Neural dynamics of reversal learning in the prefrontal cortex and recurrent neural networks

Christopher M. Kim[1,*], Carson C. Chow[1], and Bruno B. Averbeck[2]

[1]Laboratory of Biological Modeling, NIDDK/NIH, Bethesda, MD
[2]Laboratory of Neuropsychology, NIMH/NIH, Bethesda, MD
[*] Current address: Department of Mathematics, Howard University, Washington, DC

Abstract

In probabilistic reversal learning, the choice option yielding reward at higher probability switches at a random trial. To perform optimally in this task, one has to accumulate evidence across trials to infer the probability that a reversal has occurred. In this study, we investigated how this reversal probability is represented in cortical neurons by analyzing the neural activity in prefrontal cortex of monkeys and recurrent neural networks trained on the task. We found that neural trajectories encoding reversal probability had substantial dynamics associated with intervening behaviors necessary to perform the task. Furthermore, the neural trajectories were translated systematically in response to whether outcomes were rewarded, and their position in the neural subspace captured information about reward outcomes. These findings suggested that separable dynamic trajectories, instead of fixed points on a line attractor, provided a better description of neural representation of reversal probability. Near the behavioral reversal, in particular, the trajectories shifted monotonically across trials with stable ordering, representing varying estimates of reversal probability around the reversal point. Perturbing the neural trajectory of trained networks biased when the reversal trial occurred, showing the role of reversal probability activity in decision-making. In sum, our study shows that cortical neurons encode reversal probability in a family of dynamic neural trajectories that accommodate flexible behavior while maintaining separability to represent distinct probabilistic values.

Introduction

To survive in a dynamically changing world, animals must interact with the environment and learn from their experience to adjust their behavior. Reversal learning has been used for assessing ability to adapt one's behavior in such environment [1–6]. For instance, in the two-armed bandit tasks with probabilistic reward, the subject learns from initial trials that one option has higher reward probability than the other. When the reward probability of two options is reversed at a random trial, the subject must learn to reverse its preferred choice to maximize reward outcome. In these tasks, there is uncertainty in when to reverse one's choice, as reward is received stochastically even when the less favorable option is chosen. Therefore, it is essential that reward outcomes are integrated over multiple trials before the initial choice preference is reversed. Although neural mechanisms for accumulating evidence within a trial have been studied extensively [7–10], it remains unclear if a recurrent neural circuit uses a similar neural mechanism for accumulating evidence across multiple trials, while performing intervening behavior during each trial.

In this study, we merged two classes of computational models, i.e., behavioral and neural, to investigate the neural basis of multi-trial evidence accumulation. The behavior models capture subject's behavioral strategies for performing the reversal learning task. For instance, Model-free reinforcement learning (RL) [11–13] assumes that the subject learns only from choices and reward outcomes without specific knowledge about task structure. Model-based Bayesian inference [2, 14, 15], in contrast, assumes that the task structure is known to the subject, and one can infer reversal points statistically, resulting in abrupt switches in their choice preference. Model-based and model-free RL models are formal models that do not specify implementation in a network of neurons. On the other hand, neural models implemented with recurrent neural networks (RNNs) can be trained to use recurrent activity to perform the reversal learning task. In particular, attractor dynamics, in which the network state moves towards discrete [9, 16] or along continuous [8, 17] attractor states, have been studied extensively as a potential neural mechanism for decision-making and evidence accumulation [18, 19].

Here, we trained RNNs that learned from a Bayesian inference model to mimic the behavioral strategies of monkeys performing the reversal learning task [2, 4]. We found that, in the prefrontal cortex of monkeys and in trained RNNs, neural activity during a baseline hold period encoded reversal probability in a one-dimensional subspace, similar to a line attractor. However, intervening behavior during a trial, including making decisions and receiving feedback, produced substantial non-stationary neural dynamics. This observation made the attractor dynamics, which require the network state to stay close to attractor states [8, 9, 16, 17], ill-suited for explaining the neural activity associated with evidence accumulation in reversal learning.

Instead, we found that reversal probability was encoded in dynamic neural trajectories that shifted systematically across trials. Reward outcome pushed the entire trajectory in a positive (without reward) or negative (with reward) direction, separating trajectories of adjacent trials. Moreover, integrating reward outcomes across trials captured the position of a trajectory. These results suggested a neural mechanism where separable dynamic trajectories encode accumulated evidence. Around the behavioral reversal trial, reversal probabilities were represented by a family of rank-ordered trajectories that shifted monotonically. Perturbation experiment in trained RNNs demonstrated a causal link between reversal probability activity and choice outcomes.

In sum, our results show that, in a probabilistic reversal learning task that requires evidence integration across trials and execution of intervening behavior in-between trials, reversal probability is encoded in separable dynamic trajectories that allow for temporally flexible representation of accumulated evidence.

Results

1 Trained RNN's choices are consistent with monkey behavior

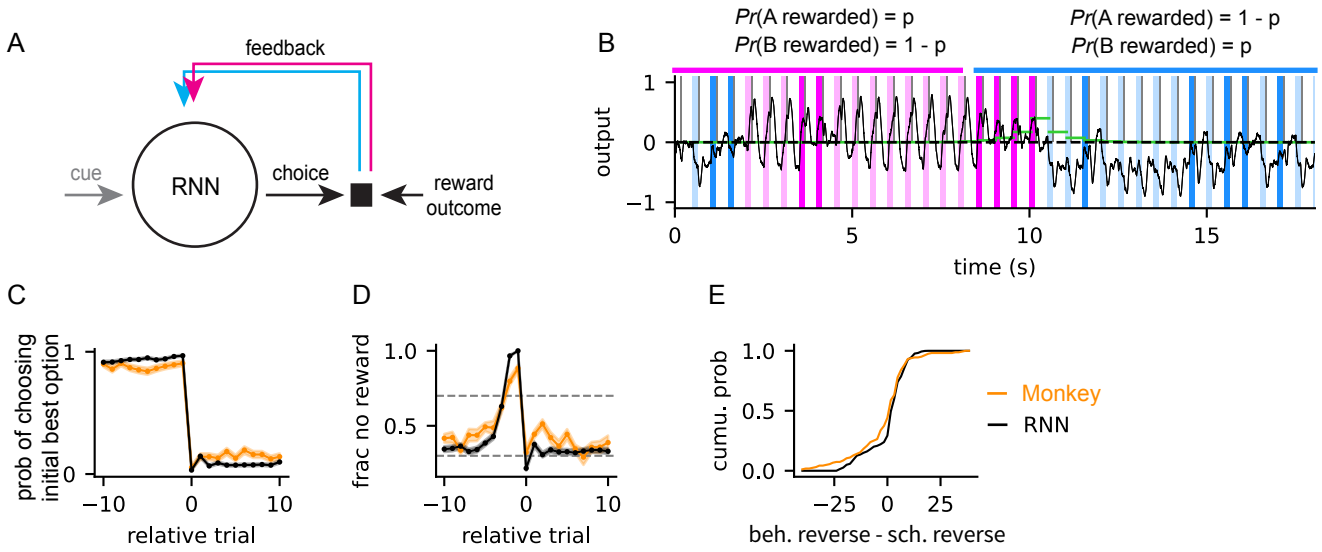

Figure 1: Comparison of the behavior of trained RNNs and monkeys. **(A)** Schematic of RNN training setup. In a trial, network makes a choice in response to a cue. Then, a feedback input, determined by the choice and reward outcome, is injected to the network. This procedure is repeated across trials. **(B)** Example of a trained RNN's choice outcomes. Vertical bars show RNN choices in each trial and the reward outcomes (magenta: choice A, blue: choice B, light: rewarded, dark: not rewarded). Horizontal bars on the top show reward schedules (magenta: choice A receiving reward is 70%, choice B receiving reward is 30%; blue: reward schedule is reversed). Black curve shows the RNN output. Green horizontal bars show the posterior of reversal probability at each trial inferred using Bayesian model. **(C)** Probability of choosing the initial best option. Relative trial indicates the trial number relative to the behavioral reversal trial inferred from the Bayesian model. Relative trial number 0 is the trial at which the choice was reversed. **(D)** Fraction of no-reward blocks as a function of relative trial. Dotted lines show 0.3 and 0.7. **(E)** Distribution of RNN's and monkey's reversal trial, relative to the experimentally scheduled reversal trial.

In the reversal learning task, in each trial, two options were available. The subject (either the monkeys or the network) chose one of the options. Rewards were delivered stochastically. The initial high-value option was rewarded 70% of the time when chosen, and the initial low value option was rewarded 30% of the time when chosen. The task was executed in blocks of trials. On a randomly chosen trial, the reward probability of the two options was switched. Because reward delivery was stochastic, the agent had to infer the reversal by accumulating evidence that a reversal had occurred. In this study, we will focus on this reversal inference process.

We began by training an RNN on the reversal learning task and comparing the performance of the network to the monkeys. This allowed us to study the solutions adopted by the network, to generate hypotheses that we could test in neural data. Therefore, we trained an RNN to choose from two options in each trial when triggered by a cue. Following the choice, feedback was provided to the network, signaling the choice it made and reward outcome (Fig.1A). The reward schedule was probabilistic and identical to the task monkeys performed. This reward schedule was reversed at a random trial, and the RNN learned to reverse its decision by mimicking the outputs of a Bayesian inference model that

captures the monkey's reversal behavior (See Methods Section 2.3 for the RNN training scheme). In a typical block consisting of 36 trials, a trained RNN selected the initial high reward options, despite receiving occasional no rewards, but abruptly switched its choice when consecutive no-reward trials persisted (Fig.1B).

The reversal behavior of trained RNNs was similar to the monkey's behavior on the same task. RNNs selected the high reward option with high probability before the behavioral reversal, at which time they abruptly switched their choice (Fig.1C). The behavioral reversal was preceded by a gradually increasing number of no-reward trials (Fig.1D). The distribution of behavioral reversal trials (i.e., trial at which preferred choice was reversed) relative to the scheduled reversal trial (i.e., trial at which reward schedule was reversed) was similar to the distribution of monkey's reversal trials (Fig.1E).

