Neural Mechanisms for Flexible Behaviors in Complex Environments

by

Mien Brabeeba Wang

B.A., Harvard University, 2018
S.M., Massachusetts Institute of Technology, 2020

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2026

Authored by:	Mien Brabeeba Wang Department of Electrical Engineering and Computer Science January 6, 2026
Certified by:	Nancy Lynch NEC Professor of Software Science and Engineering, Professor of Electrical Engineering and Computer Science, Thesis Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

THESIS COMMITTEE

Thesis Supervisor

Nancy Lynch
NEC Professor of Software Science and Engineering
Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Thesis Readers

Michael Halassa
Associate Professor of Neuroscience
Tufts University School of Medicine

Nir Shavit
Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Neural Mechanisms for Flexible Behaviors in Complex Environments

by

Mien Brabeeba Wang

Submitted to the Department of Electrical Engineering and Computer Science
on January 6, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Understanding how the brain learns the structure of the environment to support flexible behaviors in complex environments is a central question in neuroscience. This thesis presents computational frameworks that explore the neural mechanisms underlying flexibility in decision-making and navigation in different behavioral paradigms. We first introduce CogLinks, a novel class of neural architectures that solve natural tasks while revealing putative mechanisms. CogLinks integrate corticostriatal circuits for reinforcement learning and frontal thalamocortical networks for executive control—two network motifs specialized for processing distinct types of uncertainty. These modular networks enable hierarchical decision-making, efficient exploration, and optimal strategy switching by leveraging the unique specializations of these circuits. Furthermore, CogLinks establish a mathematical connection between neural mechanisms and corresponding algorithms, shedding light on the computational roles of specialized circuits. They also offer insights into computational psychiatry, connecting prefrontal dysfunction to altered reasoning patterns in schizophrenia. Notably, CogLinks generate neural and behavioral predictions in a human probability reversal task, which are subsequently validated through fMRI analysis.

Then, we turn to the brain's ability to represent and navigate complex cognitive and spatial maps. We develop a hierarchical planning circuit grounded in hippocampal replay and OFC–ACC interactions. We propose a variance-aware normative objective that maximizes the sum of within-cluster principal eigenvalues estimated from replay, derive local plasticity rules that optimize this objective, and show that the resulting circuit discovers task-relevant, multiscale structure. On these learned hierarchies, OFC implements fast subgoal valuation via hierarchical successor representations, while ACC realizes prospective path-selection attractors whose fixed points compute Gibbs marginals over legal trajectories; the joint system plans by iteratively projecting subgoal values down the hierarchy, mitigating long-range value attenuation. Finally, we generalize neural navigation to arbitrary manifolds by stitching simplex attractors into simplicial complexes, and coordinating global planning across simplex and local control within simplex to navigate on manifold.

Together, these models bridge neural mechanisms and cognitive functions, providing new insights into the neural basis of flexible behaviors.

Thesis supervisor: Nancy Lynch
Title: NEC Professor of Software Science and Engineering,
Professor of Electrical Engineering and Computer Science

Acknowledgments

I am deeply grateful to my advisor, Nancy Lynch, for supporting and believing in me as I explored the research questions that most inspired me, and for teaching me the principles of clear abstraction, rigorous modeling, and careful mathematical writing. I thank my advisor, Michael Halassa, for shaping my scientific taste and for showing me how to think, frame, and write about problems in neuroscience. I also thank my thesis committee reader, Nir Shavit, for his thoughtful feedback on my thesis and defense, and for teaching me how to communicate ideas with clarity and impact.

I thank my mentor in Taiwan, Sze-Bi Hsu, for instilling in me a philosophy of interdisciplinary work that is both mathematically elegant and biologically meaningful. I am grateful to Chi-Ning Chou, Tiancheng Yu, Jingxuan Fan, and Carlos Albors Riera for many conversations about research, problem-solving, and navigating the complexities of both science and life. I also thank my music and poker friends for bringing joy, community, passion and balance throughout my PhD.

Finally, I thank my family, Peng-Fei Wang, Li-Chen Peng and Yoyo Wang, for their patience and unconditional support, especially during the most stressful periods of this journey. I am also grateful to Sophie Lei, Wei Fang, En-Chi Cheng, and Xiaoyang Zhuang for being a second family, feeding me, encouraging me, and supporting me when I felt vulnerable during the thesis-writing process. I thank my girlfriend, Tracy Cui, for her love and steadfast support throughout my PhD; I am excited to share the next chapter of my research journey with her. To all my friends and colleagues who discussed science with me, worked alongside me, or simply spent time with me: thank you. You made my PhD years more meaningful than I could have imagined.

Contents

Chapter 1

Introduction

A defining feature of our everyday experiences, whether deciding how best to react to a colleague's terse reply or charting an efficient path through a familiar neighborhood, is that the world around us exhibits rich, underlying structure. Our ability to navigate these structures, both social and spatial, is what empowers us to behave flexibly: to adapt, explore new possibilities, and make informed choices without relying on brute force trial and error.

In a social scenario, for instance, one might wonder whether the awkwardness of a conversation with a new colleague stems from a poor choice of topic or an overarching emotional context overshadowing their engagement. Disentangling these possibilities requires a "hierarchical model" that tracks how local factors (such as discussion topics) are influenced by broader contexts (like mood or personal circumstances). Armed with such a model, we can decide whether to pivot to a more promising topic or to pause until the colleague's emotional climate becomes more receptive. Naturally, as we become more familiar with someone, past interactions reduce uncertainty about their preferences, making it easier to attribute awkwardness to mood rather than a misguided conversational approach.

Similarly, when navigating a well-known cityscape, we do not need to relearn every side street if we already possess a "cognitive map" of how roads and landmarks interconnect. With such a map, we can often chart new routes, seeking shortcuts or avoiding unexpected obstacles, without systematically testing every possible turning. Whether it is discerning how best to engage with someone socially or planning an untried path through a known neighborhood, flexible behavior hinges on learning and leveraging these underlying structures.

This thesis introduces three computational frameworks that illuminate how the brain may build and exploit structured internal representations. First, we investigate hierarchical decision making, identifying neural mechanisms that integrate uncertainty across levels to support efficient exploration and strategic switching. Second, we show how the brain could construct hierarchical models of the environment and establish subgoals to enable multiscale planning and navigation. Third, we propose circuit models for encoding and traversing complex continuous cognitive spaces, in which learned manifolds support flexible route planning. Together, these approaches suggest that specialized neural circuits underlie the learning and deployment of structured models, allowing agents to handle everything from subtle social cues to novel shortcuts by leveraging the deeper organization of their world.

The organization of this thesis is as follows. Chapter 2 examines flexible choice in hierarchical, partially observable tasks: the ability to decide whether to keep exploiting a strategy

or to switch when latent context changes. We develop CogLink, a modular, biologically grounded architecture that separates low-level uncertainty about actions and outcomes from high-level uncertainty about context or task phase. A premotor cortico–thalamo–basal ganglia loop learns to explore efficiently and assign credit, while an associative PFC–MD loop infers and maintains the current context and gates strategy switches. Rather than training with backpropagation, we derive the network analysis-first, extracting an algorithm from the neural and plasticity dynamics so that each component's function is readable from the circuit. The framework explains human behavior and fMRI signatures in probabilistic reversal and offers mechanistic hypotheses for how prefrontal disruption or psychiatric disease could produce maladaptive switching.

Chapter 3 addresses flexible, hierarchical goal-directed behavior: the capacity to reach distant goals by inserting and updating intermediate subgoals. We introduce a hierarchical planning circuit in which hippocampal replay supplies compressed transition statistics, and a variance-aware clustering rule maximizes the sum of within-cluster principal eigenvalues to discover multiscale, task-aligned structure. On this learned hierarchy, OFC computes subgoal-conditioned values via hierarchical successor representations, while ACC performs a prospective search that selects option sequences with highest subgoal value using path-selection attractors and propagates the chosen subgoal downward. This architecture links hippocampal replay, OFC value maps, and ACC prospective activity into a single circuit-level realization of hierarchical model-based reinforcement learning.

Chapter 4 turns to flexible navigation over complex continuous state spaces: the ability to move smoothly through continuous manifold such as curved terrains, mazes, or abstract relational spaces. We show that such spaces can be realized as neural simplicial complexes in which individual simplices are implemented as stable attractors with local feedback control, and explicit bifurcation conditions identify when two simplex attractors can be stably glued to form a larger complex. A planning module treats the complex as a hypergraph, computes shortest paths across simplices, and hands control back to the local controller upon reaching the goal simplex. In this way, the same circuit supports both local continuous control and global discrete routing, offering a neurally plausible account of how animals might traverse high-dimensional continuous cognitive maps built from manifold-like population activity.

In our mathematical modeling, we take a hybrid approach. To capture fast-timescale neural activity, we use continuous-time rate neurons that describe rich dynamical evolution. To capture slower synaptic plasticity, often driven by discrete environmental feedback, we update synapses in a discrete-time manner at feedback events. A similar hybrid choice appears in our action models: in Chapters 2–3, we study discrete actions, whereas in Chapter 4, we use continuous control for action execution alongside a discrete planning module for higher-level action selection.

In summary, we advance a common theme in cognition: learn the structure of the environment and then use that knowledge to generalize across tasks and adapt to changing behavioral needs. This hinges on the interplay between slow plasticity and fast neural activity. Slow plasticity extracts and learns environmental structure, while fast dynamics plan on top of it to produce flexible behavior. Our work also addresses an important gap in theoretical neuroscience by linking concrete computational algorithms to plausible neural implementations. Because the circuits are grounded in experimental data, they yield testable predictions, and we validate several in human fMRI data during a probabilistic reversal task.

Taken together, the circuit realizations for hierarchical decision making, multiscale planning, and navigation on continuous manifolds suggest how brains may assemble reusable internal models that generalize beyond immediate experience. Progress will depend on tightening the loop between theory and experiment, and we hope this work lays a clear path forward.

1.1 Collaboration Statement

1. In chapter 2, the section 2.1.9 “CogLink reveals thalamic regulation of reinforcement learning strategies across prefrontal-striatal network" is a joint work done with Bin A. Wang, Mien Norman H. Lam, Liu Mengxing, Shumei Li, Ralf D. Wimmer, Pedro M. Paz-Alonso, Michael M. Halassa and Burkhard Pleger titled “Thalamic regulation of reinforcement learning strategies across prefrontal-striatal networks" [1]. The human data collection and analysis are done by Bin A. Wang and the other authors. I performed the CogLink modeling analysis.

2. Chapter 2, from section 2.1.1-2.1.8, chapter 3 and chapter 4 are joint works done with my advisors Michael M. Halassa and Nancy Lynch. I deeply appreciate their valuable inputs throughout the entire process.

Chapter 2

Neural architectures for flexible hierarchical decisions

Adapting to a constantly changing environment requires the ability to flexibly adjust behavior in response to uncertainty. When an outcome is unexpected, the brain must determine whether it arises from random variability, a suboptimal strategy, or a fundamental change in the environment-a critical process for selecting the most effective course of action.

For example, if a conversation with a new colleague feels awkward, one might question whether the topic is poorly chosen or whether the colleague is simply having a bad day. Disambiguating these possibilities is crucial for selecting the appropriate response. If the discomfort stems from a poor topic choice, switching to a different subject might improve the interaction. However, if the colleague's disengagement reflects an underlying mood, a better approach might be to pause and revisit the conversation another day. This decision-making process relies on hierarchical inference, where lower-level variables (such as topic preferences) are interpreted within the broader context of higher-level states (such as mood or personal circumstances). This distinction becomes easier with a familiar colleague, as prior experience reduces uncertainty about their preferences, making it more likely that disengagement is attributed to mood rather than topic choice.

Since both conversational preferences and emotional states are latent variables that cannot be directly observed, the brain must infer their values while also estimating the uncertainty associated with each. For instance, one must assess both the intrinsic appeal of a topic, such as the Super Bowl, and the likelihood that a friend has emotionally recovered from a breakup after four months. A fundamental challenge in neuroscience is understanding how the brain processes and integrates uncertainty across multiple hierarchical levels to drive flexible decision-making.

Machine learning approaches have contributed to progress in addressing this question. Traditionally, normative models based on Bayesian inference have been used to solve hierarchical tasks and model the strategies animals employ [1–3]. These models estimate uncertainty at multiple levels and use it for effective credit assignment [4–6]. However, they pose significant limitations as tools for neuroscientific discovery. First, their explanatory power is constrained when the generative model of the environment is misspecified [7], leaving the challenge of specifying an accurate model unresolved. Second, their components are non-neural, making it unclear whether and how they correspond to neural circuits and computations.

To address these limitations, the field has increasingly turned to neural networks trained through deep learning, which have been proposed as models of neural computation and have demonstrated exceptional performance on a range of tasks, sometimes exceeding human capabilities [8, 9]. However, these architectures operate as frequentist prediction models that do not explicitly account for uncertainty. As a result, they cannot estimate confidence in different task components in the way humans and animals do [6, 10–12]. These shortcomings highlight the need for a new approach to understand how brains make decisions in hierarchical environments and how uncertainty processing enables this cognitive ability.

In this study, we introduce CogLink, a novel class of computational architectures that address this gap. Fundamentally, CogLink networks are dynamical systems composed of rate neurons and share structural similarities with artificial feedforward and recurrent neural networks (Figure 2.1a). However, they differ from conventional machine learning neural networks in three important ways.

First, CogLink networks are optimized using a novel multi-step procedure. Instead of using backpropagation to minimize network error, we employ an approach that leverages scale separation principles to extract a structured computational algorithm from neural dynamics, followed by mathematical analysis to determine near-optimal network connectivity parameters (Figure 2.1b, c).

Second, CogLink networks incorporate biological realism by modeling specific brain systems, including known connectivity patterns, cell types, and learning rules for dynamic control (Figure 2.1d).

Third, CogLinks offer greater interpretability. Because they are explicitly structured to approximate an algorithm, they allow us to directly map neural mechanisms to their functional roles, unlike traditional deep neural networks, which are often considered black boxes (Figure 2.1d).

Through iterative development, we construct progressively complex CogLinks, mirroring the increasing computational complexity observed in biological evolution. Specifically, the "basic" network models a premotor cortico-thalamic-basal ganglia (BG) loop, emphasizing BG circuitry's role in reinforcement learning and efficient environmental exploration, which addresses lower-level uncertainty in hierarchical environments. The "augmented" network incorporates an associative cortico-thalamic-BG loop, highlighting the mediodorsal thalamus (MD) and its interactions with the prefrontal cortex (PFC) to process higher-level uncertainty related to contextual inference and strategy switching.

In addition to demonstrating how partitioning uncertainty types in hierarchical environments is critical for the networks to reproduce animal behavior, CogLinks provide insights into neural mechanisms underlying complex decision-making. Specifically, our model explains findings from our study on human behavior and fMRI readouts in the next section 2.1.9 [13] and offers insights into perturbed dynamics in a mouse model relevant to schizophrenia [14]. To our knowledge, few existing neural frameworks simultaneously solve complex cognitive tasks while providing computational insight into neural mechanisms. We propose that CogLinks constitute an important step toward bridging this gap.

