivery on the side port entrance. After another 20-30 trials, the final training stage had begun. Here, after correct trial initiation via the center port entry, a light in one of the side ports indicated the side of the next reward. The animals performed the task daily for 45-60 minutes until they reached approximately 70% correct responses. Each side port was calibrated to deliver 2ul of distilled water.

We used sound attenuated chambers (MAC3, IAC Acoustics, Denmark) for training and testing mice behavior.

Behavioral tasks

Each trial was initiated by a poke into the center port, where the animal had to wait for a variable delay period (0.2-0.5s. in VI and VR tasks). Following the delay, a small water reward (less than 0.5µl) was delivered in the center port, and lights switched on in the side ports signaling the decision stage. Animals reported their decision by poking into one of the two side ports. If the chosen port had been assigned a reward, this reward was delivered after a variable delay of 0.2-0.5 seconds. Upon leaving the side port, a new trial could be initiated immediately without any inter-trial interval. Early withdrawals from the center port or before reward delivery, or if the animal did not make a decision within 5 seconds, were considered missed trials. We did not impose any punishments.

The task was subdivided into blocks consisting of a variable number of trials, between 35 and 200 trials. For each block, the pair of set reward probabilities were kept constant but changed randomly between blocks. Overall, the following probability pairs were tested. The bold ones indicate probability pairs that we used in all of the animals except in 5 mice that were run on VI task with spikes sorted using MClust software (Fig.1 and Fig.2) and 6 mice that were run with VR task with long delay (Fig. 5F).

0.1 : 0.4
0.25 : 0.25
0.1:0.6
0.1:0.5
0.17:0.33
0.3 : 0.3

In the VI tasks once, an option had been loaded with a reward, it remained available until the animal collected it (baiting schedule). Thus, the reward probabilities were updated in each trial according to the following rule:

$$\mathrm{Pb_t}(R|T_c, i) = 1 - (1 - \mathrm{P_{set}}(R|i))^{T_c}, \qquad 1$$

$Pb_t(R|T_c, i)$, represents the probability of receiving reward R at trial *t* for action *i*; this probability depends on the number of trials Tc since the last time action *i* was chosen and on the set reward probability $P_{set}(R|i)$ of scheduling a reward for the option *i* determined by the experimenter. A more general equation that describes the reward probabilities looks like

$$\mathrm{Pb_t}(R|T_c, i) = 1 - \left(1 - \mathrm{P_{set}}(R|i)\right)\left(1 - \left(1 - \delta_{i,t}\right)\mathrm{Pb_{t-1}}(R|i)\right). \qquad 2$$

Here $\delta_{i,t} = 1$ if animal chooses the option *i* at trial *t* and $\delta_{i,t} = 0$ is otherwise.

Microdrive design

Extracellular recordings were performed via chronically implanted microdrives that housed 8 tetrodes. The microdrive was equipped with a screw-driven shuttle system, which provided a precise means of controlling the electrode depths via manual configuration. The microdrive consisted of 3D printed skeleton (Shapeways, Inc Netherlands) housing a single screw (320 um thread pitch, SSCF-M1.4-12-A2, Accugroup, Inc, UK) that moved tetrodes. The 8 tetrodes were loaded into the polyamide tube (TSP320450-AVXA03A, Polymicro technologies,Inc, USA) with an internal diameter of 320 um. This polyamide tube was glued to 21GA stainless steel hypodermic tube (Microgroup, Inc. USA) that was epoxied to the 3D printed shuttle after loading the tetrodes. The shuttle was connecting the screw with the tetrodes. The 21GA tube was sliding inside the 18GA stainless steel hypodermic tube or the guide tube (Microgroup, Inc. USA) that was glued (Loctite superglue, Silvan,Inc. Denmark) to the microdrive skeleton.

Electrophysiology

Before implantation, the tetrodes were gold-plated using nanoZ device (Neuralynx,Inc, USA). The final impedance of the tetrodes before implantation was within the 300-600 KΩ range. The tetrodes were lowered by 40 µm after each successfully completed behavioral session. For signal acquisition, we used OpenEphys system (https://open-ephys.org). The signal was acquired at 30KHz frequency and filtered between 600-6000Hz for MClust sorted spikes and between 300-5000Hz for kilosort2 sorted spikes (see below).

Viral injection and chemogenetic inactivation

For VR and VI task with short delay and VI task with the long delay we used 38 mice at 4 weeks of age bilaterally injected with 1.5 µl of adeno-associated viruses (AAV) 2/5 serotype produced at the University of Zurich vector core facility. We injected AAV-CaMKIIa-hM$_4$Di-mCherry (titer of the virus 7.3*10^12 genome copies per ml) in 22 mice, AAV-CAG-EGFP (5.4*10^12 genome copies per ml) in 16 mice per each hemisphere. One additional mouse was injected unilaterally with the AAV-CaMKIIa-hM$_4$Di-mCherry virus for concurrent inactivation and recording experiments. For VR task with long delay, we bilaterally injected AAV-hSyn-hM$_4$Di-mCherry (titer of the virus 1*10^13 genome copies per ml, gift from Bryan Roth lab Addgene plasmid #50477) in 3 mice and AAV-hSyn-EGFP (titer of the virus 7*10^12 genome copies per ml, gift from Bryan Roth lab, Addgene plasmid #50465) in 3 mice of 4 weeks old age. The injection was performed using glass micropipette using a picospritzer (General Valve, USA). Pulses of ~10ms duration were delivered at ~2Hz intervals. The volume was equally distributed at three depths in order to achieve uniform expression across the entire medial wall. The injection sites were 1.6 mm anterior and, ±0.3 mm lateral to bregma. We injected the virus at -1.8, -1.4, and -1.0 mm depths from the brain surface. At each depth, we paused injections for 2-3 min and 5 min at the final injection site. At the end of the virus injection, we glued (Vetbond, 3M Company, USA) skin halves together.

One week after the recovery from surgery, we trained animals on VR and VI tasks with two different delays (as described in the main text). Four weeks after the virus injection, we performed mPFC inactivation. For this, mice were intraperitoneally injected with either Clozapine N-oxide (C0832, Sigma Aldrich, Switzerland) (1 mg/kg) dissolved in 0.9% saline or 0.9% saline (300µl) on alternating days. The behavioral testing began 30 minutes post-injection.

Histology

After the behavior tests were concluded, the animals were anesthetized by either 500 ul of 10% urethane (U2500, Sigma Aldrich, USA) diluted in distilled water or Ketaminol/ Xylazine mixture that was used for surgery. The mice that had microdrive implants were connected to an impedance testing device (nanoZ) and a lesion was induced on 1-3 channels by delivering a 10µA of negative current for 10 seconds through the electrode wires. Afterward, the animals were perfused with 4% paraformaldehyde, and the brains were dissected and post-fixed for 1.5 hours. The dissected brains were stored in phosphate buffer at 4°C until further use.

The brains were sliced by a vibratome (Leica VT1200, Leica Biosystems, Germany) into 70 µm thick coronal sections. The regions where the recording took place were identified either by the lesion site or by anatomical landmarks of tetrode tracks. The slices were mounted on microscope slides by Vectashield mounting medium containing DAPI (Vector Laboratories, Burlingame, USA) and imaged by a fluorescence microscope (Zeiss Axiocam 712, Zeiss, Germany, or Leica DM 4500, Leica, Germany).

Data analysis

All data analysis was performed using MATLAB software (MathWorks). All data in the main text is reported as mean ± s.e.m.

Behavioral analysis

Trials, when the animals withdrew from the center port before the required waiting time or did not make a choice within the set time limit were removed from the data. Furthermore, trials, where the animals withdrew too early from the reward port were treated as non-rewarded trials. In the case of the inactivation experiments, sessions with less than 150 trials were excluded.

Logistic regression analysis for behavior

A logistic linear regression model was used to examine the influence of the choice and reward histories on the current choices. For this analysis, the behavioral data were analyzed session by session after the removal of missed trials (as described in the above section). To calculate the regression coefficients, first, we defined the rewards and choices as follows: $RR_t = 1$ if a reward was delivered on the right side at trial t, and 0 if there was no reward delivery or the left side was chosen. The equivalent was true for the left reward vector (LR_t). The choice vector (C) was defined as $C_t = 1$ if the right option was chosen in the given trial, and $C_t = 0$ otherwise. The probability of current choice C_t at trial t can be expressed as a logistic function that takes into account linear combination of the choice and reward histories multiplied by their corresponding coefficients β_m, *M* trials back plus bias term β_0:

$$\mathrm{h_t} = \sum_{m=1}^{M} \beta_m * \mathrm{RR_{t-m}} + \sum_{m=1}^{M} \beta_{m+M} * \mathrm{LR_{t-m}} + \sum_{m=0}^{M} \beta_{m+2*M} * C_{t-m} + \beta_0 \qquad \text{3a}$$

$$\mathrm{P}(C_t) = 1/(1 + \exp(-\mathrm{h_t})) \qquad \text{3b}$$

In order to account for the collinearity between rewards and choices and to improve the interpretability of the results, elastic net regularization was used[64]. The cost function involves negative log likelihood and penalty term of the elastic net, where *β* are the coefficients to estimate, interpolates between the *L1* and *L2* norms:

$$\mathrm{Cost}(\boldsymbol{\beta}) = -\sum_{t=1}^{N} [C_t(\log \mathrm{P}(C_t)) + (1 - C_t)(1 - (\log \mathrm{P}(C_t))] + \lambda \sum_{j=1}^{p} (0.25\,\beta_j^2 + 0.5|\beta_j|) \qquad 4$$

Therefore, the coefficients that minimized the cross-validated deviance (i.e., penalized negative log-likelihood) as a function of λ in a five-fold cross-validation process were selected as $\hat{\beta}$. We used MATLAB built-in function *lassoglm* with the linker function *logit* to perform regression.....

Bias estimation

To assess the choice bias for left or right ports, we calculated it as previously described[7]. Bias = 2 *|N_R /(N_R + N_L) – 0.5|. Here N_R is the total number of choices for the right port and N_L is the total number of choices for the left port in a session.

Linear Mixed Models (LMM)

LMM was implemented using MATLAB built in function *fitlme*. When residuals met the normality assumption - passing both the Lilliefors and Jarque-Bera tests - we reported the p-values derived from the LMM's t-statistics. Otherwise, we used shuffled distribution to derive the p values. We describe for each case how shuffled distributions were generated.