2 Task-relevant neural activity evolves dynamically

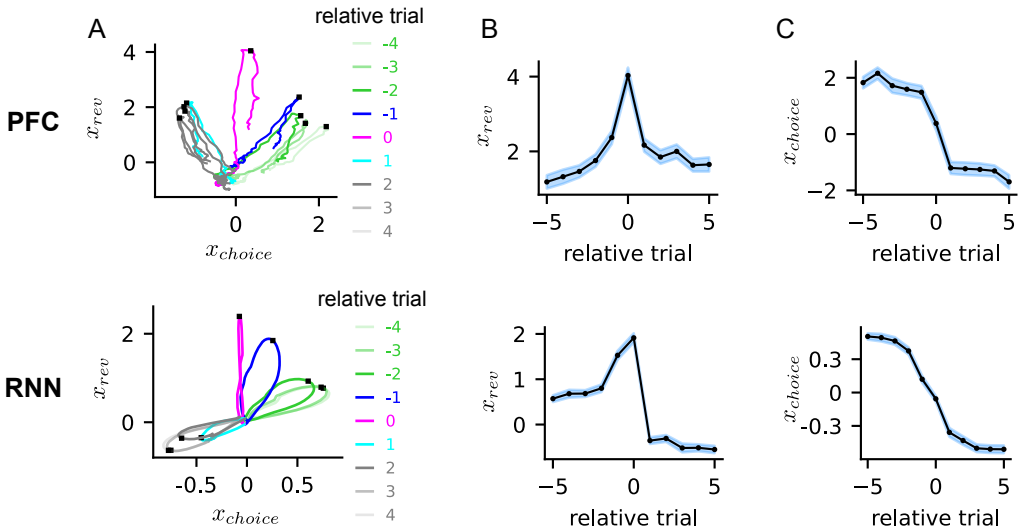

Figure 2: Neural trajectories encoding choice and reversal probability variables. **(A)** Neural trajectories of PFC (top) and RNN (bottom) obtained by projecting population activity onto task vectors encoding choice and reversal probability. Trial numbers indicate their relative position to the behavioral reversal trial. Neural trajectories in each trial were averaged over 8 experiment sessions and 23 blocks for the PFC, and 40 networks and 20 blocks for the RNNs. Black square indicates the time of cue onset. **(B-C)** Neural activity encoding reversal probability and choice in PFC (top) and RNN (bottom) at the time of cue onset (black squares in panel A) around the behavioral reversal trial. Shaded blue shows the standard error of mean over sessions (or networks) and blocks.

Next we examined the temporal dynamics of task-relevant neural activity, in particular activity encoding the choice and reversal probability. This analysis focused on trials around the reversal point in each block. To capture task-relevant neural activity, we first identified population vectors that encoded the task variables using a method called targeted dimensionality reduction [20]. It regresses the activity of individual neurons onto task variables and finds the maximal population vector of each task variable. Then, neural activity representing the task variable is obtained by projecting the population activity onto the identified task vectors (see Methods Section 3 for details).

When averaged over blocks, the neural activity associated with choices and inferred reversal probability, denoted as x_{choice} and x_{rev}, respectively, produced non-stationary dynamics in each trial (Fig.2A). Their activity level reached a maximum around the time of cue onset (black squares in Fig.2A), when the monkey and RNN were about to make a choice. The rotational neural dynamics were found both in the prefrontal cortex (PFC) of monkeys and trained RNNs.

The orientation of rotational trajectories shifted as trials progressed, indicating systematic changes in the choice and reversal probability activity across trials. When the task-relevant activity at cue onset was analyzed, we found that reversal probability activity, x_{rev}, peaked at the reversal trial in the PFC and RNN (Fig.2B). On the other hand, choice activity, x_{choice}, decreased gradually over trials reflecting the changes in choice preference (Fig.2C). The inverted-V shape of x_{rev} and the monotonic decrease in x_{choice} over trials explained the counter-clockwise shift in the rotational trajectories observed in the two-dimensional phase space (Fig.2A).

3 Integration of reward outcomes drives reversal probability activity

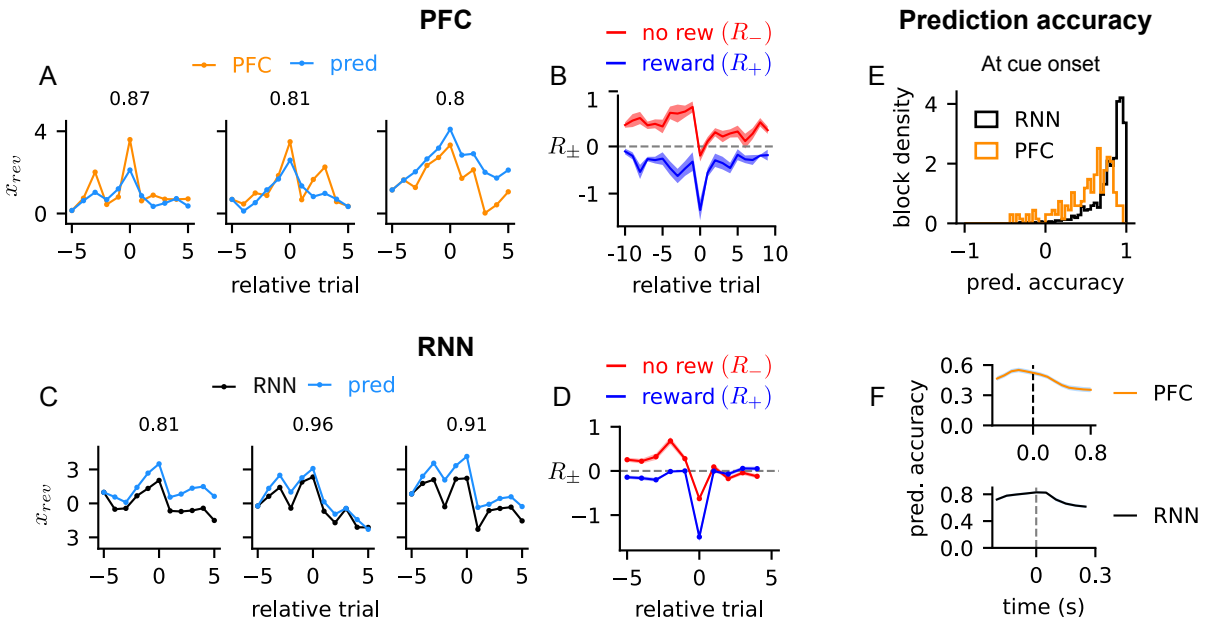

Figure 3: Integration of reward outcomes drives reversal probability activity. **(A)** The reversal probability activity of PFC (orange) and prediction by the reward integration equation (blue) at the time of cue onset across trials around the behavioral reversal trial. Three example blocks are shown. Pearson correlation between the actual and predicted PFC activity is shown on each panel. **(B)** $R^k_\pm$ of PFC estimated from the reward integration equation at cue onset. R^k_- and R^k_+ correspond to no-reward (red) and reward trials (blue), respectively. **(C-D)** Same as in panels (A) and (B) but for trained RNNs. **(E)** Prediction accuracy of the reward integration equation of all 8 PFC recording sessions and all 40 trained RNNs at cue onset. **(F)** Average prediction accuracy of the reward integration equation across time. The value at each time point shows the prediction accuracy averaged over all blocks in PFC recording sessions (top) or trained RNNs (bottom).

We asked if the changes in reversal probability activity x_{rev} across trials, as shown in Fig.2B, can be explained by integrating reward outcomes. In particular, we wondered if the reward outcomes from each trial would drive the shifts in reversal probability activity. To investigate this question, we set up a reward integration equation $x^{k+1}_{rev}(t_{onset}) = x^k_{rev}(t_{onset}) + R^k_\pm(t_{onset})$ that predicts the next trial's reversal probability activity based on current trial's reversal probability and reward outcome, therefore predicting across-trial reversal probability by integrating reward outcomes. Here, $x^k_{rev}(t_{onset})$ is the reversal probability activity at the time of cue onset t_{onset} at trial k, and $R^k_\pm(t_{onset})$ is an estimate of the shift in reversal probability activity driven by trial k's reward outcome ($+$ if rewarded and $-$ if not rewarded. See Methods Section 4 for details).

The predicted reversal probability activity was in good agreement with the actual activity of PFC and RNN (example blocks shown in Figs.3A,C; prediction accuracy of all blocks shown in Fig.3E). Moreover, we found that $x^k_{rev}(t_{onset})$, the neural activity encoding reversal probability, responded to reward outcomes consistently with how reversal probability itself would be updated. In other words, receiving no-reward at trial k increased the reversal probability activity x^{k+1}_{rev} in the next trial $k+1$ (Figs.3B, D; no reward), while receiving a reward at trial k decreased x^{k+1}_{rev} (Figs.3B, D; reward). At the behavioral reversal trial ($k = 0$), however, the reversal probability activity in the following trial ($k = 1$) decreased regardless of the reward outcome at the reversal trial. When the reward integration equation

was fitted to the reversal probability activity at other time points (i.e., $R^k_{\pm}(t)$ was estimated at each t), the prediction accuracy remained stable in time (Fig.3F).

These findings show that neural activity encoding reversal probability exhibits structured responses to reward outcomes, consistent with how reversal probability itself would respond: increase with no reward and decrease with reward. In addition, the reversal probability activity can be predicted by integrating reward outcomes, supporting that it encodes accumulation of decision-related evidence.

4 Dynamic neural trajectories encoding reversal probability are separable

Previous works have shown that accumulation of decision-related evidence can be represented as a line attractor in a stable subspace of network activity [20, 21]. One might hypothesize that reversal probability in the reversal learning task could be similarly characterized by such line attractor dynamics. A direct application of the line attractor model would imply that, across a trial when no decision-related evidence is presented, the reversal probability activity should remain constant (Fig.4A, line attractor).

However, we found that there was substantial activity in this neural subspace (Fig.2A). In particular, the non-stationary neural activity was associated with intervening behaviors during a trial. The time derivative of reversal probability activity increased rapidly at the time of cue onset, when a decision is made, followed by sharp decrease until the time of reward (dx_{rev}/dt in Fig.4B). So, instead of a static view of evidence accumulation, we explored the hypothesis that different levels of reversal probability could be encoded in dynamic neural trajectories (Fig.4A, dynamic trajectory).

To encode distinct values of reversal probability in dynamic trajectories, the trajectories representing the values must remain separated as they evolve in time. We compared trajectories at adjacent trials to examine if the reward outcome drives the next trial's trajectory away from the current trial's trajectory, thus separating them, and, if so, to what extent the trajectories are separated.

Analysis of PFC activity showed that, not receiving a reward increased the next trial's trajectory $x_{rev}^{k+1}(t)$, compared to the current trial's trajectory $x_{rev}^{k}(t)$. Within a trial, this positive shift was observed over the entire trial duration until the next trial's reward was revealed, as shown in the difference of adjacent trials' trajectories (Fig.4C, R_-). Moreover, across trials, the same trend was observed in all the trials except at the behavioral reversal trial, at which the reversal probability activity reached its maximum value and decreased in the following trial (Fig.4D, R_-). On the other hand, when a reward was received, the next trial's trajectory was decreased compared to the current trial's trajectory. This negative shift persisted until the next trial's reward, similarly to the case when reward was received (Fig.4C, R_+). Across trials, the same trend was observed in all the trials except at the trial preceding the behavioral reversal trial, at which the trajectory increased to the maximum value at the reversal trial (Fig.4D, R_+). Additional analysis on R_- and R_+ beyond the next trial's reward time can be found in Supp. Figure S1.