2.1 Results

2.1.1 Building a basic CogLink network for handling lower-level uncertainty

To illustrate lower-level uncertainty, let us revisit the example of conversing with a new colleague. Suppose you know nothing about the person and naively attribute each sigh of boredom to a suboptimal choice of topic, disregarding higher-level factors such as mood or personal circumstances. This scenario highlights two key types of lower-level uncertainty: outcome uncertainty, which may arise from factors such as variability in the person's focus (e.g., low focus might prevent them from following certain sentences), and associative uncertainty, which reflects our lack of knowledge about the person's preferences (greater unfamiliarity corresponds to higher associative uncertainty). Successfully navigating this interaction requires balancing exploration and exploitation. Persisting with the Super Bowl (exploitation) tests its suitability as a topic but risks disengagement if it proves uninteresting. Conversely, switching to a new topic, such as a shared hobby or current news (exploration), sacrifices immediate feedback on the Super Bowl but creates an opportunity to reduce associative uncertainty by learning more about the colleague's preferences.

The basal ganglia (BG) are a natural candidate for handling such lower-level uncertainties. A substantial body of research implicates the BG in learning action-outcome associations by integrating sensory inputs, motor actions, and reward feedback [15, 16]. Dopaminergic signals encoding reward prediction errors (RPEs) facilitate synaptic plasticity within the BG, enabling the adaptive adjustment of action values over time. This iterative refinement process makes the BG well-suited for encoding associative uncertainty and guiding the trade-off between exploration and exploitation. Accordingly, our basic CogLink network incorporates BG-like circuits, along with dopamine-dependent plasticity mechanisms for online learning and premotor/motor cortical areas for action selection (Figure 2.2a). Neuronal activity in these areas is modeled as rate neurons governed by:

$$\tau \frac{dx}{dt} = -x + f(x, I; W), \tag{2.1.1}$$

where x represents the neuron's firing rate, τ is the membrane time constant, f is a nonlinearity function, I is the input and W is denotes synaptic weights.

A defining feature of the basic CogLink network is its incorporation of a quantile population code in the BG-like area, which encodes associative uncertainty as a distribution over action-value beliefs. In this scheme, each neuron is associated with a fixed quantile of the probability distribution, meaning that selecting a subset of neurons corresponds to sampling specific probabilities from the encoded distribution. Random sparsification dynamics in the premotor cortex-like area leverage this property to extract uncertainty through sampling (Figure 2.2b, Methods). This use of a quantile code builds upon the broader concept of population encoding, where neuronal ensembles represent probability distributions. Established approaches to population encoding include probabilistic population codes [17], sampling codes [18], explicit probability codes [19], and quantile codes [20]. In our model, premotor corticostriatal synapses

represent the distribution of action-value beliefs using a quantile code:

$$P(v_a = V_{a,m}^{\text{alm/bg}}) = \frac{1}{M}, \forall a \in [A], m \in [M],$$

where A is number of alternatives and M is the size of a neuronal ensemble. This representation allows the network to efficiently extract uncertainty, as random sparsification in the premotor cortex-like area samples directly from the quantile-coded distribution formed by corticostriatal-like synapses. These sampled values are then relayed to the motor cortex-like area, where they inform action selection and balance exploration and exploitation during decision-making (Figure 2.2c).

The sampling mechanism thus provides a way to translate associative uncertainty into inputs for the motor-like area, supporting efficient exploration during decision-making. While biological basal ganglia (BG) circuits project to the motor cortex via the thalamus and involve intricate circuitry beyond the corticostriatal loops modeled here, we abstract these additional components as relay functions to simplify the model. In this abstraction, the sampled values are directly projected to the motor cortex-like area to focus on the computations critical for exploration.

To convert the sampled action values, which encode associative uncertainty, into motor signals for action selection, we employ a model of the action selection mechanism inspired by ramp-to-threshold circuits observed in motor-related decision-making cortical circuits [21]. Specifically, the recurrent connections in the motor cortex-like area are configured to implement mutual inhibition, enabling a Winner-Take-All (WTA) procedure (Figure 2.2d, Methods). In this setup, when the activity of a motor neuron ramps up to the threshold, the corresponding action a_t will be chosen. Thus, the motor circuit effectively chooses the action corresponding to the highest action-value sample emitted from striatal circuits (Figure 2.2d).

Importantly, because the probabilistic nature of the sampled values carries information about the uncertainty (e.g., a flat distribution with a low mean can still sample a high value to drive exploration), this allows for associative uncertainty-based exploration: when value belief distributions of different actions have large overlapping (high uncertainty on which action is optimal), the CogLink will explore more; on the other hand, when value belief distribution of different actions have small overlapping (low uncertainty on which action is optimal), the CogLink will exploit more (Figure 2.2e).

Finally, after the model chooses action a_t and receives reward r_t at trial t, the DA activities form a distributional RPE, $\delta \in \mathbb{R}^M$, given by:

$$\forall m \in [M],\ \delta_m = r_t - V_{a_t,m}^{\text{alm/bg}},$$

where $V_{a_t,m}^{\text{alm/bg}}$ represents the predicted value of action a_t for the m-th quantile. The distributional RPE is then used to update the premotor-BG synapses according to the following rule:

$$\forall m \in [M],\ V_{a_t,m}^{\text{alm/bg}} \leftarrow V_{a_t,m}^{\text{alm/bg}} + \eta_{a_t}\delta_m, \tag{2.1.2}$$

where η_{a_t} denotes the learning rate of the corticostriatal synapses for the selected action a_t.

Unlike prior work on distributional reinforcement learning [20], which focuses on learning reward distributions, our approach learns action value belief distributions. This distinction is

critical for representing uncertainty in action values, which is essential for reasoning about lower-level uncertainty. We elaborate on this distinction and its implications in the Discussion.

To evaluate the model's behavior, we use an A-alternative forced choice task (A-AFC task) (Figure 2.2f). In this task, the reward probabilities for each action at trial t are represented as $\theta_t \in \mathbb{R}^A$, where $(\theta_t)_a$ denotes the probability of receiving a reward for action a. We test the model in both stationary and dynamic environments. In the stationary environment, the reward probabilities remain constant across trials, such that $\theta_t = \theta_1$ for all $t \in [T]$. In the dynamic environment, the probabilities θ_t can vary across trials to reflect changing conditions.

To assess the model's performance, we use the standard regret metric from bandit literature [22], defined as:

$$R_T = \sum_{t=1}^{T} (\theta_t)_{a_t^*} - (\theta_t)_{a_t},$$

where a_t^* is the retrospectively optimal action and a_t is the action chosen by the model. Regret measures the cumulative difference in rewards between the model's chosen actions and the optimal actions, providing a benchmark for evaluating the model's ability to adapt and balance exploration and exploitation.

The CogLink network successfully minimized regret by balancing exploration and exploitation (Figure 2.2g). To explore the neural underpinnings of our model's performance, we examined the synaptic strength of corticostriatal connections. Intriguingly, these synaptic profiles exhibited distinct signatures indicative of efficient exploration. The ensemble of synapses tuned to the accurate choice (action 1) rapidly narrowed the distribution and converged to the correct value estimates, while the synapses tuned to less preferred choices still showed a gradient of smaller synaptic strengths distinct from those of the preferred choice. The distinct gradient of synaptic strengths tuned to the less preferred choice indicated that the less preferred choice has high associative uncertainty (i.e., the distributions remained wide and hence those ensembles tuned to exhibit a gradient of strength in Figure 2.2h) but low enough to confidently exploit the correct choice (i.e., the distributions are separated from the distribution of the optimal action and hence exhibit a gradient of synaptic strengths smaller than the synaptic strengths tuned to the optimal action in Figure 2.2e).

To test the necessity of specific mechanisms in balancing exploration and exploitation, we performed two lesion experiments: one with reduced sparsification in the premotor cortex (KO-sparseness) and another replacing the distributional RPE with a scalar RPE (KO-distributional RPE). Both lesion variants resulted in significantly higher regret (Figure 2.2i), driven by premature exploitation that led to persistent suboptimal choices (Figure 2.2j). These results provide mechanistic insights into both random sparsification and distributional RPEs in balancing exploration and exploitation.

2.1.2 Our basic CogLink model approximates a novel algorithm with nearly-optimal regret.

A mechanistic model often proves too complex to clearly illustrate the underlying computational mechanisms or to admit mathematical analysis. To address this challenge, we approximate the basic CogLink network with an algorithm by leveraging the separation of

scales, assuming that neural dynamics occur instantaneously (see Methods). This simplification enables the premotor corticostriatal-like ensemble tuned to action a to act as a sampling mechanism for the action-value distribution. Specifically, the $K-$WTA dynamics in the premotor cortex-like circuit serve to randomly select K neurons, enabling efficient sampling of the action-value distribution:

$$\mathcal{V}_a = \text{unif}\left(\left\{\frac{1}{K}\sum_{j=1}^{K} V_{a,i_j}^{\text{alm/bg}}\right\}_{1\le i_1<\cdots<i_K\le M}\right),\ \hat{v}_a \sim \mathcal{V}_a. \tag{2.1.3}$$

where $\mathcal{V}_a$ represents the distribution of value belief of action a and $\hat{v}_a$ is the sampled value. The WTA mechanism of the motor cortex-like circuit selects the action with the highest sample value:

$$a_t = \arg\max_a \hat{v}_a. \tag{2.1.4}$$

Finally, the dopamine-gated plasticity adjusts the corticostriatal synapses to refine action-value estimates over time:

$$\forall m \in [M],\ \delta_m = r_t - V_{a_t,m}^{\text{alm/bg}},\ V_{a_t,m}^{\text{alm/bg}} \leftarrow V_{a_t,m}^{\text{alm/bg}} + \eta_{a_t}\delta_m$$

where δ_m is the distributional RPE and η_{a_t} is the learning rate for the a_t-tuned synapses.

The algorithm provides an intuitive framework for understanding the functionality of our corticostriatal network model. In this framework, A posterior-like distributions, representing action-value beliefs, are sampled through random sparsification in the premotor cortex. The motor cortex then selects the action corresponding to the largest sampled value through the recurrent competitive dynamics. Following action selection, the model refines its action-value distributions based on distributional reward prediction error (RPE) signals from dopamine (DA) neurons. High associative uncertainty—indicating a lack of confidence in the value estimate—results in the significant overlap between posterior-like distributions, promoting exploration (Figure 2.2e, left). Conversely, low associative uncertainty leads to well-separated posterior-like distributions, enabling exploitation as the model confidently selects the optimal action (Figure 2.2e, right).

To assess our network's performance against the theoretical regret limit, we conducted a mathematical analysis of the algorithm. Our analysis demonstrates that by appropriately configuring the parameters in the synaptic update rule, the regret of the algorithm will be on the order of $O(\sqrt{AT\log(AT)})$, where A represents the number of actions and T denotes the number of trials (see Theorem 2.3.5 in Methods for a formal theorem).

Theorem 2.1.5 (Informal)**.** *If we select the sparsity K, the learning rate $\{\eta_{a,t}\}_{a\in[A],t\in[T]}$ and the initial synaptic weight $\{\bar{V}_{a,m}^{\text{alm/bg}}\}_{a\in[A],m\in[M]}$ appropriately, then the regret of the algorithm after T trials in a static A-*AFC *task is at most $C\sqrt{AT\log(AT)}$, where C is a constant.*

It has been demonstrated that no algorithm can achieve regret smaller than $\Theta(\sqrt{AT})$ [23]. Our algorithm, which differs by only a logarithmic factor, is therefore close to optimal in terms of regret. This result provides a theoretical foundation for the model's ability to perform efficient exploration under lower-level uncertainty, demonstrating its near-optimal balance of exploration and exploitation.

2.1.3 Relationship between basic CogLink and Bayesian inference with probability matching

We evaluated the performance of our basic CogLink model in static A-AFC tasks and compared it to Thompson Sampling (TS), a widely used algorithm that combines optimal Bayesian inference with probability matching and provides asymptotically optimal theoretical guarantees [24]. TS was chosen as the baseline because it represents a principled approach to balancing exploration and exploitation under uncertainty. Task difficulty was manipulated by varying the expected reward difference between the most and least rewarding actions (Δ) and the number of alternatives (A) (see Methods) (Figure 2.1a). Across all tested environments, CogLink consistently outperformed TS, achieving a better balance between exploration and exploitation, as evidenced by faster convergence to optimal actions and improved regret performance (Figure 2.3a-c, Figure 2.S1a-c).

To further evaluate CogLink's versatility in handling more complex decision-making scenarios, we extended its application to two generalizations: a cued A-AFC task, which incorporates state (cue) information, and a binary tree maze task, which introduces state transitions (see Methods). These tasks represent a progression from stateless bandit problems to scenarios where decisions depend on environmental states. In these tasks, we compared CogLink against TS and a neural network-based method, Deep Q Network (DQN, [25]). CogLink demonstrated robust performance across varying difficulty settings, maintaining competitive regret compared to both TS and DQN (Figure 2.S2a-d). These results underscore CogLink's ability to adapt from simple, stateless environments to more complex tasks involving state information and transitions, demonstrating its versatility in managing lower-level uncertainty during decision-making.

The robust performance of CogLink in various tasks raises questions about the underlying principles that enable its effective decision-making. To better understand these mechanisms, we examined the algorithm resulting from CogLink's approximation and observed that the resulting algorithm after approximation shares key similarities with Thompson Sampling (TS), particularly in its use of action-value distributions and probability matching-like action selection, but differs in the update rule. To investigate this relationship further, we analyzed the correspondence between the distributional RPE update in Equation 2.1.2 and Bayesian updates.

First, we initialized the corticostriatal weights to approximate a uniform prior, analogous to Bayesian inference using a uniform prior:

$$\forall a \in [A], m \in [M],\ V_{a,m}^{\text{alm/bg}} \leftarrow \bar{V}_{a,m}^{\text{alm/bg}}, \text{where } \bar{V}_{a,m}^{\text{alm/bg}} = \frac{m}{M}.$$

Next, we examined how the expectation and variance of the action-value distribution evolved under the distributional RPE update. By selecting learning rates $\eta_t \propto 1/t$ (see Methods), we found that our updates closely approximated the evolution of both the expectation and variance under optimal Bayesian inference (Figure 2.3d, e). This choice of learning rate satisfies two critical conditions: $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\lim_{t\to\infty} \eta_t = 0$, ensuring that the variance diminishes over time while the expectations converge to the true action-value distribution.

In addition to the update rule, CogLink provides flexibility in balancing exploration and exploitation by modulating the parameter K, the sparsity in premotor cortex-like

area. Larger K values result in narrower distributions after updates, favoring exploitation (Figure 2.3f). Specifically: when $K = 1$, this is probability matching and when $K = M$, this is deterministically sampling the expected value. When $1 < K < M$, the model employs generalized probability matching, where higher K values increase the emphasis on exploitation while still allowing some degree of exploration. This framework provides a continuum of strategies for balancing exploration and exploitation.

We hypothesize that K could be dynamically modulated in biological systems through neural mechanisms, such as altering the excitability of premotor cortical neurons via neuromodulation. This hypothesis aligns with prior studies demonstrating that neuromodulatory systems, including dopamine and norepinephrine signaling, play a role in adjusting the exploration-exploitation trade-off [26, 27]. This dynamic modulation offers a plausible pathway for organisms to adapt their decision-making strategies to changing environmental demands.