Reinforcement learning models of behavior

Here we describe how each of the RL models computed the *Q* value and/or choice history effects.

Indirect actor model. The indirect actor model updates the values (or state-action value) *Q* only for the chosen option,

$$Q_{i,t} = Q_{i,t-1} + \delta_{i,t-1}\alpha\left(R_{t-1} - Q_{i,t-1}\right), \qquad 6$$

$$preference_{i,t} = Q_{i,t},$$

Here $\delta_{i,t} = 1$ if animal chooses the option *i* at trial *t* and $\delta_{i,t} = 0$ is otherwise. $Q_{i,t}$ is the value of a given action *i* at time *t*. α is a learning rate and $R_{t-1} = 1$ in rewarded and $R_{t-1} = 0$ in unrewarded trials *t*. This model is the most simple version of the RL models[65].

The same notations are used for all the models below unless new variables are introduced.

Direct actor model. The value (Q) of the direct actor model is updated based on the probability of the chosen action and the reward outcome. This rule also affects the value of the unchosen actions.

$$Q_{i,t} = Q_{i,t-1} + \alpha\left(\delta_{i,t-1} - P(i_{t-1})\right)(R_{t-1} - c), \quad 7$$
$$preference_{i,t} = Q_{i,t}.$$

Here *c* is a parameter that we fit to the behavioral data and can be seen as the average reward rate in R-learning models [66].

***F-Q* down model.** The *F-Q* model is a slight modification of the indirect actor model, where the value of the unchosen actions is forgotten with the same learning rate α and vanishes to zero.

$$Q_{i,t} = Q_{i,t-1} + \alpha\left(\delta_{i,t-1}R_{t-1} - Q_{i,t-1}\right), \quad 8$$
$$preference_{i,t} = Q_{i,t}.$$

The *F-Q* model is equivalent to linear-nonlinear Poisson model [67].

***F-Q* with the choice effects model.** According to the *F-Q W/C* model, the probability of taking the next action depends not only on the reward expectation, but also on choice history *F* updated according to the following equation:

$$Q_{i,t} = Q_{i,t-1} + \alpha\left(\delta_{i,t-1}R_{t-1} - Q_{i,t-1}\right), \quad 9$$
$$F_{i,t} = F_{i,t-1} + \tau_F\left(\delta_{i,t-1} - F_{i,t-1}\right), \quad 10$$
$$preference_{i,t} = Q_{i,t} + \varphi F_{i,t}.$$

In this case, τ_F is the weighting of the choice history $F_{i,t}$, while φ characterizes the decay rate of it [68].

***F-Q* up model.** While the *F-Q down* model decreases the value (Q) of the unchosen actions, the proposed model increases the value up to a constant *C* with its own learning rate α_{up}, which acts as a positive counter for unchosen actions that increments the value.

$$Q_{i,t} = Q_{i,t-1} + \delta_{i,t-1}\alpha\left(R_{t-1} - Q_{i,t-1}\right) + \left(1 - \delta_{i,t-1}\right)\alpha_{up}\left(C - Q_{i,t-1}\right), \quad 11$$
$$preference_{i,t} = Q_{i,t}.$$

The initial value of $Q_{i,t}$ is set from a uniform random distribution constrained between [0, 1]. This model is just inverted version of the *F-Q down* model.

***Delta-F-Q* model.** The proposed model (adapted from Hattori et al. 2009) implements the different learning rates α_R and α_{NR} for rewarded and non-rewarded trials, respectively. For chosen options $Q_{i,t}$ is updated according to the following equation.

$$\boldsymbol{Q_{i,t}} = \begin{cases} \boldsymbol{Q_{i,t-1} + \delta_{i,t-1}\alpha_R * \left(R_{t-1} - Q_{i,t-1}\right) + \left(1 - \delta_{i,t-1}\right)(1 - \gamma) * Q_{i,t-1}, if\ R_{t-1} = 1} \\ \boldsymbol{Q_{i,t-1} + \delta_{i,t-1}\alpha_{NR} * \left(R_{t-1} - Q_{i,t-1}\right) + \left(1 - \delta_{i,t-1}\right)(1 - \gamma) * Q_{i,t-1}, if\ R_{t-1} = 0,} \end{cases} \quad 12$$

when an option is not chosen

$$\boldsymbol{Q_{unch,t} = (1 - \gamma) * Q_{unch,t-1}}, \qquad 13$$
$$preference_{i,t} = Q_{i,t}.$$

Here, γ is the forgetting rate of the unchosen option and is bounded between [0 1].

DT model

The value $Q_{i,t}$ for option *i* at a trial *t* is updated in the following way:

$$Q_{i,t} = Q_{i,t-1} + \alpha\left(\delta_{i,t-1}R_{t-1} - Q_{i,t-1}\right). \qquad 14$$

The fast *F* and slow choice history effects *S* are updated with the learning rate τ_F and τ_S respectively in the following way:

$$F_{i,t} = F_{i,t-1} + \tau_F\left(\delta_{i,t-1} - F_{i,t-1}\right), \qquad 15$$
$$S_{i,t} = S_{i,t-1} + \tau_S\left(\delta_{i,t-1} - S_{i,t-1}\right). \qquad 16$$
$$preference_{i,t} = Q_{i,t} + \phi * F_{i,t} + \theta * S_{i,t} \qquad 17$$

The choice history effects *F* and *S* are weighted by ϕ and θ parameters, respectively. As in previous equations $\delta_{i,t} = 1$ if the animal chooses option *i* at a trial *t* and $\delta_{i,t} = 0$ is otherwise. This model is better described in[27].

Action selection rule

Preference (which is identical to the value in simple RL models) computed by all RL models is converted to choice probability $P(a_t = i)$ for each option *i*, by softmax selection rule.

$$\mathrm{P}(a_t = i) = \frac{\mathrm{e}^{\beta\left(preference_{i,t}\right)}}{\sum_{k=1}^{A}\mathrm{e}^{\beta\left(preference_{k,t}\right)}}. \qquad 18$$

In the denominator, *k* will take up the number of available options.

Model optimization and comparisons

The modeling of behavior was conducted by maximizing the log-likelihood between the RL choice prediction and actual choice. We used the reward and choice history, $R(1{,}2 \ldots t-1)$ and $\delta_{i,t}(1{,}2{,} \ldots t-1)$ respectively, to find the best parameters of each RL model described above. We searched the parameters of each model to maximize the log-likelihood of choice prediction. For this, we sampled a combination of parameters from initial random distribution to generate choice probability by a softmax action selection rule.

The probability $P(a_t = i)$ and the action taken by the animal $\delta_{i.t}$ at trial t with the set of n parameters $\theta_{RL}(n)$ per model determines the log-likelihood l_t on each trial *t*:

$$\mathrm{l}_t(n)(\theta_{RL}(n)|x) = \delta_{i,t}\ln\Big(P\big(a_t = i\big|\theta_{RL}(n)\big)\Big) + \ln\Big(1 - \mathrm{P}\big(a_t = i\big|\theta_{RL}(n)\big)\Big). \qquad 19$$

To find the optimal parameters, first 1000 combinations of parameters were selected for each RL from a uniform distribution. The boundaries of the parameters were $\alpha \in [0{,}1], \beta \in [0{,}50], \varphi \in [-25{,}25], \vartheta \in$

$[-25,25], \iota \in [-25,25], \tau_F \in [0,1]$ and $\tau_S \in [0,1]$. Next, 1% of the combinations with the maximum mean log-likelihood was taken and a new set of 1000 combinations was drawn as follows:

$$\theta_{RL,2}(n) \sim \mathrm{N}\left(\bar{\theta}_{RL,1,1\%}, \sigma\left(\theta_{RL,1,1\%}\right)\right) \tag{20}$$

This process was repeated five times to narrow the original parameter space and search for higher log-likelihood values. The optimal parameters for the prediction of each model $\hat{\theta}_{RL}$ were then the combination of parameters with the highest mean log-likelihood of the last iteration in the optimization process.

The optimized parameters for each model and the estimated coefficients were trained and tested via five-fold cross-validation on the behavioral data in order to obtain the average of the minimum negative log-likelihood. In addition we computed the average area under the receiver operating characteristic (AUC) scores (Green DM Signal detection theory and psychophysics). To compute the AUC, the probability $P(a_t = i|\hat{Q}_{RL})$ was set as the score and the original action $\delta_{i,t}$ as the label. We used MATLAB function *perfcurve* to compute the AUC scores. These metrics were used for model selection and the goodness of prediction for each model, respectively.

Behavior of artificial agents

The random agent draws probabilities for left and right ports from uniform p ∈ $[0\ 1]$ probability distribution and selects ports with the higher probability. If the probabilities are the same, they are drawn again from the same uniform distribution and the process of port selection repeats.

The alternation agent constantly alternates between left and right choices.

Rew_prob agent selects ports with higher set reward probability. If the set reward probabilities are the same then agent randomly selects the port and sticks to the same port until change of probabilities favors alternative port.

Optimal_Baiting agent uses the baiting update equation (eq.2) to select higher reward probability port. If probabilities are the same the agent randomly (using p ∈ $[0\ 1]$ uniform distribution) selects either left or right port.

Regret

The regret of the animal or model is computed by subtracting the reward harvesting efficiency (reward rates per trial) of the animal or model from the Optimal_Baiting agent.

Selectivity of single neurons

The selectivity of individual cells for various behavioral events was computed using AUC. In brief, trial average firing rates in the decision epoch from the time animal entered the center port till the exit (minimum 0.2-0.5s) were divided into two distributions. These distributions corresponded to behavioral events such as immediate past rewards vs. no rewards, immediate past choices for the left vs. right and etc. as indicated for each selectivity analysis in the main text. Thus computed AUC scores were normalized from -1 to 1 that reflected the separability of firing rate distributions indicating the selectivity of the cells to the behavioral events. The scores close to -1 (suppression) and 1 (activation) indicated good separability and scores close to 0 indicated poor separability. For example, neurons that had a high firing rate (on average) for immediate past right rewarded vs. non -rewarded trials would have a positive AUC score. The same was true for neurons that had high firing rate (on average) for immediate past left rewarded trials vs. non- rewarded trials. Right reward selectivity and left reward selectivity were combined by inverting the sign of the left selectivity score to align it to the selectivity scores for immediate past choices. This was done because immediate past choice selectivity were computed for right vs. left choices, meaning that neurons that had high firing rate (on average) for immediate past right vs.

left choices showed positive AUC scores. Significance of AUC scores were computed by a permutation test, in which each dataset was pseudo-randomly shuffled 1,000 times to yield a *p* value.