We examined what type of activity mode the dynamic trajectories exhibited when separating away from the previous trial's trajectory. Ramping activity is often observed in cortical neurons of animals engaged in decision-making [22–26]. We found that, when no rewards were received, trajectories were separated from the previous trial's trajectory by increasing their ramping rates towards the decision time ($dR_-/dt > 0$ in Fig.4E). On the other hand, when rewards were received, trajectories were separated by decreasing their ramping rate ($dR_+/dt < 0$ in Fig.4E). The increase (or decrease) in the ramping rates was observed in consecutive no reward (or reward) trials around the reversal trial (Fig.4E, left).

Consistently with the PFC activity, the trained RNN exhibited similar activity responses to reward outcomes: neural trajectories encoding reversal probability increased when reward was not received and decreased when reward was received. The shift in trajectories persisted throughout the trial duration (Fig.4G) and ramping rates changed in agreement with the PFC findings (Fig.4H).

Since the dynamics of trained RNNs are fully known, we sought to examine the circuit dynamic motif that separates neural trajectories. We projected the differential equation governing the network dynamics onto a one-dimensional subspace and analyzed the contribution of recurrent and external inputs to reversal probability dynamics: $\tau \dot{x}_{rev} = x_{rec} + x_{ext} \equiv x_{total}$ (see Methods Section 1 for details). We found that the external input x_{ext} was positive, while the recurrent input x_{rec} was negative and curtailed the external input (Fig.4F, external and recurrent). When no reward was received, x_{ext} and x_{rec} were both amplified by approximately the same factor and resulted in increased total input: $x_{total}^{no\ rew} = \gamma^{no\ rew} x_{total}$

with $\gamma^{no\ rew} > 1$ (Fig.4F, amplification). On the other hand, when reward was received, they were both suppressed, resulting in decreased total input: $x_{total}^{reward} = \gamma^{reward} x_{total}$ with $\gamma^{reward} < 1$. This suggested a circuit dynamic motif, where external feedback balanced by recurrent inhibition drives the reversal probability dynamics. The total drive is amplified or suppressed, depending on reward outcomes, resulting in a trajectory that separates from the previous trial's trajectory.

In sum, our findings show that dynamic neural trajectories encoding reversal probability are separated from the previous trial's trajectory in response to reward outcomes, allowing them to represent distinct values of reversal probability across a trial.

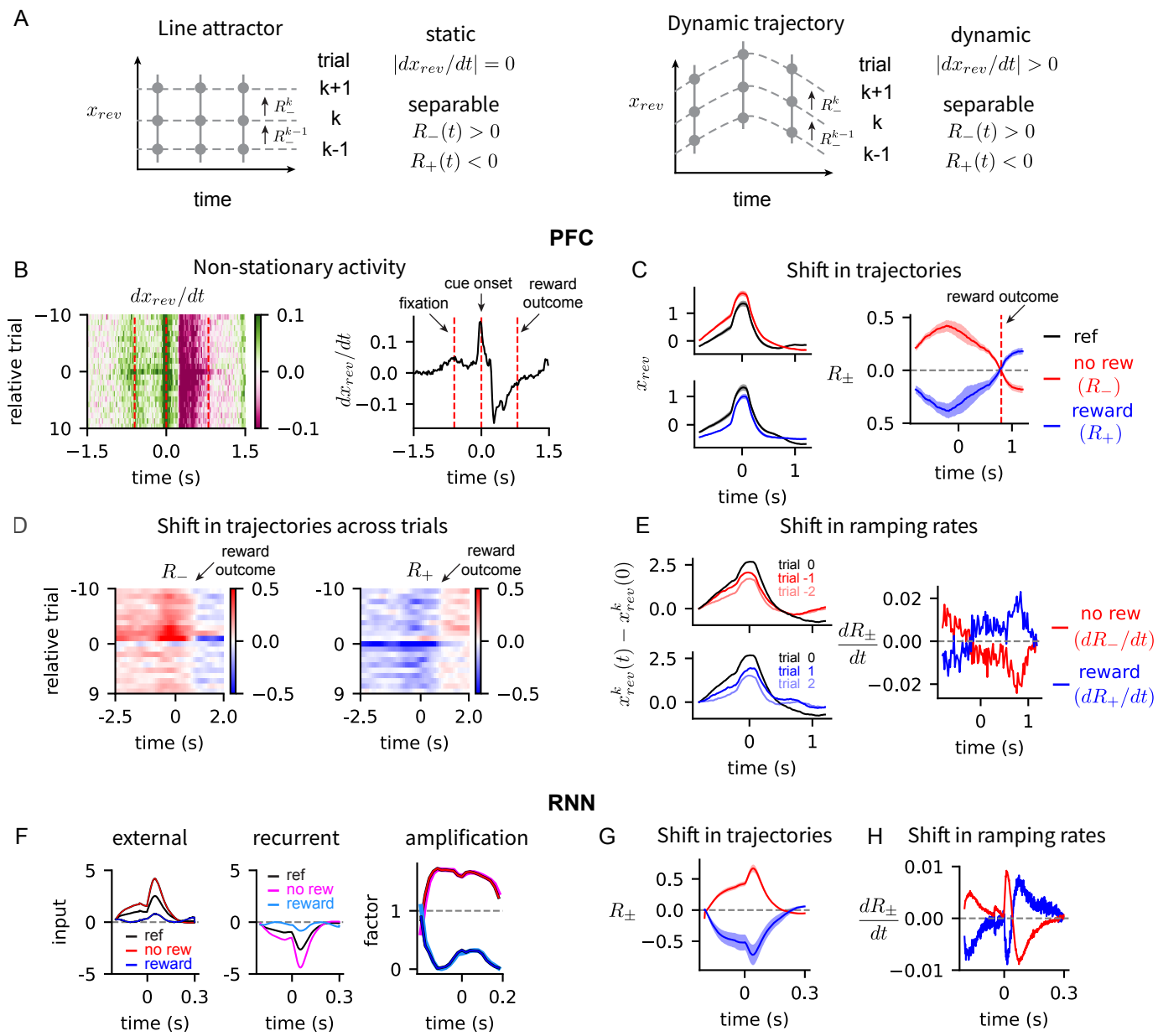

Figure 4: Dynamic neural trajectories encoding reversal probability are separated in response to reward outcomes. **(A)** Two neural models for the reversal probability dynamics. Left: Line attractor model where $x_{rev}(t)$ remains constant during a trial. Right: Dynamic trajectory model where $x_{rev}(t)$ is non-stationary. In both models, the trajectories of adjacent trials are separable if the shift due to reward R_+^k (or no reward R_-^k) is negative (or positive) throughout the trial. **(B)** Left: Block-averaged x_{rev}/dt of PFC across trial and time. Dotted red lines indicate the onset time of fixation (-0.5s), cue (0s) and reward (0.8s); same lines shown on the right. Right: Average of x_{rev}/dt over trials shown on the left panel. **(C)** Left: $x_{rev}(t)$ of PFC at current trial (black) is compared to $x_{rev}(t)$ in the next trial when reward is received (red) and not received (blue). Right: The difference of $x_{rev}(t)$ between current and next trials shown on the left panels. **(D)** Difference of x_{rev} of two adjacent trials when reward is not received (R_-) or received (R_+). The approximate time of reward outcome is shown. **(E)** Left: $x_{rev}(t)$ of PFC of consecutive no reward trials before the behavioral reversal trial (top) and consecutive reward trials after the behavioral reversal (bottom). The initial value was subtracted to compare the ramping rates of $x_{rev}(t)$. Right: Difference in the ramping rates of trajectories of adjacent trials, when reward was received (blue) and not received (red). **(F)** External (left) and recurrent (middle) inputs to the reversal probability dynamics, when reward was not received (red) and received (blue). Amplification factor (right) shows the ratio of no reward (or reward) input to the reference input. The amplification factors for both the external and recurrent inputs are shown. **(G-H)** Same as the right panels in (C) and (E) but for trained RNNs.

5 Monotonic shift of reversal probability trajectories across trials

So far, we showed that neural trajectories of two adjacent trials, encoding reversal probability, were separable. In this section, we investigated if trajectories exhibited systematic changes across multiple trials. Specifically, we quantified the mean behavior of trajectories in each trial (referred to as mean trajectory of a trial) and looked for consistent trends in the changes of mean trajectories across trials.

Since a mean trajectory was obtained by averaging over all reward outcomes, we compared how reward and no-reward blocks contributed to modifying the next trial's mean trajectory. This analysis amounted to comparing R_-^k and R_+^k shown in Fig.4D. Since $R_-^k > 0$ and $R_+^k < 0$ throughout a trial, we flipped the sign of R_+^k to $-R_+^k$ and compared the magnitudes of two positive traces, R_-^k and $-R_+^k$. We found that, before the behavioral reversal trial ($k < 0$), the contribution of no-reward R_-^k was larger than reward $-R_+^k$. This is shown as R_-^k lying above $-R_+^k$ during a trial and across pre-reversal trials (see relative trials $k = -5$ to -1 in Fig.5A). Their temporal averages over trial duration also captured this finding: $\langle R_-^k \rangle_t$, was larger than $\langle -R_+^k \rangle_t$ before the behavioral reversal trial (Fig.5B, bottom). This analysis showed that the sum $R_-^k + R_+^k$, i.e., the difference of mean trajectories between two adjacent trials, stayed positive during the pre-reversal trials (Fig.5B, top). We confirmed that the fraction of trials, for which $R_-^k + R_+^k$ is positive, was close to 0.8 in the pre-reversal phase (Fig.5C, top).

The sum $R_-^k + R_+^k$ being positive meant that the next trial's mean trajectory was increased relative to the current trial's mean trajectory. Furthermore, the sum $R_-^k + R_+^k$ of pre-reversal trials being positive was equivalent to the mean trajectories increasing monotonically across trials towards the behavioral reversal trial (Fig.5C, bottom). The monotonicity of trajectories implied that a topological structure was present in pre-reversal trajectories; namely, the rank order of the trajectories was preserved throughout the trial duration. Consistent with this observation, we found that the Spearman rank correlation of pre-reversal trajectories was stable in time (Fig.5E, pre).

After the reversal trial ($k \geq 0$), on the other hand, the contributions of no reward R_-^k and reward R_+^k were the opposite of pre-reversal trials. The traces of $-R_+^k$ were positioned above R_-^k (see relative trials $k = 0$ to 4 in Fig.5A), and the temporal average $\langle -R_+^k \rangle_t$ was larger than $\langle R_-^k \rangle_t$ (Fig.5B, bottom). This showed that $R_-^k + R_+^k$ was mostly negative during post-reversal trials. The fraction of trials, for which $R_-^k + R_+^k$ is negative, was close to 0.8 in the post-reversal phase (Fig.5D, top). The negativity of $R_-^k + R_+^k$ across post-reversal trials meant that the post-reversal trajectories were monotonically decreasing (Fig.5D, bottom). Similarly to the pre-reversal trajectories but in reversed order, the rank order of post-reversal trajectories was stable over the trial duration (Fig.5E, post).