2.1.4 Building an augmented CogLink network for handling higher level uncertainty

Returning to our example of conversing with a new colleague, the assumption that each sigh of boredom is solely due to a suboptimal topic choice or random outcome variability (e.q. low focus) is overly simplistic. Higher-level factors, such as the colleague's mood or workload, can also play a role, and these conditions are often dynamic and unobservable. In naturalistic environments, animals must contend not only with lower-level uncertainties, such as outcome and associative uncertainty, but also higher-level uncertainties, including contextual uncertainty—ambiguity about the underlying context governing the environment. To address this challenge, we designed a probabilistic reversal task in a dynamic environment (Figure 2.4a). While the basic CogLink network performed well in static environments, it struggled to adapt quickly to changing contexts in this dynamic setting (Figure 2.S3e, f).

As an initial step, we introduced explicit external contextual cues to the model, activating separate instances of the basic CogLink network depending on the provided cues (Figure 2.S3b). This modification allowed the model to achieve instantaneous behavioral switching (Figure 2.S3g, h). However, animals in natural environments rarely have access to explicit contextual cues and instead must infer the underlying context from ambiguous and incomplete observations.

The prefrontal cortex (PFC)-mediodorsal thalamus (MD) circuit is a natural candidate for enabling such contextual inference. The PFC is well-established as a key region for flexible, context-dependent behavior [28, 29], generating complex activity patterns to support such capacities. Recent studies suggest that these patterns are regulated by interactions with the MD [30–36], which encodes task context explicitly in a range of decision-making paradigms [37–40]. Inspired by these findings, we augmented CogLink by incorporating a PFC-MD-like circuit to infer and provide contextual information to the basic CogLink networks (Figure 2.4b). This augmentation enables the model to adapt to dynamic environments without relying on explicit external cues. One important assumption in our model is that disjoint basic CogLink networks are activated based on the inferred context. We discuss the biological plausibility of this mechanism further in the Discussion.

A prominent feature of the augmented CogLink model is its low-dimensional representation of contextual likelihood in the MD-like area, consistent with previous literature [37–40]. Specifically, we propose that the MD encodes the conditional likelihood $p(c|a_{\leq t}, r_{\leq t})$ of a context c, given the history of action-outcome pairs $\{a_{\leq t}, r_{\leq t}\}$. To achieve this, we hypothesize that MD activity lies on a low-dimensional simplex attractor, enabling a stable representation of contextual likelihood. Since the thalamus lacks intrinsic excitatory recurrence [35], we propose that PFC and MD form an excitatory loop, with the thalamic reticular nucleus providing local inhibition to stabilize the attractor (Figure 2.4, see Methods). In this framework, contextual likelihood is represented as an explicit probability code, where higher neural activity corresponds to a higher likelihood of the associated context. The simplex attractor structure allows MD activity to dynamically integrate inputs, enabling it to traverse the manifold in response to changing environmental conditions (Figure 2.4c, d). However, contextual inference requires that these inputs be appropriately encoded to reflect the conditional likelihood.

Bayes' rule provides a framework for determining the required input encoding by defining how contextual likelihoods are computed:

$$p(c|a_{\leq t}, r_{\leq t}) \propto \prod_{i=1}^{t} p(a_i, r_i|c)p(c). \tag{2.1.6}$$

This formalism suggests that the inputs should correspond to the single-trial contextual generative model $p(a_t, r_t|c)$ which is accumulated across trials to compute the overall likelihood. We hypothesize that PFC-MD synapses learn this single-trial generative model, while the MD's low-dimensional attractor dynamics perform the accumulation. To enable this process, we implemented a Hebbian learning rule for PFC-MD connections (Figure 2.4e, see Methods):

$$\Delta V_{c,a,r}^{\text{pfc/md}} \propto f_{\text{hebb}}(x_c^{\text{md}}) x_{a,r}^{\text{pfc}}. \tag{2.1.7}$$

Here, x_c^{md} and $x_{a,r}^{\text{pfc}}$ denote the activities of MD and PFC neurons tuned to context c and action-outcome pair (a, r), respectively, $V_{c,a,r}^{\text{pfc/md}}$ represents PFC-MD synaptic weights between these MD and PFC neurons, and f_{hebb} (Figure 2.S4b) is a sigmoidal gating function that modulates synaptic plasticity.

Naive Hebbian plasticity may incorrectly associate action-outcome pairs with the wrong context when contextual uncertainty is high, resulting in inaccurate estimates of the contextual generative model (Figure 2.4f). To mitigate this issue, we incorporate a gating mechanism f_{hebb} that modulates plasticity based on MD activity. This gating enhances learning when the MD confidently infers the context (high MD activity) and suppresses plasticity when contextual uncertainty is high (low MD activity). By doing so, the mechanism achieves two key objectives: it accelerates the learning of contextual statistics when confidence is high and prevents the misattribution of associations under high contextual uncertainty (Figure 2.4g, h).

To causally test the necessity of this mechanism, we evaluated a variant of the model with naive Hebbian plasticity (KO-nonlinear Hebb). The results show that the generative model learned by the full CogLink closely approximates the true generative model of the environment, whereas the KO-nonlinear Hebb variant deviates significantly (Figure 2.4f). This supports the critical role of f_{hebb} in enabling accurate contextual learning under uncertainty.

Another key component of the augmented CogLink is an interneuron-mediated thalamocortical projection pathway that modulates cortical activity to drive exploration under high contextual uncertainty (Figure 2.4i). When contextual uncertainty is high, animals need to explore more to gather information about the current context. To implement this pathway, we drew inspiration from experimental findings showing that the MD thalamus modulates PFC functional connectivity through distinct interneuron-mediated mechanisms. Specifically, Mukherjee et al. identified two thalamocortical pathways: one amplifies cortical connections via local disinhibition by vasoactive intestinal peptide (VIP) interneurons, and the other suppresses cortical activity through fast inhibition mediated by parvalbumin (PV) interneurons [38].

Building on these findings, we assumed that such modulation enables contextually relevant PFC populations to differentially influence downstream premotor circuits, thereby facilitating context-dependent behavior. To model this mechanism, we included both thalamic projection pathways: one amplifies effective cortical connectivity for the preferred context, while the other inhibits cortical activity related to the opposing context. This modulation adjusts the activity in the PFC-like area and influences downstream premotor corticostriatal connections (Figure 2.4i). Specifically, the projections modulate the effective strength of corticostriatal connections according to the following dynamics:

$$\forall c \in [2], a \in [A],\ m \in [M],\ \tau^{\mathrm{bg}} \frac{dx^{\mathrm{bg}}_{c,a,m}}{dt} = -x^{\mathrm{bg}}_{c,a,m} + f_{\mathrm{in}}(x^{\mathrm{vip}}_c - x^{\mathrm{pv}}_c) V^{\mathrm{alm/bg}}_{c,a,m} x^{\mathrm{alm}}_{c,a,m}. \tag{2.1.8}$$

Here, τ^{bg} is the membrane time constant of striatal neurons. $x^{\mathrm{bg}}_{c,a,m}$ and $x^{\mathrm{alm}}_{c,a,m}$ represent the activities of the mth BG and premotor neurons tuned to context c and action a, respectively, while $V^{\mathrm{alm/bg}}_{c,a,m}$ is the strength of the corticostriatal synapse connecting these neurons. f_{in} is a sigmoidal nonlinearity, and x^{vip}_c and x^{pv}_c denote the activities of VIP and PV interneurons receiving MD inputs tuned to the preferred and opposing contexts, respectively (see Methods). Since these interneurons receive contextual inputs from MD, the term $f_{\mathrm{in}}(x^{\mathrm{vip}}_c - x^{\mathrm{pv}}_c)$ encodes contextual certainty. This mechanism ensures that corticostriatal connections are weakened under high contextual uncertainty, promoting exploratory behavior.

To validate this mechanism, we systematically varied MD activity to manipulate contextual uncertainty and measured its effect on exploratory behavior. Consistent with our predictions, higher contextual uncertainty corresponded to increased exploration, confirming the role of thalamocortical projections in dynamically regulating contextual uncertainty-based exploration (Figure 2.4j).

In addition to modulating exploratory behaviors, contextual uncertainty should also regulate learning. Under high contextual uncertainty, naive dopamine-dependent plasticity risks misattributing associations to the wrong context, resulting in inaccurate action-value estimates (Figure 2.4k). To address this, we implemented a mechanism in which interneuron-mediated inputs gate the plasticity of corticostriatal synapses (Figure 2.4i). This design is inspired by experimental findings that interneuron-mediated pathways can modulate cortical plasticity [41, 42]. Specifically, our model incorporates the following update rule:

$$\forall c \in [2], m \in [M],\ \delta_{c,m} = r_t - V^{\mathrm{alm/bg}}_{c,a_t,m},\ \Delta V^{\mathrm{alm/bg}}_{c,a_t,m} \propto f_{\mathrm{in}}(x^{\mathrm{vip}}_c - x^{\mathrm{pv}}_c)\delta_{c,m}. \tag{2.1.9}$$

Here, $\delta_c \in \mathbb{R}^M$ represents the distributional dopamine activities tuned to context c, and $V^{\mathrm{alm/bg}}_{c,a_t,m}$ denotes the corticostriatal synaptic weights. The gating term $f_{\mathrm{in}}(x^{\mathrm{vip}}_c - x^{\mathrm{pv}}_c)$, a

sigmoidal nonlinearity, reflects the relative activities of VIP and PV interneurons, encoding contextual certainty. When contextual uncertainty is high (low f_{in}), the mechanism suppresses learning to avoid associating incorrect contexts with observed outcomes. Conversely, under low uncertainty, plasticity is enhanced, promoting accurate learning.

To test the necessity of this gating mechanism, we developed a variant of the model that bypasses interneuron-mediated gating and uses direct thalamocortical modulation (KO-interneuron gating). We compared the action-value estimates learned by the full CogLink model to those of the KO-interneuron gating variant. The full model closely approximated the true action values of the environment, whereas the KO-interneuron gating variant deviated significantly (Figure 2.4k). These results underscore the critical role of interneuron-mediated gating in enabling accurate and continual learning across contextual switches.

2.1.5 The MD circuit approximates an algorithm that detects environmental changes optimally

To computationally understand how CogLink achieves flexible switching, we approximate the dynamics of the MD circuit with a novel algorithm. The MD circuit is structured to accumulate contextual likelihoods, enabling robust context inference. Mathematically, by letting the dynamics of TRN and frontal neurons happen instantaneously, MD circuit can be effectively described by the following equations (see Methods):

$$\begin{cases} \tau^{\text{eff}} \dot{x_1^{\text{md}}} = -\frac{2x_1^{\text{md}}}{D} + f_{\text{md}}(wx_1^{\text{md}} - wx_2^{\text{md}}) + I_1^{\text{pfc/md}} \\ \tau^{\text{eff}} \dot{x_2^{\text{md}}} = -\frac{2x_2^{\text{md}}}{D} + f_{\text{md}}(wx_2^{\text{md}} - wx_1^{\text{md}}) + I_2^{\text{pfc/md}} \end{cases}. \tag{2.1.10}$$

where $\tau^{\text{md}} = \tau^{\text{eff}} D/2$ represents the membrane time constant of MD, τ^{eff} represents the effective time constant for accumulation dynamics, $w = \frac{1}{D}$, and $I_1^{\text{pfc/md}}, I_2^{\text{pfc/md}}$ represent the PFC inputs to MD. The nonlinearity function is defined as:

$$f_{\text{md}}(x) = \begin{cases} x+1, & \text{for } -1 \le x \le 1 \\ 2, & \text{for } x > 1 \\ 0, & \text{for } x < -1 \end{cases}.$$

Defining $X = x_1^{\text{md}} - x_2^{\text{md}}$, the dynamics simplify to:

$$\tau^{\text{eff}} \frac{dX}{dt} = I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}, \text{when } |X| < D \tag{2.1.11}$$

and

$$\tau^{\text{eff}} \frac{dX}{dt} = -\frac{2X}{D} + I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}} + 2, \text{when } |X| > D. \tag{2.1.12}$$

Equation 2.1.11 corresponds to a drift-diffusion process, while Equation 2.1.12 describes the dynamics above threshold $\pm D$. When $|I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}| \ll 1$, X stabilizes at approximately $\pm D$, resulting in thresholded drift-diffusion behavior with inputs $I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}$ (Figure 2.5a).

If the PFC inputs learn the accurate generative model from Equation 2.1.7, these dynamics align with the CUSUM algorithm, a theoretically optimal method for detecting distributional

changes [43, 44]. Specifically, this occurs when:

$$I_1^{\text{pfc/md}}(t) = \log P(a_t, r_t|c = 1) + \alpha, I_2^{\text{pfc/md}}(t) = \log P(a_t, r_t|c = 2) + \alpha.$$

Here, α denotes the baseline excitation. To illustrate, if we set $X_0 = -D$ and $S_t = X_t + D$, the evolution of X_t corresponds to

$$S_t = \min(2D, \max(0, S_{t-1} + I_1^{\text{pfc/md}}(t) - I_2^{\text{pfc/md}}(t))).$$

When $S_t < 2D$, the CogLink model functions as a CUSUM algorithm with a threshold at D for detecting distributional changes (see Methods). This alignment underscores the efficiency of the thalamocortical model in identifying environmental changes and facilitating transitions between different instances of the basic CogLink for decision-making. Consistent with our theoretical predictions, the model closely approximates the behavior of the CUSUM algorithm during the first contextual switch (Figure 2.5b).

Recognizing that real-world environments often involve multiple sequential changes, the CogLink model incorporates a capping mechanism to address the limitations of the CUSUM algorithm, which is designed for single change point detection. By capping the accumulation of evidence for each context (Equation 2.1.12), this mechanism prevents overcommitment to a single context and enables the model to reset quickly and prepare for subsequent environmental shifts (Figure 2.5a). This explains the observed deviations from the CUSUM algorithm's behavior after the first detected change and highlights the importance of this feature in maintaining adaptability.

Furthermore, the evidence-capping mechanism supports the model's independence from prior knowledge of the generative model. As long as there is sufficient time between context changes for $I^{\text{pfc/md}}$ to accurately learn the contextual generative model, the CogLink model operates effectively without requiring specific environmental assumptions. This model-agnostic property not only distinguishes it from ideal observer models, which depend on precise access to the generative model, but also underscores its versatility and robustness across diverse and dynamic environments.

2.1.6 The augmented CogLink achieves flexible decision-making and continual learning by managing hierarchical uncertainty.

To empirically evaluate CogLink's performance in dynamic environments, we compared it to a Hidden Markov Model (HMM) that has prior knowledge of the hidden generative model of the environment and uses Thompson sampling for action selection (see Methods). Despite the HMM's advantage of full prior knowledge, CogLink achieves comparable levels of regret and accuracy while learning the generative model from scratch (Figure 2.5c-f). This comparison underscores CogLink's ability to perform effectively without relying on predefined assumptions about the environment.

Analyzing the models' behaviors after a context switch reveals differences in their adaptation strategies. While both models transition rapidly to the new context, the HMM switches slightly faster but requires more trials to fully stabilize its decisions (Figure 2.5e). To quantify this, we define "trials to switch" as the number of trials needed for a model to achieve 80%

accuracy over the past 10 trials following a context change. As expected, the HMM exhibits faster switching times due to its prior knowledge, though CogLink's switching performance remains competitive (Figure 2.5g).