Linear regression analysis for neurons

Here we used events (past rewards, past choices and current choices) to regress against firing rates of neurons in the decision epoch. This analysis was done using the elastic net regularization as described above (Linear regression analysis for behavior) except that we used firing rates of neurons as a dependent variable. For trial-based analysis we used regressors (β coefficients) events up to M = 10 trials back in history and for time based analysis we used events taking within 0.2 s. windows for up to M = 40 time bins back in history with respect to current trials.

$$\mathrm{Z_t} = \sum_{m=1}^{M} \beta_m * RR_{t-m} + \sum_{m=1}^{M} \beta_{m+M} * \mathrm{LR_{t-m}} + \sum_{m=0}^{M} \beta_{m+2*M} * \mathrm{RC_{t-m}}. \tag{21}$$

RR defines the right rewards, *LR* - left rewards and *RC* right choices for *m* trials back with respect to the current trial *t*. *Z* is z-scored trial average firing rate of neurons in the decision epoch for each *n* neuron.

The linear regressors were then regularized with an elastic net.

$$\mathrm{P}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} (0.25\, \beta_j^2 + 0.5|\beta_j|). \tag{22}$$

The coefficients that had the minimum mean squared error as a function of λ in a five-fold cross-validation process were selected as $\hat{\beta}$. This process was repeated 100 times with different train/validation sets in order to compute the mean of the $\hat{\beta}$ as

$$\bar{\boldsymbol{\beta}} = \frac{1}{100} \sum_{iter=01}^{100} \hat{\beta}_{iter}\ , \tag{23}$$

and the probability that $\bar{\beta} = 0$. This probability, corresponding to a p-value, computed from the cumulative empirical distribution from $\hat{\beta}$'s. More specifically for each $\hat{\beta}$, we resampled 100 times the data and refit the elastic net model to obtain a distribution of $\hat{\beta}$ coefficients. We then estimated the probability that a coefficient was effectively zero by evaluating its empirical cumulative distribution function (ECDF) near zero. This probability was taken as the empirical *p*-value: high values indicate that the regressor's coefficient was frequently zero across resamples (weak evidence for an effect), whereas low values indicate that the coefficient was consistently non-zero (stronger evidence for an effect). We correct for the false positive discoveries using Benjamini-Hochberg procedure and only the regressors that have a probability below of 0.00001 (1 in 100,000) of being falsely discovered were selected. Without this correction almost all (97%) of neurons would show the significant regressors. Elastic net regularization also strikes a balance between being too relaxed and too restrictive.

Neural population decoding of behavioral variables

For population decoding of DT model and other behavioral variables (like current and past choices, immediate past left and right rewards) we used generalized linear model (MATLAB function *lassoglm* with binomial distribution). We used z-scored trial-averaged firing rate of neurons in the decision epoch to regress against either choices predicted by the DT model, current, immediate past choices or immediate past left and right rewards.

$$\mathrm{BV_t} = \sum_{m=1}^{M} \beta_m * N_m. \tag{24}$$

Here N stands for z-scored firing rate of a neurons from a total population of M neurons. BV_t stands for the behavioral variable (choice probability, current choice and etc) on a current trial. The regressors were regularized as described above using elastic net regularization. We used 10 fold cross validation for regressors. The same regression analysis was performed on shuffled data. For shuffling we randomly permuted the BV across the trials. The final decoding accuracy was computed as an average of the difference between unshuffled and shuffled decoding accuracy scores.

Support vector machine classification of behavioral states

The trial averaged firing rate of neurons were analyzed in the decision epoch to solve the multiclass classification problem by training many binary classifiers to discriminate pair of unique states. Binary classifiers were pooled by a majority of votes to separate one unique state from the rest of the states. Each state comprised a unique set of combinations of rewards and choices for each trial back in history plus choices for current trials. Thus, for each trial back there are 4 different states (rewards, no rewards left and right choices). There are 4^n possible states for states that contained *n* trials in history, multiplied on 2 additional states for current trials (left and right choices).

We trained support vector machine classifier (*fitcsvm* function in MATLAB) on 1/2 of the trials from each session and tested performance on the remaining 1/2 of the test data. We performed classification on a different number of states starting with 2 states (left and right current choices) and ending with 16 states (SFig.4). The performance of classifier was compared to shuffled data by shuffling the labels for only one pair of events (for rewards or for choices for specific trials in history) while holding the remaining labels the same.

Recurrent neural network model

The recurrent neural network (RNN) was implemented using MATLAB function *layerecnet*. The two inputs of the RNN at time *t* are the reward R_t and choice C_t experienced by the animal. The network consists of a fully connected hidden layer with 30 units, a sigmoid function for non-linearity and a layer with the tap delay of 1. The net has two targets as outputs, each of them correspond to the value of the DT model for the left ($Q_{left,t} + \phi * F_{left,t} + \theta * S_{left,t}$) and right ($Q_{right,t} + \phi * F_{right,t} + \theta * S_{right,t}$) ports. The training of the net weights was done by a Levenberg-Marquardt optimization. We split data into half for training and testing. The test sets give in average a correlation higher than 0.98 for the outputs of the trained network with respect to the desired targets. The z-scored firing rate of each hidden layer neuron was analyzed in the same way as real neurons.

Spike sorting using Kilosort2 based algorithm

We used Kilosort2 [37] to initially isolate and build spatial-temporal templates for each unit. Template for each unit and each channel is defined as

$$\mathbf{K} = \sum_{j=1}^{N} W_j E_j U_j.$$

Here *K* is the template defined by the sum of its *j* temporal *W* and spatial *U* components. *E* is the singular value matrix. We used up to N = 3 components.

These Kilosort2 identified units underwent further “cleaning” steps to remove the noise. Namely we performed the following steps:

1. For each unit we identified the set of channels that passed a specific criterion. Namely, we identified for each unit, “top channel” that had the highest first singular component *U* (spatial component)

across all channels. Along with the “top channel” we retained all the remaining channels whose first singular component was above 30% of the “top channel”. This resulted in the selection of channels for each unit not restricted to the tetrode configuration. Note that MClust sorting algorithm restricts single unit isolation to four channels.

2. We used Hartigan’s dip test to see if spikes projected on the “top channel’s” template (defined by the first singular component) violate unimodal distribution.
3. If Hartigan’s test was violated we applied Gaussian mixture model to partition spikes into two clusters. For this we again used projections of spikes into the template waveforms reconstructed by first two singular components.
4. Units were inspected manually and further steps were determined by the user.
5. On autocorrelograms (100 ms. window) if unit had no violation of refractory period (1ms window centered around zero) of less than of 0.05% of total counts we retained that unit.
6. We discarded all the units that passed the step 5, but had fewer than 1000 spikes or waveforms did not resemble typical spike waveforms.

We deposit the Matlab code that implements all the steps described above in github repository (https://github.com/1804MB/Kvistiani-lab_Dsort).

Data availability: All behavioral and spike sorted data are available from public repository Code Ocean https://codeocean.com/capsule/6312901/tree, DOI : 10.24433/CO. 6312901.v1. Source data are provided with this paper.

Code availability: All code is available from public repository Code Ocean https://codeocean.com/capsule/6312901/tree , DOI : 10.24433/CO. 6312901.v1

References

1. Sutton, R.S and Barto, A. G. *Reinforcement Learning: An Introduction, by Sutton, R.S. and Barto, A.G. Trends in Cognitive Sciences* (1999). doi:10.1016/S1364-6613(99)01331-5.

2. Corrado, G. & Doya, K. Understanding neural coding through the model-based analysis of decision making. *J Neurosci* **27**, 8178–80 (2007).

3. Hattori, R. & Komiyama, T. Context-dependent persistency as a coding mechanism for robust and widely distributed value coding. *Neuron* **110**, 502-515.e11 (2022).

4. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci* **7**, 404–410 (2004).

5. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* **84**, 555–79 (2005).

6. Kim, H., Sul, J. H., Huh, N., Lee, D. & Jung, M. W. Role of Striatum in Updating Values of Chosen Actions. *Journal of Neuroscience* **29**, 14701–14712 (2009).

7. Bari, B. A. *et al.* Stable Representations of Decision Variables for Flexible Behavior. *Neuron* https://doi.org/10.1016/J.NEURON.2019.06.001 (2019) doi:10.1016/J.NEURON.2019.06.001.

8. Beron, C. C., Neufeld, S. Q., Linderman, S. W. & Sabatini, B. L. Mice exhibit stochastic and efficient action switching during probabilistic decision making. *Proc Natl Acad Sci U S A* **119**, (2022).

9. Belinsky, R., González, F. & Stahl, J. Optimal behavior and concurrent variable interval schedules. *J Math Psychol* **48**, 247–262 (2004).

10. Hwang, E. J., Dahlen, J. E., Mukundan, M. & Komiyama, T. History-based action selection bias in posterior parietal cortex. *Nat Commun* **8**, 1242 (2017).

11. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-Specificity and Plasticity of History-Dependent Value Coding During Learning. *Cell* **177**, 1858-1872.e15 (2019).

12. Guo, Z. V. *et al.* Flow of Cortical Activity Underlying a Tactile Decision in Mice. *Neuron* **81**, 179–194 (2014).

13. Allen, W. E. *et al.* Global Representations of Goal-Directed Behavior in Distinct Cell Types of Mouse Neocortex. *Neuron* **94**, 891-907.e6 (2017).

14. Phillips, C. G., Zeki, S. & Barlow, H. B. Localization of function in the cerebral cortex. Past, present and future. *Brain* **107 (Pt 1)**, 327–61 (1984).

15. Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810-824.e9 (2019).

16. Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).

17. Murakami, M., Shteingart, H., Loewenstein, Y. & Mainen, Z. F. Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in Rodent Frontal Cortex. *Neuron* **94**, 908-919.e7 (2017).

18. Machens, C. K., Romo, R. & Brody, C. D. Functional, But Not Anatomical, Separation of 'What' and 'When' in Prefrontal Cortex. https://doi.org/10.1523/JNEUROSCI.3276-09.2010 (2010) doi:10.1523/JNEUROSCI.3276-09.2010.

19. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

20. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).

21. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol* **37**, 66–74 (2016).

22. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* **21**, 860–868 (2018).

23. Hattori, R. *et al.* Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience 2023 26:12* **26**, 2182–2191 (2023).

24. Liu, D. *et al.* Medial prefrontal activity during delay period contributes to learning of a working memory task. *Science (1979)* **346**, 458–463 (2014).

25. Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct Roles of Rodent Orbitofrontal and Medial Prefrontal Cortex in Decision Making. *Neuron* **66**, 449–460.

26. Akam, T. *et al.* The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection. *Neuron* **109**, 149–163 (2021).

27. López-Yépez, J. S., Martin, J., Hulme, O. & Kvitsiani, D. Choice history effects in mice and humans improve reward harvesting efficiency. *PLoS Comput Biol* **17**, e1009452 (2021).

28. Bari, B. A. & Gershman, S. J. Undermatching Is a Consequence of Policy Compression. *Journal of Neuroscience* **43**, 447–457 (2023).

29. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Neuroscience: Representation of action-specific reward values in the striatum. *Science (1979)* **310**, 1337–1340 (2005).

30. Tsutsui, K. I., Grabenhorst, F., Kobayashi, S. & Schultz, W. A dynamic code for economic object valuation in prefrontal cortex neurons. *Nature Communications 2016 7:1* **7**, 1–16 (2016).

31. The Mouse Brain in Stereotaxic Coordinates, Compact - 3rd Edition. https://shop.elsevier.com/books/the-mouse-brain-in-stereotaxic-coordinates-compact/franklin/978-0-12-374244-5.

32. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131**, 1–11 (2005).

33. Kvitsiani, D. *et al.* Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature* **498**, 363–366 (2013).

34. Seo, H. & Lee, D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *Journal of Neuroscience* **27**, 8366–8377 (2007).

35. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782–7 (2004).

36. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat Neurosci* **24**, 715–726 (2021).

37. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Harris, K. D. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. *bioRxiv* 061481 (2016) doi:10.1101/061481.

38. Killcross, S. & Coutureau, E. Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex* **13**, 400–408 (2003).

39. Jog, M. S., Kubota, Y., Connolly, C. I. & Graybiel, A. M. Building Neural Representations of Habits. *Science (1979)* **286**, 1745–1749 (1999).

40. Ozdemir, A. T. *et al.* Unexpected Rule-Changes in a Working Memory Task Shape the Firing of Histologically Identified Delay-Tuned Neurons in the Prefrontal Cortex. *Cell Rep* **30**, 1613-1626.e4 (2020).

41. Roth, B. L. DREADDs for Neuroscientists. *Neuron* **89**, 683 (2016).

42. Bolkan, S. S. *et al.* Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience 2017 20:7* **20**, 987–996 (2017).

43. Baeg, E. H. *et al.* Dynamics of population code for working memory in the prefrontal cortex. *Neuron* **40**, 177–188 (2003).

44. Kim, D. *et al.* Distinct Roles of Parvalbumin- and Somatostatin-Expressing Interneurons in Working Memory. *Neuron* **92**, 902–915 (2016).

45. Bae, J. W. *et al.* Parallel processing of working memory and temporal information by distinct types of cortical projection neurons. *Nat Commun* **12**, (2021).

46. Goldman-Rakic, P. S. Regional and cellular fractionation of working memory. *Proc Natl Acad Sci U S A* **93**, 13473 (1996).

47. Amarasekare, P. Spatial Dynamics of Foodwebs. *https://doi.org/10.1146/annurev.ecolsys.39.110707.173434* **39**, 479–500 (2008).

48. Fleshler, M. & Hoffman, H. S. A progression for generating variable-interval schedules. *J Exp Anal Behav* **5**, 529–530 (1962).

49. De Wit, S., Kosaki, Y., Balleine, B. W. & Dickinson, A. Dorsomedial prefrontal cortex resolves response conflict in rats. *Journal of Neuroscience* **26**, 5224–5229 (2006).

50. Tervo, D. G. R. *et al.* Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. *Cell* **159**, 21–32 (2014).

51. Walton, M. E., Bannerman, D. M. & Rushworth, M. F. S. The role of rat medial frontal cortex in effort-based decision making. *Journal of Neuroscience* **22**, 10996–11003 (2002).

52. Ostlund, S. B. & Balleine, B. W. Lesions of Medial Prefrontal Cortex Disrupt the Acquisition But Not the Expression of Goal-Directed Learning. *Journal of Neuroscience* **25**, 7763–7770 (2005).

53. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J Neurosci* **29**, 9861–9874 (2009).

54. Waelti, P., Dickinson^2, A. & Schultz, W. *Dopamine Responses Comply with Basic Assumptions of Formal Learning Theory*. *NATURE* vol. 412 www.nature.com (2001).

55. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature 2019 566:7743* **566**, 212–217 (2019).

56. Seo, H. & Lee, D. Temporal Filtering of Reward Signals in the Dorsal Anterior Cingulate Cortex during a Mixed-Strategy Game. *Journal of Neuroscience* **27**, 8366–8377 (2007).

57. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience 2008 11:8* **11**, 966–973 (2008).

58. Kreitzer, A. C. & Malenka, R. C. Striatal Plasticity and Basal Ganglia Circuit Function. *Neuron* **60**, 543–554 (2008).

59. Beier, K. T. *et al.* Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. *Cell* **162**, 622–634 (2015).

60. Yoshizawa, T., Ito, M. & Doya, K. Neuronal Representation of a Working Memory-Based Decision Strategy in the Motor and Prefrontal Cortico-Basal Ganglia Loops. *eNeuro* **10**, (2023).

61. Grabot, L., Kayser, C. & van Wassenhove, V. Postdiction: When Temporal Regularity Drives Space Perception through Prestimulus Alpha Oscillations. *eNeuro* **8**, (2021).

62. Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cognition* **204**, (2020).

63. Heidbreder, C. A. & Groenewegen, H. J. The medial prefrontal cortex in the rat: evidence for a dorso-ventral distinction based upon functional and anatomical characteristics. *Neurosci Biobehav Rev* **27**, 555–579 (2003).

64. James, G., Witten, D., Hastie, T. & Tibshirani, R. Springer Texts in Statistics An Introduction to Statistical Learning. http://www.springer.com/series/417.

65. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience*. (Peter Dayan and L.F. Abbott, 1991).

66. Schwartz, A. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. https://doi.org/10.1016/B978-1-55860-307-3.50045-9 (1993) doi:10.1016/B978-1-55860-307-3.50045-9.

67. Corrado, G. S., Sugrue, L. P., Sebastian Seung, H. & Newsome, W. T. Linear-Nonlinear-Poisson Models of Primate Choice Dynamics. *J Exp Anal Behav* **84**, 581–617 (2005).

68. Katahira, K. The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *J Math Psychol* **66**, 59–69 (2015).

Acknowledgments: We express our gratitude to the members of the Kvitsiani lab, including Sophie Seidenbecher, Madeny Belkhiri, and Jesper Hagelskaer, for their valuable feedback on both the analysis and the writing of the manuscript. We thank Ashok Litwin Kumar and Larry Abbott for their assistance with the neural data analysis. We also appreciate Naoshige Uchida for providing critical feedback on the manuscript.

This study was supported by the Lundbeck Foundation grant: DANDRITE-R248-2016-2518 https://lundbeckfonden.com and startup funds from Southern Illinois University at Carbondale.

Author contributions statement: D.K. conceived and designed the project. A.B., J.M., M.M., T.-F.W. and E.D. performed the experiments. J.S.L.-Y., A.B., J.M. and D.K. analyzed the data. J.S.L.-Y. developed the computational modeling of behavior. O.H. contributed to modeling and interpretation. J.S.L.-Y., A.B., O.H. and D.K. wrote the manuscript with input from all authors. E.D. and D.K. revised the manuscript and addressed reviewers' comments.

Competing interests statement: Authors declare no competing interests.

Figure Legends

Figure 1. Choice- and reward-history exert opposing effects on animals decisions. A) Discrete version of the VI task illustrating trial structure and behavioral epochs. The upper panel shows task-relevant events. The lower panel shows the event timeline; dashed lines mark variable intervals and solid lines fixed intervals. **B)** Top: performance of animal D004 in one session. Black vertical lines indicate block transitions and the change in set reward probabilities. Choices were convolved with a Gaussian filter (length = 10 trials, s.d. = 5). Right choices were coded as 1 and left choices as −1. Rewards were convolved identically. Bottom: left panel shows experienced reward rates across block transitions; right panel shows filtered choices across the same transitions. **C)** Relationship between reward ratio and choice ratio across sessions. A significant positive association was observed (two-sided linear regression: slope = 0.65, 95% CI = 0.578–0.724, *t*(df = 127) = 17.69, *r* = 0.89, *P* = 2.25×10^{-29}). No adjustment for multiple comparison tests (NAMCT) were applied. **D)** Mean ± s.e.m. reward-harvesting efficiency for artificial agents and animals. Agents include Random, Alternation, Rew_Prob, and Optimal_Baiting. Performance from animal D004 (left) and all animals combined (right; *n* = 4 animals, 76 sessions) is shown. Animals were compared to agents using two-sided Mann–Whitney U tests; *** indicates *P* < 0.001. Artificial agents were evaluated on the same reward schedules as animals. **E)** Logistic regression for past right rewards, past left rewards, and past choices predicting current choice (mean ± s.e.m.; *n* = 82 sessions). **F)** Alternation rate (mean ± s.e.m.) as a function of the difference in set reward probabilities (low = 0; medium = 0.3; high = 0.5–0.6). Differences across the three conditions were assessed using a linear mixed-effects model (LMM) with alternation rate as the dependent variable, reward-probability difference as a fixed effect, and animal identity as a random intercept. Two-sided *P*-values from the LMM were FDR-corrected using the Benjamini–Hochberg procedure; ** indicates *P* < 0.01 (*P* = 0.003 and 0.009). Data were collected from *n* = 2 animals across 45 sessions. Source data are provided as a Source Data file.

