Serotonin as a State Prediction Error

A state prediction error model explains serotonin activity and its contribution to cognitive flexibility.

Abstract

Growing evidence suggests that serotonin promotes cognitive flexibility, the ability to quickly adapt to changes in the causal structure of the environment. We devise a mechanistic model of serotonin's contribution to this process by taking inspiration from model-based reinforcement learning, where learning a predictive model of its environment allows an agent to improve its performance and flexibility. We propose that serotonin broadcasts a state prediction error (SPE) signal across the brain, which is the teaching signal in self-supervised predictive learning. To test this, we develop a neural network model of the interactions between predictive learning and goal-directed behavior and simulate a wide range of decision-making experiments. We find that the SPE signal of our network bears a striking resemblance with dorsal raphe nucleus serotonergic phasic activity, sharing core features such as its response to unpredictability and its modulation by behavioral relevance. Furthermore, by emulating the effect of SPE manipulation on behavioral flexibility, we are able to explain a broad repertoire of behavioral phenomena induced by manipulations of serotonergic activity in mice and humans, such as impairments of reversal learning, model-based behavior, and behavioral persistence. Our work shines a new light on the role of serotonin in the brain by casting it as a state prediction error, accounting for its activity patterns and its contribution to cognitive flexibility.

Introduction

Cognitive flexibility is the ability to rapidly adapt to changes in the causal structure of the environment, a hallmark of biological intelligence. It has been extensively studied in the field of psychology and neuroscience, where it has long been proposed that the brain evolved specialized computational processes to support adaptive behavior (Miller and Cohen 2001). Recent advances in reinforcement learning have provided new insights into what these processes might be. In particular, auxiliary learning objectives have emerged as crucial tools for promoting flexibility in deep reinforcement learning (Jaderberg et al. 2016). These additional objectives shape internal representations to support other learning goals beside the primary task of reward maximization.

Among these, predictive learning objectives, where an agent learns to predict the consequences of its actions, stand out as especially powerful. It has been shown to support flexibility in several ways. First, predictive learning encourages agents to maintain an internal model of their environment. This world model can be used to simulate the outcomes of future actions, providing planning abilities that are critical for fast adaptation (Sutton 1991; Jensen et al. 2024). Indeed, when environmental dynamics change, agents with world models need only to update their beliefs about new transitions rather than re-learn complete state-action mappings from scratch. Second, as an auxiliary objective, predictive learning shapes neural representations: it organizes them into meaningful, disentangled features that are easier to predict (Recanatesi et al. 2021) , while preserving information not directly relevant to the current

task, which helps prevent overfitting and representation collapse (Dabney et al. 2021). Such representation shaping enables rapid adaptation to environmental changes and generalization to new tasks (Anand et al. 2020). Lastly, a predictive world model naturally tracks contextual information about its environment such as context and uncertainty. This uncertainty can be used to dynamically modulate goal-directed processes in a meta-learning fashion (Daw et al. 2005).

The rich and structured representations produced by predictive world models are strikingly similar to neural activity patterns in the hippocampus (Fang and Stachenfeld 2024), which may be specialized for predicting future states (Stachenfeld et al. 2017), as well as in the prefrontal cortex (Soltani and Koechlin 2022). Evidences suggest that the prefrontal cortex and hippocampus play key roles in planning (Pfeiffer and Foster 2013; Mattar and Daw 2018; Mattar and Lengyel 2022), representation shaping (Behrens et al. 2018; Whittington et al. 2020), and contextual modulation (Mante et al. 2013; Soltani and Izquierdo 2019), supporting the idea that they jointly maintain a predictive model of the environment that supports adaptive behavior. Furthermore, the ubiquity of predictive coding across cortical and subcortical circuits suggests that predictive learning may represent a general computational principle of the brain, extending far beyond these structures (Rao 2022; Gabhart et al. 2025)).

The learning of a predictive model depends on a *state prediction error (SPE)* signal, defined as the difference between predicted and actual next state. This teaching signal is to predictive learning what the reward prediction error (RPE) is to reinforcement learning, a more extensively studied class of reward-maximization algorithms that have been successfully mapped to different neural structures. In particular, dopamine activity is thought to support reinforcement learning by signaling an RPE across the brain. Neuromodulators are well suited to act as teaching signals, as they broadcast information widely across the brain and influence neural networks structure by modulating synaptic plasticity (Dayan 2012).

Among neuromodulators, Serotonin emerges as an especially compelling candidate for a neural substrate of state prediction error. Serotonin has widespread axonal projections throughout the forebrain, including the prefrontal cortex and hippocampus (Hamada et al. 2024; Beliveau et al. 2016), and exerts a strong influence on the neural plasticity of these areas (Lesch and Waider 2012; Hyun et al. 2023), features consistent with a teaching signal for predictive learning in the brain. Supporting this idea, serotonergic activity has been shown to correlate with various surprising events (Matias et al. 2017), including unexpected reward, punishment and even neutral outcomes, suggesting that it may encode a state prediction error. Indeed, state prediction error provides a formal definition of surprise, understood as an error of prediction, and may best account for serotonergic activity patterns. Recent technological advances have enabled precise recording and manipulation of serotonin neurons, leading to a wave of findings that firmly establish serotonin’s crucial role in cognitive flexibility (Matias et al. 2017; Grossman et al. 2022; Hyun et al. 2023; Kanen et al. 2021; Clarke et al. 2004). Casting serotonin as SPE, the teaching signal of predictive learning in the brain, could thus explain the causal link between serotonin signaling and flexibility. Akin to dopamine signalling a reward prediction error that supports reinforcement learning, we propose that serotonin signals a state prediction error that supports predictive learning in the brain - thereby enabling cognitive flexibility.

To test this hypothesis, we take inspiration from a series of works which demonstrates the power of neural network modeling to test hypotheses about the brain's computations (Zipser and Andersen 1988; Lehky and Sejnowski 1988; Mante et al. 2013; Jensen et al. 2023). We

construct a mechanistic model of the possible interactions between (unsupervised) predictive learning and (goal-directed) reinforcement learning with a neural architecture consisting of a policy network trained to act in a way that maximizes reward, and a predictive model network trained to perform next-state prediction given a state and an action. We use this model to exhibit the normative properties of a SPE signal, such as its interpretable multi-dimensionality, its response to surprising events and its modulation by behavioral relevance. Using simulation, we find that predictive learning can contribute to cognitive flexibility through background planning, and also through a wide diversity of SPE-driven mechanisms, such as representation shaping, learning rate adaptation and model-free / model-based arbitration.

This mechanistic model allows us to provide compelling evidence that (1) serotonin activity correlates with a SPE signal, and that (2) serotonin causally drives flexible behavior through the signaling of a SPE. First, we simulate our model on a new experimental task design and find that the DRN serotonergic activity shares the same features as a state prediction error signal, in particular its response and adaptation to novel and surprising events, and its modulation by behavioral relevance. Second, we simulate activation and inhibition of SPE signaling within our model and find that these manipulations strikingly reproduce the behavioral effects observed in causal serotonin manipulation studies, suggesting that serotonin acts as a teaching signal for predictive learning in the brain.

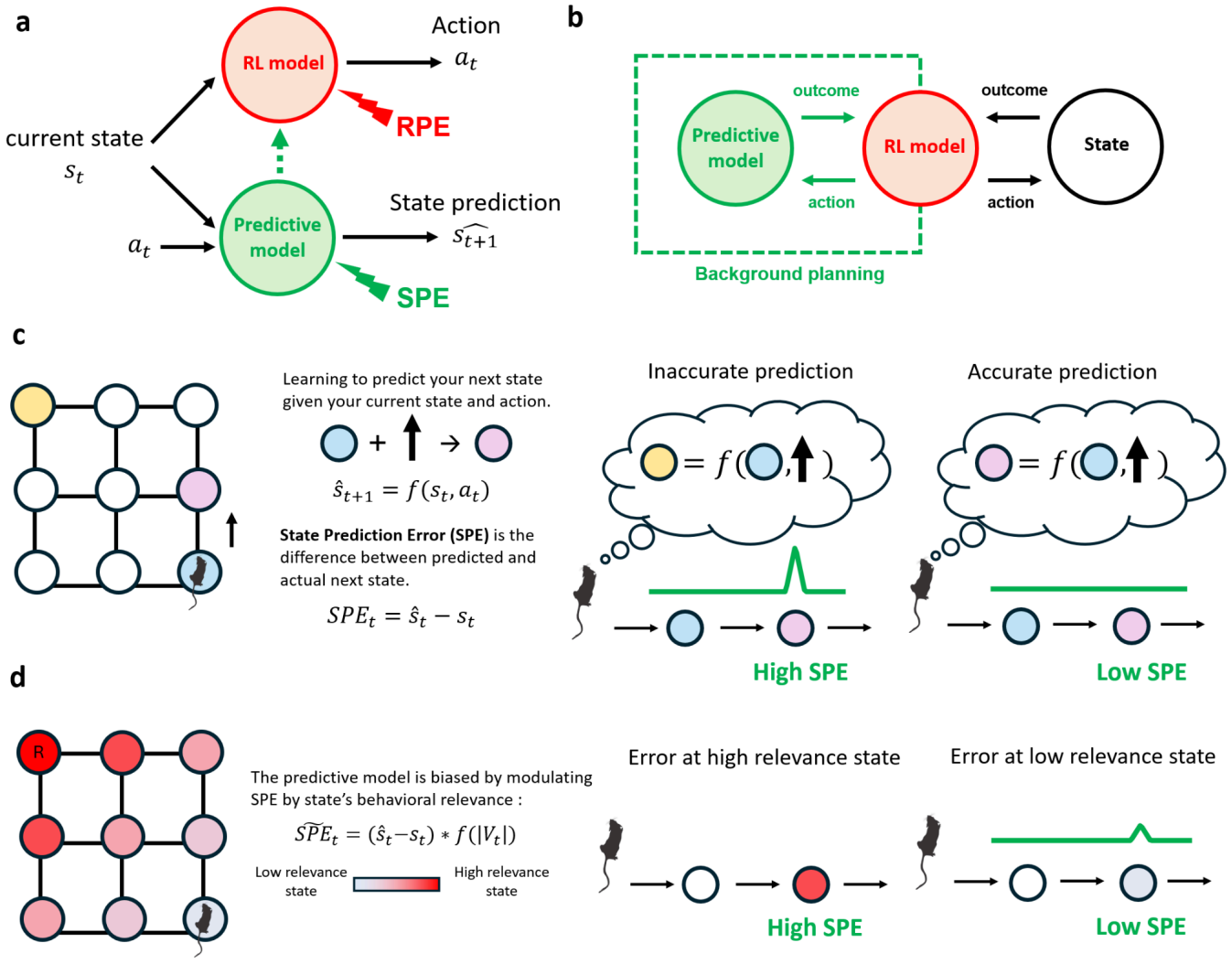


Figure 1. Modeling the interaction between predictive learning and goal-directed behavior. **a.** Model architecture. A Q-learning network estimates action values from the state, while a predictive network learns to predict the next latent state and reward given the current state. RPE is the teaching signal of the Q-learning network, while SPE is the teaching signal of the predictive network. **b.** Background planning: the predictive model simulates environment transitions that are used by the Q-learning network to update offline. **c.** SPE provides a formal definition of surprise : inaccurate prediction causes a high SPE transient, while accurate prediction causes a low SPE transient. **d.** Modulation by behavioral relevance. SPE is high for surprising, behaviorally relevant events but suppressed for irrelevant ones.

A model of predictive learning interaction with goal-directed behavior

We model the interaction between predictive learning and goal-directed behavior using a neural architecture composed of a Q-learning network and a predictive network (Fig. 1a). This framework allows us to examine in detail how predictive learning contributes to cognitive flexibility, what state prediction error (SPE) activity patterns emerge during goal-directed behavior, and how manipulations of SPE signals influence flexibility.

In simple settings where linear function approximation is enough to learn optimal behaviors, the Q-learning network is a linear decoder which takes as input a state s_t and output an estimation of the Q-values for pairs of states and actions $Q(s_t, a) = W_a \cdot s_t$. Its weights are updated using the TD-error $\delta_t = r_t + \gamma\, max(Q(s_{t+1}, \cdot)) - Q(s_t, a_t)$ following a TD(0) algorithm : $W_a \leftarrow W_a + \alpha_q \delta_t s_t$, with γ the discount rate, and α_q the Q-learner learning rate. Similarly, the predictive network is a linear decoder which takes as input a state s_t and an action a_t and outputs a prediction of the next state $\hat{s}_{t+1} = W^s_a \cdot s_t$. Crucially, the state representation includes reward onset as one of its components, treating reward as an observable event in the environment rather than a separate value signal. The state prediction weights are updated using a State Prediction Error $SPE_t = \hat{s}_t - s_t$ with the update rule $W^s_a \leftarrow W^s_a + \alpha_p SPE_t * s_t$, where α_p is the predictive model learning rate. Our base architecture simultaneously learns a behavioral policy, and a next state prediction, with the Q-learner weights and the predictive weights being updated at each step in the environment.

The predictive model can support goal-directed behavior through background planning (Sutton 1991, 91, Sutton and Barto, n.d.), where it is used to simulate experiences, allowing the Q-learner policy to be updated in an offline fashion (Fig. 1b). After each step in the environment, the Q-learner weights are updated by one-step transition simulations, using the predictive model. K starting states $(s_1, \ldots, s_K)$ are randomly sampled from a memory buffer. For each state s_i, an action a_i is randomly sampled and the predictive model output a predicted next state s_i^{next} from which the reward component r_i^{next} is extracted. For each simulated transition $(s_i, a_i, s_i^{next}, r_i^{next})$, we perform an offline Bellman backup update of the Q-learner weights : $W_{a_i} \leftarrow W_{a_i} + \alpha_q \delta_i s_i$ with $\delta_i = r_i^{next} + \gamma\, max_{a'} Q(s_i^{next}, a') - Q(s_i, a_i)$. This mechanism allows the predictive model to shape the agent behavioral policy. We focus on background planning as the

primary interaction mechanism between predictive learning and goal-directed behavior ; alternative mechanisms including representation shaping (Ha and Schmidhuber 2018; Recanatesi et al. 2021), learning rate adaptation (Gershman 2015; Grossman et al. 2022), and model-based/model-free adaptation (Daw et al. 2011, 2005; Gläscher et al. 2010) are compared later in the paper.

Acting as the teaching signal of the predictive learning network, the *State Prediction Error (SPE)* is an error vector of the same dimension as the state space, with different dimensions of the error being indicative of predictions errors along specific features of the environment. SPE is thus highly dependent on the way states are encoded from sensory observations. In what follows, we focus on the norm of the SPE vector. When the agent enters a new environment, it has not yet learned to make accurate predictions about its next state and therefore exhibits large SPEs, which gradually diminish as an accurate internal model is learned (Fig. 1c). Similarly, changes in the causal structure of a familiar environment also produce high SPEs before adaptation occurs. In this way, SPE provides a formal definition of surprise.

Naturalistic environments are often too large and complex for biological agents with limited representational resources to learn an accurate predictive model of the full state space. Maintaining such a model has an associated "wiring" cost, the resources necessary to sustain learned synaptic connectivity, which for our predictive model can be expressed as the norm of its weight matrices $C = \sum_a \|W_a\|$. We model pressure to minimize this cost by adding a weight decay term to the predictive weight update equation : $W_a^s \leftarrow W_a^s + \alpha_p SPE_t * s_t - \varepsilon W_a^s$, with ε the regularization strength. Under this constraint, it is not optimal for an agent to represent all state transitions with equal fidelity. Instead, limited resources should be preferentially allocated to states that matter most for guiding behavior. Consistent with this idea, empirical work has shown trade-offs in the fidelity of cortical representations as a function of behavioral relevance (Stroud et al. 2025) .

Which states matter most? From the perspective of reward maximization, states with high absolute value, whether strongly rewarding or strongly punishing, are most important for decision-making. We therefore modulate the SPE teaching signal according to state relevance, defined as the deviation of a state's value from the average:

$$\widetilde{SPE}(t) = (\hat{s}_t - s_t)(1 + \lambda(|V(t)| - E[|V(t)|]))$$

where $|V(t)|$ denotes the absolute value of the current state, estimated by the Q-learning network as $V(t) = max_a Q(s_t, a)$, and $E[|V(t)|]$ is the expectation of this quantity over previously visited states. The parameter λ controls the strength of this bias, with λ = 0 corresponding to an unbiased predictive model.

This formulation implements a resource allocation mechanism: unpredictable transitions involving high-relevance states generate larger SPE and are encoded with higher fidelity, whereas equally unpredictable transitions involving low-relevance states generate smaller SPE and are learned more weakly (Fig. 1d).The predictive model thus concentrates its limited capacity on behaviorally important aspects of the environment. This relevance modulation will be important for explaining why serotonergic activity is modulated by reward expectation in our photometry data.