Consistently with the PFC findings, in the trained RNNs, the effects of reward outcomes on mean trajectories were characterized by $R_-^k + R_+^k$ being positive and negative before and after the reverse trial, respectively (Fig.5F). Consequently, trained RNNs exhibited monotonic increase and decrease in the pre- and post-reversal phases, respectively (Figs. 5G, H). Also, the rank order of trajectories was stable over the trial duration (Fig.5I)).

Our analyses show that the mean behavior of dynamic neural trajectories, encoding reversal probability, is to shift monotonically across trials near the behavioral reversal. It suggests that a family of graded neural trajectories, with a temporally stable rank order, could represent varying estimates of the probability that a reversal has occurred.

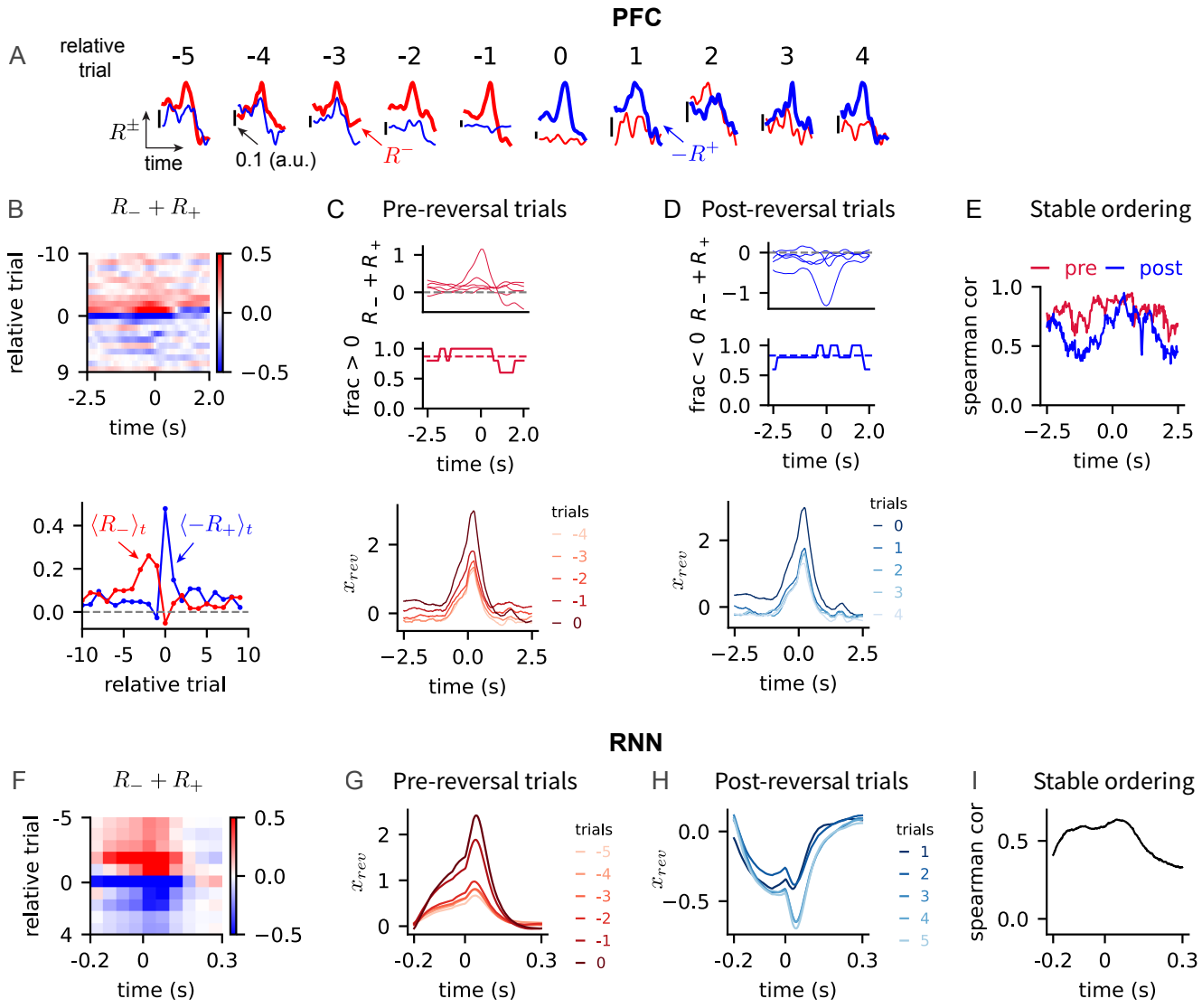

Figure 5: Mean trajectories encoding reversal probability shift monotonically across trials. **(A)** Traces of R_-^k and $-R_+^k$ around the behavioral reversal trial. Note the sign flip in $-R_+^k$, which was introduced to compare the magnitudes of R_-^k and R_+^k. **(B)** Top: $R_-^k + R_+^k$ across trial and time. Bottom: Temporal averages of R_-^k and $-R_+^k$ over the trial duration. **(C)** Top: Traces of $R_-^k + R_+^k$ of pre-reversal trials (relative trial $k = -5$ to -1), and the fraction of trials at each time point that satisfy $R_-^k + R_+^k > 0$. Bottom: Mean PFC reversal probability trajectories of pre-reversal trials. **(D)** Same as in panel (C), but for post-reversal trials (relative trial $k = 0$ to 4). **(E)** Spearman rank correlation between trial numbers and the mean PFC reversal probability trajectories across pre-reversal (red) and post-reversal (blue) trials at each time point. For the post-reversal trials, Spearman rank correlation was calculated with the trial numbers in reversed order to capture the descending order. **(F)** $R_-^k + R_+^k$ of trained RNNs across trial and time. **(G-I)** Trained RNNs' block-averaged x_{rev} before and after the reversal trial and their average Spearman correlation at each time point.

6 Perturbing neural activity encoding reversal probability biases choice outcomes

Next we turned to the RNN to see if we could perturb activity within the reversal probability space, and consequently perturb the network's choice preference. Previous experimental works demonstrated that perturbing neural activity of medial frontal cortex [27], specific cell types [28, 29] or neuromodulators [5, 6] affect the performance of reversal learning. In our study, the perturbation was tailored to be within the reversal probability space by applying an external stimulus aligned (v_+) or opposite (v_-) to the reversal probability vector. An external stimulus in a random direction was also applied as a control (v_{rnd}). All the stimuli were applied before the time of choice at the reversal trial or at preceding trials (Fig.6A).

We found that the deviation of perturbed reversal probability activity from the unperturbed activity peaked at the end of perturbation duration and decayed gradually (Fig.6B, red solid). The perturbed choice activity, however, deviated more slowly and peaked during the choice duration (Fig.6B, black solid). This showed that perturbation of the reversal probability activity had its maximal effect on the choice activity when the choice was made. The strong perturbative effects on the reversal probability and choice activity were not observed in the control (Fig.6B, dotted).

The perturbation in the aligned (v_+) and opposite (v_-) directions shifted the reversal probability activity along the same directions as the perturbation vector, as expected (Fig.6C, left). The choice activity, on the other hand, increased when the perturbation was in the opposite direction (v_-) and decreased when the perturbation was in the aligned direction (v_+) (Fig.6C, right). This finding showed that the choice activity could be biased (1) towards pre-reversal choices if the perturbation decreases the reversal probability activity and (2) towards the post-reversal choices if the perturbation increases the reversal probability activity.

We further analyzed if perturbing within the reversal probability space could affect the choice outcomes, specifically the behavioral reversal trial. We found that the reversal trial was delayed when v_- stimulus was applied to reduce the reversal probability activity (Fig.6D, left). The effect of v_- stimulus increased gradually with the stimulus strength and was significantly stronger than the v_+ or v_{rnd} stimuli in delaying the reversal trial. Perturbation had the strongest effect when applied to the reversal trial, while perturbations on trials preceding the reversal showed appreciable but reduced effects (Fig.6D, right). When the v_+ stimulus was applied to trials preceding the reversal trial, the reversal was accelerated (Fig.6E, left). The effect of v_+ stimulus also increased with stimulus strength and was significantly stronger than the v_- or v_{rnd} stimuli in accelerating the reversal trial (Fig.6E, right).

We asked if perturbation of neural activity in PFC could exhibit similar responses. In other words, does increase (or decrease) in reversal probability activity lead to decrease (or increase) in choice activity in PFC? Although PFC activity was not perturbed by external inputs, we considered the residual activity of single trials, i.e., deviation of single trial neural activity around the trial-averaged activity, to be "natural" perturbation responses. We fitted a linear model to the residual activity of reversal probability and choice and found that they were strongly negatively correlated (i.e., negative slope in Fig.6F)) at the trial preceding the behavioral reversal trial. This analysis demonstrated the correlation between perturbation responses of reversal probability and choice activity. However, it remains to be investigated, through perturbation experiments, whether reversal probability activity is causally linked to choice activity in PFC and, moreover, to animal's choice outcomes.