To understand the mechanisms underlying CogLink's performance, we analyzed the evidence accumulation dynamics in MD, as predicted by the theoretical framework in the previous section. The model rapidly and accurately detects context switches after each block, leveraging these dynamics to adapt effectively (Figure 2.5a). Furthermore, CogLink demonstrates robust continual learning by accurately updating action values and the contextual generative model, even as environmental statistics shift across blocks. These learned estimates remain stable across switches, retaining prior block information while enabling adaptation to new contexts (Figure 2.4f, k).

To further explore this adaptability, we examined how CogLink leverages contextual uncertainty encoded in MD populations to support continual learning. Contextual uncertainty peaks immediately after context switches, reflecting the model's need to gather information during transitions (Figure 2.4g). This uncertainty modulation directly influences Hebbian learning rates of PFC-MD synapses (Equation 2.1.7), which rapidly decrease for the previous context after a switch. This reduction prevents the model from incorrectly learning generative models in the wrong context (Figure 2.4h). Similarly, VIP- and PV-mediated learning rates (Equation 2.1.9) are modulated to ensure that action-outcome associations are appropriately attributed to the current context (Figure 2.S4c, d).

Interestingly, uncertainty modulation operates bidirectionally between hierarchical levels. High associative uncertainty, arising from insufficient knowledge of action-outcome associations, slows the model's updates to contextual uncertainty, reflecting the difficulty in attributing evidence to the correct hierarchical process. This behavior manifests in longer switching times when CogLink encounters a novel block (Figure 2.5h). Conversely, in dynamic environments with low outcome uncertainty (e.g., reward probabilities of 90%/10%), CogLink switches contexts much more rapidly (Figure 2.5i). This suggests that reduced variability in outcomes enables the model to more readily attribute failure to context changes, thereby facilitating faster contextual updates. Together, these findings indicate that both associative and outcome uncertainty shape the dynamics of contextual uncertainty. By orchestrating these interactions across hierarchical uncertainty levels, CogLink achieves flexible decision-making and robust continual learning, even in complex and dynamic environments.

2.1.7 The model explains experimental findings showing MD causal engagement in decision-making involving changing but not stationary environments.

A number of studies have shown that MD lesions or inactivation perturbs behavioral adjustment when the environment changes, but doesn't necessarily impact behavior when conditions are stable [37, 38, 45–47]. To test whether our model exhibits these features, we performed perturbation studies by suppressing model MD neural activity (see Methods). In agreement with the corpus of experimental findings, we found that the MD-suppressed model took significantly longer to switch compared to the normal model (Figure 2.6a-f) [37]. Specifically, following a block switch, the MD-suppressed model exhibits a gradual

increase in exploration of the alternate action until commitment (Figure 2.6c). Moreover, the model provided a unique perspective on why this happens with an experimentally testable prediction: in the other model component, the M1-BG component, analysis of corticostriatal connection strength revealed fluctuating value estimates across blocks, indicating unlearning of value estimates from the previous context to adapt to the current one (Figure 2.6g, h). This is consistent with the idea that without the MD, animals may default to lower-level or model-free strategies to solve tasks that they would otherwise be able to solve with frontal control.

A natural question then arises: is MD necessary for efficient exploration in the stationary environment? To answer this question, we evaluated our models in various stationary 2-AFC tasks. In contrast to the results above, MD inhibition model still has comparable behavioral performance with Thompson sampling across various environments and only slightly degrades its performance from the full model (Figure 2.6i, Figure 2.S8). This further indicates that MD is not directly involved in simple associative learning, but rather serves as a central hub to orchestrate the learning of contextual models and modulation of downstream associative learning through learned contextual models (see Discussion).

2.1.8 Hyperactivation of striatal D2 receptors induces Schizophrenia like behaviors and MD stimulation can rescue these deficits.

There is increasing evidence that schizophrenia patients exhibit impaired belief updating processes [48–51], which may be related to susceptibility to delusional thinking [48, 52, 53]. Separately, resting state functional connectivity between the MD thalamus and PFC is altered in schizophrenia patients [54–61]. A recent study using mouse models carrying schizophrenia-relevant mutation showed both perturbed MD function and belief updating, and optogenetic MD stimulation led to a normalization of the belief updating process [14]. While striking, these findings leave open the question of what the mechanistic links are between MD perturbation and the belief updating decision-making process are.

Inspired by the fact that most antipsychotics targeting D2 receptors (D2Rs) are dopamine antagonists [62–64] and the most Schizophrenia patients show an elevated level of striatal D2Rs [65, 66], we consider a model with hyperactivation of striatal D2Rs. Since a hyperactive striatum is expected to inhibit the MD thalamus [67], we model our impaired model with decreased MD excitability (see Methods, Figure 2.7a).

$$\begin{cases} \tau^{\text{eff}} \dot{x_1^{\text{md}}} = -\frac{2x_1^{\text{md}}}{D} + \beta_{\text{d2}} f_{\text{md}}(wx_1^{\text{md}} - wx_2^{\text{md}}) + \beta_{\text{d2}} I_1^{\text{pfc/md}} \\ \tau^{\text{eff}} \dot{x_2^{\text{md}}} = -\frac{2x_2^{\text{md}}}{D} + \beta_{\text{d2}} f_{\text{md}}(wx_2^{\text{md}} - wx_1^{\text{md}}) + \beta_{\text{d2}} I_2^{\text{pfc/md}} \end{cases}.$$

Here, β_{d2} is the decreased excitability from D2R hyperactivation. It is observed that the impaired model has suboptimal regret and accuracy and never fully commits to accurate choices after a block switch (Figure 2.7b-e). Moreover, the model also exhibits longer exploration after a switch (Figure 2.7f, g), showing impaired of cognitive flexibility [14, 68]. On the other hand, the impaired model also shows an elevated win-switch rate (Figure 2.7h), suggesting a perception of environmental instability leading to this erratic behavior. These two seemingly contradictory behaviors of slow switching and high win-switch are consistent with experimental findings in both patients and animal models [14, 49, 68, 69].

To investigate the neural mechanisms behind these two behaviors, we first examine the drift process formed by the difference in activities of two contextual MD populations. Compared to the normal drift process, which is well-separated across contexts, the impaired drift process saturates its evidence at a much lower threshold, inducing a strong prior for the volatility of the environment (Figure 2.7i). To understand its underlying dysfunctions, we leverage CogLink's capacity to approximate an algorithm and show that the threshold of the accumulation dynamics becomes smaller. Moreover, the impaired normative model exhibits leaky evidence integration, further reinforcing its prior belief in the environment's volatility (See Methods). By examining the contextual uncertainty the impaired model decoded, we can also observe its strong belief on environmental volatility (Figure 2.S7b).

On the other hand, the corticostriatal strengths exhibit more homogeneous profiles (Figure 2.S7c), indicating low associative uncertainty. Moreover, its learning rate modulated by VIP/PV interneurons is much lower than the normal model (Figure 2.7m, Figure 2.S4d). This suggests that although the impaired model has a strong prior on the volatility of the environment, it also updates its belief at a much slower rate within a single context, potentially contributing to slow switching.

Numerous studies have demonstrated alterations in PFC-MD coupling in schizophrenia patients [54–61]. Given that our model suggests PFC-MD connections are involved in learning contextual generative models, we aim to investigate whether the impaired model also exhibits deficits in model learning. Compared to the normal model, the impaired model struggles to learn the correct contextual generative model of the environments (Figure 2.7k). To probe the mechanism, we examine the learning rate of PFC-MD connections. Indeed, lower excitability results in neuronal activities insufficient to induce Hebbian plasticity (Figure 2.7l).

To restore the model's learning capacity, we introduce a small excitatory current into the MD neurons (Figure 2.7a). This intervention reduces both regret and exploratory behaviors after a switch (Figure 2.7b-g). Additionally, although the rescue model does not reduce the win-switch rate (Figure 2.7h), the drift process of the rescue model exhibits a higher threshold for evidence accumulation, indicating a weaker prior on environmental volatility (Figure 2.7i). Moreover, the rescue model learns a more accurate generative model of the world (Figure 2.7k) and reinstates proper learning in PFC-MD connections (Figure 2.7l). These findings are consistent with the recent MD activation experiments on Schizophrenia-relevant mouse models [14].

2.1.9 CogLink reveals thalamic regulation of reinforcement learning strategies across prefrontal-striatal network

Under conditions of uncertainty, human choice behavior is controlled by habitual and goal-directed systems, which can be well formalized by the algorithmic framework of model-free and model-based reinforcement learning (RL) [70, 71]. Model-free strategies prioritize computational efficiency, determining the value of each action and guiding behavior based on reward prediction errors. Conversely, model-based control strategically computes optimal actions by incorporating contextual details, thereby facilitating adaptable, outcome-specific behaviors [72–74]. Although the two systems have been considered as competitors for behavioral control [70], evidence indicates their cooperative interactions in response to

concurrent cognitive demands [75, 76].

The prefrontal cortico-striatal circuits are considered to be the neural substrates underlying distinct components of the RL process [77]. Within in this system, model-free control is most strongly associated with the dorsolateral striatum [78–80], while the ventromedial (vmPFC) and dorsolateral prefrontal cortex (dlPFC) is essential for model-based processes as it combines rewards with contextual information [81, 82]. Focal brain lesions targeting dlPFC can shift the control of behavior from one RL system to another [73], indicating that each system has its own distinct representation. Disruptions in the prefrontal-striatal circuits are associated with recurring pathological behaviors, such as those seen in obsessive-compulsive disorders [83], and with the behavioral rigidity observed in schizophrenia [68]. Despite the recognized importance of RL strategies in adaptive behavior, their precise neural implementation and mechanisms governing their arbitration remain incompletely understood.

Unlike first-order thalamic nuclei that primarily relay peripheral sensory information to the cortex, the mediodorsal thalamus (MD) regulates excitatory/inhibitory balance and effective connectivity within and across frontal areas [35, 84]. This regulation is thought to mediate executive functions that underlie adaptive behavior [30, 34, 85]. Research in non-human animals has clarified the microcircuit substrates involved in these processes, demonstrating, for example, that flexible switching of attention upon cue changes involves a subset of MD neurons that activate task-relevant prefrontal ensembles, while another MD subset suppresses task-irrelevant ones [36, 37]. More recently, experiments have shown that covert rule reversals engage a subset of MD neurons that encode context prediction errors, which switch PFC state underlying strategy updating [39]. In human neuroimaging studies, the MD has been shown to integrate inputs from various prefrontal regions when dynamically selecting between competing behavioral strategies, with more widespread interactions between the MD and large-scale learning networks as task demands increase [86–88]. However, the precise role of the human MD in adaptive behavior generally and RL strategies specifically remain largely unknown.

Here, we address this critical gap by pursuing a serendipitous finding. Specifically, in a human probabilistic reversal learning task, we found robust MD activation upon strategy updating as well as the encoding of the new strategy. Because subjects learned the initial strategy and detected the environmental changes to various degrees, they showed variability in their switching behavior that was well-fit by a model-based RL strategy on some reversals and model-free on others. fMRI revealed that dorsal prefrontal cortex engaged in the former process while striatum in the latter. Remarkably, the MD engaged in both, but with its lateral subdivision showing model-based activation and medial division model-free. Causal connectivity analysis showed that transthalamic processing was progressively recruited as subject transitioned from deploying a stable strategy to updating it utilizing a model-based strategy, with model-free updating exhibiting intermediate values. This tantalizing finding was explained by CogLinks, a novel class of biologically plausible mechanistic models of forebrain networks capable of solving complex tasks. CogLinks uniquely demonstrated that under certain conditions, model-free RL is not instantiated as a distinct algorithm but instead an outcome variation of the same model-based RL algorithm. Specifically, our modeling explains the emergence of a model-free strategy in a bottom-up manner, where prefrontal-thalamic mechanisms of context inference temporarily fail, resulting in the inability to rapidly adjust prefrontal dynamics underlying new contextual learning. This results in the slow over-writing

of prefrontal strategy substrates, which our fMRI decoding confirms empirically. Overall, our findings reveal an unexpected role of transthalamic processing in human cognitive flexibility and highlight the value of biologically plausible modeling in showing how complex algorithms are implemented in the human brain.

Human behavioral and fMRI data on a probability reversal task

We leveraged a probabilistic rule reversal task in humans, in which the associations between two tactile stimuli and responses are initially learned and then reversed (Figure 2.8a, b). In each trial, one out of the two tactile patterns was applied to the right index fingertip. Participants had to find out, by trial and error, whether the applied tactile pattern was associated to a "Go"-response, in which participants should press a button with the left index finger, or to "NoGo", in which they should refrain from pressing the button (Figure 2.8a). For one tactile pattern, one response option (e.g., 'Go') had a higher reward probability ($p = 0.7$) than the other ('NoGo', $p = 0.3$). For the alternative tactile pattern, probabilistic reward associations for 'Go' and 'NoGo' were reversed ('Go', $p = 0.3$; 'NoGo', $p = 0.7$). Each block consisted of 45 trials and was divided into two phases: the initial learning phase and the reversal phase. During the initial learning phase, participants learned the stimulus-response association. At a variable time-point, ranging between the twentieth to the twenty-fifth trial, the probabilistic associations between both tactile patterns and corresponding Go/NoGo responses were reversed, marking the start of the reversal phase (Figure 2.8b). In the reversal phase, participants had to switch their decision strategy by reversing cue-response associations. A novel pair of tactile patterns were randomly selected from eight alternative patterns for each new block, which were presented to the participants at the beginning of each block.

We then plotted the averaged proportion of correct strategy across participants aligning the reversal phase starting from the rule reversal point (Figure 2.8c). During the initial learning phase, the strategy about the stimulus-outcome association was quickly learned and was then held in the exploitation state to guide decisions across the next trials (i.e., Steady State, Figure 2.8c). Following rule reversal, the performance dropped as participants shifted to the new strategy governed by the new rule (i.e., Switch, Figure 2.8c). Based on this group performance, we defined the 10 trials immediately following the reversals as the Switch (SW) period, as this window effectively encompasses both the exploratory phase and the successful transition to the new rule content. The 10 trials before the reversals were defined as the Steady State (SS). As expected, the proportion of correct strategy in SS was significantly higher than in SW ($t(31) = 13.20$, $p < 0.0001$, two-tailed, Figure 2.8d).

In our human reversal learning task, rule reversals forced the updating of the decision strategy to maximize reward. The underlying process can be understood to largely rest on RL strategies that are either MF or MB. The MF system learns the value of different behaviors solely based on reward prediction error (RPE), while the MB system builds an intrinsic model about state transitions in the decision-making process, taking state transition relationships into consideration (state prediction error, SPE) [89, 90]. To explore this further, we calculated the distances between human behavioral performance and the behaviors simulated by the MF and MB models. For each participant, we assigned blocks to either the MF or MB category based on which model provided a better fit. First, we compared behavioral performance between the MF and MB blocks. During the SW period, the averaged proportion of correct

strategies in MB blocks was significantly higher than in MF blocks (linear mixed-effect test, $t(62) = 4.961$, $p = 5.7710^{-6}$, Figure 2.8e). On average, MF blocks required more than 5 trials (mean±SD $= 5.3 \pm 2.2$) to reach chance-level performance following a rule reversal, whereas MB blocks required fewer than 4 trials (mean±SD $= 3.7 \pm 1.5$). This difference suggests that the MF strategy involves more exploratory behavior after environmental changes.