Predictive learning support cognitive flexibility

To demonstrate how predictive learning contributes to cognitive flexibility, we simulate our model in a gridworld environment where at each trial the agent appears on one side of a wall while a reward is located at a fixed position on the other side. An opening (door) in the wall allows the agent to reach the reward, but the door location changes at a reversal trial, requiring the agent to relearn its policy. The speed at which an agent recovers its maximal reward rate after reversal serves as a marker of cognitive flexibility.

The agent is initially trained for 750 episodes, after which the door is moved to a different location for an additional 750 episodes (Fig. 2a). We use a gridworld of size 6×6 with a one-hot encoding vector of size N² for each state. We first examine model performances for a relevance modulation of $\lambda = 0$, learning rates of $\alpha_q = \alpha_p = 0.3$, and $K = 1$ planning steps. We measure reversal performance by the average reward in the post-reversal episodes.

During initial training, our predictive agent learns to maximize reward faster than a baseline model-free agent, though both converge to similar asymptotic performance (Fig. S1a). After reversal, the predictive agent adapts quicker to the change of door location(Fig. 2b), achieving higher reversal performance than a model-free agent with same α_q (Fig. 2c) (N=25 simulation runs per agent, t-test t(48)=5.82, p=4.67e-7). By running a parameter grid search, we find the optimal learning rate for the model-free agent in this setting ($\alpha_q^* = 1$, with optimal reward $R_{avg}^* = 0.68$), and compare the performance of our predictive agent against this optima for different values of α_p and α_q. We find that the predictive agent outperforms this baseline across a wide range of learning rates (see Fig. 2d and Fig. S1b for a comparison with baseline agents with different α_{MF}). These results show that the interaction of predictive learning with behavior through background planning supports cognitive flexibility, with optimal contribution depending on task-specific parameter settings.

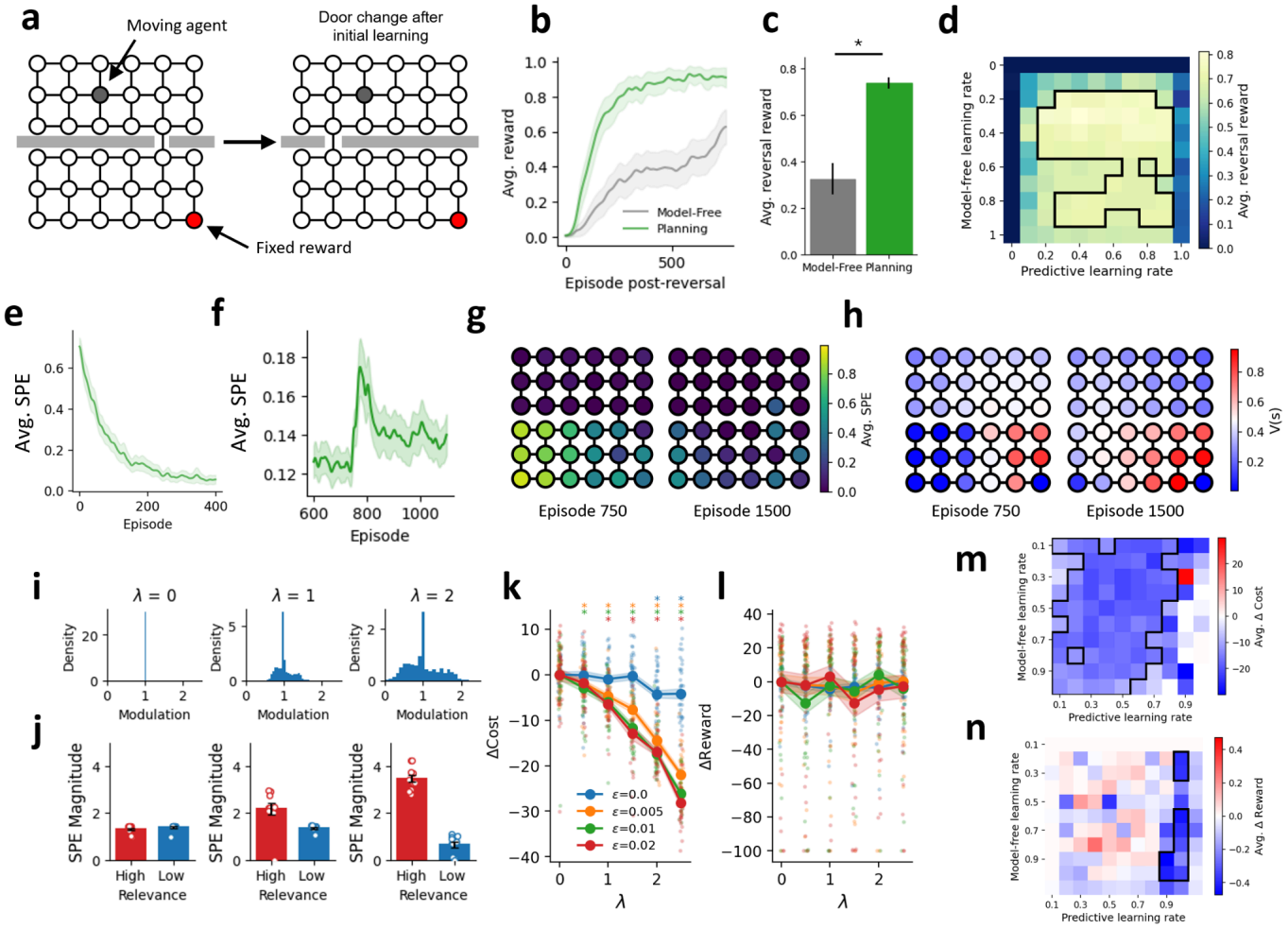


Figure 2. Predictive learning supports flexible behavior through background planning. **a.** Gridworld environment and door reversal paradigm. The agent navigates to a fixed reward location; after 750 episodes, the door position changes. **b.** post-reversal learning curves showing faster adaptation for the planning agent (green) versus model-free agent (grey). Shaded regions indicate ± SEM across N=25 runs with different random seeds. **c.** Average post-reversal reward; the planning agent significantly outperforms the model-free agent (N=25, t-test t(48)=5.82, p=4.67e-7). Error bars indicate ± SEM. **d.** Average reversal reward as a function of predictive and model-free learning rates; black contour indicates regions that significantly outperform the optimized model-free baseline (one-sample t-test, threshold at p<0.05). **e.** Average SPE per episode during initial learning. Shaded region indicates ± SEM. **f** Same as in *e* but around reversal. **g.** SPE per state at episode 750 (pre-reversal) and episode 1500 (post-reversal); SPE is lowest near frequently visited states. **h.** State values at episode 750 and 1500, showing value concentration around reward location. **i.** Distribution of relevance modulation factors for λ = 0, 1, and 2. **j.** SPE magnitude for unexpected transitions to high-relevance versus low-relevance states across λ values; higher λ amplifies SPE for high-relevance states. Error bars indicate ± SEM. **k.** Change in wiring cost as a function of λ for different weight decay values ε. λ significantly reduces wiring cost for non-zero values of ε (mean ± SEM for N=25 runs). Stars indicate significant cost difference between relevance-modulated and unmodulated SPE agents for each (ε, λ), Wilcoxon test with threshold at p<0.01). **l.** Change in reward as a function of λ; relevance modulation does not impair performance (mean ± SEM for N=25 runs). There's no significant reward difference between relevance-modulated and unmodulated SPE agents for every (ε, λ), Wilcoxon test with threshold at p<0.01). **m–n.** Change in wiring cost (m) and reward (n) across learning rate combinations for λ = 2 and ε=0.02 ; black contour indicates region with significant changes (one-sample t-test, threshold at p<0.05). Relevance modulation reduces wiring cost without affecting performance across most parameter settings.

The SPE of the predictive agent exhibits the qualities of a surprise signal. SPE is high during initial episodes and decreases as the agent learns the environment structure (Fig. 2e). The change of door location causes a transient SPE burst that disappears as the new structure is learned (Fig. 2f). Examining SPE across states, we find that SPE is lowest near the reward location where visits are frequent, and highest for rarely visited states (Fig. 2g).

We then investigated the effect of relevance modulation for different values of λ. In this task, state relevance corresponds to state value, which is learned across the session (Fig. 2h). Higher λ produces greater spread in the relevance modulation factor (Fig. 2i). If the agent encounters an unpredictable state prediction (due to a sudden change in the structure of the environment), SPE magnitude is amplified when arriving at high-relevance states and attenuated for low-relevance states (Fig. 2j).

The effect of relevance modulation on wiring cost and performance depends on the weight decay strength term ε : higher weight decay leads to greater cost savings for relevance-modulated agents (Fig. 2k). Critically, relevance modulation reduces wiring cost without impairing performance (Fig. 2l), a finding that holds across learning rates $(\alpha_{MB},\alpha_{MF})$ combinations (Fig. 2m–n). This suggests that relevance modulation optimizes representational capacity while preserving behavioral performance, an advantage we expect to be more pronounced in complex environments with many irrelevant features.

Serotonin broadcasts a state prediction error across the brain

Recording Dorsal Raphe serotonin activity during cue-reward learning

To test whether serotonin activity exhibits the features of a state prediction error, we applied our model to a novel cue-guided instrumental task. Head-restrained mice were trained to run on a treadmill to navigate through a virtual reality (VR) corridor. Two visual cues are displayed on the wall, one after the other, followed by a gray wall indicating the reward zone. The corridor can be of four types, with either ambiguous or unambiguous first cues (Fig. 3a). The second cue predicted reward or no reward depending on its identity. The first cue predicted reward or no reward on 80% of trials, but on 20% of trials was ambiguous with respect to the second cue and therefore the reward. To be rewarded in the rewarded corridors, the mice must enter the reward zone with a speed inferior to a 30cm/s threshold. Across 10 sessions, mice (N=6) learned these cue-reward contingencies, as reflected in anticipatory licking and reduced running speed before reward zone onset (Fig. 3b).

Dorsal raphe serotonin activity was recorded using fiber photometry in double-transgenic mice expressing GCaMP6 under the SERT-Cre promoter. To isolate task-related signals, we regressed out variance explained by running speed and focused on the motion-independent residual (see Fig. S2a).In parallel, we trained our model in a simulation of the same task. We used a microstimulus representation of the cues (Ludvig et al. 2008) to build the features of the state space (see methods) and discretize the action space into four speed levels for simplicity (speed level 1 to 4, with a threshold for the reward zone entry at 2). We trained our model with hand-picked parameters for a number of trials equivalent to the average mice and looked at its behavior during training. Over time, it learned to differentially slow-down before cue onset to maximize its reward rate, as observed in mice (Fig. 3c). We then examined whether

motion-independent serotonergic activity qualitatively tracked the norm of the model's SPE during learning, particularly its sensitivity to novelty, surprise, and behavioral relevance.

Dorsal Raphe Nucleus serotonin responses to Novelty and Surprise

In the first session, we observed strong serotonin responses to the initial presentations of novel cues (Fig. 3d). This matches the model's prediction that SPEs peak when the cues are first encountered and then diminishes over repeated exposure (Fig. 3e). Over the course of learning, serotonergic responses to cues decreased in amplitude (Wilcoxon signed-rank test, N=6, W=0, p=0.0312), consistent with our model in which SPE at cue onset declined as the cues became predictable.

By the last sessions, once cue-reward contingencies had been learned, the serotonin response to the second cue depended on the predictability of the first cue (Fig. 3f). When the first cue was ambiguous, the response to the second cue was stronger than in trials where the first cue predicted the second (Wilcoxon signed-rank test, N=6, W=0, p=0.0312). Our model captured this enhanced response in ambiguous trials, when the predictive model is "surprised" by cue identity and therefore generates a larger SPE (Fig. 3g).

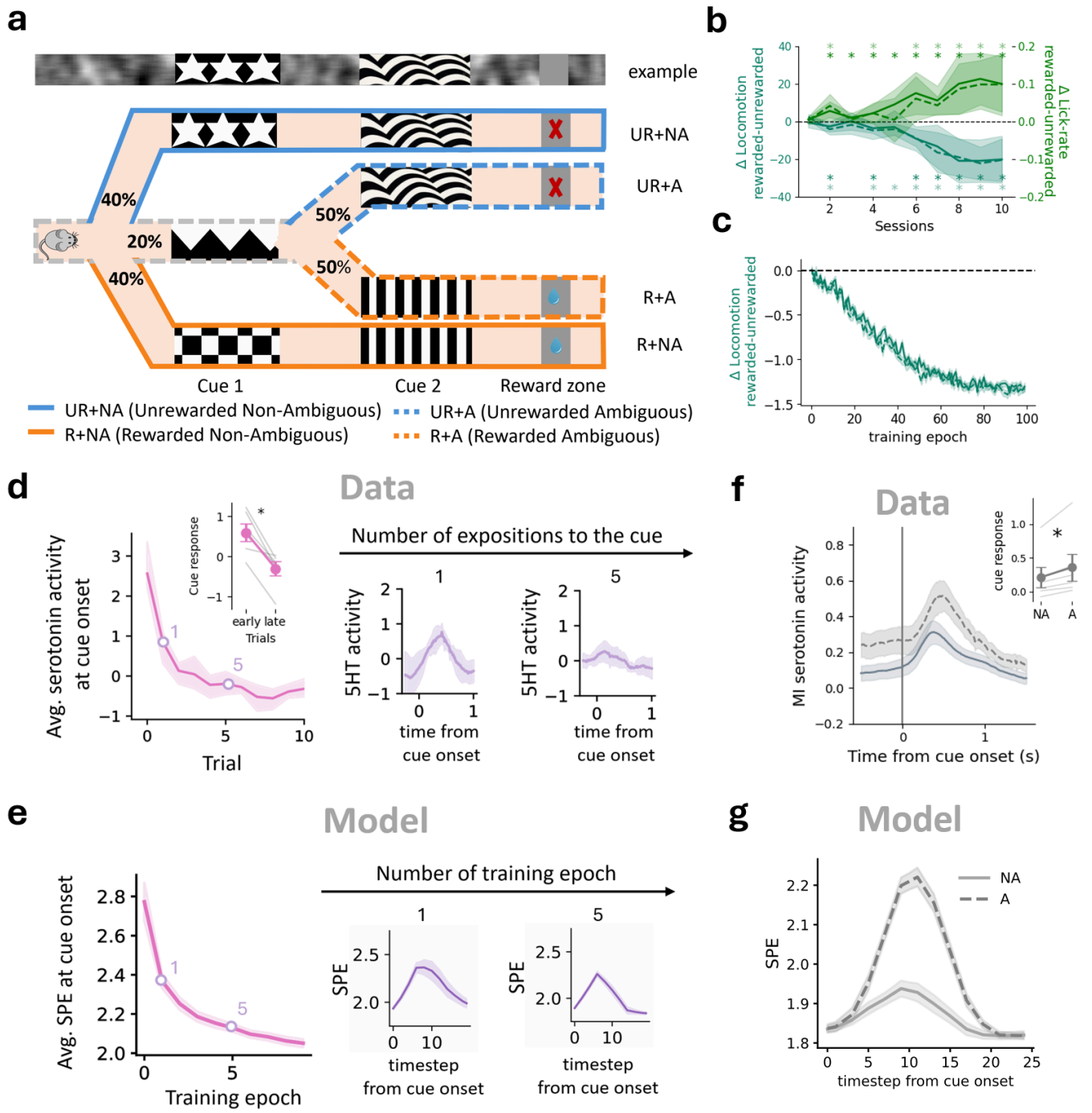


Figure 3. Dorsal raphe serotonergic phasic activity reflects novelty and surprise. **a.** Task design. Head-restrained mice run in a virtual reality (VR) circular corridor with two sequential cues followed by a reward zone. The second cue predicts reward or no reward depending on its identity, while the first cue is predictive on 80% of trials and ambiguous on 20% (P = punished; UR = unrewarded; NA = Non-ambiguous; A = ambiguous). **b.** Behavioral learning curves. Across 10 sessions (N = 6 mice), anticipatory licking increased and speed before reward zone decreased in rewarded versus unrewarded trials, indicating that the mice learned the cue-reward association (full line for non-ambiguous corridor and dotted-line for ambiguous corridors). Stars indicate significant differences between conditions for a session (Wilcoxon paired test on N=6 mice, threshold at p<0.05, dark shade for unambiguous and light shade for ambiguous corridors). **c.** Model learning curve. The difference in locomotion between rewarded and unrewarded trials becomes increasingly negative across training, indicating the model learns to slow down on rewarded trials. **d.** Dorsal raphe serotonin responses to initial onsets of cues during the first trials of the first sessions. Inset: serotonin cue response decreased between early and late trials (Wilcoxon paired test on N=6 mice,w=0, p=0.031). **e.** Model prediction. State prediction error (SPE)

peaks initial onsets of cues during the first training epoch, consistent with experimental serotonin responses. **f.** Response to ambiguity. By late training, serotonergic responses to the second cue were stronger when the first cue was unpredictable (A) compared to predictable ones (NA). Inset: average cue response to the second cue was significantly larger for ambiguous (A) than non-ambiguous (NA) trials (Wilcoxon paired test on N=6 mice, w=0, p=0.031). **g.** Model prediction. SPE at second cue onset is larger on ambiguous (A) than non-ambiguous (NA) trials, consistent with experimental observations.