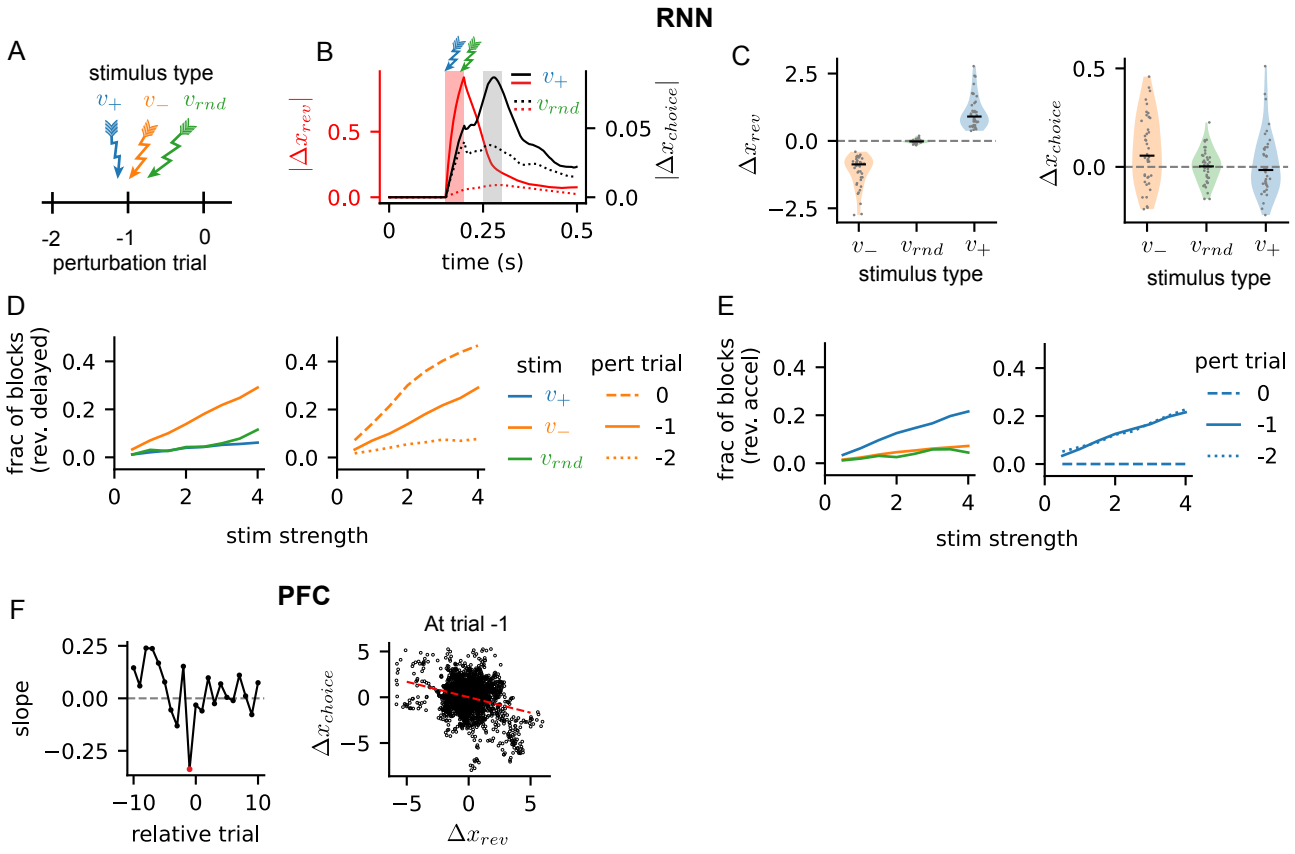

Figure 6: Perturbing RNN's neural activity encoding reversal probability biases choice outcomes. **(A)** RNN perturbation scheme. Three perturbation stimuli were used; v_+, population vector encoding the reversal probability; v_-, negative of v_+; v_{rnd}, control stimulus in random direction. Perturbation stimuli were applied at the reversal (0) and two preceding (-2, -1) trials. **(B)** Deviation of reversal probability activity Δx_{rev} and choice activity Δx_{choice} from the unperturbed activity. Perturbation was applied at the reversal trial during a time interval the cue was presented (shaded red). Choice was made after a short delay (shaded gray). Perturbation response along the reversal probability vector v_+ (solid) and random vector v_{rnd} (dotted) are shown. **(C)** Perturbation of reversal probability activity (left) and choice activity (right) in response to three types of stimulus. Δx_{rev} shows the activity averaged over the duration of perturbation , and Δx_{choice} shows the averaged activity over the duration of choice. **(D-E)** Fraction of blocks in all 40 trained RNNs that exhibited delayed or accelerated reversal trials in response to perturbations of the reversal probability activity. Perturbations at trial number -1 by three stimulus types are shown on the left panels, and perturbations at all three trials by the stimulus of interest (v_- in D and v_+ in E) are shown on the right panels. **(F)** Left: The slope of linear regression model fitted to the residual activity of reversal probability and choice. The residual activity at each trial over the time interval [0, 500]ms was used to fit the linear model. Red dot indicates the slope at trial number -1. Right: Each dot is the residual activity of a block at trial number -1. Red line shows the fitted linear model.

Discussion

Reversal learning. Reversal learning has been a behavioral framework for investigating how the brain supports flexible behavior [1–6] and for elucidating neural mechanisms underlying mental health issues [30, 31]. It has been shown that multiple brain regions (cortical [3, 4, 14, 27–29, 32] and subcortical [3, 33]), neuromodulators [2, 5, 6] and different inhibitory neuron types [28, 29] are involved in reversal learning.

Our results. Despite these recent advances, the dynamics of neural activity in cortical areas during a reversal learning task have not been well characterized. In this study, we investigated how reversal probability is represented in cortical neurons by analyzing neural activity in the prefrontal cortex of monkeys and recurrent neural networks performing the reversal learning task. Reversal probability was encoded in dynamically evolving neural trajectories that shifted in response to reward outcomes. Neural trajectories were translated in the direction consistent with how reversal probability would be updated by reward outcomes, and their position could be estimated by integrating reward outcomes across trials. These suggested a neural mechanism where separable dynamic trajectories represent reversal probability by accumulating reward outcomes. Around the behavioral reversal, the average effects of reward outcomes became monotonic, resulting in graded neural representation of reversal probabilities. Perturbation experiments in trained networks demonstrated a potential causal link between reversal probability activity and choice outcomes.

Attractor dynamics. RNNs with attractor dynamics have been investigated in various contexts as a neural implementation of normative models of decision-making and evidence integration [34–38]. One perspective is to consider decision variables as discrete or continuous attractor states of an RNN. Then, the network activity converges to an attracting state as a decision is made. Biologically plausible network models [7, 16] and neural recordings in cortical areas have been shown to exhibit discrete [9, 10, 39] and continuous [40] attractor dynamics. Another perspective, more closely related to our study, is to consider evidence integration as a movement of network state along a one-dimensional continuous attractor, as demonstrated in [8, 21, 41] (see also continuous attractor dynamics in spatial mapping [17, 42–44]).

In most of the studies, decision-related evidence was presented without significant interruption until the decision point [9, 10, 20, 39]. However, this was not the case in a reversal learning task with probabilistic rewards, as reward outcomes were revealed intermittently over multiple trials while intervening behavior must be performed in-between trials. We showed that such multi-trial evidence integration promoted substantial non-stationary activity in the neural subspace encoding reversal probability. Therefore, the continuous attractor dynamics, in which the network state stays close to the attracting states, did not fully account for the observed neural dynamics. Instead, our findings suggest that separable dynamic trajectories could serve as a neural mechanism for representing accumulated evidence in a temporally flexible way.

Related work. Recent studies showed that intervening behaviors, such as introducing an intruder [21] or accumulating reward across trials [41], could produce neural trajectories that deviate from and retract to a line attractor. In our study, we focused on characterizing the neural representation of reversal probability but did not investigate it from dynamical systems perspective. It remains as future work to characterize if and how the separable dynamic trajectories observed in our study could be augmented to the continuous attractor model and compare it to previous works [21, 41].

In a related work [45], RNNs were trained to perform a change point detection task designed by the International Brain Laboratory [46]. Authors showed that trained RNNs exhibited behavior outputs consistent with an ideal Bayesian observer, as found in our study. However, their trained RNN exhibited line attractor dynamics in contrast to ours. One possible reason for this discrepancy is that their network model stepped through only a few time points in a trial, which limited the possible range of temporal dynamics RNNs can exhibit. This suggests that the setup of a task RNNs learn can shape the trained RNN dynamics. Moreover, it needs to be investigated whether such attractor dynamics are present in the neural recordings from mice performing the change point detection task.

Although RNNs in our study were trained via supervised learning, animals learn a reversal learning task from reward feedback, making it into a reinforcement learning (RL) problem. Neuromodulators play a key role in mediating RL in the brain. In a recent study, dopamine-based RL was used to train artificial RNNs to conduct reversal learning tasks. It was shown that neural activity in RNNs and mice performing the same tasks were in good agreement [47]. In addition, projections of serotonin from dorsal raphe nuclei [6, 48] and norepinephrine from the locus coeruleus [5] to the cortical areas were shown to be involved in reversal learning. Further studies with biologically plausible network models including neuromodulatory effects [49, 50] or formal RL theories incorporating neuromodulators [51] could provide further insights into the role of neuromodulators in reversal learning.

Conclusion. Our findings show that, when performing a reversal learning task that requires evidence integration across trials, a cortical circuit adopts a dynamic neural representation of accumulated evidence to accommodate non-stationary activity associated with intervening behaviors. Such neural mechanism demonstrates the temporal flexibility of cortical computation and opens the opportunity for extending existing neural model for evidence accumulation by augmenting temporal dynamics.

Methods

1 Recurrent neural network

Network model. We trained a recurrent neural network with purely inhibitory synaptic connections. A baseline excitatory external input was applied to neurons, without which the network activity became quiescent. Such inhibitory network operated in a balanced regime where the recurrent inhibitory inputs were balanced with the external excitatory inputs [52]. Neurons were connected sparsely with connection probability p. Throughout network training, the signs of synaptic weights were preserved, resulting in a trained network that had only inhibitory synaptic connections.

The network dynamics were governed by the following equation

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W^{rec}\phi(\mathbf{u}) + \mathbf{I}_{base} + \mathbf{I}_{cue} + \mathbf{I}_{feedback} \tag{1}$$

and the network readout was

$$z = \mathbf{w}^{out} \cdot \phi(\mathbf{u}). \tag{2}$$

Here, $\mathbf{u} \in \mathbb{R}^N$ is the neural activity of population of N neurons, W^{rec} is an $N \times N$ recurrent connectivity matrix with inhibitory synaptic weights: $W^{rec}_{ij} < 0$ is connection from neurons j to i. The activation function was sigmoidal, $\phi(x) = 1/(1+\exp[-(ax+b)])$, and was applied to $\mathbf{u}$ elementwise in $\phi(\mathbf{u})$. The baseline input $\mathbf{I}_{base}$ was constant in time and same for all neurons, the cue $\mathbf{I}_{cue}$ was turned on to signal the RNN to make a choice, and the feedback $\mathbf{I}_{feedback}$ provided information about the previous trial's choice and reward outcome (see Table 1).

The duration of a trial was $T = 500$ms. The feedback $\mathbf{I}_{feedback}$ was applied on the time interval $[0, T_{feedback}]$, and the cue $\mathbf{I}_{cue}$ was applied on the the time interval $[T^{start}_{cue}, T^{end}_{cue}]$ where $T_{feedback} = 300$ms and $T^{start}_{cue} = 250$ms, $T^{end}_{cue} = 300$ms. The network choice was defined using the average of the readout z on the time interval $[T^{start}_{choice}, T^{end}_{choice}]$ where $T^{start}_{choice} = 350$ms and $T^{end}_{choice} = 400$ms.