To dissociate the neural regions associated with MF and MB RL, we characterized participants' RL behavior using the reward prediction error (RPE) and state prediction error (SPE), respectively. These RPE and SPE signals were then modeled in the GLMs of the fMRI data as parametric modulators, allowing us to investigate how distinct brain regions contribute to the two RL strategies. We found that during Switch, the RPE was significantly correlated with activity in CN ($x = 16$, $y = 2$, $z = 20$, $t(31) = 4.17$, pFWE-SVC $= 0.027$, Figure 2.8f) and MD ($x = -2$, $y = -22$, $z = 10$, $t(31) = 4.55$, pFWE-SVC $= 0.011$, Figure 2.8f), but not dmPFC. In contrast, activity in dmPFC ($x = 14$, $y = 12$, $z = 64$, $t(31) = 6.39$, pFWE-SVC < 0.001, Figure 2.8g) and MD ($x = 4$, $y = -20$, $z = 12$, $t(31) = 4.05$, pFWE-SVC $= 0.036$, Figure 2.8g), but not CN, was significantly correlated with the SPE.

Despite the fact that the MD was involved in both RPE and SPE, we found that the model-based MD engagement was localized to its lateral subdivision, whereas the model-free MD to its medial one (Figure 2.8h). These fundamental differences in MD representations were corroborated by their structural tractography and functional connectivity (i.e., psychophysiological interaction, PPI). MD-tractography from an independent human cohort ($n = 113$) indicated, that the MB-related part of the MD dominantly projected to the dorsal PFC, whereas the MF-related MD to ventral PFC (i.e., vmPFC/OFC, Figure 2.8i). The PPI findings are consistent with this, showing MB-related MD connectivity with dorsal PFC ($x = 22$, $y = 22$, $z = 42$, $t(31) = 6.01$, pFWE-SVC $= 0.036$) and MF-related MD connectivity with the OFC ($x = 46$, $y = 22$, $z = -14$, $t(31) = 6.38$, pFWE-SVC $= 0.029$, Figure 2.8j).

To explore network-level computations relevant to strategy updating, we applied another DCM that focused on the strategy-related MD subdivisions (medial, MDm and lateral, MDl), along with their cortical counterparts (OFC and dlPFC, respectively). Given the lack of recurrent connectivity in the thalamus, direct connections between the two MD nodes were discounted [35]. Bayesian parameter averaging (BPA) revealed significant connectivity patterns in Steady State (SS), and Switch (SW) trials that we separated into MF and MB updates (SW_{MF}, SW_{MB}, Figure 2.8k). We found a peculiar pattern of effective connectivity changes, which strikingly varied from the SS to SW_{MB} with the SW_{MF} exhibiting intermediate values. Specifically, cortico-cortical connections (i.e., reciprocal connections between OFC and dPFC) progressively decreased (one-way Anova, $p < 0.0001$), whereas thalamocortical connections (i.e., outputs from two MDs) increased as subjects relied on a MB strategy (one-way Anova, $p < 0.0001$, Figure 2.8l). These findings suggest the role of the MD in influencing multiple prefrontal areas and providing indirect transthalamic communication routes between directly connected cortical areas.

CogLink reveals thalamic regulation of reinforcement learning strategies across prefrontal-striatal network

To investigate the circuit mechanisms underlying this network pattern, we employed CogLink modeling, a mechanistic framework of the forebrain network designed to solve contextual

decision-making27 (Figure 2.9a). CogLink is distinct in its integration with normative modeling, enabling a principled approach to building a cognitive mechanistic model27. The network architecture includes recurrent circuits representing the executive dlPFC, where past actions and outcomes are encoded, and the OFC, where contextual values are stored and updated. Additionally, two MD networks represent the lateral and medial mediodorsal thalamus (MDl and MDm). MDl is responsible for inferring task context and generating update signals, while MDm processes and relays RPEs (see Methods; Figure 2.9a).

Our model successfully solved the probability reversal task, achieving 80% accuracy before the reversal ($82.6 \pm 1.7\%$, Mean±SEM; $n = 500$) and recovering to 80% accuracy before the block ended ($79.4 \pm 1.8\%$, Mean±SEM; $n = 500$). MDl neurons exhibited strong contextual encoding (Figure 2.9b), with their activity closely linked to MB behaviors, mirroring patterns observed in human subjects ($r = 0.56$, $p = 2.0010^{-5}$, Figure 2.9c). To further assess this relationship, we asked whether MDl activity can be used to distinguish between MB and MF strategies that the model may exhibit. Indeed, this approach reproduced the behavioral differences observed in human subjects: switch times were longer in MDl activity-derived MF blocks compared to MB blocks (Figure 2.9d). Furthermore, and more critically, causal connectivity analysis of the CogLink network replicated findings from DCM of fMRI data. Specifically, steady-state behavior was associated with high cortico-cortical and low thalamocortical connectivity, while model-based switching showed the opposite pattern, and model-free switching exhibited intermediate values (Figure 2.9e).

Our CogLink network enabled us to investigate the neural mechanisms underlying MB and MF switching which were not accessible by fMRI alone. Analysis of MDl neural activity during MB and MF switches revealed that new contextual signals emerged only in MB blocks, whereas MF blocks failed to generate such representations (Figure 2.9f). In the CogLink network, distinct contextual populations of MDl neurons selectively modulate disjoint populations in the orbitofrontal cortex (OFC), regulating both their activity and plasticity (see Methods; Supplementary Figure 2.S8a, b)17. Consequently, during MF switches, the absence of an alternative contextual representation in MD caused the same OFC population previously engaged in the initial context to remain active (Figure 2.S8a, b). As a result, switching behavior relied on overwriting the existing mapping in the same OFC population. Specifically, in Fig. Figure 2.9g, we observed a "crossing" in the activity of two OFC populations in response to cue 1: the population tuned to "Go", which initially exhibited high activity, decreased its activity after the switch, while the population tuned to "NoGo", which initially exhibited low activity, increased its activity.

The network modeling results in a critical prediction: model-based switching involves frontal networks simultaneously encoding multiple strategies that can be rapidly toggled, while model-free switching relies on the slower overwriting of the same strategy in frontal networks. To test this prediction, we revisited our original human finding of strategy decoding from frontal networks (Figure 2.8e-j). With this new insight, we separated switching behavior into model-based and model-free blocks as Figure 2.8e (see Methods). Strategy decoding using RSA showed a striking confirmation to the CogLink prediction: MB switching exhibited clear evidence of toggling between distinct strategy representations, whereas MF switching did not. In the dmPFC, for instance, MB behavior was linked to a significantly larger difference between the two strategy representations ($t(31) = 4.15$, $p = 2.5 \times 10^{-4}$), with similar effect in the dlPFC and OFC ($t(31) = 2.42$, $p = 0.02$; , $t(31) = 2.79$, $p = 0.009$ respectively,

Figure 2.9h-i). These findings suggest that, during MB switching, the original ('Staying') representation is suppressed while the network engages a new ('Switching') representation, resulting in a pronounced difference in activation patterns. By contrast, MF switching involves overwriting the same representation, leading to less differentiation between old and new strategies. Thus, across frontal networks, model-based strategy updates are associated with the ability to maintain and rapidly toggle between multiple representations, whereas model-free updates lack this capability. These results provide a novel 'bottom-up' perspective on the neural substrates of these two algorithmic processes.

These findings show that model-free switching depend on slow modification of an existing strategy. However, if model-free switching relies on overwriting rather than encoding a new representation, what prevents MDl from forming an alternative contextual representation in MF blocks? To understand why MDl fails to form an alternative contextual representation in MF blocks, we examined the strengths of dlPFC-MDl model synapses. In the CogLink model, these synapses learn the contextual generative model through Hebbian plasticity, enabling accurate contextual inference (see Methods). Our analysis revealed that in MB blocks, the generative model, learned by these synapses, closely approximated the true generative model of the environment. In contrast, in MF blocks, the learned generative model deviated significantly (Figure 2.S8c, d). Specifically, synaptic strengths during SS in MB blocks encoded larger value differences between correct and incorrect strategies compared to MF blocks (Figure 2.9j). This failure in MF blocks likely results from the stochasticity of actions and rewards, which can provide conflicting evidence about the current context, resulting in insufficient learning of the environmental generative model before the rule reversal (Figure 2.9j, Figure 2.S8c, d). As a result, the model struggles to distinguish new contexts in MF blocks because the generative model from the previous context is not different enough from the post-reversal environment.

These CogLink results lead to a key behavioral prediction. Since larger estimated value differences between correct and incorrect strategies increase the likelihood of selecting the correct strategy, and MB blocks exhibit larger value differences compared to MF blocks (Figure 2.9j), we predict that the ratio of correct strategy selection during SS should be higher in MB blocks. In other words, pre-reversal behavioral performance directly influences post-reversal strategy selection. Consistent with this prediction, we observed a significant difference in SS performance between MF and MB blocks in both the CogLink model and human data (Figure 2.9k). These findings suggest that MF behaviors emerge as a consequence of incomplete learning of the pre-reversal generative model.

Taken together, these modeling results demonstrate that model-free strategy does not arise from a distinct algorithm but instead emerges as a variation of model-based mechanisms. Specifically, our findings suggest that model-free behavior results from a temporary failure in prefrontal-thalamic mechanisms of context inference due to insufficient learning of the environmental generative model before the reversal. This bottom-up perspective provides a mechanistic explanation for how variations in neural substrates shape decision-making flexibility, highlighting that, under certain conditions, model-free RL strategies reflect a temporary limitation rather than a fundamentally separate process.

2.2 Discussion

2.2.1 Biological plausibility of the CogLink

CogLink incorporates diverse, biologically inspired mechanisms to model associative and contextual uncertainty processing in frontal networks. To ensure computational tractability while maintaining biological plausibility, the model includes certain assumptions and simplifications, which we discuss below.

To address hierarchical uncertainty, CogLink models hierarchical cortico-thalamic-basal ganglia (BG) loops. In animals, these loops process different types of information—such as motor, limbic, and associative—through parallel streams [91, 92]. In CogLink, we specifically model the motor and associative components of the cortico-thalamic-BG loop to process low- and high-level uncertainty, respectively.

The basic CogLink focuses on the premotor cortico-thalamic-basal ganglia (BG) loop, emphasizing BG's role in associative learning under uncertainty. Instead of modeling the full complexity of BG circuitry, including direct and indirect pathways [93], we adopt a simplified actor-critic structure, commonly used to capture BG's associative learning capacity [15]. Importantly, we assume that striatal neurons encode a distribution of action value beliefs, updated by dopamine neurons through distributional reward prediction errors (RPEs). This assumption is suggested by evidence that dopamine neurons exhibit inhomogeneous responses forming distributional RPEs [20]. To implement sampling of these distributions, we hypothesize random input sparsification in the premotor cortex, inspired by evidence of variable sparse firing in supragranular cortical layers (layers 2/3) [94–96], which are parts of the cortical circuit known to generate striatal inputs.

The augmented CogLink extends this framework to an associative cortico-thalamic-BG loop to address contextual uncertainty. Although thalamic-striatal connections also have been implicated in flexible behavior [97, 98], we focus on the well-established role of the prefrontal cortex (PFC) and mediodorsal thalamus (MD) in cognitive flexibility [14, 36–39, 99] and choose not to model those connections. Another key assumption in this model is the use of disjoint cortical representations that are contextually activated. This modular strategy representation is supported theoretically [100, 101] and experimentally, as Kim et al. demonstrated context-dependent stable representations of the same action in mice [102].

Extensive studies have shown that the MD thalamus explicitly encodes task contexts across various decision-making paradigms [37, 38, 40], with recent findings highlighting its role in encoding contextual uncertainty [39]. Based on this evidence, we propose that MD acts as a central hub for encoding contextual uncertainty. In CogLink, thalamocortical projections serve dual roles: driver-like projections maintain stable contextual representations via recurrent PFC-MD loops (Equation 2.1.10), while modulatory projections influence cortical connectivity and plasticity through interneurons (Equation 2.1.8, Equation 2.1.9). These dual roles align with the classical distinction between core (driver) and matrix (modulatory) thalamic projections [103, 104], as well as recent findings on the diverse functions of the thalamus in modulating cortical dynamics [36, 37, 99, 105–111].

Additionally, we hypothesize that PFC populations modulated by these interneuron-mediated thalamocortical projections activate downstream premotor circuits in a context-dependent manner. Supporting this hypothesis, Wang and Sun [112] demonstrated that PFC

sends context-encoding inputs to the premotor cortex to initiate movement.

Together, these features make CogLink a biologically plausible framework for understanding how hierarchical uncertainty shapes decision-making and cognitive flexibility.

2.2.2 Neural representation, computation and usage of uncertainty

Even though it is well established that uncertainty profoundly impacts behavior [113–115], an overarching computational framework for how different forms of uncertainty are encoded, represented and decoded to drive behavioral adjustment is lacking [116, 117]. On the encoding front, uncertainty may be represented at the single neuron level, which empirical studies have found in basal ganglia [118] and frontal cortex [119, 120]. Another encoding strategy is in the form of a distribution at the neural population level [17, 121]. A distributional code is more computationally demanding but may offer flexibility through differential decoding (e.g. different parts of the distribution can be selectively weighted based on other state variables). There are different frameworks to represent distributions in a neural population such as probabilistic population codes [17, 122, 123], sampling-based codes [18, 121, 124], explicit probabilistic codes [19, 125] and quantile codes [20]. Our model used two distinct ways to encode uncertainty as distribution, which are motivated by empirical findings that we explain below.

For associative uncertainty, which is computed in the BG component of our model, it is encoded in a quantile code similar to distributional reinforcement learning (RL) [20]. This is consistent with the fact that the striatum is a major output for dopaminergic neurons and indeed our model shows that it is quite straightforward for dopamine-gated plasticity to update this form of BG distribution. In addition, our simulations show that it is easy to sample from a quantile distribution through recurrent competitive dynamics because sampling neurons corresponds to sampling the corresponding probabilistic quantity in such code. We should note, however, that even though we use a quantile code similar to the distributional RL literature, our model implementation differs from that of previous distributional RL models in two crucial ways: first, instead of varying the optimism of each synapse, we vary the initial strength of each synapse. Second, we add a mechanism to couple this representation to behavioral adjustment through sampling. This is a conceptually-motivated deviation from previous distributional RL implementations which focus on the notion that learning a distribution allows for better learning and generalization, without a clear link to how the representation is behaviorally decoded. In our model, we posit that sampling, which can be efficiently done on its quantile code, is a mechanism to couple uncertainty to efficient exploration.