Dorsal Raphe Nucleus serotonergic activity is modulated by behavioral relevance

In addition to novelty and surprise, serotonin responses are also modulated by behavioral relevance, i.e., by the predicted value of the cue. Indeed, over training, serotonin responses to cues became modulated by corridor identity: it significantly increased between early (days<4) and late (days>7) sessions for reward-predicting first and second cues (diff=+0.3431 p=0.0048 for cue 1 (R+NA) and diff=+0.3628 with p=0.019 for cue 2 (R+NA), statistical significance assessed with a hierarchical bootstrap test, see methods for details), while it did not significantly change for non-rewarding ones (diff=+0.0362 with p=0.6740 for cue 1 (UR+NA) and diff=+0.0298 with p=0.7406 for cue 2 (UR+NA)) (Fig. 4a-d). The same qualitative evolution is seen in SPE response to rewarded first and second cues, which increase over training after sharply decreasing during the first trials (Fig. 4e-f). This is due to SPEs being weighted by the estimated values of the cues, which gradually increase for rewarded cues and diminish for unrewarded ones across training (Fig. S3a) We see that for unrewarded cues, the SPE response decreases, which is not seen in the data.

In late sessions, serotonergic responses around cues onset were significantly modulated by corridor identity. For the first cue, responses were largest when it predicted reward, smallest when it predicted no reward, and intermediate for ambiguous cues (R+NA vs UR+NA, diff=+0.3227, p<1e-5 ; R+NA vs R+A, diff=+0.1639, p=0.0430 ; UR+NA vs UR+A, diff=-0.1942, p=0.0030) (Fig. 4g & 4k). The second cue response showed a similar modulation by predicted reward (R+NA vs UR+NA, diff=+0.3082, p<1e-5 ; R+A vs UR+A, diff=+0.2822, p=0.0334) (Fig. 4h & 4l). Furthermore, the response to unpredictable second cues were stronger than for predictable second cues, for both rewarded and unrewarded conditions (R+NA vs R+A, diff=-0.1376, p=0.0496 ; UR+NA vs UR+A, diff=-0.1636, p<1e-5). The model reproduced this pattern, predicting stronger SPEs for reward-predictive cues and intermediate responses for ambiguous ones for cue 1 (Fig. 4k-n), while predicting the same modulation by value and predictability for cue 2 (Fig 4i-j & 4m-n). What appears to be a code for value is explained by our model as a modulation of state prediction error by the behavioral relevance of current states.

Overall, our model closely captures the features of dorsal raphe serotonergic activity during cue–reward learning. These results, showing modulation of serotonin by novelty, surprise, and behavioral relevance, are consistent with previous studies that emphasized response to novelty (Matias et al. 2017) or to reward and punishment (Cohen et al. 2015; Li et al. 2016), both of which are naturally predicted by an SPE account of serotonin. Responses to novelty, uncertainty, and behavioral relevance were also present in the motion-dependent serotonin signal, though less consistently due to speed-related variability at cue onset (Fig. S2b-k).

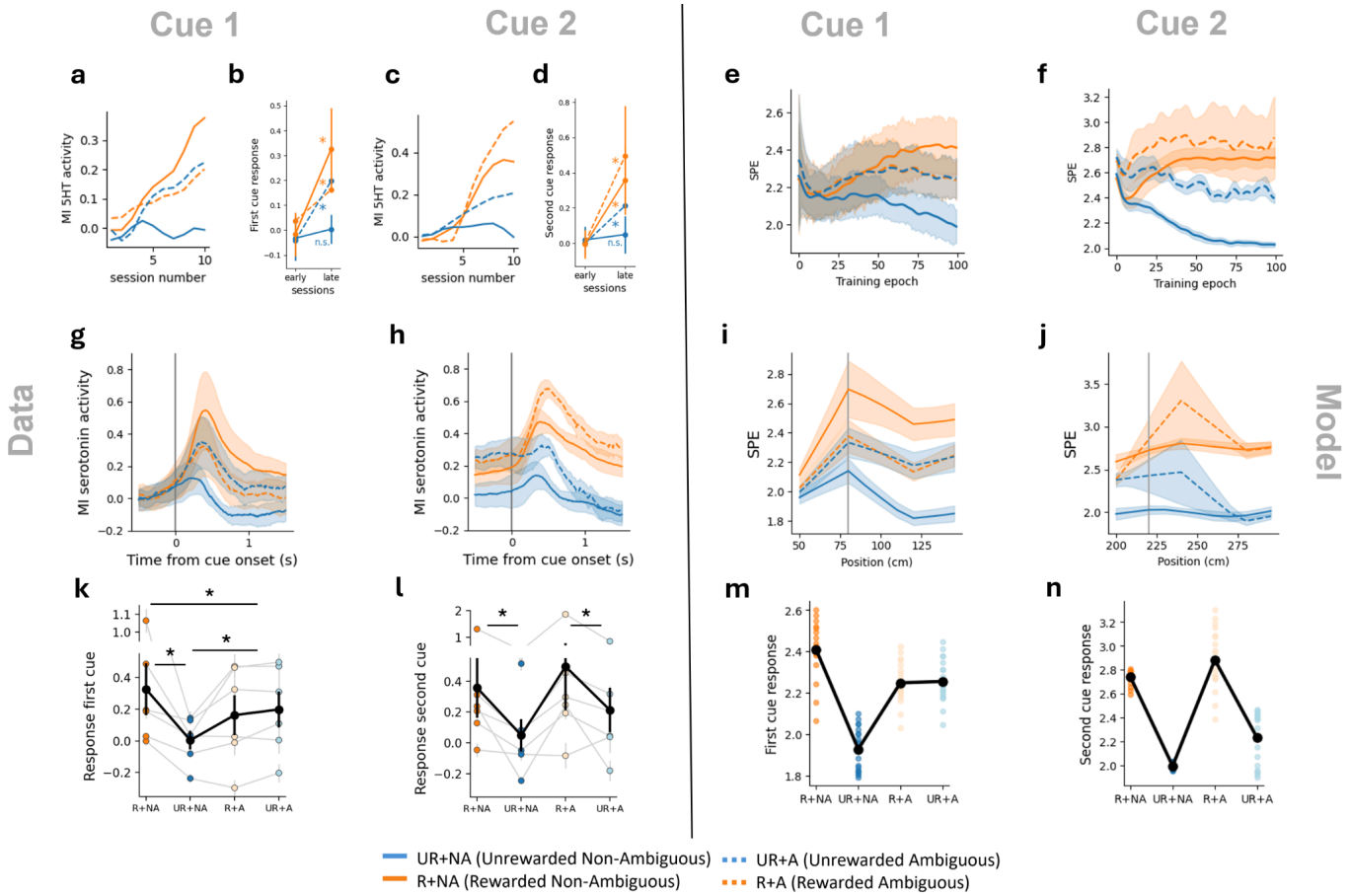


Figure 4. Dorsal raphe serotonergic phasic activity is modulated by state behavioral relevance.

a. Evolution of mean serotonin responses at cue 1 (average activity between 0 and 1 s after cue onset) across the ten experimental sessions for each trial type (N=6 mice, mean). **b.** Comparison of serotonin responses at cue 1 between early (sessions 1–3) and late (sessions 7–10) sessions. Serotonergic activity increased significantly for reward-predicting cues (R+NA: N=6, diff=+0.3431, p=0.0048) but not for unrewarded cues (UR+NA: N=6, diff=+0.0362, p=0.6740). **c.** Evolution of mean serotonin responses at cue 2 across sessions. **d.** Comparison of serotonin responses at cue 2 between early and late sessions, following the same pattern as cue 1 (R+NA: N=6, diff=+0.3628, p=0.019 UR+NA: diff=+0.0298, p=0.7406; R+A: diff=+0.4907, p=0.0090; UR+A: diff=+0.2132, p=0.0478). **e.** Model SPE at cue 1 (average SPE between positions 80–150 cm) across training epochs. **f.** Model SPE at cue 2 (average SPE between positions 220–290 cm) across training epochs. **g.** Time course of serotonin activity (mean ± SEM, N=6 mice) aligned to cue 1 onset for each trial type during late sessions (sessions 7–10). **h.** Time course of serotonin activity aligned to cue 2 onset during late sessions. **i.** Model SPE as a function of corridor position around cue 1. **j.** Model SPE as a function of corridor position around cue 2. **k.** Mean serotonin responses to cue 1 (average activity 0–1 s after onset, late sessions) across trial types. Individual mice are shown as colored dots connected by gray lines. Stars indicate significant pairwise differences (N=6 mice; R+NA vs UR+NA, diff=+0.3227, p<1e-5 ; R+NA vs R+A, diff=+0.1639, p=0.0430 ; UR+NA vs UR+A, diff=-0.1942, p=0.0030; threshold at p<0.05). ***l.*** Mean serotonin responses to cue 2 across trial types. Same conventions as panel k (R+NA vs UR+NA, diff=+0.3082, p<1e-5 ; R+A vs UR+A, diff=+0.2822, p=0.0334, R+NA vs R+A, diff=-0.1376, p=0.0496 ; UR+NA vs UR+A, diff=-0.1636, p<1e-5). **m.** Model SPE responses to cue 1 across trial types. Large black dots represent the mean across simulations; small colored dots represent individual simulation runs. **n.** Model SPE responses to cue 2 across trial types. Same conventions as panel m.

Comparison with other models of serotonin activity

Several alternative theories have been proposed for the nature of serotonin's phasic activity. One view, grounded in the observation that dorsal raphe serotonin neurons respond to both positive and negative prediction errors, holds that serotonin encodes a general surprise signal, formally an unsigned prediction error (Matias et al. 2017; Grossman et al. 2022). In our framework, this corresponds to the norm of an SPE vector without modulation by behavioral relevance. Another line of work has proposed that serotonin signals a reward prediction error (RPE), based on theoretical considerations of dopamine–serotonin opponency (Daw et al. 2002) and empirical evidence that dorsal raphe neurons signal RPEs during decision-making tasks (Feng et al. 2024). Finally, several studies have emphasized that serotonin tracks reward and state value more generally (Bromberg-Martin et al. 2010; Li et al. 2016; Cohen et al. 2015). Building on this, (Harkin et al. 2023) recently proposed that serotonin encodes a predictive code for value, a biologically constrained predictive signal for future reward, which quantitatively accounts for population serotonin activity better than previous theories centred on surprise or reward alone.

In this experiment, we find that our model explains serotonin activity patterns better than these alternatives. We first fit five models (Surprise, SPE, value, RPE, predictive code for value, see methods for the implementation of each model) on the average serotonin activity per corridor during the late sessions (days 7-10, average across trials and across mice), using a cross-validation procedure (Fig. 5a). We summarize the qualitative match to the data by plotting the fitted model's predicted activity at cues onsets for the different corridors (Fig. 5b). We find that our SPE model, a value model, and predictive code for value model do qualitatively account for serotonin activity at cue onset.

When it comes to predicting the full serotonin signal across corridors, we find that our SPE model best matches the population-averaged serotonin signal, which we assess using the cross-validated R^2 score (Fig. 5c). Notably, our model accounted for more variance in the data than the predictive code for value model (R^2=0.42±0.01 for our SPE model, R^2=0.35±0.01 for the predictive code for value model). We then used the same cross-validation procedure to fit the models on individual mouse serotonin activity during late sessions (see Fig. S4a for parameter estimation per mouse). We found that our SPE model provided the best match to individual mice serotonin activity (Fig. 5d). It best predicts serotonin activity for 3 out of 5 mice (excluding the mouse 5 for which no model manages to meaningfully fit the data), while the predictive code for value best explains 1 out of 5 mice and the RPE model 1 out of 5 mice. The models have different numbers of parameters, but this is mitigated by the use of cross-validation. Furthermore, we compute AIC and BIC for the population-level fit and individual-level-fits to penalize the effect of model complexity and find that our SPE model still best explains the data (see Fig. S4b-c).

Importantly, our model provides advantages beyond statistical fit on this experiment. Unlike other theories such as RPE or value, our model outputs a multidimensional signal (whose norm we have used for comparison to photometry data) which offers a potential explanation for the known heterogeneity of serotonergic responses. Moreover, casting serotonin as a SPE allows us to generate predictions about the role of serotonin in behavior, including effects of serotonergic manipulations - something that most existing theories of serotonin activity do not address.

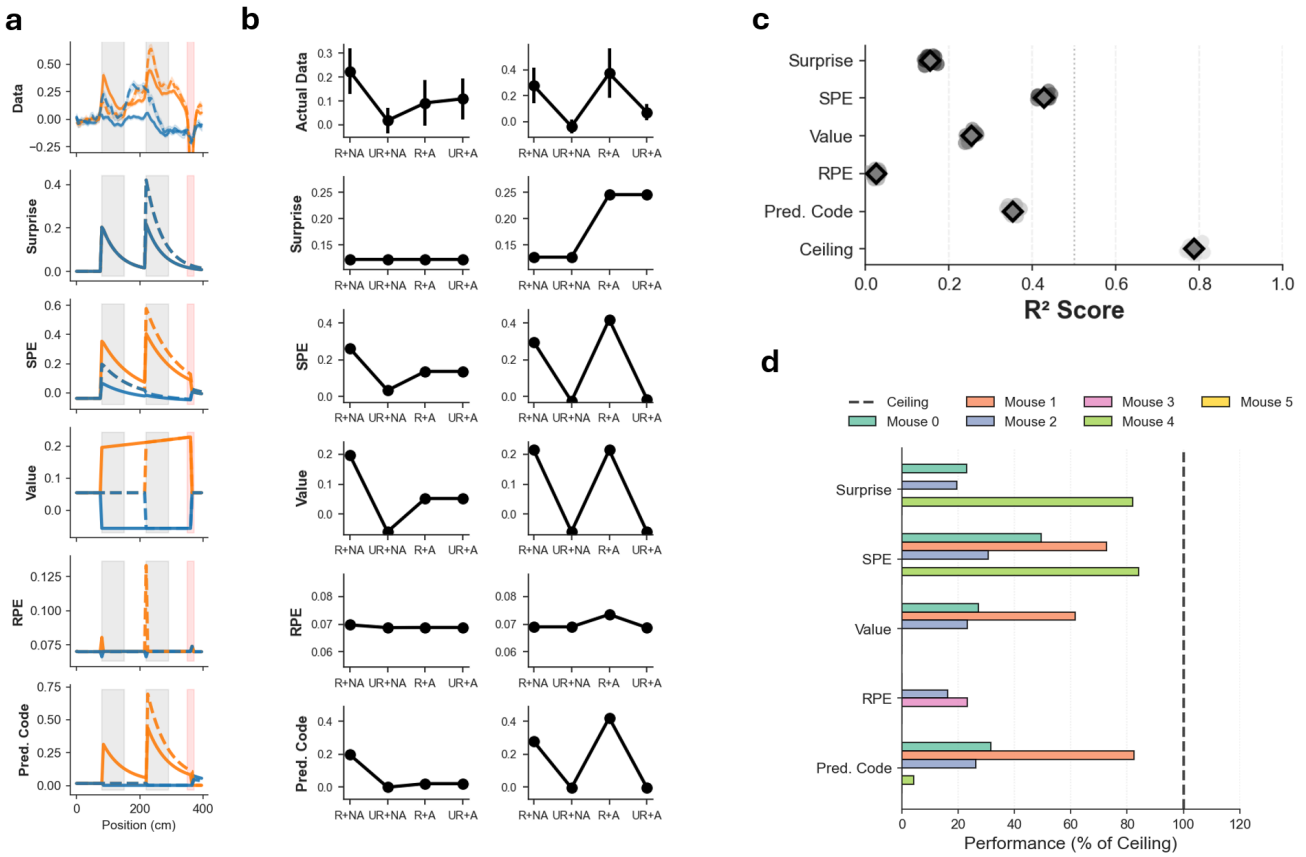


Figure 5. Comparison with alternative models of serotonin activity. **a.** Population-averaged MI serotonin signal during late sessions (≥7) (top row) and predicted serotonin activity for five candidate models: Surprise, SPE, Value, RPE, and Predictive Code for Value, each fitted to the population-averaged signal. Grey shaded regions indicate cue 1 and cue 2 zones, respectively, while the red shaded region indicates reward zone. **b.** Mean serotonin responses to cue 1 (left column) and cue 2 (right column) across trial types (top row), alongside model-predicted responses (rows below). Cue responses are computed as the average signal within the cue zones (80–150 cm for cue 1, 220–290 cm for cue 2). **c.** Cross-validated R^2 scores for each model fitted to the population-averaged MI serotonin signal. Individual validation fold scores are shown as grey dots and the mean across folds as a diamond. The ceiling corresponds to the maximum achievable R^2, obtained using the mean signal of the training folds as predictor. **d.** Model performance fitted to individual average serotonin signals for each of the six learner mice. Performance is expressed as the ratio of cross-validated R^2 to the mouse-specific ceiling R^2. Scores below 0 are not shown.