Reduced model. One-dimensional reduction of the network dynamics in a subspace defined by a task vector, $\mathbf{v}$, was derived as follows (see Fig.4). The projection of network activity onto the task vector was

$$x = \langle \phi(\mathbf{u}), \mathbf{v} \rangle. \tag{3}$$

Then, the dynamics of the projected activity is governed by

$$\tau \frac{dx}{dt} = \left\langle \nabla_{\mathbf{u}} \phi(\mathbf{u}) \cdot \frac{d\mathbf{u}}{dt}, \mathbf{v} \right\rangle = x_{rec} + x_{ext} \tag{4}$$

where

$$x_{rec} = \left\langle \nabla_{\mathbf{u}} \phi(\mathbf{u}) \cdot [-\mathbf{u} + W^{rec}\phi(\mathbf{u})], \mathbf{v} \right\rangle \tag{5}$$

$$x_{ext} = \left\langle \nabla_{\mathbf{u}} \phi(\mathbf{u}) \cdot \mathbf{I}, \mathbf{v} \right\rangle. \tag{6}$$

Here x_{rec} includes both the decay and recurrent terms, and x_{ext} accounts for all external inputs $\mathbf{I} = \mathbf{I}_{base} + \mathbf{I}_{cue} + \mathbf{I}_{feedback}$.

2 Reversal learning task

Overview. Each block consisted of $T = 24$ trials during network training. The reversal trial r was sampled randomly and uniformly from 10 trials around the midtrial:.

$$r \in Unif[T_m - 5, T_m + 5] \quad \text{where} \quad T_m = T/2.$$

The network made a choice in each trial: A or B. To model which choice was rewarded, we generated a "rewarded" choice for each trial. One of the choices was more likely to be rewarded than the other. The network's choice was compared to the rewarded choice, and the network received a feedback that signaled its choice and reward outcome (e.g., chose A and received a reward). The option that yielded higher reward prior to the reversal trial was switched to the other option at the reversal trial.

To train the network to reverse its preferred choice, we used the output of an ideal Bayesian observer model as teaching signal. Specifically, we first inferred the scheduled reversal trial (i.e., the trial at which reward probability switched) using the Bayesian model. Then, the network was trained to flip its preferred choice a few trials after the inferred scheduled reversal trial, such that network's behavioral reversal trial occurred a few trials after the scheduled reversal trial.

Note that, although we refer to "rewarded" choices, there were no actual rewards in our network model. The "rewarded" choices were set up to define feedback inputs that mimic the reward outcomes monkey received.

2.1 Experiment variables.

The important variables for training the RNN were network choice, rewarded choice and feedback.
Network choice. To define network choice, we symmetrized the readout $\mathbf{z}^{sym} = (z, -z)$ and computed its log-softmax $f(\mathbf{z}^{sym}) = (\frac{e^z}{s}, \frac{e^{-z}}{s}) \equiv (f_0, f_1)$ where $s = e^z + e^{-z}$. The network choice was

$$c = \arg\max(f_0, f_1) = \begin{cases} A & \text{if} \quad f_0 > f_1 \\ B & \text{if} \quad f_0 < f_1 \end{cases} \tag{7}$$

where

$$A = 0, B = 1. \tag{8}$$

Rewarded choice. To model stochastic rewards, rewarded choices were generated probabilistically for each trial k:

$$\begin{aligned} Pr(\text{rewarded choice} = A) &= q_k \\ Pr(\text{rewarded choice} = B) &= 1 - q_k. \end{aligned} \tag{9}$$

The reversal of reward schedule was implemented by switching the target probability at the scheduled reversal trial of the block, denoted by r_{sch}.

$$\begin{aligned} \text{(before reversal)} \quad q_k &= p \quad \text{for} \quad k < r_{sch} \\ \text{(after reversal)} \quad q_k &= 1 - p \quad \text{for} \quad k \geq r_{sch}. \end{aligned} \tag{10}$$

Feedback. We considered that reward is delivered when the network choice agreed with the rewarded choice, and no reward is delivered when they disagreed. This led to four types of feedback inputs show in Table 1.

network choice	rewarded choice	feedback
A	A	A1 (rewarded)
	B	A0 (no reward)
B	A	B0 (no reward)
	B	B1 (rewarded)

Table 1: Four types of feedback inputs

2.2 Bayesian inference model

Here we formulate Bayesian models that infer the scheduled reversal trial or the behavior reversal trial.

Ideal observer model. The ideal observer model, developed previously [2, 4], inferred the scheduled reversal trial and assumed that (a) the target probability was known (Eq. 9) and (b) it switched at the reversal trial (Eq. 10).

The data available to the ideal observer were the choice $y_k \in \{A, B\}$ and the reward outcome $z_k \in \{0, 1\}$ at all the trials $k \in [1, T]$. We inferred the posterior distribution of scheduled reversal at trials $k \in [1, T]$. By Bayes' rule

$$p(r|y, z) = p(y, z|r)p(r)/p(y, z). \tag{11}$$

We evaluated the posterior distribution of r when data were available up to any trial $t \leq T$. The likelihood function $f_{IO}(r) = p(y_{1:t}, z_{1:t}|r)$ of the ideal observer was defined by

$$f_{IO}(r) = \prod_{k=1}^{t} q_k.$$

For $k < r$,

$$\begin{aligned} q_k &= p && \text{if} \quad y_k = A, \ z_k = 1 && (12)\\ &= 1 - p && \text{if} \quad y_k = A, \ z_k = 0 && (13)\\ &= 1 - p && \text{if} \quad y_k = B, \ z_k = 1 && (14)\\ &= p && \text{if} \quad y_k = B, \ z_k = 0. && (15) \end{aligned}$$

For $k \geq r$,

$$\begin{aligned} q_k &= 1 - p && \text{if} \quad y_k = A, \ z_k = 1 && (16)\\ &= p && \text{if} \quad y_k = A, \ z_k = 0 && (17)\\ &= p && \text{if} \quad y_k = B, \ z_k = 1 && (18)\\ &= 1 - p && \text{if} \quad y_k = B, \ z_k = 0. && (19) \end{aligned}$$

To obtain the posterior distribution of r (Eq. 11), the likelihood function $f_{IO}(r)$ was evaluated for all $r \in [1, t]$, assuming flat prior $p(r)$ and normalizing by the choice and reward data $p(y_{1:t}, z_{1:t})$.

Behavioral model. To infer the trial at which choice reversed, i.e., behavior reversal, we used a likelihood function that assumed the preferred choice probability switched at the behavior reversal. Here, the reward schedule was not known.

The data available to the behavioral model were the choice $y_k \in \{A, B\}$ at all the trials $k \in [1, T]$. We inferred the posterior distribution of behavior reversal at trials $k \in [1, T]$. By Bayes' rule

$$p(r|y) = p(y|r)p(r)/p(y). \tag{20}$$

The likelihood function for the behavioral model was

$$f_{BM}(r) = \prod_{k=1}^{t} q_k.$$

For $k < r$,

$$\begin{aligned} q_k &= p \quad \text{if} \quad y_k = A && (21)\\ &= 1 - p \quad \text{if} \quad y_k = B. && (22) \end{aligned}$$

For $k \geq r$,

$$\begin{aligned} q_k &= 1 - p \quad \text{if} \quad y_k = A && (23)\\ &= p \quad \text{if} \quad y_k = B. && (24) \end{aligned}$$

To obtain the posterior distribution of r, we assumed flat prior $p(r)$, as in the ideal observer, and normalized by the choice data $p(y_{1:t})$.

2.3 Training scheme

Overview. The ideal observer successfully inferred a scheduled reversal trial, which occurred randomly around the mid-trial. To learn to switch its preferred choice, we trained the network to learn from scheduled reversal trials inferred from the ideal observer. In other words, in a block consisting of T trials, the network choices and reward outcomes were fed into the ideal observer model to infer the randomly chosen scheduled reversal trial. Then, the network was trained to switch its preferred choice a few trials after the inferred reversal trial. This delay in the behavior reversal from the scheduled reversal was observed in monkey's reversal behavior [4] and a running estimate of the Maximum a Posterior of the reversal probability (see Step 3 below). As the inferred scheduled reversal trial varied across blocks, the network learned to reverse its choice in a block-dependent manner.

Below we described the specific steps taken to train the network.

Step 1. Simulate the network and store the network choices and reward outcomes.

Step 2. Apply the ideal observer model to network's choice and reward data to infer the scheduled reversal.

Step 3. Identify the trial t^* at which network choice should be reversed.

The main observation is that the running estimate of Maximum a Posterior (MAP) of the reversal probability converges a few trials past the MAP estimate. In other words, let

$$\begin{aligned} \text{MAP estimate :} \quad & r_T^* = \underset{k \in [1,T]}{\text{argmax}}\, p(r = k|D_{1:T}) \\ \text{Running estimate :} \quad & r_t^* = \underset{k \in [1,t]}{\text{argmax}}\, p(r = k|D_{1:t}) \end{aligned}$$

then,

$$r_t^* \to r_T^* \quad \text{if} \quad t > r_T^*$$

where the convergence occurs around

$$t^* = r_T^* + \delta \quad \text{with} \quad \delta = 4.$$

Step 4. Construct the choice sequences the network will learn.

$$\text{choice}_{IO}(k) = A \quad \text{if} \quad k < t^*$$
$$\text{choice}_{IO}(k) = B \quad \text{if} \quad k \geq t^*$$

Step 5. Define the loss function of a block.

$$loss = \sum_{k=1}^{T} CrossEnt\Big(\text{choice}_{actual}(k), \text{choice}_{IO}(k)\Big)$$

Step 6. Train the recurrent connectivity weights W^{rec} and the readout weights $\mathbf{w}^{out}$ with backpropagation using Adam optimizer with learning rate 10^{-2}. The learning rate was decayed by a factor 0.9 every 3 epochs. The batch size (i.e., the number of networks trained) was 256. The training was continued until the fraction of rewarded trials was close to reward probability p of the preferred option.

3 Targeted dimensionality reduction

Targeted dimensionality reduction (TDR) identifies population vectors that encode task variables explicitly or implicitly utilized in the experiment the subject or RNN performs [20]. In this study, we were interested in identifying population vectors that encode choice preference and reversal probability. Once those task vectors were identified, we analyzed the neural activity projected to those vectors to investigate neural representation of task variables.