For contextual uncertainty, which is computed in the MD thalamus, it is encoded as an explicit probabilistic code inspired by past experimental works showing MD encodes context [37]. This representation has the distinct advantage of contextually modulating local learning (Figure 2.4g, h). However, the detailed mechanism of how this representation arises is poorly understood. To investigate how to compute such representation, we include two mechanisms in PFC-MD circuits. First, PFC-MD connections learn the contextual model of the environment at a single trial level via Hebbian learning. Second, recurrent dynamics in PFC-MD circuit accumulate the single trial likelihoods from corticothalmic inputs to calculate the current likelihood of contexts conditioned on previous experiences. Based on

recent evidence showing that thalamus modulates both activities and plasticity of downstream cortical networks [38, 39, 41, 42], we include interneuron-mediated pathways to allow the contextual MD representation to accomplish these functions and explain how contextual uncertainty can impact exploration and learning through these mechanisms.

2.2.3 Thalamocortical interaction as a system-level solution for flexible behaviors and model-based learning.

Both animals and humans rely on a delicate coordination between model-free and model-based learning processes to adapt flexibly to their environments [74, 126–129]. BG has traditionally been associated with model-free learning, while PFC has emerged as a locus for model-based learning and the mediation between the two systems [15, 130–134]. However, the intricate mechanisms underlying the coordination of these learning types remain poorly understood. In our study, we propose the thalamus as a potential communication hub orchestrating this coordination, hypothesizing a detailed circuit mechanism to achieve this integration.

The thalamus is well-known for its topographic and reciprocal connections with the neocortex, as well as its projections to the BG [104, 135]. While traditionally viewed as a relay station for sensory information, recent research has revealed its involvement in diverse functions across sensory [99, 107, 136, 137], cognitive [36–39, 106] and motor domains [105, 109]. The convergence of inputs onto the thalamus and its diverse modulation of cortical and BG circuits position it ideally as a locus of plasticity for learning contextual states in model-based learning to coordinate between model-free and model-based systems.

In our model, PFC-MD circuits learn the contextual model of the environment and represent contexts in MD. This model-based learning component then modulates both plasticity and activities of downstream model-free learning component, corticostriatal circuits, based on estimation and uncertainty of current contexts in MD. Lesioning MD disrupts this coordination, impairing the model's ability to flexibly switch behaviors in dynamic environments. However, the lesioned model can still perform in a stationary environment, indicating MD is not involved directly in pure model-free learning. These observations underscore the pivotal role of PFC-MD circuits as the locus of model-based learning, utilizing the learned model of the world to modulate corticostriatal model-free learning and achieve flexible behaviors.

2.2.4 Brains provide different levels of specialized mechanisms for credit assignment.

The role of dopamine innervation in the basal ganglia is well-established in carrying reward prediction error (RPE) signals to reinforce behaviors associated with unexpected rewards through synaptic plasticity mechanisms [130, 138–140]. However, decision-making in animals involves navigating through multiple cues, actions, and contexts, posing the challenge of appropriately assigning credits to the corresponding synaptic connections responsible for the unexpected rewards—a phenomenon termed credit assignment [141–144].

Traditional machine learning approaches, such as backpropagation, attempt to reinforce internal activity states leading to unexpected rewards [143]. However, backpropagation

relies on symmetric feedback weights and a separation of errors and activities, which are not observed in biological brains [144]. Additionally, traditional artificial neural networks often struggle with crediting sensorimotor associations to the correct context across different contexts, leading to catastrophic forgetting [145–149].

To address these challenges, researchers have proposed a plethora of cellular, circuit, and system-level mechanisms for proper credit assignment [150–160]. In our work, we integrate mechanisms at multiple levels to facilitate credit assignment.

At the cellular level, Hebbian-like learning in thalamocortical connections enables credit assignment by crediting associations to specific contexts only when the model is confident in its context inference. Circuit-level credit assignment is exemplified by dopamine-gated plasticity in the basal ganglia, where only corticostriatal connections corresponding to the chosen action undergo plasticity changes. This can be implemented by maintaining an eligibility trace from a motor action's efference copy back to corticostriatal synapses.

Moreover, thalamocortical interactions via interneurons offer a system-level solution for credit assignment. In our model, the thalamus modulates cortical learning through cortical interneurons to correctly attribute sensorimotor associations to the appropriate context. PV neurons inhibit context-irrelevant cortical ensembles to prevent learning in the wrong context, while VIP neurons facilitate downstream learning when the model is confident in its inferred context.

These examples illustrate the brain's utilization of diverse mechanisms operating at different levels to perform credit assignments effectively in complex natural environments.

2.2.5 CogLink Network as a way to link molecular and behavioral changes in Schizophrenia.

Genetics are recognized as significant risk factors in schizophrenia [161], and computational modeling has highlighted deficits in belief updating as a key aspect of the disorder [48–51]. However, the intricate mechanisms bridging these genetic risk factors and belief updating deficits remain poorly understood. Our CogLink Network, capable of linking mechanisms with normative behavior, offers a novel avenue to explore these connections.

In constructing our schizophrenia model, we focus on a striatal D2R overexpression model because most antipsychotics targeting D2 receptors (D2Rs) are dopamine antagonists [62–64] and most SZ patients showed an elevated level of striatal D2Rs [65, 66]. We focus on the effects of striatal D2R overexpression on the PFC-MD circuit, given mounting evidence implicating alterations in these regions in schizophrenia pathology [54–61]. Since the abundance of D2Rs increases the inhibition from BG to thalamus, we model schizophrenia by reducing the excitability of MD neurons to mimic high level of BG inhibition.

Our schizophrenia model replicates experimental findings in both patients and animal models such as exploratory behavior following contextual switches and an elevated win-switch rate [14, 49, 68, 69]. CogLink Network further explains how circuit-level perturbation connects to these specific cognitive impairments. In particular, by examining the corresponding normative model, we can show that the model exhibits a much lower threshold for evidence accumulation and the accumulation dynamic becomes leaky, indicating a strong bias toward environmental volatility. Additionally, decreased excitability in MD compromised the ability

of PFC-MD connections to accurately learn the environmental model. To address this impairment, we applied current injections to MD to restore activity levels to a range conducive to Hebbian plasticity. Remarkably, the rescue model demonstrated reduced exploratory behavior following switches and exhibited a higher threshold for MD activity switching, indicative of a diminished bias towards environmental volatility. Moreover, the rescue model exhibited improved learning of the environmental model within its PFC-MD connections. These findings validate recent experiments in Schizophrenia-related animal models [14] and demonstrate the utility of CogLink Network in computational psychiatry.

2.2.6 CogLink Network vertically integrates and describes neural phenomena from different perspectives.

Different modeling approaches offer distinct perspectives on understanding brain computation [162, 163]. Normative theories have traditionally elucidated animal behaviors and neural coding but often lack direct connections to lower-level neural correlates. In contrast, mechanistic models provide such links, allowing for testable predictions through symmetric perturbations on models and animals. However, understanding mechanistic models at the computational level can be challenging due to their complexity.

In this paper, our CogLink Network aims to bridge the gap by constructing a mechanistic model capable of approximating normative models. By incorporating observed neural mechanisms into our model, we establish a direct connection to neural circuits. Simultaneously, approximating normative theories enables mathematical analysis, offering both quantitative and qualitative computational insights. Furthermore, our CogLink Network provides distinct advantages in providing a model hypothesis. On the one hand, neural mechanisms provide strong biological prior to a normative model; on the other hand, connections to a normative model provide a guide to adjust the mechanistic parameters to achieve complex cognitive behaviors. We view this modeling approach as the initial step toward integrating Marr's three levels of analysis—computational, algorithmic, and implementation levels. Furthermore, many neurological diseases have genetic origins along with cognitive symptoms. Since our modeling approach contains both mechanistic details and computational insights into behaviors, it can serve as a lens to study these neurological diseases.

Recently, population dynamic approaches have proven potent in uncovering underlying computations from electrophysiological data [164]. However, these approaches often lack integration with connectivity and functional data information, limiting their ability to provide insights at the circuit level. In the future, we aim to develop a CogLink Network that incorporates electrophysiological data as well as connectivity and functional data.

2.3 Methods

2.3.1 Model overview.

Our model is specified by a differential equation governing the evolution of the neural activities (Equation 2.1.1), a set of synaptic weights, and synaptic update rules (Equation 2.1.2, 2.1.7, 2.1.9). In the subsequent section, we provide a more detailed specification of our model.

2.3.2 Basic CogLink model.

This section presents the details of the basic CogLink model. Let A denote the number of alternatives, M represent the size of a premotor cortex ensemble, and K indicate the sparsity of cortical activities. In the basic CogLink model, there are A ensembles of premotor neurons. Within the ath ensemble, the premotor cortex activities $x_a^{\text{alm}} \in \mathbb{R}^M$ evolve according to the following equation:

$$\tau^{\text{alm}} \frac{dx_a^{\text{alm}}}{dt} = -x_a^{\text{alm}} + W_a^{\text{alm}} g(x_a^{\text{alm}}) + I + 0.2 \frac{dB_t}{dt}. \tag{2.3.1}$$

Here, the membrane time constant $\tau^{\text{alm}} = 1/6$, excitatory inputs $I = K - 0.25$, recurrent synaptic weights $W_a^{\text{alm}} \in \mathbb{R}^{M \times M}$ are defined as:

$$W_a^{\text{alm}} = \begin{bmatrix} 0.75 & -1 & \cdots & -1 \\ -1 & 0.75 & \cdots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \cdots & 0.75, \end{bmatrix}.$$

The nonlinearity function $g : \mathbb{R} \to \mathbb{R}$ is defined as:

$$g(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } 1 \geq x > 0 \\ 0 & \text{otherwise} \end{cases}$$

B_t represents a standard Brownian motion with unit variance. The selection of the recurrent weights W_a^{alm} and inputs I is designed to implement K-WTA dynamics [165].

The premotor cortex then projects to the BG. The activities of the BG at the ath ensemble, $x_a^{\text{bg}} \in \mathbb{R}^M$, evolve according to the following equation:

$$\tau^{\text{bg}} \frac{dx_a^{\text{bg}}}{dt} = -x_a^{\text{bg}} + V_a^{\text{alm/bg}} \circ g\left(x_a^{\text{alm}}\right) \tag{2.3.2}$$

Here, the membrane time constant $\tau^{\text{bg}} = 0.1$, and the premotor cortex-BG synapses, $V_a^{\text{alm/bg}} \in \mathbb{R}^M$, are initialized with:

$$\forall a \in [A], m \in [M],\ V_{a,m}^{\text{alm/bg}} \leftarrow \bar{V}_{a,m}^{\text{alm/bg}},\ \text{where } \bar{V}_{a,m}^{\text{alm/bg}} = \frac{m}{M}.$$

The BG then projects to the motor cortex, and the recurrent competitive dynamics of the motor cortex determine the action a_t at trial t. Specifically, the activities of the motor cortex, denoted as $x^{\text{mct}} \in \mathbb{R}^A$, evolve according to the following equations:

$$\forall a \in [A],\ I_a^{\text{mct}} = W_a^{\text{bg/mct}} x_a^{bg} \tag{2.3.3}$$

and

$$\forall a \in [A],\ \tau^{\text{mct}} \frac{dx_a^{\text{mct}}}{dt} = -x_a^{\text{mct}} + g\left(\sum_{i=1}^{A} W_{a,i}^{\text{mct}} x_i^{\text{mct}}\right) + I_a^{\text{mct}}. \tag{2.3.4}$$

Here, the membrane time constant $\tau^{mct} = 1$, and BG-motor cortex synapses that tuned to action a, $W_a^{\text{bg/mct}} \in \mathbb{R}^M$, where $\forall m \in [M], a \in [A]$, $W_{a,m}^{\text{bg/mct}} = 1/K$. The recurrent synaptic weights, $W^{\text{mct}} \in \mathbb{R}^{A \times A}$, are defined as:

$$W^{\text{mct}} = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ -1 & 1 & \cdots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \cdots & 1 \end{bmatrix}.$$

The action a is chosen as a_t if either x_a^{mct} reaches the threshold $= 1$ within 5 seconds after the trial starts or chosen stochastically from a softmax distribution, $a_t \sim \text{softmax}(30x^{\text{mct}})$, after that.

Once the model receives the reward r_t, it forms a distributional reward prediction error (RPE), denoted as $\delta \in \mathbb{R}^M$:

$$\forall m \in [M],\ \delta_m = r_t - V_{a_t,m}^{\text{alm/bg}}.$$

This RPE is then used to update the premotor cortex-BG synapses, $V_{a_t}^{\text{alm/bg}}$, according to the equation:

$$\forall m \in [M],\ V_{a_t,m}^{\text{alm/bg}} \leftarrow V_{a_t,m}^{\text{alm/bg}} + \eta_{a_t}\delta_m.$$

Here, the learning rate, η_a, is defined as follows: $\eta_a = \frac{1}{7+N_a}$, where N_a represents the count of the number of times a was chosen up to trial t.

For the KO-sparseness model, we let $K = 80$ and for KO-distributional-RPE model, we let $M = 1$.

The simulation is conducted by discretizing the differential equation using $dt = 0.005$.

2.3.3 Approximation of basic CogLink model to an algorithm with an analysis of the algorithm.

In this section, we approximate the basic CogLink as an algorithm and conduct a mathematical analysis of its performance.

The stable fixed points S_a^{alm} of the premotor cortex dynamics at the ath ensemble (see Equation 2.3.1) are defined as:

$$S_a^{\text{alm}} = \left\{x : x \in \mathbb{R}^M,\ |\text{supp}(x)| = K,\ \forall i \in [M], x_i = 1.5, \text{if } x_i \neq 0\right\}.$$

Here, $\text{supp}(x) = \{x_i | x_i \neq 0\}$. Assuming the K-WTA sampling dynamic at the premotor cortex occurs instantaneously (i.e., τ^{alm} is small), the premotor cortex dynamic x_a^{alm} converges to one of the fixed points above for each ensemble. As the network is symmetric, it uniformly converges to one of the fixed points $\hat{x}_a^{\text{alm}}$, i.e., $\hat{x}_a^{\text{alm}} \sim \text{unif}(S_a^{\text{alm}})$.

Similarly, assuming the BG dynamic (see Equation 2.3.2) occurs instantaneously (i.e., τ^{bg} is small), we obtain:

$$x_a^{\text{bg}} = V_a^{\text{alm/bg}} \circ g(x_a^{\text{alm}}) = V_a^{\text{alm/bg}} \circ g(\hat{x}_a^{\text{alm}}).$$

Algorithm 1 Algorithmic form of CogLink model

Input parameters $A, K, M, \{\eta_{(t)}\}_{t\in[T]}, \left\{\bar{V}_{a,m}^{\text{alm/bg}}\right\}_{a\in[A],m\in[M]}$
For all $a \in [A]$, for $m \in [M]$, initialize $N_{1a} = 1$ and $v_{1am} = \bar{V}_{a,m}^{\text{alm/bg}}$
for trial $t = 1, \ldots, T$ **do**
    for action $a = 1, \ldots, A$ **do**
        Let V_{ta} be the uniform distribution over $\left\{\frac{1}{K}\sum_{j=1}^{K} v_{tai_j}\right\}_{1\le i_1<\ldots<i_K\le M}$
        Sample $\hat{v}_{ta} \sim V_{ta}$
    Output action $a_t \leftarrow \arg\max_a \hat{v}_{ta}$
    Receive reward r_t
    $\delta_t \leftarrow r_t - v_{ta_t}$
    if $a = a_t$ **then**
        $N_{(t+1)a} \leftarrow N_{ta} + 1$
        $v_{(t+1)a} \leftarrow v_{ta} + \eta_{(N_{ta})}\delta_t$
    else
        $N_{(t+1)a} \leftarrow N_{ta}$
        $v_{(t+1)a} \leftarrow v_{ta}$

From Equation 2.3.26, we also have

$$\forall a \in [A],\ I_a^{\text{mct}} = V_a^{\text{bg/mct}} \cdot V_a^{\text{alm/bg}} \circ f(\hat{x}_a^{\text{alm}}) = \frac{1}{K} \sum_{m\in\text{supp}(\hat{x}_a^{\text{alm}})} V_{a,m}^{\text{alm/bg}}.$$

I_a^{mct} is then the sample value $\hat{v}_a$ in our algorithm (Equation 2.1.3).