Serotonin supports cognitive flexibility through the signalling of a state prediction error

Having established that dorsal raphe serotonergic activity correlates with SPE, we next tested whether manipulating this signal produces the behavioral effects predicted by our framework. If serotonin acts as a teaching signal for a predictive model, then disrupting it should impair cognitive flexibility.

Our network model enables us to simulate such manipulations by scaling the SPE signal: a factor below 1 mimics serotonergic inhibition, while a factor above 1 mimics stimulation. Long-term pharmacological inhibition is modeled by applying the scaling throughout the task, effectively lowering the predictive learning rate and disrupting predictive model acquisition. Short-term optogenetic manipulation is modeled by scaling the SPE only during specific trials,

disrupting predictive model updating when environmental contingencies change during that period.

We investigate the causal role of serotonin in cognitive flexibility by simulating three behavioral tasks and find that our model captures the observed effects of serotonergic manipulation across these paradigms.

SPE manipulation effects on adaptation to structure change

We first test the effect of SPE manipulation on adaptive behavior in the door reversal task (Fig. 6a). In this task, accurate SPE signaling after the door location change is crucial for updating the predictive model and allowing background planning to adequately update state Q-values. We start to alter the SPE at the reversal episode, such that a predictive model has already been learned during the initial phase, but its update is now affected by SPE manipulation.

Inhibition of the SPE lowers the effective predictive learning rate, leading to progressively worse performance as the scaling factor decreases (Fig. 6b). The effect of SPE stimulation depends on the baseline predictive learning rate: if not already optimal, stimulation improves performance; if already optimal, stimulation has no effect or may even impair it.

Serotonin manipulation effects on reward reversal learning

These results parallel findings from studies of serotonin manipulation in reversal learning paradigms, a classical setting for assessing how quickly an agent adapts to changed reward contingencies. In mice, optogenetic inhibition of dorsal raphe serotonergic activity impairs reversal learning, while stimulation accelerates it (Hyun et al. 2023). Pharmacological serotonergic depletion similarly impairs reversal learning in both mice and humans (Clarke et al. 2004; Kanen et al. 2021).

Because our predictive model treats reward onset as a component of the state representation, changes in reward location generate SPE just as changes in environmental structure do. Serotonergic manipulation of the SPE signal therefore affects reversal learning directly.

We test this in a classical reversal learning setting within the same gridworld environment: at the reversal episode, the reward location shifts to a different corner of the grid (Fig. 6c). SPE inhibition is simulated by multiplying the SPE by a factor of 0.01 from reversal onset; stimulation is simulated with a factor of 3. We find that reversal speed is significantly impaired by SPE inhibition and accelerated by SPE stimulation (Fig. 6d) (two-sample t-test, Inhib. vs control : $t(48)=-24.89$, $p=4.31e-29$; Stim. vs control : $t(48)=3.72$, $p = 5.17e-4$). Under inhibition, the predictive model fails to update efficiently, preventing rapid adaptation. Under stimulation, the boosted learning rate allows the model to relearn the new reward location faster.

Serotonin manipulation effects on model-based behavior

Recent works highlighted serotonin's role in model-based behavior, which enables greater flexibility than pure model-free reactive behavior (Taira and Sharpe 2025). In mice, optogenetic inhibition of dorsal raphe nucleus serotonergic activity impairs the animals ability to perform prospective inference, a hallmark of model-based behavior (Ohmura et al. 2021). Similarly, in humans, serotonergic depletion reduces the contribution of model-based strategy in a classical two-step task (Worbe et al. 2016). In this task, each trial begins with a first-stage choice that

probabilistically leads to one of two second-stage states, where another choice is made for reward, with reward probabilities drifting slowly over time (Fig. 6e). Behavior following rewarded and unrewarded trials reveals the extent to which subjects rely on model-free versus model-based strategies.

Our framework offers a mechanistic account of these findings. Because the predictive model learns from SPE and shapes goal-directed behavior through background planning, disrupting SPE should impair the update of its world model. Because the agent still relies on a now-inaccurate model for planning, its behavior appears less model-based, not because the world model's contribution is reduced but because the model itself is unreliable. To test this, we trained our agents in a simulation of the two-step task and examined how SPE inhibition affects their behavior. We simulated serotonergic inhibition by multiplying the SPE signal in each agent by 0.01 from the start of training. Using a behavioral metric from the original study (Daw et al., 2011,see Methods), we computed a model-based coefficient for the predictive agent behavior, with and without SPE inhibition. We found that SPE inhibition substantially reduced the model-based coefficient (see Fig. 6f) : inhibiting SPE impaired accurate learning of the environment's transition structure, making the agent's behavior less model-based (two-sample t-test, t(48) = 4.55, p=2.66e-5).

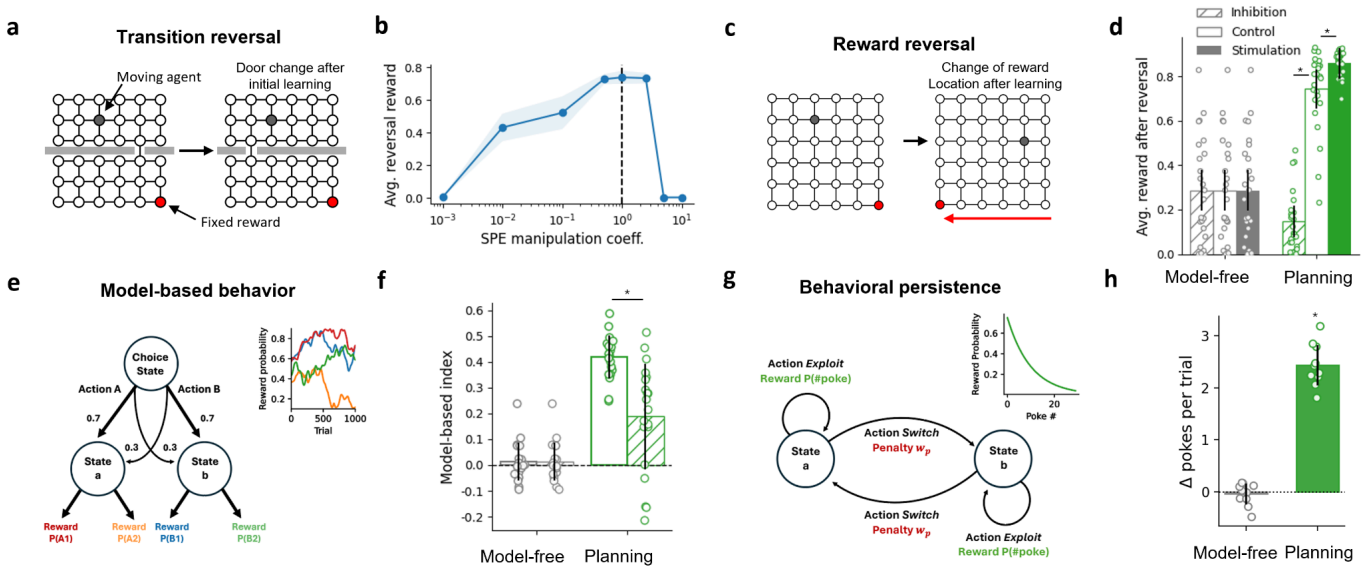


Figure 6. A SPE-driven mechanism of cognitive flexibility accounts for the behavioral effects of serotonin manipulation. **a.** Transition reversal task: the agent navigates to a fixed reward; after learning, the door location changes. **b.** Average reversal reward as a function of SPE manipulation coefficient; inhibition impairs performance, stimulation can improve it. **c.** Reward reversal task: the reward location changes after learning. **d.** Average reward after reversal under SPE inhibition, control, and stimulation for model-free and planning agents. Error bars indicate ± SEM. SPE inhibition significantly lowers avg. post-reversal reward and stimulation increases it (two-sample t-test, Inhib. vs control : t(48)=-24.89, p=4.31e-29; Stim. vs control : t(48)=3.72, p = 5.17e-4) **e.** Two-step task: first-stage choices lead probabilistically to second-stage states where reward probabilities drift over time. **f.** Model-based index for model-free and planning agents; SPE inhibition reduces model-based index for the planning agent (two-sample t-test, t(48) = 4.55, p=2.66e-5). **g.** Probabilistic foraging task: reward probability decays with each poke; switching incurs a cost. **h.** Change in pokes per trial under SPE stimulation; the predictive agent shows increased persistence, with significantly more pokes per trial under stimulation (one-sample t-test, t(24) = 12.50, p=5.42e-7).

Serotonin manipulation effects on behavioral persistence

The previous tasks involved serotonergic manipulations over relatively long timescales (hundreds of trials for reversal learning, full sessions for the two-step task). However, optogenetics studies have also reported effects from short-time within-trial serotonin manipulation on behavior. The most consistent finding is that short-term serotonergic activation promotes behavioral persistence in a context-dependent manner. Activation can promote waiting for future rewards (Fonseca et al. 2015) or prolong active exploitation of a reward site (Lottem et al. 2018). This suggests that at fast timescales, serotonin flexibly regulates behavioural strategies (Ahmadlou et al. 2025).

To investigate whether our model can account for this effect, we simulated the probabilistic foraging task from Lottem et al. (Lottem et al. 2018). In this task, the mouse can exploit a reward site where reward probability decays exponentially with each poke, or switch to another site at a subjective cost e (Fig. 6g). On 50% of trials, dorsal raphe serotonergic activity is stimulated, which increases the animal's willingness to continue foraging compared to non-stimulated trials.

We simulated serotonergic stimulation by multiplying the SPE signal by a factor of 20 during all steps of randomly selected stimulation trials (50% probability). As in the original experiment, we assessed persistence by comparing the number of nose pokes between control and stimulated trials. We found that our predictive agent reproduced the persistence effect of within-trial serotonin activation with an increased number of pokes per trial under stimulation (see Fig. 6h) (one-sample t-test, t(24) = 12.50, p=5.42e-7).

This result follows directly from our framework. Because reward onset is a component of the predicted state, the reward prediction error is part of the SPE. Stimulating SPE therefore effectively amplifies perceived reward for the planning agent, leading it to persist longer at the current site (see Fig. S5 for comparison with a model where the reward feature is excluded from the SPE).

Comparison with other models of SPE-driven flexible behavior

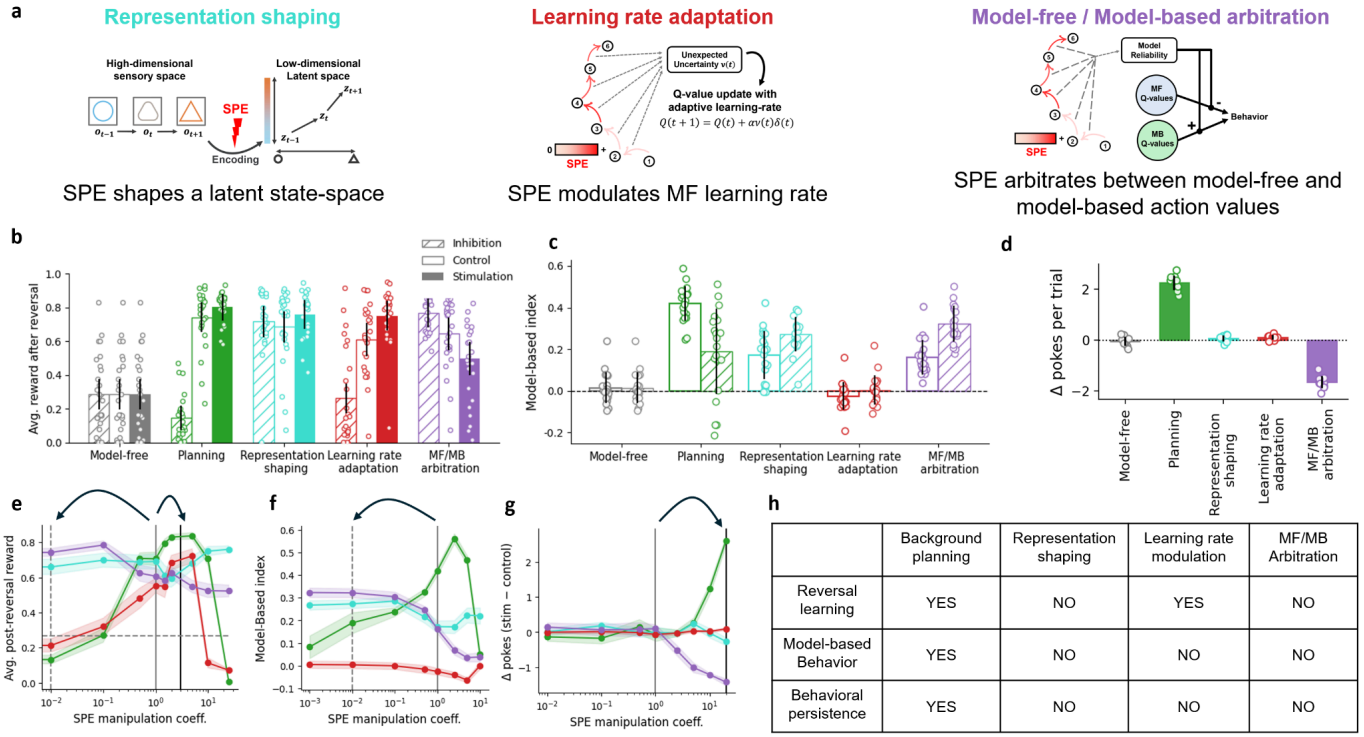
We proposed that serotonin, by signaling a state prediction error (SPE) throughout the brain, could support cognitive flexibility by acting as a teaching signal for a predictive model that is used for background planning. However, there are other ways by which a SPE signal could be used to support adaptive behavior (see Fig. 7a for illustration and methods for description of their implementation) :

Representation shaping. The predictive model loss can be backpropagated to the latent representation that serves as input to the model-free Q-learner. In such cases, SPE shapes latent representation to make them more easily predictable, potentially imbuing the model-free Q-learner with useful information about the structure of its environment (Ha and Schmidhuber 2018; Recanatesi et al. 2021).

Learning rate adaptation. By signaling the estimated uncertainty of the environment, the SPE of the predictive model can modulate the learning rate of a model-free agent : high uncertainty should lead to higher learning rate, while low uncertainty should lead to lower learning rate (Gershman 2015; Grossman et al. 2022).

Model-free(MF)/Model-based(MB) arbitration. By signaling the reliability of the predictive model, its SPE can be used to regulate the balance between model-free and model-based control (Daw et al. 2011, 2005; Gläscher et al. 2010).

Crucially, these alternative mechanisms make a different claim about serotonin's role: that it broadcasts an SPE signal which other adaptive systems read out, but is not itself the teaching signal for the predictive model. To test this claim against our proposal, we simulate SPE manipulation as affecting only the adaptive mechanisms (learning rate adaptation, representation shaping, arbitration weight), while leaving the predictive model's learning intact. If these mechanisms are enough to account for serotonin's effects on behavior, it would suggest that serotonin's role could be limited to driving adaptive mechanisms by signaling a SPE. If they fail, it supports the proposal that serotonin serves as a teaching signal for world model learning. Of course, serotonin could fulfill both roles simultaneously and we examine this possibility after testing each mechanism in isolation.



	Background planning	Representation shaping	Learning rate modulation	MF/MB Arbitration
Reversal learning	YES	NO	YES	NO
Model-based Behavior	YES	NO	NO	NO
Behavioral persistence	YES	NO	NO	NO

Figure 7. Background planning accounts for serotonin manipulation effect better than other SPE-driven adaptive mechanisms. **a.** Alternative SPE-driven adaptive mechanisms: representation shaping (left), learning rate adaptation (center), and MF/MB arbitration (right). **b.** Average reward after reversal under SPE inhibition, control, and stimulation across mechanisms for the reward reversal task. **c.** Model-based index during the two-step task across mechanisms, for control and serotonin inhibition. **d.** Change in pokes per trial across mechanisms during the foraging task. **e–g.** Effect of SPE manipulation coefficient on post-reversal reward in the reward reversal task (e), model-based index in the two-step task (f), and poke per trial in the foraging task (g) for each mechanism. **h.** Summary table: only background planning captures serotonin manipulation effects across all three paradigms.