We describe how TDR was performed in our study (see [20] for the original reference). First we regressed the neural activity of each neuron at each time point onto task variables of interest. Then we used the matrix of regression coefficients (i.e., neuron by time) to identify the task vector. Let $y_{it}(k)$ be the spiking rate of neuron i at time t on trial k where we have N neurons and M time points. We regressed the spiking activity on task variables of interest $z^v(k)$ where the task variables were $v \in$ {reversal probability, choice preference, direction, object, block type, reward outcome, trial number}. For each neuron-time pair, (i, t), we performed linear regression over all trials $k \in [0, T]$ with a bias:

$$y_{it}(k) = \sum_v \beta_{it}^v z^v(k) + bias_{it}. \tag{25}$$

This regression analysis yielded an $N \times M$ coefficient matrix β_{it}^v for each task variable, v. We considered this coefficient matrix as a population vector evolving in time: $\boldsymbol{\beta}_t^v = (\beta_{1t}^v, ..., \beta_{Nt}^v)$. Then, a task vector was defined as the population vector $\mathbf{w}^v \in \mathbb{R}^N$ at which the L_2-norm $\|\boldsymbol{\beta}_t^v\|_2$ achieved its maximum:

$$\mathbf{w}^v = \boldsymbol{\beta}_{t_{\max}}^v \tag{26}$$
$$t_{\max} = \arg\max_t \|\boldsymbol{\beta}_t^v\|_2. \tag{27}$$

We performed QR-decomposition on the matrix of task vectors $W = [\mathbf{w}_{rev}, \mathbf{w}_{choice}, ...]$ to orthogonalize the task vectors. Then, the population activity was projected onto each (orthogonalized) task vector to obtain the neural activity encoding each task variable:

$$x_t^v(k) = \mathbf{w}^v \cdot \mathbf{y}_t(k) \tag{28}$$

where $\mathbf{y}_t(k) = (y_{1t}(k), ..., y_{Nt}(k))$ is the population activity at time t on trial k.

4 Reward integration equation

To derive the reward integration equation shown in Figure 3, we considered the neural activity in a subspace encoding the reversal probability:

$$x^k_{rev}(t) = \mathbf{w}^{rev} \cdot \mathbf{y}^k(t). \tag{29}$$

We analyzed the neural activity at the time of cue onset $t = t_{on}$ and obtained a sequence of reversal probability activity across trials: $x^0_{rev}(t_{on}), ..., x^K_{rev}(t_{on})$. To set up the reward integration equation

$$x^{k+1}_{pred}(t_{on}) = x^k_{pred}(t_{on}) + R^{\pm}_k(t_{on}) \quad \text{with} \quad x^0_{pred}(t_{on}) = x^0_{rev}(t_{on}), \tag{30}$$

we estimated the update $R^{\pm}_k(t_{on})$ driven by reward outcomes at each trial k. Specifically, the update term was defined as the block-average of the difference of reversal probability activity at adjacent trials:

$$\begin{aligned} \Delta x^k_{rev}(t_{on}) &= x^{k+1}_{rev}(t_{on}) - x^k_{rev}(t_{on}) \\ R^+_k(t_{on}) &= \left\langle \Delta x^k_{rev}(t_{on}) \right\rangle_{b^+_k} \quad \text{if rewarded at trial } k \\ R^-_k(t_{on}) &= \left\langle \Delta x^k_{rev}(t_{on}) \right\rangle_{b^-_k} \quad \text{if not rewarded at trial } k. \end{aligned} \tag{31}$$

Here, b^+_k denotes all the blocks across sessions (or networks) in which reward was received at trial k. Similarly, b^-_k denotes all the blocks in which reward was not received at trial k.

To predict $x^k_{rev}(t_{on})$, we set the initial value $x^0_{pred}(t_{on}) = x^0_{rev}(t_{on})$ at trial 0 and sequentially predicted the following trials using Eq. (30) with the update term from Eq. (31). The same analysis was performed at different time points t. We derived integration equations for each time and assessed its prediction accuracy as shown in Figure 3F.

To evaluate the contribution of reward and no-reward outcomes on the average responses of $\Delta x^k_{rev}(t)$ over blocks, we computed

$$\langle \Delta x^k_{rev}(t) \rangle_{b_k} = f^+_k \langle \Delta x^k_{rev}(t) \rangle_{b^+_k} + f^-_k \langle \Delta x^k_{rev}(t) \rangle_{b^-_k} = f^+_k R^+_k(t) + f^-_k R^-_k(t) \tag{32}$$

where

$$f^+_k = \frac{|b^+_k|}{|b_k|}, \quad f^-_k = \frac{|b^-_k|}{|b_k|} \tag{33}$$

with $b_k = b^+_k \cup b^-_k$ denote the fractions of reward and no-reward blocks at trial k. In Figure 4D and Figure 5A, the weighted responses, i.e., $f^+_k R^+_k(t)$ and $f^-_k R^-_k(t)$, were shown.

5 Decoding monkey's behavioral reversal trial

The PFC activity $x^{rev}_t(k)$ encoding reversal probability was used to decode the behavioral reversal trial at which monkey reversed its preferred choice (see Supp. Fig.S2). Our analysis is similar to the Linear Discriminant Analysis (LDA) performed in a previous study [4] at a fixed time point. Here, we applied LDA to time points across a trial.

For training, 90% of the blocks were randomly selected to train the decoder and remaining 10% of the blocks were used for testing. This was repeated 20 times. Input data to LDA was $x^{rev}_t(k)$ of Δk trials around the reverse trial, i.e., $k \in [k_{rev} - \Delta k, ..., k_{rev} + \Delta k]$ with $\Delta k = 10$. At each trial k, we took the activity vector $\mathbf{x}^{rev}_{t_0}(k) = (x^{rev}_{t_0-\Delta t}(k), ..., x^{rev}_{t_0+\Delta t}(k))$ around time t_0 with $\Delta t = 160$ms. The target output

of LDA was a one-hot vector $\mathbf{y}^{target}$, whose element was 1 at the reversal trial k_{rev} and 0 at other trials. The following input-output shows dataset of a block used for training:

$$\text{Input:} \quad \left[\mathbf{x}_{t_0}^{rev}(k_{rev} - \Delta k), ..., \mathbf{x}_{t_0}^{rev}(k_{rev}), ..., \mathbf{x}_{t_0}^{rev}(k_{rev} + \Delta k)\right] \tag{34}$$

$$\text{Output:} \quad \mathbf{y}^{target} \tag{35}$$

Here $\mathbf{x}_{t_0}^{rev}(k) \in \mathbb{R}^{T_{dec}}$ where $T_{dec} = 2\Delta t/\Delta h + 1$ denotes the number of time points around t_0 with time increment $\Delta h = 20$ms, and the one-hot vector $\mathbf{y}^{target} = [0, ..., 1, ..., 0] \in \mathbb{R}^{K_{dec}}$ where $K_{dec} = 2\Delta k + 1$ denotes the number of trials around the reversal trial. As mentioned above, this analysis was repeated for time point t_0 across a trial.

6 Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health: the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the National Institute of Mental Health (NIMH).

References

[1] Charles M Butter. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in macaca mulatta. *Physiology & Behavior*, 4(2):163–171, 1969.

[2] Vincent D Costa, Valery L Tran, Janita Turchi, and Bruno B Averbeck. Reversal learning and dopamine: a bayesian perspective. *Journal of Neuroscience*, 35(6):2407–2416, 2015.

[3] Stephanie M Groman, Colby Keistler, Alex J Keip, Emma Hammarlund, Ralph J DiLeone, Christopher Pittenger, Daeyeol Lee, and Jane R Taylor. Orbitofrontal circuits control multiple reinforcement-learning processes. *Neuron*, 103(4):734–746, 2019.

[4] Ramon Bartolo and Bruno B Averbeck. Prefrontal cortex predicts state switches during reversal learning. *Neuron*, 106(6):1044–1054, 2020.

[5] Zhixiao Su and Jeremiah Y Cohen. Two types of locus coeruleus norepinephrine neurons drive reinforcement learning. *bioRxiv*, pages 2022–12, 2022.

[6] Jung Ho Hyun, Patrick Hannan, Hideki Iwamoto, Randy D Blakely, and Hyung-Bae Kwon. Serotonin in the orbitofrontal cortex enhances cognitive flexibility. *bioRxiv*, 2023.

[7] Xiao-Jing Wang. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968, 2002.

[8] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[9] Hidehiko K Inagaki, Lorenzo Fontolan, Sandro Romani, and Karel Svoboda. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 566(7743):212–217, 2019.

[10] Thomas Zhihao Luo, Timothy Doyeon Kim, Diksha Gupta, Adrian G Bondy, Charles D Kopec, Verity A Elliot, Brian DePasquale, and Carlos D Brody. Transitions in dynamical regime and neural mode underlie perceptual decision-making. *bioRxiv*, pages 2023–10, 2023.

[11] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

[12] Rescorla Ra. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Classsical conditioning II: Current research and theory*, pages 64–99, 1972.

[13] Bruno B Averbeck. Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–5. IEEE, 2017.

[14] Anthony I Jang, Vincent D Costa, Peter H Rudebeck, Yogita Chudasama, Elisabeth A Murray, and Bruno B Averbeck. The role of frontal cortical and medial-temporal lobe brain areas in learning a bayesian prior belief on reversals. *Journal of Neuroscience*, 35(33):11751–11760, 2015.

[15] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

[16] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.

[17] H Sebastian Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, 1996.

[18] Anil Bollimunta, Douglas Totten, and Jochen Ditterich. Neural dynamics of choice: single-trial analysis of decision-related activity in parietal cortex. *Journal of Neuroscience*, 32(37):12684–12701, 2012.

[19] Carlos D Brody and Timothy D Hanks. Neural underpinnings of the evidence accumulator. *Current opinion in neurobiology*, 37:149–157, 2016.

[20] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[21] Aditya Nair, Tomomi Karigo, Bin Yang, Surya Ganguli, Mark J Schnitzer, Scott W Linderman, David J Anderson, and Ann Kennedy. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell*, 186(1):178–193, 2023.

[22] Nuo Li, Tsai-Wen Chen, Zengcai V Guo, Charles R Gerfen, and Karel Svoboda. A motor cortex circuit for motor planning and movement. *Nature*, 519(7541):51–56, 2015.

[23] Nuo Li, Kayvon Daie, Karel Svoboda, and Shaul Druckmann. Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, 532(7600):459–464, 2016.

[24] Arseny Finkelstein, Lorenzo Fontolan, Michael N Economo, Nuo Li, Sandro Romani, and Karel Svoboda. Attractor dynamics gate cortical information flow during decision-making. *Nature neuroscience*, 24(6):843–850, 2021.

[25] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

[26] David M Zoltowski, Kenneth W Latimer, Jacob L Yates, Alexander C Huk, and Jonathan W Pillow. Discrete stepping and nonlinear ramping dynamics underlie spiking responses of lip neurons during decision-making. *Neuron*, 102(6):1249–1258, 2019.

[27] Huriye Atilgan, Cayla E Murphy, Hongli Wang, Heather K Ortega, Lucas Pinto, and Alex C Kwan. Change point estimation by the mouse medial frontal cortex during probabilistic reward learning. *bioRxiv*, pages 2022–05, 2022.

[28] Jee Hyun Yi, Young Ju Yoon, Huijeong Jeong, Seo Yeon Choe, and Min Whan Jung. Selective engagement of prefrontal vip neurons in reversal learning. *bioRxiv*, pages 2024–04, 2024.