Finally, assuming the WTA motor dynamic (see Equation 2.3.27) occurs instantaneously (i.e., τ^{mct} is small), the motor cortex dynamic outputs action A_t:

$$a_t = \arg\max_a I_a^{\text{mct}}$$

This corresponds to Equation 2.1.4 in the algorithm.

2.3.4 The regret of the algorithm is nearly optimal.

In this section, we analyze the performance of the algorithm. To simplify the notation for analysis, we present the pseudo-code of the algorithm and introduce a few notation changes (see Algorithm 1).

Let μ_i represent the probability of receiving rewards for choosing action i. Without loss of generality, let action 1 denote the optimal action. Define $\Delta_i = \mu_1 - \mu_i$ and $D = \frac{\Delta_1}{\Delta_2}$. The primary objective of this section is to establish the following theorem:

Theorem 2.3.5. *Let $K = 1$, and for all $a \in [A]$ and $m \in [M]$, $\bar{V}_{a,m}^{\text{alm/bg}} = \frac{Cm}{M}$, where $C = \frac{16\log(2ATD\Delta_2^2\log T)}{\Delta_2}$. Additionally, let $\eta_{(t)} = \frac{1}{1+t}$ for all $t \in [T]$. Under these conditions, the regret of this algorithm is bounded by $\sqrt{324ATD\log(2ATD\Delta_2^2\log T)}$.*

Let M denote the number of neurons in each ensemble, and let v_{tam} represent the synaptic strength of the mth neuron at the ath ensemble at trial t. The learning rate after an action is chosen t times is denoted by η_t. Additionally, N_{ta} denotes the number of times action a has been chosen at the end of trial t, and T_{na} represents the trial when action a has been chosen for the nth time. If action $\hat{a}$ is chosen at trial t, we employ the standard reward prediction error update:

$$v_{(t+1)\hat{a}m} = v_{t\hat{a}m} + \eta_{(N_{t\hat{a}})}\delta_t,\ \ \delta_t = r_t - v_{t\hat{a}m}. \tag{2.3.6}$$

And for $a \neq \hat{a}$, $v_{(t+1)am} = v_{tam}$. One can conceive of this ensemble of synapses as a quantile distribution V_{ta} representing the values for each action. Each ensemble randomly samples $\hat{v}_{ta}$ from this quantile distribution V_{ta} by selecting K synaptic strengths uniformly at random and averaging them:

$$\hat{v}_{ta} \sim V_{ta}.$$

The action is then chosen based on the values of the samples through a mutual competition process:

$$a_t = \arg\max_a \hat{v}_{ta}.$$

By recursively expanding Equation 2.3.6, we obtain:

$$v_{(t+1)am} = \left(\prod_{n=1}^{N_{ta}}(1-\eta_{(n)})\right)\left(v_{0am} + \sum_{j=1}^{N_{ta}}\eta_{(j)}\prod_{n=1}^{j}(1-\eta_{(n)})^{-1}r_{T_{na}}\right).$$

For the theoretical analysis of this circuit, we consider the following simple setting: let $K = 1$, and for all $a \in [A]$, $v_{0am} = \frac{Cm}{M}$, where $C > 0$ is a constant we will define later. For all $m \in [M]$ and $t \in [T]$, let $\eta_{(t)} = \frac{1}{1+t}$. By substituting these conditions into the equation, we obtain:

$$v_{(t+1)am} = \frac{Cm}{(N_{ta}+1)M} + \frac{1}{N_{ta}+1}\sum_{n=1}^{N_{ta}} r_{T_{na}}. \tag{2.3.7}$$

Now, we aim to bound the expectation of N_{ta} for $a \neq 1$. Demonstrating that the model selects suboptimal actions infrequently implies small regret. Given any $\varepsilon \in \mathbb{R}$, we define $E_a(t) = \{\hat{v}_{ta} \leq \mu_1 - \varepsilon\}$. We establish the following stopping time to capture the event when rewards are concentrated around the mean:

$$\tau = \inf\left\{t : \exists a \in [A],\ \left|\frac{1}{N_{ta}}\sum_{n=1}^{N_{ta}} r_{T_{na}} - \mu_a\right| > \sqrt{\frac{1}{N_{ta}}\log\frac{2A\log T}{\delta}}\right\}. \tag{2.3.8}$$

Applying the maximal Hoeffding inequality yields:

$$P\left(\exists t \in [T], \left|\frac{1}{N_{ta}}\sum_{n=1}^{N_{ta}} r_{T_{na}} - \mu_i\right| > \sqrt{\frac{1}{N_{ta}}\log\frac{2}{\delta'}}\right) < \delta'.$$

By union bounding over all intervals and actions:

$$P\left(\exists t \in [T], \exists a \in [A], \left|\frac{1}{N_{ta}}\sum_{n=1}^{N_{ta}} r_{T_{na}} - \mu_a\right| > \sqrt{\frac{1}{N_{ta}}\log\frac{2}{\delta'}}\right) < \delta' A\log T.$$

Setting $\delta' = \frac{\delta}{A \log T}$, we obtain:

$$P(\tau < T) < \delta.$$

Now, let's bound the expectation of N_{ta} using this stopping time. We have:

$$\begin{aligned}
\mathbb{E}[N_{Ta}] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T\wedge\tau} \mathbf{1}\{a_t = a\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T\wedge\tau} \mathbf{1}\{a_t = a\}\mathbf{1}\{\tau \geq T\}\right] + \mathbb{E}\left[\sum_{t=1}^{T\wedge\tau} \mathbf{1}\{a_t = a\}\mathbf{1}\{\tau < T\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\mathbf{1}\{\tau \geq T\}\right] + TP(\tau < T) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\mathbf{1}\{\tau \geq T\}\right] + T\delta. \qquad (2.3.9)
\end{aligned}$$

Now, let's decompose the first term as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\mathbf{1}\{\tau \geq T\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a, E_a(t)\}\mathbf{1}\{\tau \geq T\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{a_t = a, E_a^c(t)\}\mathbf{1}\{\tau \geq T\}\right].
\end{aligned}$$

Let $a'_t = \arg\max_{a\neq 1} \hat{v}_{ta}$, and let F_{ta} be the cumulative distribution function of V_{ta} conditioning on $\tau \geq t$. To bound the first term, let's examine:

$$\begin{aligned}
P(a_t = a, E_a(t) | \tau \geq t, \mathcal{F}_{t-1}) &\leq F_{t1}(\mu_1 - \varepsilon) P(a'_t = a, E_a(t) | \tau \geq t, \mathcal{F}_{t-1}) \\
&\leq \frac{F_{t1}(\mu_1 - \varepsilon)}{1 - F_{t1}(\mu_1 - \varepsilon)} P(a_t = 1, E_a(t) | \tau \geq t, \mathcal{F}_{t-1}).
\end{aligned}$$

Now, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{a_t=a,E_a(t)\}\mathbf{1}\{\tau\geq T\}\right]\\
=&\mathbb{E}\left[\sum_{t=1}^{T}P\left(a_t=a,E_a(t)|\tau\geq t,\mathcal{F}_{t-1}\right)\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\frac{F_{t1}(\mu_1-\varepsilon)}{1-F_{t1}(\mu_1-\varepsilon)}P\left(a_t=1,E_a(t)|\tau\geq t,\mathcal{F}_{t-1}\right)\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\frac{F_{T_{t1}1}(\mu_1-\varepsilon)}{1-F_{T_{t1}1}(\mu_1-\varepsilon)}\right].
\end{aligned}$$

Note that from Equation 2.3.7 and Equation 2.3.8, we have, conditioning on $\tau\geq t$:

$$\begin{aligned}
\frac{N_{ta}}{N_{ta}+1}\left(\frac{Cm}{N_{ta}M}+\mu_a-\sqrt{\frac{1}{N_{ta}}\log\frac{2A\log T}{\delta}}\right)\leq v_{tam}\\
\leq\frac{N_{ta}}{N_{ta}+1}\left(\frac{Cm}{N_{ta}M}+\mu_a+\sqrt{\frac{1}{N_{ta}}\log\frac{2A\log T}{\delta}}\right).
\end{aligned}\tag{2.3.10}$$

Notice that if $N_{t1}\geq\frac{4\log\frac{2A\log T}{\delta}}{\varepsilon^2}$, then we have

$$\forall m,\ v_{t1m}\geq\mu_1-\varepsilon.$$

If $N_{t1}<\frac{4\log\frac{2A\log T}{\delta}}{\varepsilon^2}$ and $C\geq\frac{8\log\frac{2A\log T}{\delta}}{\varepsilon}$, then

$$\forall m\geq\frac{M}{2},\ \frac{Cm}{N_{t1}M}-\sqrt{\frac{1}{N_{t1}}\log\frac{2A\log T}{\delta}}>0.$$

This implies that $F_{T_{t1}1}(\mu_1-\varepsilon)\leq\frac{1}{2}$, hence

$$\frac{F_{T_{1t}1}(\mu_1-\varepsilon)}{1-F_{T_{1t}1}(\mu_1-\varepsilon)}\leq 1.$$

Consequently,

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{F_{T_{1t}1}(\mu_1-\varepsilon)}{1-F_{T_{1t}1}(\mu_1-\varepsilon)}\right]\leq\frac{4\log\frac{2A\log T}{\delta}}{\varepsilon^2}.\tag{2.3.11}$$

Similarly, we can bound the second term:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{a_t=a,E_a^c(t)\}\mathbf{1}\{\tau\geq T\}\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}(1-F_{ta}(\mu_1-\varepsilon))\mathbf{1}\{a_t=a\}\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}(1-F_{T_{ta}a}(\mu_1-\varepsilon))\right].
\end{aligned}$$

By Equation 2.3.10, if $N_{ta} > \frac{16D\log\frac{2A\log T}{\delta}}{(\Delta_a-\varepsilon)^2}$ and $C \le \frac{8D\log\frac{2A\log T}{\delta}}{\Delta_a-\varepsilon}$, then $F_{T_{ta}a}(\mu_1-\varepsilon)=1$. Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}(1-F_{T_{ta}a}(\mu_1-\varepsilon))\right] \le \frac{16D\log\frac{2A\log T}{\delta}}{(\Delta_a-\varepsilon)^2}. \tag{2.3.12}$$

Now, let's set $\varepsilon=\frac{\Delta_a}{2}$, $\delta=\frac{1}{TD\Delta_2^2}$, and $C=\frac{16\log\frac{2A\log T}{\delta}}{\Delta_2}$. This satisfies the condition for C:

$$\frac{4D\log\frac{2A\log T}{\delta}}{\Delta_a-\varepsilon} \ge C \ge \frac{8\log\frac{2A\log T}{\delta}}{\varepsilon}.$$

Combining Equation 2.3.9, Equation 2.3.11, and Equation 2.3.12, we find:

$$\begin{aligned}\mathbb{E}[N_{Ta}] &\le \frac{(64D+16)\log(2ATD\Delta_2^2\log T)}{\Delta_a^2}+\frac{1}{D\Delta_2^2}\\ &\le \frac{81D\log(2ATD\Delta_2^2\log T)}{\Delta_a^2}.\end{aligned}$$

Now, let's bound the regret:

$$\begin{aligned}R_T &= \sum_{a=1}^{A}\Delta_a\mathbb{E}[N_{Ta}]\\ &\le \sum_{a=1}^{A}\frac{81D\log(2ATD\Delta_2^2\log T)}{\Delta_a}.\end{aligned}$$

For any $\Delta>0$, we can divide the sum as follows.

$$\begin{aligned}&= \sum_{a:\Delta_a<\Delta}^{A}\frac{81D\log(2ATD\Delta_2^2\log T)}{\Delta_a}+\sum_{a:\Delta_a\ge\Delta}^{A}\frac{81D\log(2ATD\Delta_2^2\log T)}{\Delta_a}\\ &\le T\Delta+\frac{81AD\log(2ATD\Delta_2^2\log T)}{\Delta}.\end{aligned}$$

By geometric inequality, we have

$$\le \sqrt{324ATD\log(2ATD\Delta_2^2\log T)},$$

as desired.

Specifically, when $A=2$, we can present the following simplified theorem.

Theorem 2.3.13. *Let $K=1$ and $\Delta=|\mu_1-\mu_2|$ and $\forall a\in[2]$, $v_{0am}=\frac{Cm}{M}$ where $C=\frac{16\log(4T\Delta^2\log T)}{\Delta}$. For all $m\in[M], t\in[T]$, let $\eta_{(t)}=\frac{1}{1+t}$. Then the regret of this algorithm is bounded by $36\sqrt{T\log(4T\Delta\log T)}$.*

2.3.5 Details on the augmented CogLink.

This section provides details of the augmented CogLink model. At its core, the model comprises the PFC-MD-like circuit for contextual inferences and copies of basic CogLink models for dynamically switching behavioral strategies based on the inferred context.

At trial t, the prefrontal cortex activities $x^{\text{pfc}} \in \mathbb{R}^{A\times 2}$ jointly encode actions and rewards at the last trial, with the following formulation:

$$x^{\text{pfc}}_{a,r} = \begin{cases} 1, \text{if } a = a_{t-1}, r = r_{t-1} \\ 0, \text{otherwise} \end{cases}.$$

To form a line attractor in MD, we consider the following thalamocortical loop:

The MD activities, denoted as $x^{\text{md}} \in \mathbb{R}^2$, evolve according to the equation:

$$\tau^{\text{eff}} \frac{dx^{\text{md}}}{dt} = -\frac{2x^{\text{md}}}{D} + \beta_{\text{d2}} \left(f_{\text{md}} \left(\frac{1}{2} x^{\text{fc}} - \frac{1}{4} x^{\text{trn}} \right) + I^{\text{pfc/md}} \right) + I^{\text{rescue}},$$

the frontal cortex activities, denoted as $x^{\text{fc}} \in \mathbb{R}^2$, evolve according to the equation:

$$\tau^{\text{fc}} \frac{dx^{\text{fc}}}{dt} = -x^{\text{fc}} + x^{\text{md}},$$

and the TRN activity, denoted as $x^{\text{trn}} \in \mathbb{R}$, evolve according to the equation:

$$\tau^{\text{trn}} \frac{dx^{\text{trn}}}{dt} = -x^{\text{trn}} + \sum_{i=1}^{2} x^{\text{md}}_i.$$

Here, $\tau^{\text{eff}} = 5$ represents the effective time constant for accumulation dynamics, $\tau^{\text{fc}} = \tau^{\text{trn}} = 0.1$ represents the membrane time constant of frontal neurons and TRN neurons and $D = 4$ signifies the threshold for accumulation dynamics.