We implement and simulate these SPE-driven adaptive mechanisms in the manipulation experiments. We first verify that, for appropriate parameters, these mechanisms can improve reward rate compared to a pure model-free baseline in the reward reversal, two-step, and foraging tasks (see Fig. S6a). We then examine whether they capture the qualitative effects of serotonergic manipulation. None of the alternative mechanisms reproduce the pattern of

manipulation effects across all three paradigms (Fig. 7b–d). Importantly, these failures are not due to particular choices of manipulation factor: plotting performance across a range of SPE scaling values reveals that these mechanisms simply do not show the expected sensitivity to SPE manipulation (Fig. 7e–g). Background planning was the only mechanism that captured the effects of serotonin manipulation across all three paradigms (Fig. 7h).

Because multiplexing is a ubiquitous phenomenon in biological systems, serotonin may support multiple SPE-driven mechanisms simultaneously by broadcasting the same signal to different brain areas. On the reversal learning task, we add representation shaping, arbitration or learning rate modulation on top of background planning and find that these additions preserve the qualitative match to behavioral data while increasing reward rate (Fig. S6b), suggesting that biological agents may combine these mechanisms for optimal performance.

Overall, we find that the SPE-driven mechanism that best accounts for the effects of serotonin manipulation is its role as a teaching signal for a predictive model that shapes behavior through background planning. This mechanism also accounts qualitatively for other observations, including planning impairment (Huys et al. 2012), spatial memory effects (Teixeira et al. 2018) and learning rate adaptation (Grossman et al. 2022; Iigaya et al. 2018).

Discussion

Serotonin as a state prediction error

In this work, we proposed that serotonin signals a state prediction error (SPE) that serves as the teaching signal for a predictive world model, and that this world model shapes goal-directed behavior through background planning. Using neural data from a novel experimental paradigm, we found that dorsal raphe serotonergic activity strikingly correlates with a SPE signal, sharing key properties such as modulation by uncertainty and behavioral relevance. Furthermore, by simulating causal manipulation of this signal in our network model, we found that disrupting SPE impaired cognitive flexibility across different behavioral paradigms, mirroring the effects of serotonergic manipulation observed experimentally.

We also considered alternative ways by which a SPE signal could support adaptive behavior, including representation shaping, learning rate adaptation, and model-free/model-based arbitration. None of these mechanisms, when tested in isolation, captured the full pattern of serotonin manipulation effects. Background planning was necessary. Taken together, these results suggest that serotonin may be best understood as a teaching signal for predictive learning in the brain, accounting for both its activity patterns and its contribution to cognitive flexibility.

Relevance modulation and value coding

Our model includes a relevance modulation mechanism, where SPE is amplified for states with high absolute value and attenuated for low absolute value states. This mechanism was introduced to account for resource constraints on predictive learning, and it successfully explains why serotonergic activity in our photometry data was modulated by reward expectation, an observation that might otherwise suggest serotonin encodes value. Under our framework, what appears to be value coding is instead a relevance-weighted state prediction error.

In our simulations, relevance modulation reduced wiring cost without impairing behavioral performance, suggesting it optimizes representational capacity. However, its behavioral role remained limited in our simple gridworld environment. We expect relevance modulation to become more important in complex, naturalistic environments with many irrelevant features, a prediction that remains to be tested.

Reward as a component of state

A key theoretical commitment of our framework is that reward onset is treated as a component of the state representation, rather than a separate reward signal. This choice has important implications: it means that changes in reward contingencies generate SPE, allowing our model to account for serotonin's effects on pure reward reversal learning and behavioral persistence, and not just structural changes in the environment.

This approach aligns with a broader shift in thinking about prediction errors in the brain. Traditionally, sensory prediction errors and reward prediction errors have been treated as qualitatively distinct signals. However, recent work suggests this distinction may be less sharp than previously assumed. For example, dopamine neurons appear to signal generalized prediction errors encompassing both sensory and reward features (Takahashi et al. 2017; Gershman et al. 2024), not just pure reward prediction errors. Our treatment of reward as a predicted feature of the world, rather than a special teaching signal, fits within this emerging framework.

Toward a general computational principle for neuromodulation

Our theory of serotonin invites comparison to the classical view of dopamine. Dopamine is widely believed to signal a scalar reward prediction error (RPE), the teaching signal of reinforcement learning algorithms, for which many neural substrates have been identified (Niv 2009). By contrast, we propose that serotonin signals a *vectorial* state prediction error, a teaching signal for predictive learning.

A closer parallel can be drawn with a recent theory of dopamine function by (Wang et al. 2018). Similarly to our theory, their model explains the function of dopamine across both long and short timescales. This fits well with known properties of neuromodulatory systems: they shape network structure via synaptic plasticity over long timescales, while also providing context-dependent flexibility to otherwise structurally fixed circuits over shorter timescales (Dayan 2012). In their work, dopamine shapes the prefrontal cortex into a meta-reinforcement learning system over long timescales, while also supplying it with a RPE signals that drive its computations over a shorter timescale.

We propose that serotonin shapes the structure of the brain over long timescales to support predictive learning at different levels of the brain hierarchy, while it could also flexibly modulate behavior over short timescales to promote cognitive flexibility. This dual-timescales function may reflect a more general computational principle of neuromodulation. Applying this principle to other neuromodulatory systems might be a fruitful research direction, with the potential to provide mechanistic explanation to both neuromodulators activity and their causal effects on behavior.

Multiplexing of SPE-driven mechanisms

Our results indicate that background planning is necessary to account for the behavioral effects of serotonin manipulation. However, we also found that adding other SPE-driven mechanisms, such as representation shaping or MB/MF arbitration on top of planning preserved the qualitative match to behavioral data while improving overall reward rate. This suggests that serotonin may serve multiple functions simultaneously by broadcasting the same SPE signal to different brain areas, each implementing a distinct adaptive mechanism.

This multiplexing view is consistent with the known heterogeneity of serotonin's projections and effects across brain regions. It also suggests that debates about serotonin's "true" function may be partly misguided: serotonin may not have a single function, but rather provide a common teaching signal that is used differently by different circuits.

Effect of drugs that act on the serotonergic system

Our framework offers a unified lens through which to interpret the effects of drugs that act on the serotonergic system. Classical psychedelics such as psilocybin and LSD are potent 5-HT2A agonists that directly activate postsynaptic serotonin receptors. This sustained receptor activation could functionally mimic a state in which phasic SPE signals are compressed against an elevated baseline, much as elevated tonic dopamine compresses phasic reward prediction error signaling (Grace 1991). The brain's predictive model would then receive a weakened teaching signal, causing priors to dominate over sensory evidence. This is consistent with the structured nature of psychedelic hallucinations, which reflect prior-driven pattern completion rather than random noise, and resonates with the REBUS model (Carhart-Harris and Friston 2019) while grounding it in a more concrete computational mechanism. Moreover, if predictive structures require ongoing error-driven recalibration to be maintained, sustained SPE suppression could gradually destabilize them, potentially enabling their subsequent revision. This could underlie the lasting therapeutic benefits observed in depression and anxiety, where maladaptive predictive patterns are thought to be central to pathology.

Other serotonergic drugs alter tonic and phasic serotonin dynamics more directly, though the consequences for SPE computation remain unclear. MDMA, which acts primarily by triggering massive serotonin release, has been shown to reopen critical periods of social reward learning (Nardou et al. 2019), suggesting large-scale reorganization of predictive circuits through prolonged increase of serotonergic signaling, a plasticity phenomenon compatible with our framework.

Limitations

Multidimensional nature of a SPE signal. One limitation of our study is that fiber photometry only captures a scalar signal, allowing us to compare serotonergic activity to the norm of the SPE vector but not its full multidimensional structure. This collapses the rich content of the SPE, whose individual dimensions should map to relevant features of the environment. Our model predicts that in single-cell recordings of dorsal raphe serotonergic neurons, distinct subpopulations should be tuned to different environmental features. Preliminary evidence for this exists in prior single-cell studies (Paquelet et al. 2022; Ranade and Mainen 2009), which report serotonergic tuning to sensory, reward, or motor events. However, those studies were not

designed to directly test our hypothesis. We plan to design new experiments in multisensory decision-making tasks, where different sensory modalities predict different rewards, providing an ideal setting to observe a multidimensional SPE encoded across the dorsal raphe nucleus serotonergic population activity.

Leverage properties of the SPE signal to inform adaptive mechanisms. In our current framework, fast-timescale adaptive mechanisms such as model-based/model-free arbitration depend only on the norm of the SPE signal. Future work should extend this framework to leverage the full information contained in this error vector for dynamic behavioral modulation. Rather than hand-designing the interaction between the predictive model and reinforcement learner, a promising direction would be to let the system learn how information should flow between modules. It would be interesting to see whether identifiable, SPE-driven adaptive mechanisms emerge naturally in more unconstrained architectures.

Heterogenous role of serotonin in the brain. While our work focuses on serotonin's role in predictive model learning and cognitive flexibility, serotonin is known to have a wide range of functions including roles in development, sleep, emotion regulation, and many other processes (Bacqué-Cazenave et al. 2020; Ligneul and Mainen 2023), that our model does not aim to explain. To rigorously test our framework alongside alternative models of serotonergic function, specific "strong inference" experiments should be designed in collaboration between laboratories. Though no single theory is likely to capture the full breadth of serotonin's function, each new hypothesis can provide valuable insight into brain computations.

References

Valence-dependent influence of serotonin depletion on model-based choice strategy | Molecular Psychiatry

Ahmadlou, Mehran, Maryam Yasamin Shirazi, Pan Zhang, et al. 2025. "A Subcortical Switchboard for Perseverative, Exploratory and Disengaged States." *Nature* 641 (8061): 151–61. https://doi.org/10.1038/s41586-025-08672-1.

Anand, Ankesh, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R. Devon Hjelm. 2020. "Unsupervised State Representation Learning in Atari." arXiv:1906.08226. Preprint, arXiv, November 5. https://doi.org/10.48550/arXiv.1906.08226.

Bacqué-Cazenave, Julien, Rahul Bharatiya, Grégory Barrière, et al. 2020. "Serotonin in Animal Cognition and Behavior." *International Journal of Molecular Sciences* 21 (5): 5. https://doi.org/10.3390/ijms21051649.

Behrens, Timothy E. J., Timothy H. Muller, James C. R. Whittington, et al. 2018. "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior." *Neuron* 100 (2): 490–509. https://doi.org/10.1016/j.neuron.2018.10.002.

Beliveau, Vincent, Melanie Ganz, Ling Feng, et al. 2016. "A High-Resolution in Vivo Atlas of the Human Brain's Serotonin System." Research Articles. *Journal of Neuroscience*, ahead of print, November 17. https://doi.org/10.1523/JNEUROSCI.2830-16.2016.

Bromberg-Martin, Ethan S., Okihide Hikosaka, and Kae Nakamura. 2010. "Coding of Task Reward Value in the Dorsal Raphe Nucleus." Articles. *Journal of Neuroscience* 30 (18): 6262–72. https://doi.org/10.1523/JNEUROSCI.0015-10.2010.

Carhart-Harris, R. L., and K. J. Friston. 2019. "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics." Review Article. *Pharmacological Reviews* 71 (3): 316–44. https://doi.org/10.1124/pr.118.017160.

Clarke, H. F., J. W. Dalley, H. S. Crofts, T. W. Robbins, and A. C. Roberts. 2004. "Cognitive Inflexibility After Prefrontal Serotonin Depletion." *Science* 304 (5672): 878–80. https://doi.org/10.1126/science.1094987.

Cohen, Jeremiah Y., Mackenzie W. Amoroso, and Naoshige Uchida. 2015. "Serotonergic Neurons Signal

Reward and Punishment on Multiple Timescales." *eLife* 4 (February): e06346. https://doi.org/10.7554/eLife.06346.
Dabney, Will, André Barreto, Mark Rowland, et al. 2021. "The Value-Improvement Path: Towards Better Representations for Reinforcement Learning." arXiv:2006.02243. Preprint, arXiv, January 4. https://doi.org/10.48550/arXiv.2006.02243.
Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. "Model-Based Influences on Humans' Choices and Striatal Prediction Errors." *Neuron* 69 (6): 1204–15. https://doi.org/10.1016/j.neuron.2011.02.027.
Daw, Nathaniel D., Sham Kakade, and Peter Dayan. 2002. "Opponent Interactions between Serotonin and Dopamine." *Neural Networks* 15 (4): 603–16. https://doi.org/10.1016/S0893-6080(02)00052-7.
Daw, Nathaniel D., Yael Niv, and Peter Dayan. 2005. "Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control." *Nature Neuroscience* 8 (12): 12. https://doi.org/10.1038/nn1560.
Dayan, Peter. 2012. "Twenty-Five Lessons from Computational Neuromodulation." *Neuron* 76 (1): 240–56. https://doi.org/10.1016/j.neuron.2012.09.027.
Fang, Ching, and Kimberly L. Stachenfeld. 2024. "Predictive Auxiliary Objectives in Deep RL Mimic Learning in the Brain." arXiv:2310.06089. Preprint, arXiv, October 29. https://doi.org/10.48550/arXiv.2310.06089.
Feng, Yang-Yang, Ethan S. Bromberg-Martin, and Ilya E. Monosov. 2024. "Dorsal Raphe Neurons Integrate the Values of Reward Amount, Delay, and Uncertainty in Multi-Attribute Decision-Making." *Cell Reports* 43 (6): 114341. https://doi.org/10.1016/j.celrep.2024.114341.
Fonseca, Madalena S., Masayoshi Murakami, and Zachary F. Mainen. 2015. "Activation of Dorsal Raphe Serotonergic Neurons Promotes Waiting but Is Not Reinforcing." *Current Biology* 25 (3): 306–15. https://doi.org/10.1016/j.cub.2014.12.002.
Gabhart, Kaitlyn M., Yihan (Sophy) Xiong, and André M. Bastos. 2025. "Predictive Coding: A More Cognitive Process than We Thought?" *Trends in Cognitive Sciences* 0 (0). https://doi.org/10.1016/j.tics.2025.01.012.
Gershman, Samuel J. 2015. "A Unifying Probabilistic View of Associative Learning." *PLOS Computational Biology* 11 (11): e1004567. https://doi.org/10.1371/journal.pcbi.1004567.
Gershman, Samuel J., John A. Assad, Sandeep Robert Datta, et al. 2024. "Explaining Dopamine through Prediction Errors and Beyond." *Nature Neuroscience* 27 (9): 1645–55. https://doi.org/10.1038/s41593-024-01705-4.
Gläscher, Jan, Nathaniel Daw, Peter Dayan, and John P. O'Doherty. 2010. "States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning." *Neuron* 66 (4): 585–95. https://doi.org/10.1016/j.neuron.2010.04.016.
Grace, A. A. 1991. "Phasic versus Tonic Dopamine Release and the Modulation of Dopamine System Responsivity: A Hypothesis for the Etiology of Schizophrenia." *Neuroscience* 41 (1): 1–24. https://doi.org/10.1016/0306-4522(91)90196-u.
Grossman, Cooper D., Bilal A. Bari, and Jeremiah Y. Cohen. 2022. "Serotonin Neurons Modulate Learning Rate through Uncertainty." *Current Biology* 32 (3): 586-599.e7. https://doi.org/10.1016/j.cub.2021.12.006.
Ha, David, and Jürgen Schmidhuber. 2018. *World Models*. March 28. https://doi.org/10.5281/zenodo.1207631.
Hamada, Hiro Taiyo, Yoshifumi Abe, Norio Takata, Masakazu Taira, Kenji F. Tanaka, and Kenji Doya. 2024. "Optogenetic Activation of Dorsal Raphe Serotonin Neurons Induces Brain-Wide Activation." *Nature Communications* 15 (1): 4152. https://doi.org/10.1038/s41467-024-48489-6.
Harkin, Emerson F., Cooper D. Grossman, Jeremiah Y. Cohen, Jean-Claude Béïque, and Richard Naud. 2023. *Serotonin Predictively Encodes Value*. Preprint. Neuroscience. https://doi.org/10.1101/2023.09.19.558526.
Huys, Quentin J. M., Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. 2012. "Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees." *PLOS Computational Biology* 8 (3): e1002410. https://doi.org/10.1371/journal.pcbi.1002410.
Hyun, Jung Ho, Patrick Hannan, Hideki Iwamoto, Randy D. Blakely, and Hyung-Bae Kwon. 2023. "Serotonin in the Orbitofrontal Cortex Enhances Cognitive Flexibility." Preprint, bioRxiv, March 9. https://doi.org/10.1101/2023.03.09.531880.
Iigaya, Kiyohito, Madalena S. Fonseca, Masayoshi Murakami, Zachary F. Mainen, and Peter Dayan.