Figure 2. mPFC representations are consistent with the behavioral effects of reward and choice history. A) Example neuron showing spike raster and PETH aligned to past rewarded and non-rewarded trials (top) and to past right and left choices (bottom). PETH is computed across all trials and smoothed with a Gaussian kernel (2 ms window, s.d. = 50 ms). Lines and shaded bands indicate mean ± s.e.m. **B)** Selectivity (AUC) scores for immediate past rewards vs no reward and for past choices. Associations between reward-related and choice-related AUCs were tested using a two-sided linear mixed-effects model with AUC_choice as the dependent variable, AUC_reward as a fixed effect, and animal identity as a random intercept. A significant negative effect was observed (slope = −0.30, 95% CI = −0.554 to −0.054, $P = 0.037$, two-sided permutation test, NAMCT). **C)** DT model description. Q-values incorporate reward input R(t). Fast (F) and Slow (S) choice traces integrate past choices with parameters ϕ and θ. Q, F, and S jointly determine choice probability P(c|t) via a softmax function. **D)** Relationship between regret (difference between Optimal_Baiting and DT model performance) and DT model fit (AUC) across sessions. A significant negative association was found using two-sided Pearson correlation ($r = -0.351$, 95% CI [−0.545, −0.121], $t(65) = -3.02$, $P = 0.0036$, NAMCT). **E)** Pearson correlations between reward-history/value signals (Q) and choice-history signals ($\phi F + \theta S$) for left and right options (one point per session). Confidence intervals and significance, using two-sided permutation **tests** (1,000 shuffles of session labels); $P < 0.001$ indicates that fewer than 0.1% of shuffled correlations exceeded the observed absolute correlation. **F)** Trial-by-trial correlations between firing rate and DT-model variables for left-port and right-port trials. Numbers of neurons per quadrant, confidence intervals, and significance derived using two-sided Pearson correlations and two-sided permutation tests (1,000 shuffles; $P < 1\times10^{-15}$). **G)** Regression coefficients relating DT-model variables to trial-averaged firing rates during the decision epoch. Coefficients were estimated using two-sided elastic-net regression, with empirical P-values derived from resampling of coefficient estimates and corrected using the Benjamini–Hochberg FDR procedure ($P < 1\times10^{-5}$ threshold). Source data are provided as a Source Data file.

Figure 3. mPFC neurons encode history-specific rewards and choices. A) Prediction accuracy of the neural population for DT-model choice probability (x-axis) vs. prediction accuracy for immediate past choice (top left), the animal’s current choice (top right), immediate past right reward (bottom left), and immediate past left reward (bottom right). Each point is one session. **B)** Example neuron’s spike raster and PETH aligned to center-port entry. Trials are grouped by left (upper row) or right (lower row) choices and by rewarded or unrewarded outcomes n trials back (shown in columns). PETHs were smoothed with a Gaussian kernel (2 ms window, s.d. = 50 ms). Lines indicate mean; shading shows s.e.m.**C)** Heatmap of regression coefficients obtained by regressing each neuron’s trial-averaged firing rate during the decision epoch against rewards and choices up to 10 trials back. Only coefficients that remained significant after cross-validation (held-out 1/5 of trials) and false-discovery-rate correction are shown. Neurons are sorted by the lag with the largest absolute coefficient and grouped by sensitivity to past right rewards, past left rewards, current choice, or past choices (separated by dashed blue lines). **D)** Stability of preferred lag across the session. For neurons whose dominant coefficient retained the same sign in both halves of the session, preferred lag was computed separately for the first and second halves (defined as the predictor with the largest absolute coefficient and non-zero in ≥85% of 100 bootstrap resamples). Each point shows paired lag indices for one neuron; marker size and gray shade indicate how many neurons share that value (scale bar, right). Diagonal points reflect stable lag preference; off-diagonal points denote shifts across the session. **E)** Percentage of neurons with significant regression coefficients to any past event. **F)** Architecture of the recurrent neural network (RNN) trained to reproduce DT-derived Q-values and choice-history effects. Inputs encode reward and choice outcomes; the recurrent layer contains 30 units; the output layer computes Q and history signals for each option. **G)** Regression coefficients for recurrent-layer units computed as in (C). **H)** Percentage of recurrent-layer units with significant history-based coefficients as in (E). Source data are provided as a Source Data file.

Figure 4. Behavioral and mPFC representations adjust to VI and VR task structure. A) Alternation rate and reward rate difference from one animal performing VI and VR tasks. Signals were smoothed with Gaussian filter (window = 75 trials, s.d. = 25 trials). **B)** Left: reward rates in VI and VR across all animals (*n* = 5) and sessions (*n* = 129). Right: alternation rates. Two-sided LMMs, task type (VI vs VR) fixed effect and animal identity - random intercept revealed for reward rate (β = 0.178, 95% CI 0.143–0.214, *t*(127) = 9.89, $P = 1.9\times10^{-17}$) and alternation rate (β = 0.247, 95% CI 0.205–0.289, *t*(127) = 11.73, $P = 5.5\times10^{-22}$). NAMCT was applied. **C)** Regressors (mean ± s.e.m.) were estimated using elastic-net regression. Lag-wise VI–VR differences were tested with two-sided LMMs and within-animal permutation tests (1,000 shuffles), followed by Benjamini–Hochberg FDR correction. Asterisks mark FDR-corrected $P < 0.05$. **D)** Correlations between values (Q) and choice-history effects ($\phi F + \theta S$) for VI and VR sessions. Task differences were evaluated with two-sided LMM (task fixed; animal ID random) and a two-sided within-animal permutation test (1,000 shuffles; $P = 0.0009$), NAMCT was applied . **E)** Selectivity (AUC) for immediate past rewards and immediate past choices. Selectivity in VI was negative but not significant (top; $P = 0.578$), whereas VR showed a significant positive effect (bottom; $P = 1.2\times10^{-5}$). **F)** Trial-by-trial correlations between neuronal firing rate and DT-model variables ($Q, \phi F + \theta S$) for VR and VI sessions. **G)** Schematic of intra-task and cross-task evaluation of the DT model. Intra-task testing uses parameters derived within the same task (VI→VI or VR→VR). Cross-task testing applies VI-derived parameters to VR sessions and vice versa. **H)** DT-model performance (mean ± s.e.m.) under intra-task and cross-task testing (*n* = 5 animals, 65 sessions, $P = 0.005$, two-sided Mann–Whitney U test). **I)** Left: cross-task performance vs neural decoding accuracy (two-sided Pearson correlation, $P = 0.0009$). Right: decoding accuracy regressed on mean ITI, DT-model fit, mean RT, and cross-task performance using two-sided t-tests; NAMCT was applied. Source data are provided as a Source Data file.

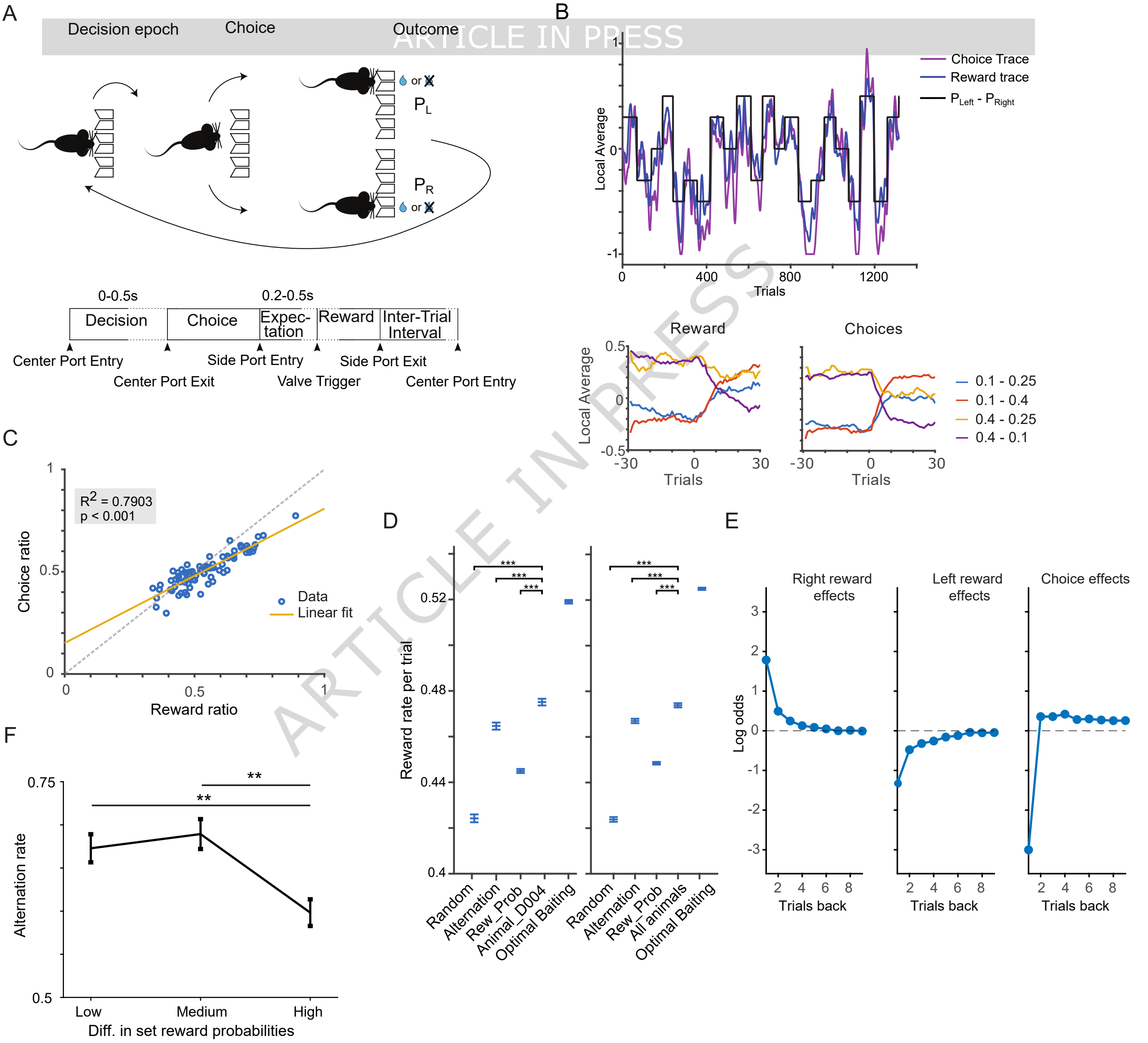
Figure 5. mPFC inactivation results in behavioral effects that exhibit both task and temporal dependency. A) Coronal brain section showing representative hM4D(Gi)-mCherry expression (red) with DAPI counterstaining (blue) in the mPFC (replicated in 19 animals). **B)** Schematic of the behavioral timeline, indicating epochs of task performance and drug (CNO) or saline injections. **C)** Influence of past rewards (top) and past choices (bottom) up to 5 trials back on current choice, shown as logistic regression coefficients in VR (left) and VI (right) tasks (mean ± s.e.m.; *n* = 16 GFP and *n* = 22 hM$_4$D(Gi) animals). For each lag, coefficients were compared using two-sided LMM (fixed effects: genotype, drug, genotype × drug; random intercept: animal ID). Significance was determined using within-animal permutation tests (1,000 shuffles) and Benjamini–Hochberg FDR correction across lags. Green asterisks denote FDR-corrected *P* < 0.05, 0.01, or 0.001 for GFP animals; red asterisks denote the same thresholds for hM4D(Gi) animals. **D)** Inter-choice intervals (ICI) in VR and VI tasks for alternation (top) and perseverance (bottom) across GFP (*n* = 16) and hM4D(Gi) (*n* = 22) animals, drug conditions combined. Differences between VI and VR were tested with a two-sided LMM (task fixed; animal ID random). ** indicates FDR-corrected *P* < 0.01 (*P* = 0.007 for alternation; *P* = 0.002 for perseverance). Box-plots: center line = median; box = 25th–75th percentiles (IQR); whiskers = full data range. **E)** Schematic of the delayed-decision version of the task indicating when the delay period was introduced. **F)** Same analysis as in (C), applied to the delay version of the VR (left; *n* = 3 GFP, *n* = 3 hM4D(Gi)) and VI (right; *n* = 5 GFP, *n* = 9 hM4D(Gi)) tasks. Two-sided LMMs (genotype × drug fixed effects; animal ID random) and 1,000-shuffle within-animal permutation tests were used at each lag, followed by Benjamini–Hochberg FDR correction. Asterisks indicate FDR-corrected *P* < 0.05, 0.01, or 0.001. **G)** Effect sizes of past-reward and past-choice regressors in short- and long-delay versions of the VR (top) and VI (bottom) tasks, computed from logistic regression coefficients. Source data are provided as a Source Data file.