[29] Huijeong Jeong, Dohoung Kim, Min Song, Se-Bum Paik, and Min Whan Jung. Distinct roles of parvalbumin-and somatostatin-expressing neurons in flexible representation of task variables in the prefrontal cortex. *Progress in Neurobiology*, 187:101773, 2020.

[30] Stephanie M Groman, Summer L Thompson, Daeyeol Lee, and Jane R Taylor. Reinforcement learning detuned in addiction: integrative and translational approaches. *Trends in neurosciences*, 45(2):96–105, 2022.

[31] Geoffrey Schoenbaum, Matthew R Roesch, and Thomas A Stalnaker. Orbitofrontal cortex, decision-making and drug addiction. *Trends in neurosciences*, 29(2):116–124, 2006.

[32] Robert C Wilson, Yuji K Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279, 2014.

[33] Bruno Averbeck and John P O'Doherty. Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, 47(1):147–162, 2022.

[34] Roger Ratcliff. A theory of memory retrieval. *Psychological review*, 85(2):59, 1978.

[35] John Palmer, Alexander C Huk, and Michael N Shadlen. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision*, 5(5):1–1, 2005.

[36] Michael N Shadlen and William T Newsome. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of neurophysiology*, 86(4):1916–1936, 2001.

[37] Mark E Mazurek, Jamie D Roitman, Jochen Ditterich, and Michael N Shadlen. A role for neural integrators in perceptual decision making. *Cerebral cortex*, 13(11):1257–1269, 2003.

[38] Roger Ratcliff, Anil Cherian, and Mark Segraves. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of neurophysiology*, 90(3):1392–1407, 2003.

[39] Mikhail Genkin, Krishna V Shenoy, Chandramouli Chandrasekaran, and Tatiana A Engel. The dynamics and geometry of choice in premotor cortex. *BioRxiv*, 2023.

[40] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431–439, 2014.

[41] Emily L Sylwestrak, YoungJu Jo, Sam Vesuna, Xiao Wang, Blake Holcomb, Rebecca H Tien, Doo Kyung Kim, Lief Fenno, Charu Ramakrishnan, William E Allen, et al. Cell-type-specific population dynamics of diverse reward computations. *Cell*, 185(19):3568–3587, 2022.

[42] Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, 2022.

[43] Ben Sorscher, Gabriel C Mel, Samuel A Ocko, Lisa M Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111(1):121–137, 2023.

[44] Brad K Hulse and Vivek Jayaraman. Mechanisms underlying the neural computation of head direction. *Annual review of neuroscience*, 43:31–54, 2020.

[45] Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Brain Laboratory International, and Ila Fiete. Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4584–4596. Curran Associates, Inc., 2020.

[46] Charles Findling, Felix Hubert, International Brain Laboratory, Luigi Acerbi, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Matteo Carandini, Joana A Catarino, et al. Brain-wide representations of prior information in mouse decision-making. *BioRxiv*, pages 2023–07, 2023.

[47] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

[48] Sara Matias, Eran Lottem, Guillaume P Dugué, and Zachary F Mainen. Activity patterns of serotonin neurons underlying cognitive flexibility. *Elife*, 6:e20552, 2017.

[49] Emerson F Harkin, Michael B Lynn, Alexandre Payeur, Jean-François Boucher, Léa Caya-Bissonnette, Dominic Cyr, Chloe Stewart, André Longtin, Richard Naud, and Jean-Claude Béïque. Temporal derivative computation in the dorsal raphe network revealed by an experimentally driven augmented integrate-and-fire modeling framework. *Elife*, 12:e72951, 2023.

[50] Carlos Wert-Carvajal, Melissa Reneaux, Tatjana Tchumatchenko, and Claudia Clopath. Dopamine and serotonin interplay for valence-based spatial learning. *Cell Reports*, 39(2), 2022.

[51] Emerson F Harkin, Cooper D Grossman, Jeremiah Y Cohen, Jean-Claude Béïque, and Richard Naud. Serotonin predictively encodes value. *bioRxiv*, pages 2023–09, 2023.

[52] Carl van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724, 1996.

A Supplementary material

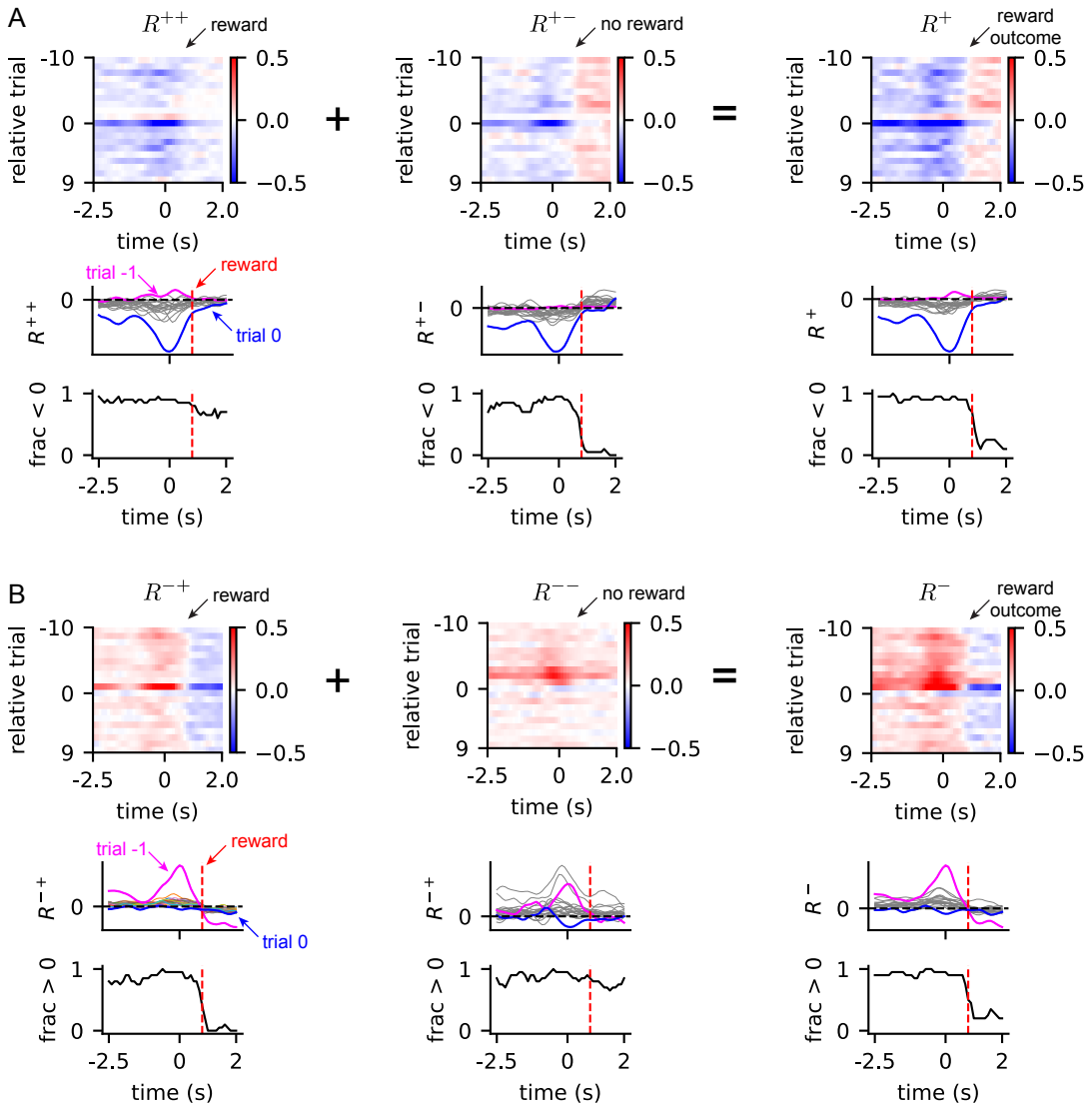

Figure S1: Break down of R^+, R^- by the reward outcomes of two consecutive trials. **(A)** R^+ was decomposed into two components $R^+ = R^{++} + R^{+-}$, where R^{++} indicates two consecutive reward trials and R^{+-} indicates a reward followed by no reward. Left: R^{++} across trial and time (top). Traces of R^{++} at individual trials and the fraction of trials whose traces are negative (bottom). Middle: Same as the left panel but for R^{+-}. Right: Same as the other panels but for R^+. **(B)** R^- was decomposed into two components $R^- = R^{-+} + R^{--}$, where R^{-+} indicates no reward followed by a reward and R^{--} indicates two consecutive no rewards. Same analysis as in panel (A) was performed.

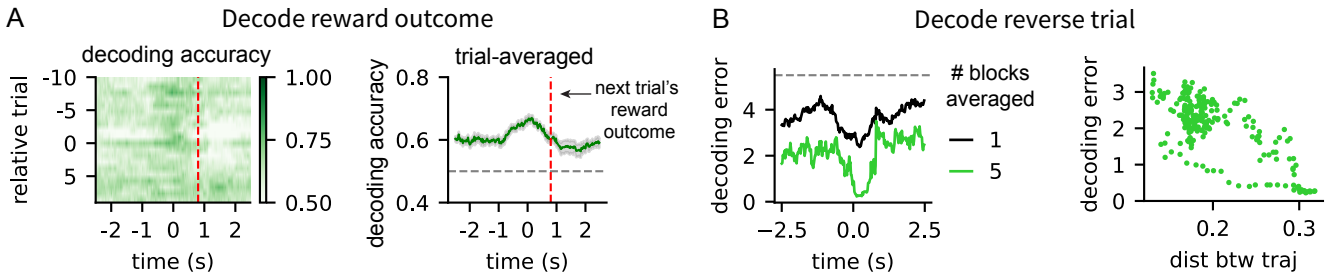

Figure S2: Decoding reward outcome and the behavioral reversal trial using neural trajectories encoding reversal probability. **(A)** Left: Decoding the reward outcome (i.e., reward or no reward) of every trial at each time point, given the difference of neural trajectories of two adjacent trials. At each time point, 300ms segment of the trajectories were used for decoding. Right: Decoding accuracy is averaged over all trials shown on the left panel. Red dotted line shows the approximate time of next trial's reward. Gray dotted line shows the chance level performance. **(B)** Left: Decoding the behavioral reversal trial using neural trajectories of 20 trials around the reversal trial. Decoding error shows the position of predicted reversal trial relative to the actual reverse trial. At each time point, 300ms segment of each trajectory was used for decoding. Black shows the decoding error when single trial trajectories were used, and green shows the result when randomly chosen 5 blocks of trajectories were averaged before decoding. Gray dotted line shows the chance level performance. Right: Distance between trajectories was measured by taking the average of normalized mean-squared-error of adjacent trajectories at all trials. Each dot corresponds to a time point shown on the left panel.