2018. “An Effect of Serotonergic Stimulation on Learning Rates for Rewards Apparent after Long Intertrial Intervals.” *Nature Communications* 9 (1): 1. https://doi.org/10.1038/s41467-018-04840-2.
Jaderberg, Max, Volodymyr Mnih, Wojciech Marian Czarnecki, et al. 2016. “Reinforcement Learning with Unsupervised Auxiliary Tasks.” arXiv:1611.05397. Preprint, arXiv, November 16. https://doi.org/10.48550/arXiv.1611.05397.
Jensen, Kristopher T., Guillaume Hennequin, and Marcelo G. Mattar. 2023. “A Recurrent Network Model of Planning Explains Hippocampal Replay and Human Behavior.” Preprint, bioRxiv, January 19. https://doi.org/10.1101/2023.01.16.523429.
Jensen, Kristopher T., Guillaume Hennequin, and Marcelo G. Mattar. 2024. “A Recurrent Network Model of Planning Explains Hippocampal Replay and Human Behavior.” *Nature Neuroscience* 27 (7): 1340–48. https://doi.org/10.1038/s41593-024-01675-7.
Kanen, Jonathan W., Annemieke M. Apergis-Schoute, Robyn Yellowlees, et al. 2021. “Serotonin Depletion Impairs Both Pavlovian and Instrumental Reversal Learning in Healthy Humans.” *Molecular Psychiatry* 26 (12): 12. https://doi.org/10.1038/s41380-021-01240-9.
Lehky, Sidney R., and Terrence J. Sejnowski. 1988. “Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields.” *Nature* 333 (6172): 452–54. https://doi.org/10.1038/333452a0.
Lesch, Klaus-Peter, and Jonas Waider. 2012. “Serotonin in the Modulation of Neural Plasticity and Networks: Implications for Neurodevelopmental Disorders.” *Neuron* 76 (1): 175–91. https://doi.org/10.1016/j.neuron.2012.09.013.
Li, Yi, Weixin Zhong, Daqing Wang, et al. 2016. “Serotonin Neurons in the Dorsal Raphe Nucleus Encode Reward Signals.” *Nature Communications* 7 (1): 10503. https://doi.org/10.1038/ncomms10503.
Ligneul, Romain, and Zachary F. Mainen. 2023. “Serotonin.” *Current Biology* 33 (23): R1216–21. https://doi.org/10.1016/j.cub.2023.09.068.
Lottem, Eran, Dhruba Banerjee, Pietro Vertechi, Dario Sarra, Matthijs oude Lohuis, and Zachary F. Mainen. 2018. “Activation of Serotonin Neurons Promotes Active Persistence in a Probabilistic Foraging Task.” *Nature Communications* 9 (1): 1000. https://doi.org/10.1038/s41467-018-03438-y.
Ludvig, Elliot A., Richard S. Sutton, and E. James Kehoe. 2008. “Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System.” *Neural Computation* 20 (12): 3034–54. https://doi.org/10.1162/neco.2008.11-07-654.
Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. “Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex.” *Nature* 503 (7474): 7474. https://doi.org/10.1038/nature12742.
Matias, Sara, Eran Lottem, Guillaume P. Dugué, and Zachary F. Mainen. 2017. “Activity Patterns of Serotonin Neurons Underlying Cognitive Flexibility.” *eLife* 6 (March): e20552. https://doi.org/10.7554/eLife.20552.
Mattar, Marcelo G., and Nathaniel D. Daw. 2018. “Prioritized Memory Access Explains Planning and Hippocampal Replay.” *Nature Neuroscience* 21 (11): 11. https://doi.org/10.1038/s41593-018-0232-z.
Mattar, Marcelo G., and Máté Lengyel. 2022. “Planning in the Brain.” *Neuron* 110 (6): 914–34. https://doi.org/10.1016/j.neuron.2021.12.018.
Miller, E. K., and J. D. Cohen. 2001. “An Integrative Theory of Prefrontal Cortex Function.” *Annual Review of Neuroscience* 24: 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167.
Nardou, Romain, Eastman M. Lewis, Rebecca Rothhaas, et al. 2019. “Oxytocin-Dependent Reopening of a Social Reward Learning Critical Period with MDMA.” *Nature* 569 (7754): 116–20. https://doi.org/10.1038/s41586-019-1075-9.
Niv, Yael. 2009. “Reinforcement Learning in the Brain.” *Journal of Mathematical Psychology*, Special Issue: Dynamic Decision Making, vol. 53 (3): 139–54. https://doi.org/10.1016/j.jmp.2008.12.005.
Ohmura, Yu, Kentaro Iwami, Srikanta Chowdhury, et al. 2021. “Disruption of Model-Based Decision Making by Silencing of Serotonin Neurons in the Dorsal Raphe Nucleus.” *Current Biology* 31 (11): 2446-2454.e5. https://doi.org/10.1016/j.cub.2021.03.048.
Paquelet, Grace E., Kassandra Carrion, Clay O. Lacefield, Pengcheng Zhou, René Hen, and Bradley R. Miller. 2022. “Single-Cell Activity and Network Properties of Dorsal Raphe Nucleus Serotonin Neurons during Emotionally Salient Behaviors.” *Neuron* 110 (16): 2664-2679.e8. https://doi.org/10.1016/j.neuron.2022.05.015.
Pfeiffer, Brad E., and David J. Foster. 2013. “Hippocampal Place-Cell Sequences Depict Future Paths to

Remembered Goals.” *Nature* 497 (7447): 74–79. https://doi.org/10.1038/nature12112.
Ranade, Sachin P., and Zachary F. Mainen. 2009. “Transient Firing of Dorsal Raphe Neurons Encodes Diverse and Specific Sensory, Motor, and Reward Events.” *Journal of Neurophysiology* 102 (5): 3026–37. https://doi.org/10.1152/jn.00507.2009.
Rao, Rajesh P. N. 2022. “A Sensory-Motor Theory of the Neocortex Based on Active Predictive Coding.” Preprint, bioRxiv, December 31. https://doi.org/10.1101/2022.12.30.522267.
Recanatesi, Stefano, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. 2021. “Predictive Learning as a Network Mechanism for Extracting Low-Dimensional Latent Space Representations.” *Nature Communications* 12 (1): 1. https://doi.org/10.1038/s41467-021-21696-1.
Saravanan, Varun, Gordon J. Berman, and Samuel J. Sober. 2020. “Application of the Hierarchical Bootstrap to Multi-Level Data in Neuroscience.” *Neurons, Behavior, Data Analysis and Theory* 3 (5): https://nbdt.scholasticahq.com/article/13927-application-of-the-hierarchical-bootstrap-to-multi-level-data-in-neuroscience.
Soltani, Alireza, and Alicia Izquierdo. 2019. “Adaptive Learning under Expected and Unexpected Uncertainty.” *Nature Reviews Neuroscience* 20 (10): 10. https://doi.org/10.1038/s41583-019-0180-y.
Soltani, Alireza, and Etienne Koechlin. 2022. “Computational Models of Adaptive Behavior and Prefrontal Cortex.” *Neuropsychopharmacology* 47 (1): 58–71. https://doi.org/10.1038/s41386-021-01123-1.
Stachenfeld, Kimberly L., Matthew M. Botvinick, and Samuel J. Gershman. 2017. “The Hippocampus as a Predictive Map.” *Nature Neuroscience* 20 (11): 1643–53. https://doi.org/10.1038/nn.4650.
Stroud, Jake Patrick, Michal Wojcik, Kristopher Torp Jensen, et al. 2025. “Effects of Noise and Metabolic Cost on Cortical Task Representations.” *eLife* 13 (January): RP94961. https://doi.org/10.7554/eLife.94961.
Sutton, Richard S. 1991. “Dyna, an Integrated Architecture for Learning, Planning, and Reacting.” *ACM SIGART Bulletin* 2 (4): 160–63. https://doi.org/10.1145/122344.122377.
Sutton, Richard S., and Andrew G. Barto. n.d. *Reinforcement Learning: An Introduction*.
Taira, Masakazu, and Melissa J. Sharpe. 2025. “Complementary Roles of Serotonin and Dopamine in Model-Based Learning.” *Current Opinion in Behavioral Sciences* 61 (February): 101464. https://doi.org/10.1016/j.cobeha.2024.101464.
Takahashi, Yuji K., Hannah M. Batchelor, Bing Liu, Akash Khanna, Marisela Morales, and Geoffrey Schoenbaum. 2017. “Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards.” *Neuron* 95 (6): 1395-1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025.
Teixeira, Catia M., Zev B. Rosen, Deepika Suri, et al. 2018. “Hippocampal 5-HT Input Regulates Memory Formation and Schaffer Collateral Excitation.” *Neuron* 98 (5): 992-1004.e4. https://doi.org/10.1016/j.neuron.2018.04.030.
Wang, Jane X., Zeb Kurth-Nelson, Dharshan Kumaran, et al. 2018. “Prefrontal Cortex as a Meta-Reinforcement Learning System.” *Nature Neuroscience* 21 (6): 860–68. https://doi.org/10.1038/s41593-018-0147-8.
Whittington, James C. R., Timothy H. Muller, Shirley Mark, et al. 2020. “The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation.” *Cell* 183 (5): 1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024.
Worbe, Y., S. Palminteri, G. Savulich, et al. 2016. “Valence-Dependent Influence of Serotonin Depletion on Model-Based Choice Strategy.” *Molecular Psychiatry* 21 (5): 624–29. https://doi.org/10.1038/mp.2015.46.
Zipser, David, and Richard A. Andersen. 1988. “A Back-Propagation Programmed Network That Simulates Response Properties of a Subset of Posterior Parietal Neurons.” *Nature* 331 (6158): 679–84. https://doi.org/10.1038/331679a0.

Supplementary Figures

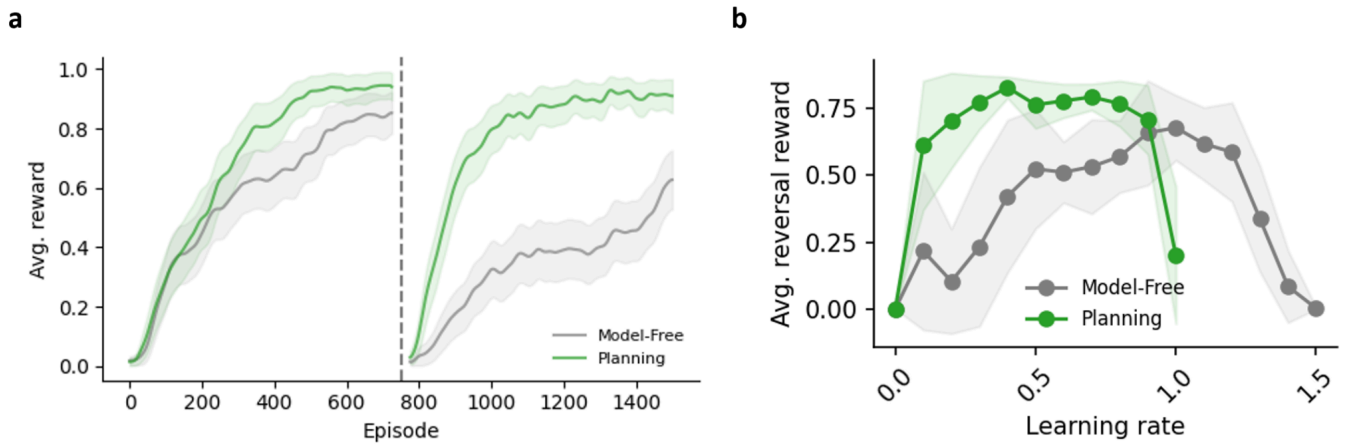


Figure S1. Additional details on the interaction between predictive learning and goal-directed behavior in the door reversal task.. **a.** Learning curves for the predictive agent and the model-free agent during the 750 episodes of initial learning and the 750 episodes of reversal learning. **b.** Average post-reversal reward the predictive agent with a fixed $\alpha_q = 0.5$ and varying α_p, superimposed with the reversal performance for a model-free agent with different α_q values.

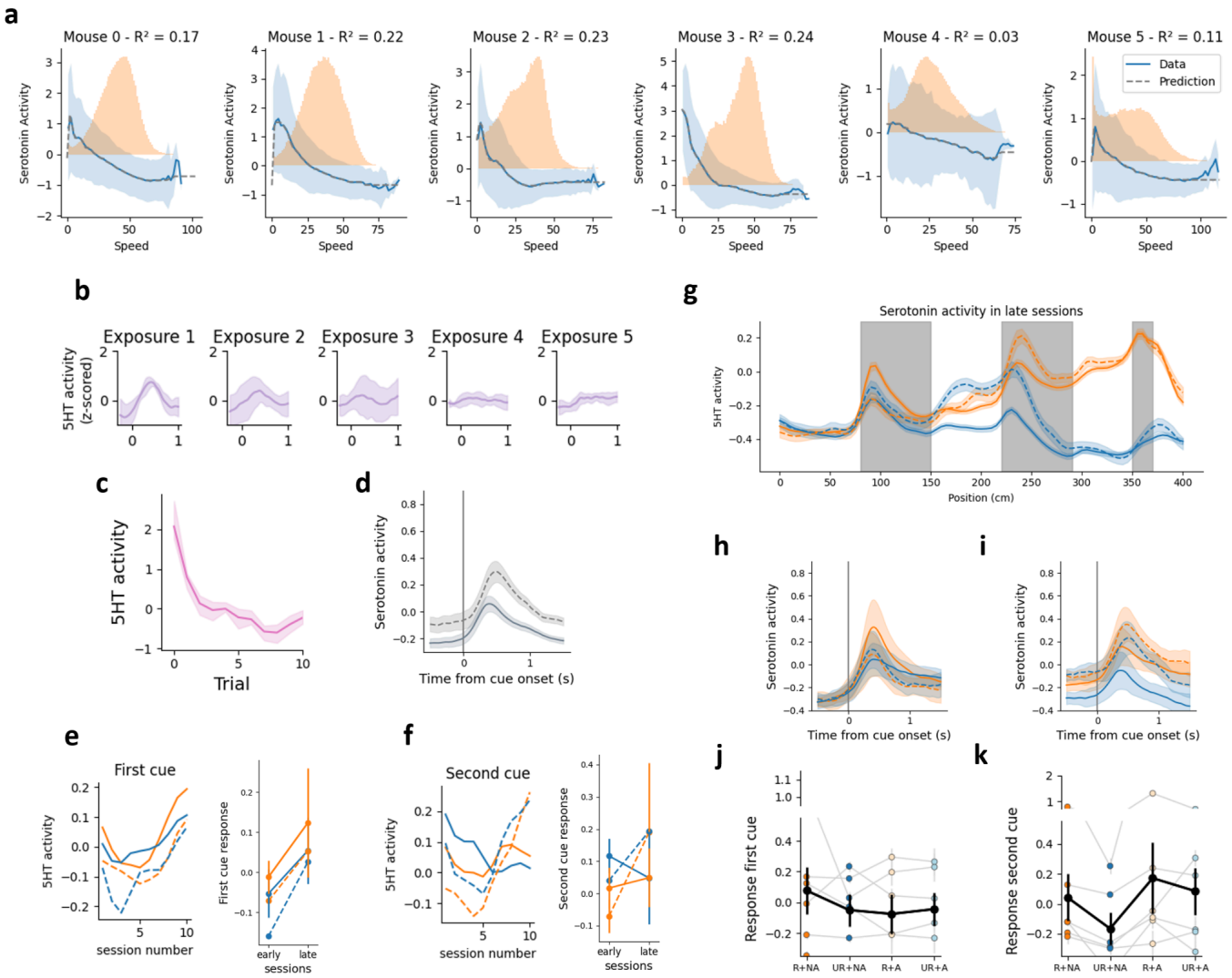


Figure S2. Relationship between speed and motion-dependent serotonin activity in a cue-reward association task. **a.** Serotonin activity as a function of running speed for each mouse (N = 6). Blue: raw data; dashed line: linear regression prediction. Orange histogram : density of mouse speed during the task. R^2 values indicate the variance in serotonin activity explained by running speed. **b.** Motion-dependent (MD) serotonin responses to the first five exposures of novel cues during the first session (mean ± SEM across mice). **c.** Trial-by-trial decay of the MD serotonin response to cue onset across the first trials of the first session, pooled across all cues and corridors. **d.** MD serotonin responses aligned to second cue onset in late sessions, shown separately for ambiguous (dashed) and non-ambiguous (solid) trials. **e.** Left: evolution of mean MD serotonin response at cue 1 across sessions for each trial type (color code as in main Fig. 4). Right: comparison of MD cue 1 responses between early (sessions 1–3) and late (sessions 7–10) sessions. **f.** Same as (e) for cue 2. **g.** Time course of MD serotonin activity as a function of corridor position during late sessions (mean ± SEM across mice) for each corridor type. Grey shaded regions indicate cue 1, cue 2 and reward zones. **h.** MD serotonin activity aligned to cue 1 onset during late sessions for each trial type. **i.** MD serotonin activity aligned to cue 2 onset during late sessions for each trial type. **j.** Mean MD serotonin responses to cue 1 across trial types during late sessions. Individual mouse data are overlaid as colored dots. **k.** Same as (j) for cue 2.