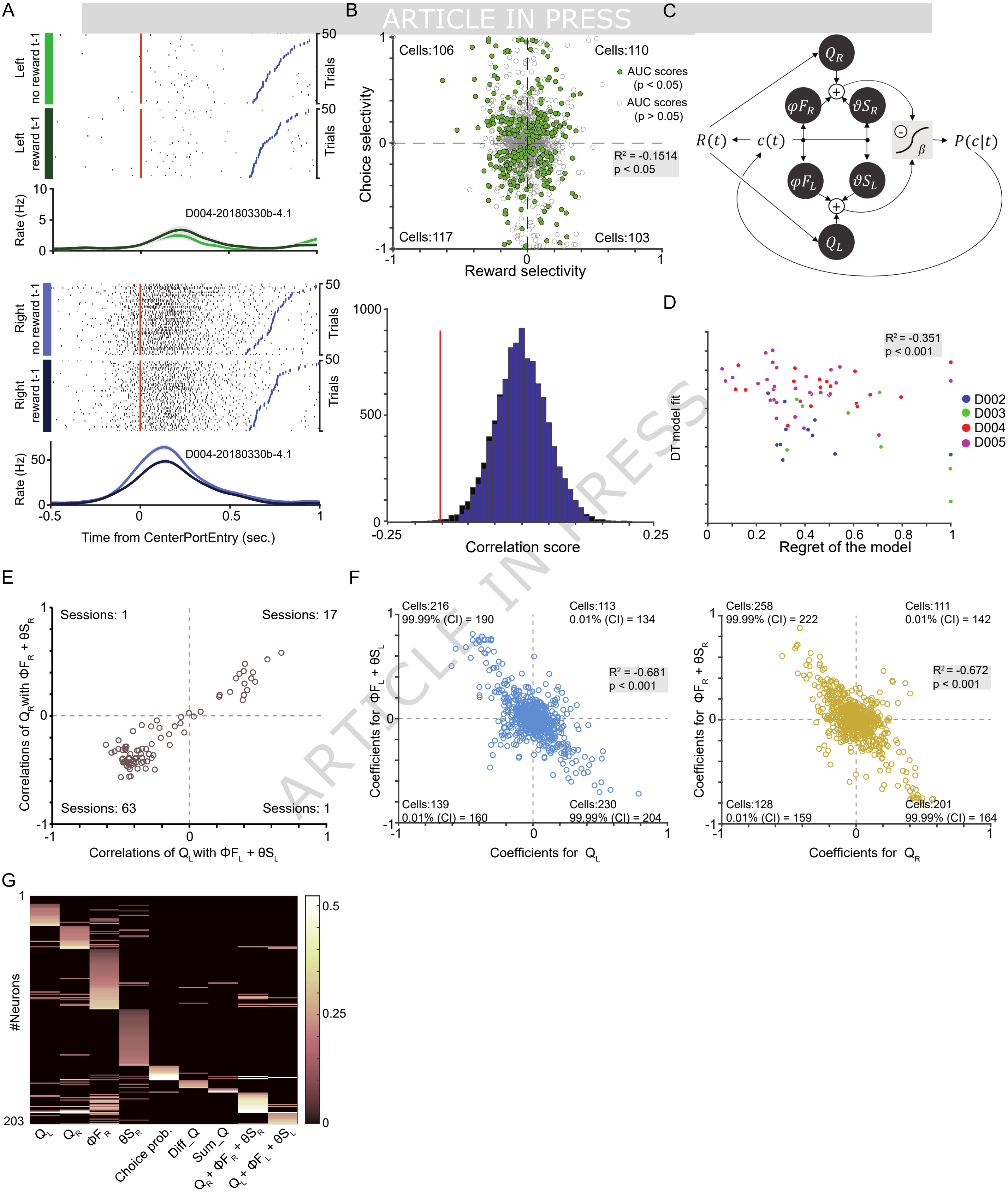
Figure 6. mPFC representations adjust to temporal delays in VI task. A) Schematic of the VI task incorporating short (0.2–0.5 s) and long (2–2.5 s) waiting periods (colored segment) in the center port. **B)** Past rewards and past choices up to 10 trials back regressed onto current choice (pooled across *n* = 4 mice). Logistic regression coefficients for past right/left rewards and choices are shown (mean ± s.e.m.) for short- and long-delay blocks. For each lag, coefficients were compared using two-sided linear mixed-effects models with delay (short vs long) as a fixed effect and animal identity as a random intercept. Significance was assessed using within-animal permutation tests (1,000 shuffles of delay labels), followed by Benjamini–Hochberg FDR correction across lags. Asterisks denote FDR-corrected *P* < 0.05 (exact *P* = 0.03). **C)** Correlations between value signals (Q) and choice-history effects ($\phi F + \theta S$) for left and right options in short- and long-delay VI sessions. Each point represents one session. **D)** Regret (computed as in Fig. 2D) for short- and long-delay blocks across sessions (*n* = 94). Error bars show mean ± s.e.m. Short vs long delay was compared using a two-sided Wilcoxon rank-sum test (*U* = 10729, *z* = 4.95, *P* = 7.53×10^{-7}). No multiple-comparison correction was applied. **E)** Absolute regression coefficients of all neurons in short- and long-delay periods (mean ± s.e.m.). For each regressor, delay effects were tested with two-sided LMMs (delay fixed; animal ID random). Significance was assessed using within-animal permutation tests (1,000 shuffles), and *P*-values were corrected across regressors using the Benjamini–Hochberg FDR procedure. Asterisks indicate FDR-corrected *P* < 0.05 or < 0.01. **F)** Contingency table of neuron counts in short-delay (rows) versus long-delay (columns) blocks, grouped by their strongest regression coefficient (past rewards, past choices, current choice, or untuned). For each category, short–long differences were tested using a two-sided permutation test (1,000 within-neuron label shuffles). *P*-values were taken directly from the permutation null (uncorrected). Stars denote permutation-derived significance (*P* < 0.05, 0.01, or 0.001). NAMCT was applied. Source data are provided as a Source Data file.

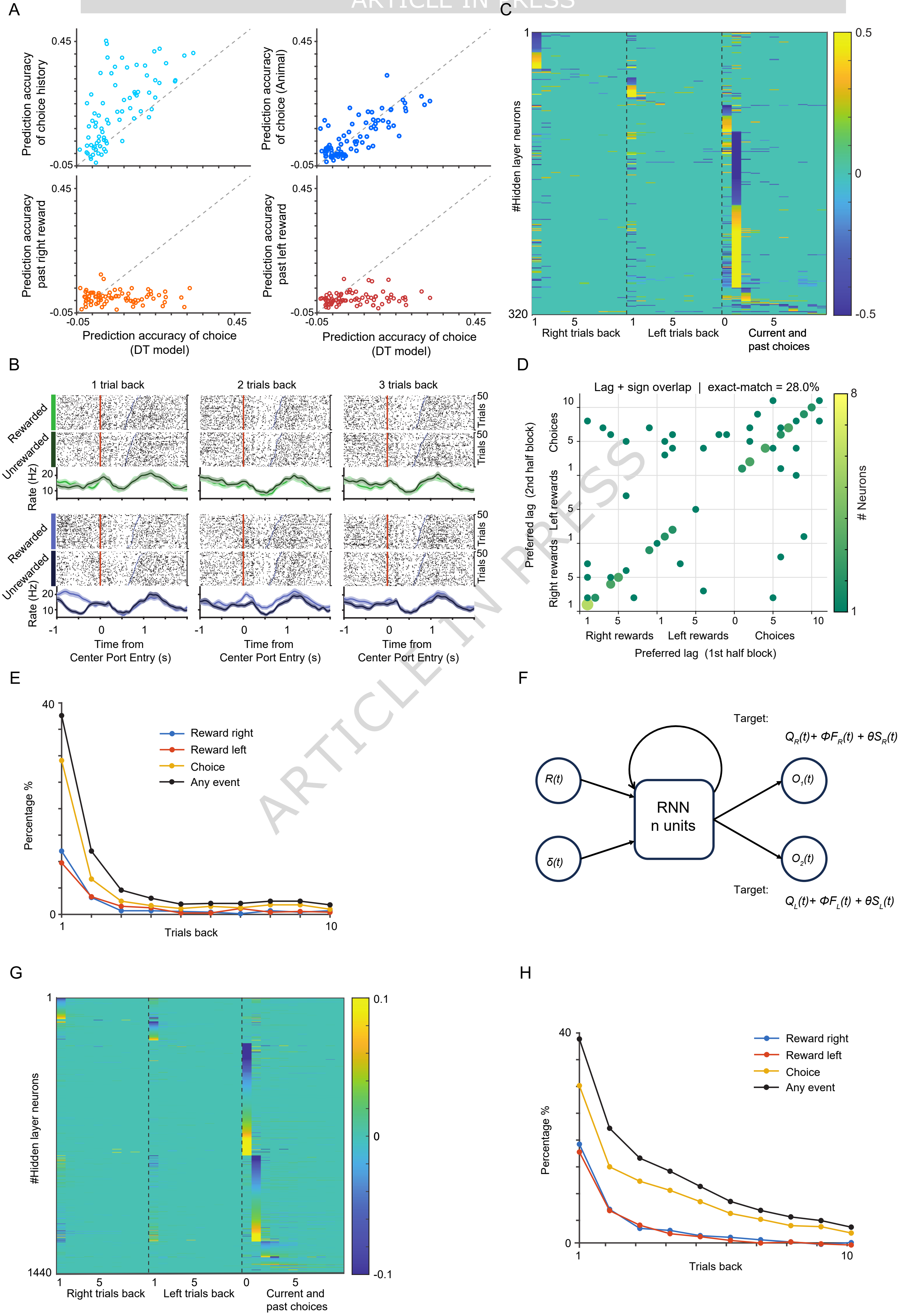
Editor's summary:

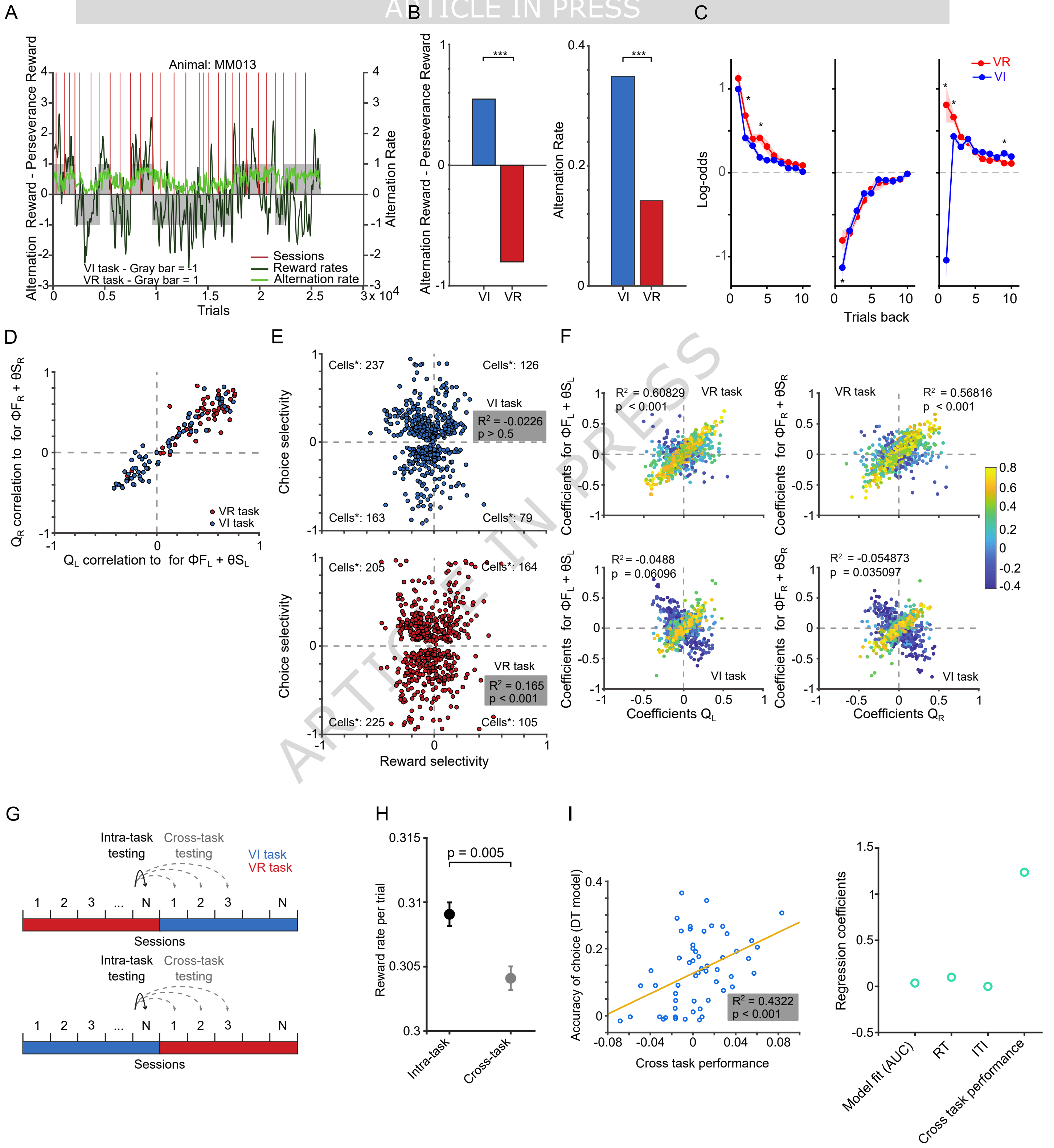
Neural basis of decision-making is not fully understood. Here authors show that mouse prefrontal neurons encode history-specific rewards and choices. However, their influence is gated by task structure and timing, affecting decisions primarily in variable interval tasks and when temporal delays separate events.

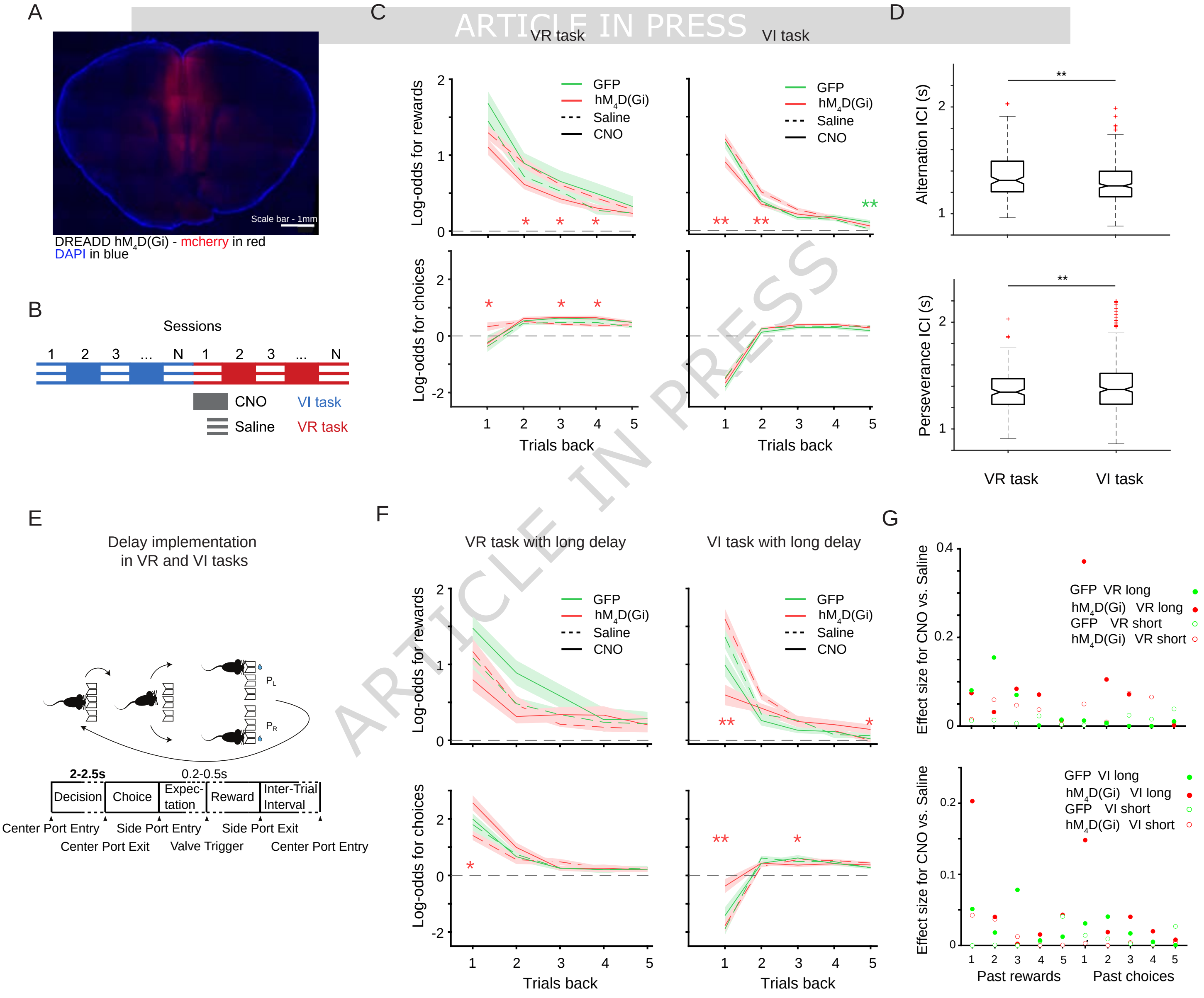
Peer review information: *Nature Communications* thanks Michael Halassa and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