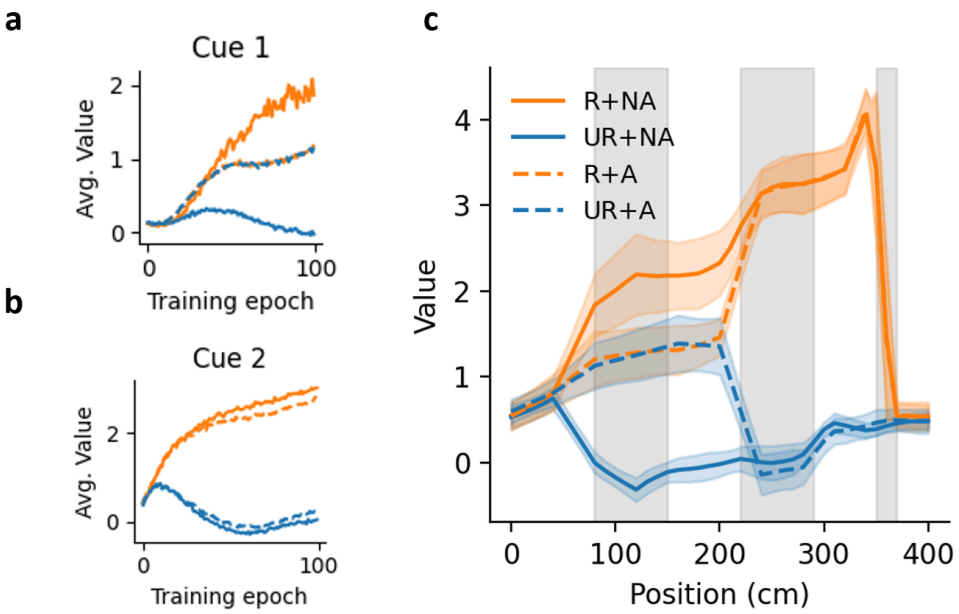


Figure S3. Learning of cue values during the task. **a.** Evolution of the value signal at cue 1 (averaged between 80cm and 150cm) during the learning of the task for the four corridors. **b.** Evolution of the value signal at cue 2 (averaged between 220cm and 290cm). **c.** Final value signal after learning.

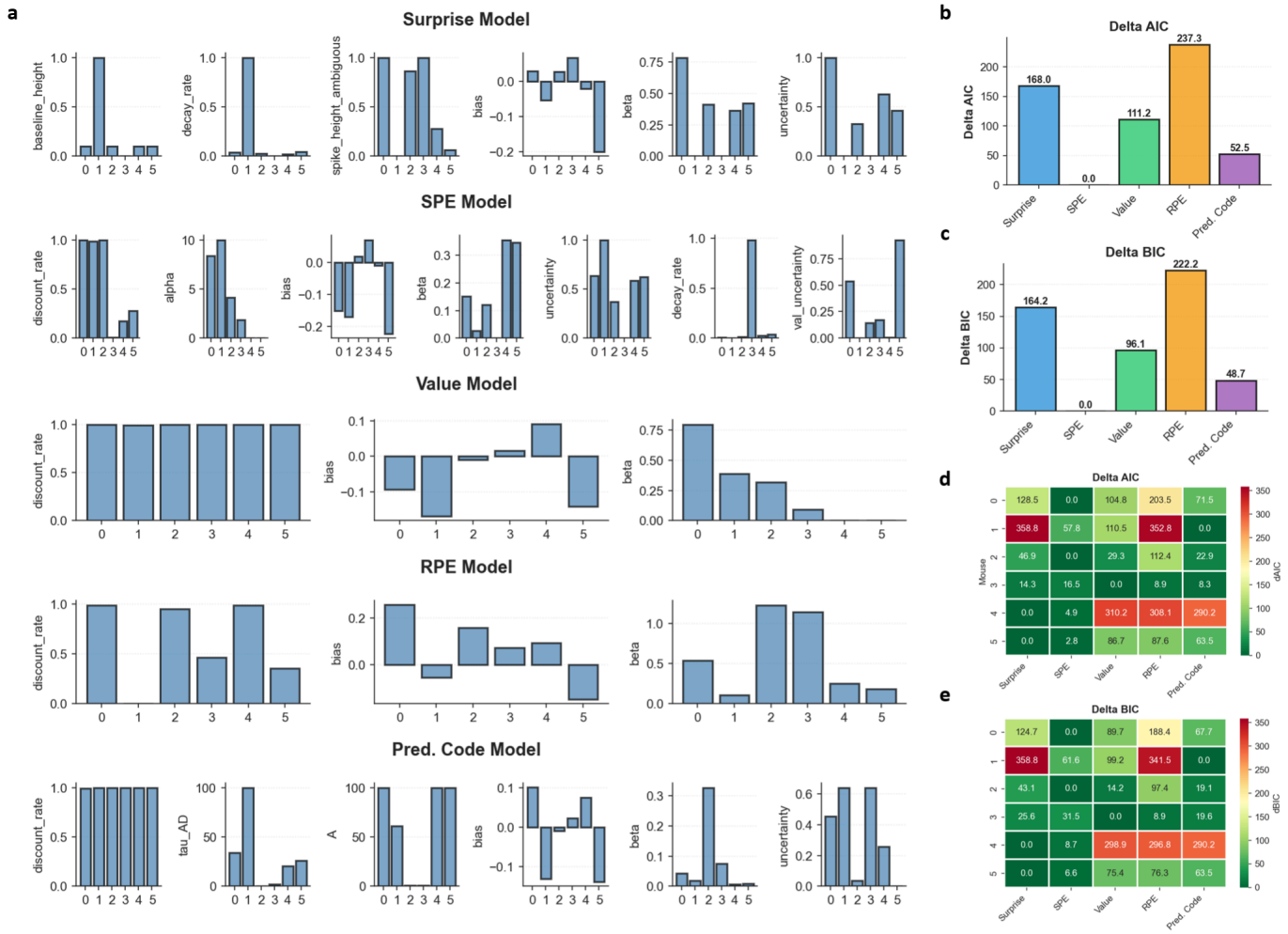


Figure S4. Additional details on model fitting to serotonin data. **a.** Parameter estimate for each model, for each of the six mice. **b.** Difference between the AIC for the SPE model and the other models for the population-averaged fit. **c.** Same for the BIC. **d.** For each mouse, difference between the lowest AIC of the models and the other models AIC. **e.** Same for the BIC.

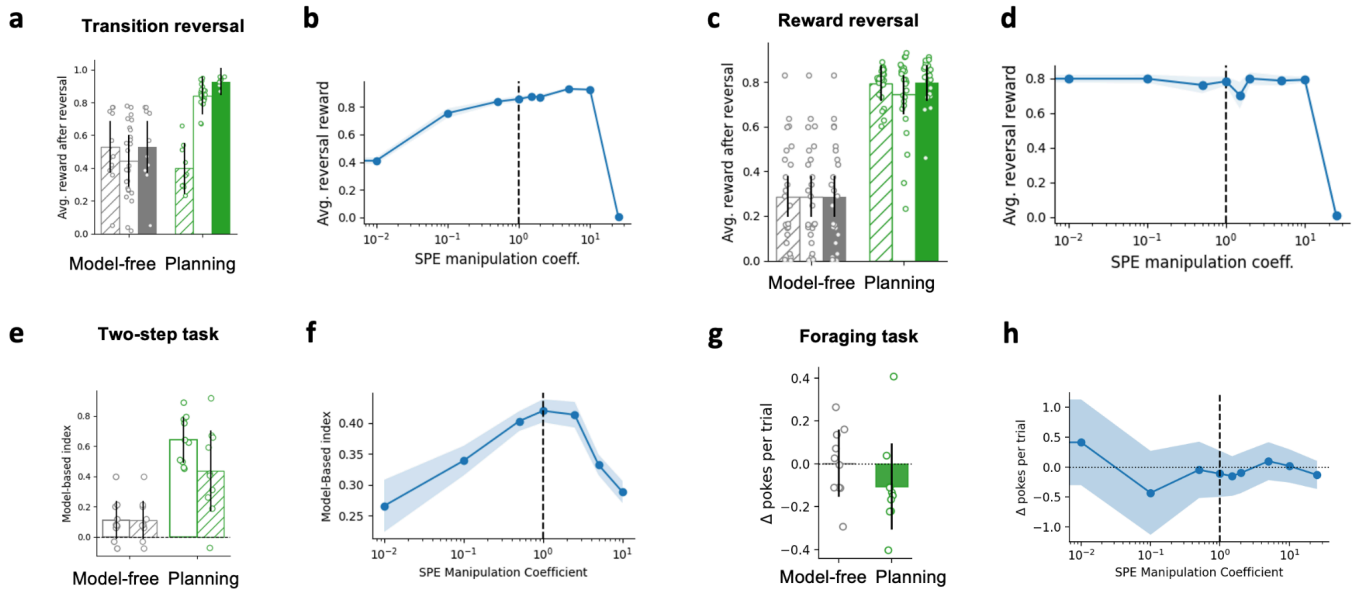


Figure S5. Effect on SPE manipulation when reward is not a feature of the state space. **a.** Trials to criterion in the transition reversal task for model-free and planning agents under SPE inhibition (hatched), control (solid gray/dark green), and stimulation (solid light green). **b.** Average reversal reward as a function of SPE manipulation coefficient. The reversal performance remains sensitive to manipulations of the SPE scaling. **c.** Trials to criterion in the reward reversal task for model-free and planning agents under SPE inhibition (hatched), control (solid gray/dark green), and stimulation (solid light green). **d.** Average reversal reward as a function of SPE manipulation coefficient. The reversal performance becomes largely sensitive to manipulations of the SPE scaling. **e.** Model-based index in the two-step task for model-free and planning agents under SPE inhibition (hatched) and control (solid). Individual runs of the model are overlaid as dots. **f.** Model-based index as a function of SPE manipulation coefficient. The model-based contribution remains sensitive to manipulations of the SPE scaling. **g.** Change in pokes per trial (stimulated minus control) in the probabilistic foraging task for model-free and planning agents. The persistence effect observed in Figure 6h is abolished when reward is not part of the state space. **h.** Change in pokes per trial as a function of SPE manipulation coefficient. No consistent increase in persistence is observed with SPE stimulation. Shaded areas indicate SEM across agents. Dashed vertical lines in (b), (d), and (f) indicate the unmanipulated control condition (coefficient = 1).

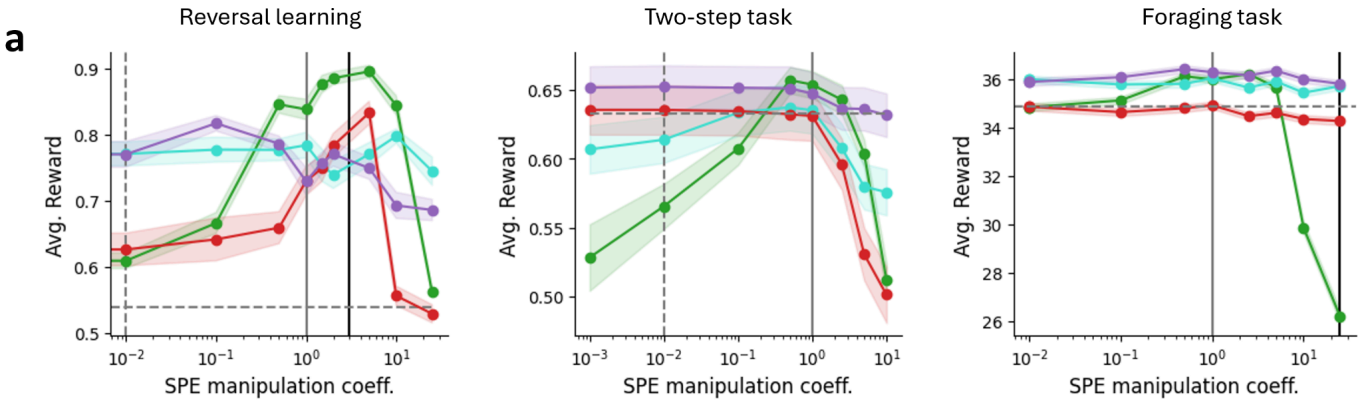


Figure S6. SPE-driven adaptive mechanisms optimize average reward for the different tasks. **a.**

Methods

SPE-driven adaptive mechanisms

Background planning

The memory buffer, from which background planning samples K starting states, is a replay buffer of fixed size 1000 with FIFO, and the state sampling is uniform.

Representation shaping

We train a single-hidden-layer encoder network with ReLU nonlinearity jointly with the predictive model. The encoder maps the raw state observation s_t to a latent representation $z_t = ReLU(W_{enc} \cdot s_t + b_{enc})$. A per-action linear predictor maps z_t and action a to a predicted next state $\hat{s}_{t+1} = W^a_{pred} \cdot z_t + b^a_{pred}$. The predictive loss $L_{pred} = ||\hat{s}_t - s_t||^2 = ||SPE_t||^2$ is minimized by gradient descent, with gradients flowing through the encoder to shape the latent representations. Separately, a linear Q-head reads out action values from z_t : $Q(s_t, a) = W^a_q \cdot z_t + b^a_q$. Crucially, the Q-learning loss gradient is not backpropagated through the encoder, ensuring that representation shaping is driven entirely by the predictive objective.

Learning rate adaptation

Theoretical works (ref) have exhibited optimal ways to flexibly adjust the learning rate of an RL agent over time, to account for the stability of the environment. In a stable environment, the learning rate should be small, allowing for a finer estimation of state values. In a changing environment, the learning rate should be high, allowing for faster adaptation of behavior. To estimate the stability of the environment, an agent can rely on the state prediction errors of its internal world model. If the errors suddenly become large, it means that the environment is unstable, and the learning rate should be increased. Conversely, if the errors are low, the environment is stable and the learning rate should decrease. Following this principle, we use the state prediction error spe_t to dynamically modulate the learning rate α_t of our Q-learning network. The agent maintains a predictive model trained on real transitions to provide the SPE signal. We define the recent average state prediction error by $\overline{SPE_t} = \frac{1}{N_{hist}} \sum_{i=1}^{N_{hist}} SPE_{t-i}$ with N_{hist} the size of the averaging window, fixed to 100 in all simulations, such that $\alpha_t = \alpha * (1 + \overline{SPE_t} * \beta_{mod})$, with β_{mod} a modulation factor which depends on the simulation.

Model-free / model-based strategy arbitration

The brain contains multiple systems for behavioral choice (ref), which are often partitioned in two : a model-free (MF) system and a model-based (MB) system. One relies on state values computed from past experiences to make decisions in a reactive way, which allows for quick and efficient behavior, while the latter relies on an internal model of the world to perform inference and simulations, which results in a more flexible behavior. If MB behavior appears more desirable in situations that do not require reactivity, it still depends on the accuracy of the internal world model that is used by the agent. If the model is accurate, which would be reflected by low state prediction errors, MB behavior should be preferred. However, if the model is inaccurate, as reflected by high state prediction error, MF behavior should be preferred until the world model becomes accustomed to the new structure of the environment and that its errors decrease. The SPEs of our predictive model give us an elegant way to arbitrate between MF and

MB behavior. We train jointly our base MF Q-learner with a MB Q-learner which only learns from predictive model simulations, akin to the offline training by background planning described in a previous section. We define the recent average state prediction error by $\overline{spe}_t = \frac{1}{N_{hist}} \sum_{i=1}^{N_{hist}} spe_{t-i}$ with N_{hist} the size of the averaging window, fixed to 100 in all simulations, and we use this recent SPE average to compute an arbitration weight between MF and MB : $\lambda = exp(-\beta_{arb} * \overline{spe}_t)$. The policy of the agent is given by the combined MB and MF Q-values : $Q(s,a) = \lambda Q_{MB}(s,a) + (1-\lambda) Q_{MF}(s,a)$.

SPE manipulation

We compare four SPE-driven mechanisms by which predictive learning can support cognitive flexibility. For background planning and representation shaping, SPE manipulation directly affects learning (by scaling the teaching signal before weight updates). For learning rate adaptation and MF/MB arbitration, SPE manipulation affects only the adaptive readout while leaving predictive model learning intact. This distinction tests whether serotonin's behavioral effects arise from its role as a teaching signal for world model or representation learning, or from its use as an informational signal by downstream adaptive systems.

Simulations

Door reversal learning in a gridworld environment

Reward reversal learning

We used a two-dimensional grid environment to study spatial learning and flexible behavior under environment changes. The environment consists of a discrete grid of size $W * H$ (that we fix at 6×6 for our analysis). At each time step, the agent occupies a single grid cell and can move in one of four cardinal directions: up, down, left, or right. Movement is bounded by the grid edges, and invalid actions (e.g., moving left from the leftmost column) result in no change in position. The agent's observation is a one-hot encoding of its current location, represented as a flattened $W * H$ binary vector. A single reward location is defined at position $(r, c) = (5, 1)$ by default. For the 750 first episodes, the agent receives a reward of 1 upon reaching the cell corresponding to this position and 0 otherwise. Episodes terminate upon reaching the rewarded location or upon reaching a maximal number of steps $T_{max} = 15$. After the initial learning, reward location is moved to position $(r', c') = (5, 5)$ and reversal learning occurs over 500 more episodes.

Two-step task

In the two-step task (Daw et al. 2011), each trial consists of two stages : an initial decision (stage 1) leading probabilistically to one of two states $\{a, b\}$ (stage 2), followed by a final choice that yields a stochastic reward. At the first stage, the agents select one of two actions $a_1 \in \{0, 1\}$. Each action leads commonly $(p = 0.7)$ to one second-stage state and rarely $(p = 0.3)$ to the alternative state. $a_1 = 0$ commonly transitions to state a while action $a_1 = 1$ commonly leads to state b. At the second stage, the agent again selects between two actions $a_2 \in \{0, 1\}$ which yield a binary reward $r \in \{0, 1\}$ according to slowly drifting reward probabilities parametrized by

the probability vector :
$p_t = (P(r = 1|s_1 = a, a_2 = 0), P(r = 1|s_1 = a, a_2 = 1), P(r = 1|s_1 = b, a_2 = 0), P(r = 1|s_1 = b, a_2 = 0$
such that $p_0 = (0.8, 0.2, 0.2, 0.8)$ and that at every trial , the reward probabilities drift according to a Gaussian random walk $p_{t+1} = clip(p_t + N(0, \sigma_{drift}), 0.1, 0.9)$ where σ_{drift} is the drift rate parameter controlling the volatility of the environment (default at 0.01). At each step in the environment, the agents receive as input a one-hot encoded observation vector of dimension 3 : $o_t = [1, 0, 0]$ during the first stage, $o_t = [0, 1, 0]$ in second-stage state a and $o_t = [0, 0, 1]$ in second stage state b.

Probabilistic foraging

We model a two-port foraging task in which the probability of reward on the currently attended port (left vs. right) decays with the number of pokes at that port, and resets after a port switch. On each step t, the agent chooses between two actions : Poke ($a_t = 0$) or Switch ($a_t = 1$). A session terminates after a fixed number of 300 steps. Within a trial, the reward probability decays exponentially with the number of pokes n attempted on the current side : $p_{reward}(n) = P_0 * exp(-\frac{n}{\tau})$ where τ controls the decay scale and $P_0 = 0.75$. On a Poke action, a binary reward $r \in \{0, 1\}$ is drawn according to $p_{reward}(n)$ then $n \leftarrow n + 1$. On a Switch action, n is reset to 0 and a switch cost of -1 is delivered. At each step, the agents receive a 4-dimensional vector $o_t = [1_{left}, 1_{right}, n, R]$ where n is the poke count at current site and R is the cumulative reward at the current site, equal to the sum of reward obtained since the last Switch action.

Hyperparameters table

We need a big table with all the parameters used for all the simulated tasks.

Parameters	Transition Reversal	Reward Reversal	Two-step task	Foraging task	Cue-reward association
α_q	0.3	0.1	0.5	0.1	0.005
γ	0.9	0.9	0.9	0.9	1
α_p	0.3	0.1	0.1	0.01	0.1
K	1	5	10	10	1
ε	variable	0	0	0	0
λ	variable	0	0	0	0.1
$\alpha_{q.latent}$		0.1	0.01	0.01	
$\alpha_{p.latent}$		0.05	0.1	0.01	

H		32	32	32	
β_{mod}		10	1	1	
β_{arb}		5	2	5	

Cue-reward association task

Task

Head-restrained mice were trained to run on a treadmill to navigate through a circular virtual reality (VR) corridor of 400cm. Two visual cues are displayed on the wall (between 80-150cm and 220-290cm), one after the other, followed by a gray wall indicating the reward zone (350-370cm). The second cue predicted reward or no reward depending on its identity. The first cue predicted reward or no reward on 80% of trials, but on 20% of trials was ambiguous with respect to the second cue and therefore the reward. The corridor can thus be of four types : rewarded non-ambiguous (R+NA), unrewarded non-ambiguous (UR+NA), rewarded ambiguous (R+A), and unrewarded ambiguous (UR+A). Before the start of neural recording, the mice are trained to run in corridors with no cues on the wall, only a gray wall indicating the reward zone. To be rewarded, the mice must enter the reward zone with a speed inferior to a 30cm/s threshold. Cues are introduced on day 1 of the neural recordings sessions, and the mice must similarly decelerate in the rewarded corridors to be rewarded. But in order to maximize reward rate, the optimal behavior in an unrewarded corridor is to not decelerate before the reward zone. Ten sessions are recorded over ten consecutive days, with each session consisting of (m±std) trials (a trial corresponds to the crossing of a corridor), and stopping when the average speed of the mice goes below xxcm/s.

Behavioral criteria for mice inclusion in the analysis

We include the mice for which there is significant learning of the cue-reward contingency, assessed by a significant difference in anticipatory licking and in running speed before reward zone onset between the rewarded and unrewarded condition. This corresponds to 6 out of 11 mice.
> how do we compute these quantities ?

Neural recordings

[Here it would be really nice if we could just reference Solène's paper]

Assessing statistical significance

Unless mentioned otherwise, statistical significance was assessed using hierarchical bootstrapping (10,000 iterations), resampling at the level of animals (n = 6) and then trials within animals, to preserve the nested data structure (Saravanan et al. 2020). We report the mean difference between conditions and two-sided p-values based on the percentile method.

Task simulation

We simulated a virtual corridor of 400 discrete positions (0–399), matching the 400 cm experimental corridor. At each timestep, the agent selects one of four discrete speed levels (1, 2, 3, or 4), which advances its position by $speed * 10$ positions. The corridor is circular: an episode terminates when the agent reaches position 399 or beyond.

State representation. We use a microstimulus representation (Ludvig et al. 2008) to encode the agent's sensory observations as it traverses the corridor. The state vector has 280 dimensions, divided into 7 channels of 40 dimensions each. Each channel represents a distinct sensory feature: two channels for the two possible first cues (rewarded and unrewarded), two for the two possible second cues, one for the ambiguous first cue, and one for spatial position. On a given trial, only the channels corresponding to the cues actually present in that corridor are activated.

Within each active cue channel, cue onset is represented as a Gaussian bump that travels across the 40 dimensions as the agent progresses through the cue zone, mimicking the temporal unfolding of a sensory stimulus. As the agent moves from the start to the end of a cue zone, the bump shifts from the first to the last dimension of that channel. This means that two positions within the same cue zone that are close together have similar but not identical representations, while distant positions have low overlap, providing a smooth, distributed temporal code that the predictive model can learn from.

The first cue zone spans positions 80–220 cm and the second cue zone spans positions 220–400 cm. A separate spatial channel encodes the agent's coarse position throughout the corridor, with a single active dimension corresponding to the agent's binned location. Gaussian noise (σ = 0.1) is added to all observations at each timestep.

Corridor types and trial sampling. Four corridor types are defined by the combination of first and second cue identity: rewarded non-ambiguous (R+NA, probability 4/10), rewarded ambiguous (R+A, probability 1/10), unrewarded non-ambiguous (UR+NA, probability 4/10), and unrewarded ambiguous (UR+A, probability 1/10). In ambiguous corridors, the first cue activates the ambiguous channel rather than the rewarded/unrewarded channel, making the second cue unpredictable from the first.

Reward structure. In rewarded corridors (R+NA, R+A), the agent receives a reward of 5 upon crossing the reward zone (position ≥ 360) if its speed is below 3 (i.e., speed ∈ {1, 2}). At every timestep, the agent incurs a waiting cost of 0.5 × (4 − speed), penalizing slow movement. This creates a speed-accuracy tradeoff: the agent must learn to slow down before the reward zone in rewarded corridors while maintaining high speed elsewhere.

Relevance modulation. During training, the SPE teaching signal is modulated by state relevance following the formulation in the main text. The relevance modulation parameter is set to λ = 0.1.

Training protocol. Similarly to the experiment, Each agent is first pretrained for 1000 episodes on a neutral corridor (no cues, reward zone only) to learn the basic speed-reward contingency. Cues are then introduced, and the agent is trained for 100 sessions of 10 trials each. Results are averaged across N = 10 agents with different environment random seeds.

Model-fitting

We compared five different computational models to explain serotonin activity in the late sessions of learning (days 7-10), where value and environment structure have been learned : surprise (norm of the SPE), relevance-modulated SPE, RPE, value and predictive code for value (following (Harkin et al. 2023)). Models were fit to late sessions serotonin signals across the four types of corridors.

Analytical expression of value

Rather than simulating the full reinforcement and predictive learning dynamics for each parameter set during model fitting, we derived closed-form analytical expressions for the value and surprise signals at asymptotic convergence.This approach is justified by the use of late training sessions, during which animals have extensively learned the task structure and reward contingencies, such that internal value and next-state estimates can be assumed to have reached steady-state.

We derived an analytical approximation of the value signal at convergence under the assumption of stationary task statistics and exponential temporal discounting. Time was discretized with a step $\Delta t = 50ms$ and the discount factor γ was converted into a continuous-time discounting process with time constant $\tau = -\frac{\Delta t}{log(\gamma)}$. Each trial consists of an inter-trial interval ITI, a cue period, a delay, and a reward period, with corridor specific onset and duration. At convergence, the value function V(t) reflects the discounted expectation of future reward conditioned on the current trial phase. During the ITI, the value signal was assumed to be constant and equal to the discounted expected value of the upcoming trial, averaged across corridors: $V_{ITI} = \langle \frac{\tau}{\tau + L_{ITI}} exp(-\frac{L_{cue} + L_{delay}}{\tau}) r \rangle$, where r denotes the corridor-specific reward magnitude and the angle brackets indicate averaging across corridors. During the cue and delay period, $t_{cue} < t < t_{reward}$, the value signal followed an exponential approach to the expected reward time: $V(t) = r\, exp(\frac{t - t_{reward}}{\tau})$ where t_{reward} is the expected time of reward delivery. During the reward period itself, the value transitioned smoothly back to the ITI baseline, ensuring continuity at reward offset: $V(t) = A(1 - exp(\frac{t - t_{end}}{\tau})) + V_{ITI}$ with A chosen to match boundary conditions at reward onset and offset. Following the task design, we set $L_{cue} = 0.7$, $t_{end} = 3.7$ and $L_{ITI} = 4$ .The parameters (t_{cue}, L_{delay}, r) are corridor specific, equal to $(0.8, 2.1, 1)$ for R+NA condition, $(0.8, 2.1, 0)$ for UR+NA, $(2.2, 0.8, 1)$ for R+A and $(2.2, 0.8, 0)$ for UR+A. To capture imperfect learning or state uncertainty, we add a parameter to optionally interpolate the value signal toward the ITI baseline between the first and second cue for the unambiguous corridors : $V(t) \leftarrow (1 - \eta) V(t) + \eta\, V_{ITI}$. This produced corridor-specific value trajectories reflecting converged value estimates under temporal discounting and uncertainty.

Analytical expression of SPE norm

We modeled the non-modulated SPE surprise-like signal as a converged prediction-error–like response to salient task events, here cue onsets, followed by an exponential decay. The surprise signal was defined over position within a trial and consisted of a constant baseline plus transient cue-locked components. For each corridor, surprise was initialized to a baseline level S_0. At each onset of a cue i at time t_i, a transient surprise spike of amplitude h_i was added and followed by an exponential decay $S(t) = S_0 + h_i\, exp(-\omega(t - t_i))$ for $t > t_i$. For unpredictable cues, the same amplitude h_0 is used, and to capture imperfect learning, we add the option for predictable cue to still induce a surprise signal of amplitude $h_1 \geq 0$.

Model 1 : Pure surprise

The pure surprise signal is defined by $S(t)$, and is parametrized by $\theta = \{S_0, h_0, h_1, \omega\}$.

Model 2 : relevance modulated SPE

The relevance-modulated signal SPE signal is defined by $S(t) * (1 + \lambda V(t))$, and parametrized by $\theta = \{S_0, h_0, h_1, \omega, \lambda, \gamma\}$.

Model 3 : Value

The value signal if defined by $V(t)$ and is parametrized by $\theta = \{\gamma, \eta\}$.

Model 4 : RPE

The RPE signal was computed from the value signal : $\delta(t) = r(t) + \gamma V(t + \Delta t) - V(t)$, with $r(t)$ the reward signal. The RPE signal is parametrized by $\theta = \{\gamma, \eta\}$.

Model 5 : Predictive code for value

This model consist of two dynamic variables : a prediction signal $U(t)$ which tracks the expected value, and a rectified prediction error $\rho(t) = max(0, (1 + A)V(t) - AU(t))$, where A controls the weighting between the instantaneous value input and the internal prediction.

The prediction signal $U(t)$ evolves according to a first-order low-pass filter driven by the prediction error $\tau_{AD} \frac{dU(t)}{dt} = \rho(t) - U(t)$, where τ_{AD} is the adaptation time constant. The predictive code for value is $\rho(t)$, computed the analytical expression for value $V(t)$, and is parametrized by $\theta = \{\gamma, \eta, A, \tau_{AD}\}$.

For all models, two link parameters $\{\alpha, \beta\}$ are added as a multiplicative and bias term to allow a better fit to the signal.

Cross validation procedure

Model parameters were estimated using 10-fold cross-validation to prevent overfitting and assess generalization performance. For each fold, individual trials were randomly partitioned into training (90%) and validation (10%) sets, with independent splits for each of the four corridor conditions. Trial-averaged signals were computed separately for each condition within the training set. Parameters were optimized by minimizing the sum of squared errors between model predictions and the training averages across all four conditions using scipy's curve_fit function. Model predictions were evaluated against held-out trial averages using the coefficient of determination (R^2). The cross-validation procedure was repeated 10 times with different random splits to assess robustness. Final model performance was reported as the mean R^2 across all runs and folds.

Score ceiling estimation

To establish an upper bound on explainable variance, we computed a score ceiling using the same cross validation framework. For each fold, the training set average served as the

prediction for the validation set. This ceiling reflects the maximum achievable R^2 given trial-to-trial variability in the data.

Statistical analysis

Model fitting was performed at two levels : (1) population-level, pooling all trials across six mice (N = 1779, 1706, 468, and 394 trials for R+NA, UR+NA, R+A, and UR+A respectively), and (2) individual mouse level with 5-fold cross-validation to accommodate smaller sample sizes. Only data from late training days (day > 7) were included to ensure stable learned behavior.

Stat tests detailed results

Figure 3

Fig 3b. Significant difference in locomotion between rewarded and unrewarded after a few days. We perform a Wilcoxon test on N=6 mice, for every session.

N=6 mice	**P-value Locomotion (R vs UR, non-ambiguous)**	**P-value Licking (R vs UR, non-ambiguous)**	**P-value Locomotion (R vs UR, ambiguous)**	**P-value Licking (R vs UR, ambiguous)**
Session				
1	0.156250	0.069005	1.000000	0.342915
2	0.015625	0.031250	0.031250	0.015625
3	0.500000	0.015625	0.218750	0.109375
4	0.031250	0.031250	0.031250	0.015625
5	0.281250	0.015625	0.015625	0.218750
6	0.015625	0.015625	0.015625	0.031250
7	0.015625	0.015625	0.015625	0.015625
8	0.015625	0.015625	0.015625	0.021557
9	0.015625	0.015625	0.015625	0.015625
10	0.015625	0.031250	0.015625	0.021557

Fig 3d. Significant drop of serotonin during first exposures to a cue.

For each mouse, we compute the average cue serotonin response for the i-th exposition to the cue for i=1 to i=10, and average across all possible cues. Then we assess the significance of the difference between early trial response and later trial response by averaging the earlier trial (<6) response and late trial (>=6) response per mice (N=6) and performing a Wilcoxon signed-rank test between the two. Wilcoxon signed-rank test: p=0.0312. Early mean: 0.595, Late mean: -0.300.

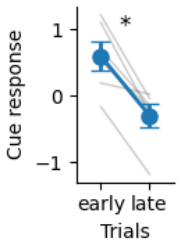


Fig 3f. Significant bump for unpredicted cues compared to predicted cue.

For each mouse we compute the average serotonin response (0-1s) to the second cue conditioned on whether it is an ambiguous (A) or nan-ambiguous (NA) corridor. We perform a Wilcoxon signed-rank test between the two conditions (N=6 mice). statistic=1.000, p-value=0.031.