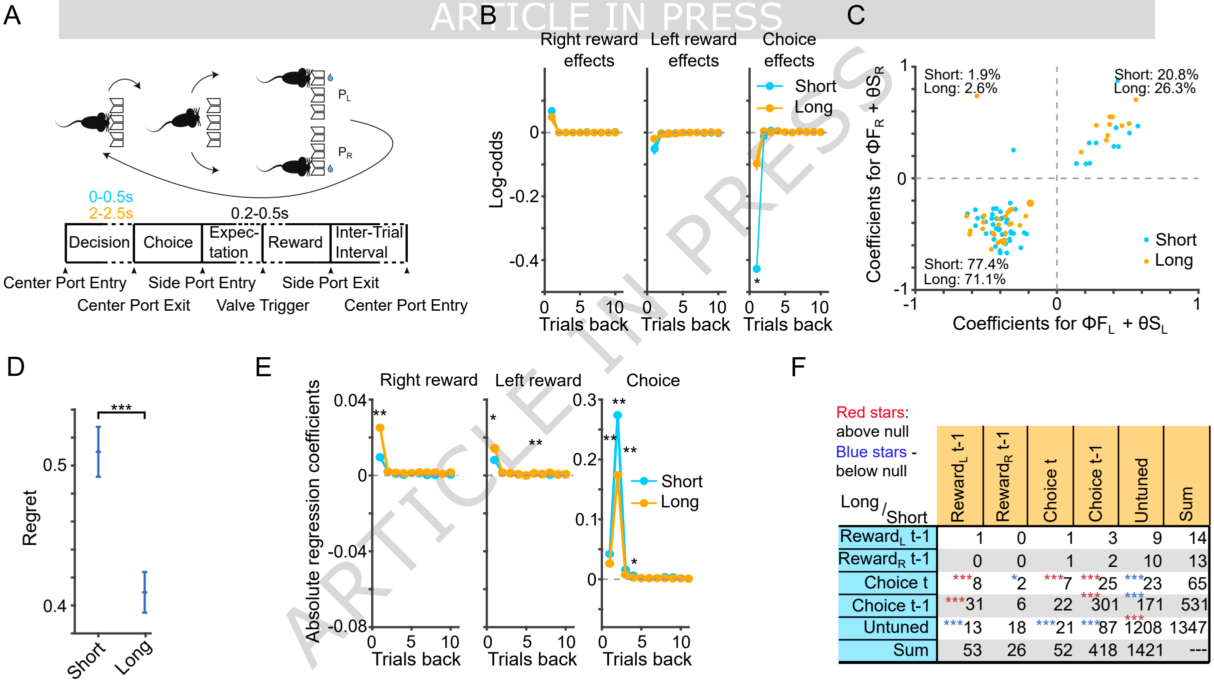

A
Decision epoch
Choice
Outcome
or
P_L
P_R
or
0-0.5s
0.2-0.5s
Decision
Choice
Expec-tation
Reward
Inter-Trial Interval
Center Port Entry
Center Port Exit
Side Port Entry
Valve Trigger
Side Port Exit
Center Port Entry
B
Choice Trace
Reward trace
P_Left - P_Right
Local Average
1
0
-1
0
400
800
1200
Trials
Reward
Choices
Local Average
0.5
0
-0.5
-30
0
30
Trials
-30
0
30
Trials
0.1 - 0.25
0.1 - 0.4
0.4 - 0.25
0.4 - 0.1
C
R^2 = 0.7903
p < 0.001
Choice ratio
1
0.5
0
0
0.5
1
Reward ratio
Data
Linear fit
D
Reward rate per trial
0.52
0.48
0.44
0.4
***
***
***
Random
Alternation
Rew_Prob
Animal_D004
Optimal Baiting
***
***
***
Random
Alternation
Rew_Prob
All animals
Optimal Baiting
E
Right reward effects
Left reward effects
Choice effects
Log odds
3
2
1
0
-1
-2
-3
2
4
6
8
Trials back
2
4
6
8
Trials back
2
4
6
8
Trials back
F
Alternation rate
0.75
0.5
**
**
Low
Medium
High
Diff. in set reward probabilities


A
Left no reward t-1
Left reward t-1
50
Trials
50
Trials
Rate (Hz)
10
5
0
D004-20180330b-4.1
Right no reward t-1
Right reward t-1
50
Trials
50
Trials
Rate (Hz)
50
0
D004-20180330b-4.1
-0.5
0
0.5
1
Time from CenterPortEntry (sec.)
B
Cells:106
Cells:110
Cells:117
Cells:103
AUC scores (p < 0.05)
AUC scores (p > 0.05)
R^2 = -0.1514
p < 0.05
Choice selectivity
Reward selectivity
1
0
-1
1000
500
0
-0.25
0.25
Correlation score
C
Q_R
φF_R
ϑS_R
φF_L
ϑS_L
Q_L
$R(t)$
$c(t)$
β
$P(c|t)$
D
R^2 = -0.351
p < 0.001
DT model fit
D002
D003
D004
D005
0
0.2
0.4
0.6
0.8
1
Regret of the model
E
Sessions: 1
Sessions: 17
Sessions: 63
Sessions: 1
Correlations of Q_R with $\Phi F_R + \theta S_R$
Correlations of Q_L with $\Phi F_L + \theta S_L$
F
Cells:216
99.99% (CI) = 190
Cells:113
0.01% (CI) = 134
Cells:139
0.01% (CI) = 160
Cells:230
99.99% (CI) = 204
R^2 = -0.681
p < 0.001
Coefficients for $\Phi F_L + \theta S_L$
Coefficients for Q_L
Cells:258
99.99% (CI) = 222
Cells:111
0.01% (CI) = 142
Cells:128
0.01% (CI) = 159
Cells:201
99.99% (CI) = 164
R^2 = -0.672
p < 0.001
Coefficients for $\Phi F_R + \theta S_R$
Coefficients for Q_R
G
1
203
#Neurons
0.5
0.25
0
Q_L
Q_R
ΦF_R
θS_R
Choice prob.
Diff_Q
Sum_Q
$Q_R + \Phi F_R + \theta S_R$
$Q_L + \Phi F_L + \theta S_L$


A
Prediction accuracy of choice history
Prediction accuracy of choice (Animal)
Prediction accuracy past right reward
Prediction accuracy past left reward
0.45
-0.05
Prediction accuracy of choice (DT model)
C
#Hidden layer neurons
1
320
Right trials back
Left trials back
Current and past choices
0.5
0
-0.5
B
1 trial back
2 trials back
3 trials back
Rewarded
Unrewarded
Rate (Hz)
Trials
50
20
10
Time from Center Port Entry (s)
-1
0
1
D
Lag + sign overlap | exact-match = 28.0%
Preferred lag (2nd half block)
Right rewards
Left rewards
Choices
Preferred lag (1st half block)
# Neurons
8
1
10
5
E
Percentage %
40
0
Trials back
1
10
Reward right
Reward left
Choice
Any event
F
R(t)
δ(t)
RNN n units
O1(t)
O2(t)
Target:
QR(t)+ ΦFR(t) + θSR(t)
QL(t)+ ΦFL(t) + θSL(t)
G
#Hidden layer neurons
1
1440
Right trials back
Left trials back
Current and past choices
0.1
0
-0.1
H
Percentage %
40
0
Trials back
1
10
Reward right
Reward left
Choice
Any event


A
Animal: MM013
Alternation Reward - Perseverance Reward
Alternation Rate
4
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3
x 10^4
Trials
Sessions
Reward rates
Alternation rate
VI task - Gray bar = -1
VR task - Gray bar = 1
B
Alternation Reward - Perseverance Reward
***
VI
VR
Alternation Rate
0.4
0.2
C
Log-odds
VR
VI
Trials back
*
5
10
D
Q_R correlation to for ΦF_R + θS_R
Q_L correlation to for ΦF_L + θS_L
VR task
VI task
E
Choice selectivity
Reward selectivity
Cells*: 237
Cells*: 126
Cells*: 163
Cells*: 79
VI task
R² = -0.0226
p > 0.5
Cells*: 205
Cells*: 164
Cells*: 225
Cells*: 105
VR task
R² = 0.165
p < 0.001
F
Coefficients for ΦF_L + θS_L
Coefficients for ΦF_R + θS_R
R² = 0.60829
p < 0.001
VR task
R² = 0.56816
p < 0.001
R² = -0.0488
p = 0.06096
VI task
R² = -0.054873
p = 0.035097
Coefficients Q_L
Coefficients Q_R
0.8
0.6
0.4
0.2
0
-0.2
-0.4
G
Intra-task testing
Cross-task testing
VI task
VR task
1 2 3 ... N 1 2 3 N
Sessions
H
p = 0.005
Reward rate per trial
0.315
0.31
0.305
0.3
Intra-task
Cross-task
I
Accuracy of choice (DT model)
Cross task performance
R² = 0.4322
p < 0.001
-0.08
-0.04
0
0.04
0.08
Regression coefficients
1.5
1
0.5
0
-0.5
Model fit (AUC)
RT
ITI
Cross task performance


A
Scale bar - 1mm
DREADD hM4D(Gi) - mcherry in red
DAPI in blue
B
Sessions
1 2 3 ... N 1 2 3 ... N
CNO
Saline
VI task
VR task
C
VR task
VI task
GFP
hM4D(Gi)
Saline
CNO
Log-odds for rewards
Log-odds for choices
*
**
Trials back
1 2 3 4 5
D
Alternation ICI (s)
Perseverance ICI (s)
**
VR task
VI task
E
Delay implementation in VR and VI tasks
PL
PR
2-2.5s
0.2-0.5s
Decision
Choice
Expec-tation
Reward
Inter-Trial Interval
Center Port Entry
Center Port Exit
Side Port Entry
Valve Trigger
Side Port Exit
Center Port Entry
F
VR task with long delay
VI task with long delay
GFP
hM4D(Gi)
Saline
CNO
Log-odds for rewards
Log-odds for choices
Trials back
G
GFP VR long
hM4D(Gi) VR long
GFP VR short
hM4D(Gi) VR short
GFP VI long
hM4D(Gi) VI long
GFP VI short
hM4D(Gi) VI short
Effect size for CNO vs. Saline
Past rewards
Past choices



Red stars: above null
Blue stars - below null

Long / Short	$Reward_L$ t-1	$Reward_R$ t-1	Choice t	Choice t-1	Untuned	Sum
$Reward_L$ t-1	1	0	1	3	9	14
$Reward_R$ t-1	0	0	1	2	10	13
Choice t	***8	*2	***7	***25	***23	65
Choice t-1	***31	6	22	***301	***171	531
Untuned	***13	18	***21	***87	***1208	1347
Sum	53	26	52	418	1421	---