The nonlinearity function, $f_{\text{md}} : \mathbb{R} \to \mathbb{R}$, is defined as

$$f_{\text{md}}(x) = \begin{cases} x + 1 & \text{if } -1 \le x \le 1 \\ 2 & \text{if } x > 1 \\ 0 & \text{otherwise} \end{cases},$$

and the PFC-MD inputs, denoted as $I^{\text{pfc/md}} \in \mathbb{R}^2$, are given by

$$\forall c \in [2],\ I^{\text{pfc/md}}_c = f_{\text{pfc}}(W^{\text{pfc/md}}_c \cdot x^{\text{pfc}}).$$

Here, $W^{\text{pfc/md}}_c \in \mathbb{R}^{A\times 2}$ represents PFC-MD connections projecting to MD neurons tuned to context c, and $\cdot$ signifies the matrix inner product. Additionally, $f_{\text{pfc}}(x) = [2.7 + \log(x)]_+$, $\beta_{\text{d2}} = 0.85$ in the D2R hyperactivation model and the rescue model, $\beta_{\text{d2}} = 1$ otherwise, and $I^{\text{rescue}} = 0.45$ in the rescue model and $I^{\text{rescue}} = 0$ otherwise. We set $x^{\text{md}} = 0$ for the MD inhibition model throughout the experiment.

We update the PFC-MD connections through Hebbian learning as follows:

$$\forall c \in [2], a \in [A], r \in [2],\ \Delta W^{\text{pfc/md}}_{c,a,r} = \eta_c \left(f_{\text{hebb}}(x^{\text{md}}_c) x^{\text{pfc}}_{a,r} - f_{\text{hebb}}(x^{\text{md}}_c) W^{\text{pfc/md}}_{c,a,r} \right)$$

Here, $W^{\text{pfc/md}}_{c,a,r}$ represents the synpaes $f_{\text{hebb}} : \mathbb{R} \to \mathbb{R}$ denotes the sigmoidal nonlinearity function,

$$f_{\text{hebb}}(x) = \left[\frac{1 - e^{8-4(x-4)}}{1 + e^{8-4(x-4)}}\right]_+$$

The learning rate is determined by $\forall c \in [2], a \in [A], \eta_c \in \mathbb{R}^2$, given by $\eta_c = \frac{f_{\text{hebb}}(x^{\text{md}}_c)}{4+N_c}$, where N_c represents a rolling sum of $f_{\text{hebb}}(x^{\text{md}}_c)$, and is updated as $N_c \leftarrow N_c + f_{\text{hebb}}(x^{\text{md}}_c)$. For the KO-nonlinear Hebb, we replace f_{hebb} with a linear function

$$f_{\text{KO-hebb}}(x) = \left[\frac{x-4}{4}\right]_+.$$

MD neurons then modulate the downstream d-CS models via interneuron-mediated pathways. Specifically, the interneuron activities are defined as

$$\forall c \in [2],\ \tau_{\text{vip}} x^{\text{vip}}_c = -x^{\text{vip}}_c + x^{\text{md}}_c$$

and

$$\forall c \in [2],\ \tau_{\text{pv}} x^{\text{pv}}_c = -x^{\text{pv}}_c + x^{\text{md}}_{\bar{c}}$$

Here, the interneuron membrane time constant is $\tau_{\text{vip}} = \tau_{\text{pv}} = 0.1$, and $\bar{c}$ represents the context different from c. These activities modulate the downstream d-CS models as follows:

$$\forall c \in [2], a \in [A],\ \tau^{\text{bg}} \frac{dx^{\text{bg}}_{c,a}}{dt} = -x^{\text{bg}}_{c,a} + f(x^{\text{alm}}_{c,a}) \circ f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c) V^{\text{alm/bg}}_{c,a}.$$

where

$$f_{\text{in}}(x) = \left[\frac{1 - e^{-2(x+0.25)}}{1 + e^{-2(x+0.25)}}\right]_+.$$

For KO-interneuron gating, we replace f_{in} with direct MD modulation

$$\forall c \in [2], a \in [A],\ \tau^{\text{bg}} \frac{dx^{\text{bg}}_{c,a}}{dt} = -x^{\text{bg}}_{c,a} + f(x^{\text{alm}}_{c,a}) \circ f_{\text{KO-in}}(x^{\text{md}}_c) V^{\text{alm/bg}}_{c,a}.$$

where

$$f_{\text{KO-in}} = \left[\frac{x-4}{4}\right]_+.$$

These interneuron-mediated pathways also modulate plasticity. Specifically,

$$\forall c \in [2], \delta_c = r_t - V^{\text{alm/bg}}_{c,a_t},\ \Delta V^{\text{alm/bg}}_{c,a_t} = \eta_{c,a_t} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c)\delta_c.$$

where $\eta_{c,a} = \frac{1}{4+N_{c,a}}$ and $N_{c,a}$ is the rolling sum of $0.5 * \mathbf{1}_{a=a_t} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c)$, with $N_{a,c} \leftarrow N_{a,c} + 0.5 * \mathbf{1}_{a=a_t} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c)$. The remainder of the model consists of two copies of the basic CogLink models.

2.3.6 Approximation of augmented CogLink model to an algorithm.

In this section, we approximate the thalamocortical model as a novel algorithm and demonstrate its connection to the CUSUM algorithm [43, 44]. Additionally, we illustrate that the D2R hyperactive impaired model corresponds to a leaky evidence integrator.

We recall that the MD circuit can be described by (Section 2.3.5, Section 2.3.5, Section 2.3.5). Notice by letting the dynamic of x^{fc} and x^{trn} instantaneous, we can describe the effective MD circuit dynamic as follows:

$$\begin{cases} \tau^{\text{eff}}\dot{x_1^{\text{md}}} = -\frac{2x_1^{\text{md}}}{D} + f_{\text{md}}(wx_1^{\text{md}} - wx_2^{\text{md}}) + I_1^{\text{pfc/md}} \\ \tau^{\text{eff}}\dot{x_2^{\text{md}}} = -\frac{2x_2^{\text{md}}}{D} + f_{\text{md}}(wx_2^{\text{md}} - wx_1^{\text{md}}) + I_2^{\text{pfc/md}} \end{cases}.$$

Here, $D = 4$ and we will show that D represents the threshold of accumulation dynamics. $\tau^{\text{eff}} = 5$ represents the effective time constant for accumulation dynamics, $\tau^{\text{md}} = \tau^{\text{eff}}D/2$ represents the membrane time constant, $w = \frac{1}{D}$, and $I_1^{\text{pfc/md}}, I_2^{\text{pfc/md}}$ represent the PFC inputs to MD. The nonlinearity function is defined as:

$$f_{\text{md}}(x) = \begin{cases} x + 1, & \text{for } -1 \le x \le 1 \\ 2, & \text{for } x > 1 \\ 0, & \text{for } x < -1 \end{cases}.$$

Let $X = x_1^{\text{md}} - x_2^{\text{md}}$. Then, we have:

$$\tau^{\text{eff}}\frac{dX}{dt} = I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}, \text{when } |X| < D$$

and

$$\tau^{\text{eff}}\frac{dX}{dt} = -\frac{2X}{D} + I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}} + 2, \text{when } |X| > D.$$

If $|I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}| \ll 1$, then at the stationary point, X remains approximately $\pm D$. This corresponds to a drift-diffusion process with a threshold at $\pm D$ and inputs of $I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}$. To be precise, we discretize the differential equation and threshold X at ± 4 to derive the following algorithm:

$$X_t = \min\left\{D, \max\left\{-D, X_{t-1} + \frac{dt}{\tau^{\text{eff}}}(I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}})\right\}\right\}. \tag{2.3.14}$$

We prove the following theorem:

Theorem 2.3.15. *Let $dt = \tau^{\text{eff}}$. If we set $X_0 = -D$, $S_t = X_t + D$ and assume $|I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}| \ll 1$, the evolution of X_0 approximates to*

$$S_t = \min(2D, \max(0, S_{t-1} + I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}))$$

We can prove the theorem by substituting the variable into Equation 2.3.14 and adding D to both sides:

$$S_t = \min(2D, \max(0, S_{t-1} + I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}))$$

as desired. Notably, when $I_1^{\text{pfc/md}}(t) = \log P(a_t, r_t|c=1) + \alpha, I_2^{\text{pfc/md}}(t) = \log P(a_t, r_t|c=2) + \alpha$ for any $\alpha > 0$ and $S_n < 2D$, this corresponds exactly to the CUSUM algorithm:

$$S_t = \max(0, S_{t-1} + \log P(a_t, r_t|c=1) - \log P(a_t, r_t|c=2)).$$

To analyze the impaired model, we recall the equations:

$$\begin{cases} \tau^{\text{eff}} \dot{x_1^{\text{md}}} = -\frac{2x_1^{\text{md}}}{D} + \beta_{\text{d2}} f_{\text{md}}(wx_1^{\text{md}} - wx_2^{\text{md}}) + \beta_{\text{d2}} I_1^{\text{pfc/md}} \\ \tau^{\text{eff}} \dot{x_2^{\text{md}}} = -\frac{2x_2^{\text{md}}}{D} + \beta_{d2} f_{\text{md}}(wx_2^{\text{md}} - wx_1^{\text{md}}) + \beta_{\text{d2}} I_2^{\text{pfc/md}} \end{cases}.$$

Let $X = x_1^{\text{md}} - x_2^{\text{md}}$. Then, we have:

$$\tau^{\text{eff}} \frac{dX}{dt} = \frac{2}{D}(\beta_{\text{d2}} - 1)X + \beta_{\text{d2}}(I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}}), \text{when } |X| < D.$$

This indicates that the evidence accumulation dynamic is a leaky integrator. At the stationary point, we have

$$|X| = \frac{\beta_{\text{d2}} D}{2(1-\beta_{\text{d2}})} |\langle I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}} \rangle|.$$

By plugging in the model learned by the impaired model in Figure 2.7k, we have

$$\begin{aligned} \langle I_1^{\text{pfc/md}} - I_2^{\text{pfc/md}} \rangle_{c=0} \approx & \mathbb{E}_{r_t|c=0,a_t=0} [\log p(a_t=0, r_t, |c=0) - \log p(a_t=0, r_t, |c=1)] \\ \approx & 0.7(\log 0.57 - \log 0.49) + 0.3(\log(1-0.57) - \log(1-0.49)) \\ \approx & 0.055. \end{aligned}$$

So we have the threshold $X \approx \frac{0.055\beta_{\text{d2}} D}{2(1-\beta_{\text{d2}})} = 0.62 \ll 4 = D$, consistent with Figure 2.7i. This demonstrates that the impaired model has a much lower evidence accumulation threshold compared to the normal model, thereby inducing a strong prior on environmental volatility.

2.3.7 Details of other models.

This section contains details on other models used in the paper. For Thompson sampling, let γ to be the discounted factor. Initialize for all $a \in [A], \alpha_a = \beta_a = 0$. Then, we sample from the posterior

$$\hat{v}_a \sim \text{Beta}(\alpha_a + 1, \beta_a + 1).$$

We output action

$$a_t \leftarrow \arg\max_a \hat{v}_a.$$

and receive reward r_t. We then update the parameter

$$\forall a \in [A], \alpha_a = \gamma\alpha_a + \mathbf{1}_{a=a_t}\mathbf{1}_{r_t=1}, \ \beta_a = \gamma\beta_a + \mathbf{1}_{a=a_t}\mathbf{1}_{r_t=0}$$

In all simulations in the paper, we use $\gamma = 1$ for Thompson sampling and $\gamma = 0.93$ for discounted Thompson sampling.

For Hidden Markov model with Thompson sampling, Initialize for all $a \in [A], c \in [2], \alpha_{a,c} = \beta_{a,c} = 0$. we use HMM with known environmental parameters to infer the current contextual likelihood, $p_c \in \mathbb{R}^2$

$$p_c = \text{HMM}(a_{<t}, r_{<t}).$$

And then we sample from the posterior,

$$\hat{v}_a \sim p_1 * \text{Beta}(\alpha_{a,1} + 1, \beta_{a,1} + 1) + p_2 * \text{Beta}(\alpha_{a,2} + 1, \beta_{a,2} + 1).$$

We output action

$$a_t \leftarrow \arg\max_a \hat{v}_a.$$

and receive reward r_t. We then update the parameter

$$\forall c \in [2],\ \alpha_{a_t,c} = \alpha_{a_t,c} + p_c \mathbf{1}_{1=r_t},\ \beta_{a_t,c} = \beta_{a_t,c} + p_c \mathbf{1}_{0=r_t}.$$

For the Deep Q-Network (DQN) [25], we use a multilayer perceptron with a hidden layer size of 10 and an ε-greedy exploration strategy. To balance exploration and exploitation, at trial t, given state s_t, we define $N_\varepsilon \in \mathbb{R}^S$, where S is the total number of states. The visit count for each state is updated as: $N_s \leftarrow N_s + \mathbf{1}_{s=s_t} 0.2$. The model explores uniformly at random with probability $\varepsilon = \frac{1}{N_{s_t}}$ and otherwise selects the action with the highest Q-value.

2.3.8 Tasks.

A-AFC task. This section contains details for the stationary A-AFC task. The task contains two parameters, the expected difference in reward probability between the most and the least rewarding actions (Δ) and the number of alternatives (A). The reward probability θ_a of action $a \in [A]$ is specified by

$$\forall a \in [A],\ \theta_a = (0.7 - \Delta) + \frac{a-1}{A-1}\Delta.$$

Each session contains 500 trials and for each simulation, we run the task for 50 sessions.

Cued 2-AFC task. This section contains details for the cue 2-AFC task (Figure 2.S2a). The task contains one parameter, the expected difference in reward probability between the most and the least rewarding actions (Δ). Each trial with uniform probability the model will be presented with cue 1 or cue 2. The reward probability of action 1 after seeing cue 1 is 70% while action 2 is $(70 - \Delta)$%. On the other hand, The reward probability of action 1 after seeing cue 2 is $(70 - \Delta)$% while action 2 is 70%. Each session contains 500 trials and for each simulation, we run the task for 50 sessions.

Binary tree maze task. This section contains details for the binary tree maze task (Figure 2.S2c). The task consists of a depth 2 binary tree maze with 4 end locations. Upon reaching each end location, the model will receive a reward with probability $1, \frac{2}{3}, \frac{1}{3}, 0$ respectively. The task contains one parameter a; at the start, if the model chooses left, it receives a reward with probability a and if the model chooses right, it receives a reward with probability $(1 - a)$. Each session contains 500 trials and for each simulation, we run the task for 50 sessions.

Probabilistic reversal task. This section contains details for the probabilistic reversal task. There are two alternatives, left or right, in the task and the reward probability in the context 1 is $\theta_R = 0.3$, $\theta_L = 0.7$ and the reward probability in the context 2 is $\theta_R = 0.7$, $\theta_L = 0.3$. The task starts with context 1 and switch to alternative context for every 200 trials. The task consists of 1000 trials and for each simulation, we run the task for 50 sessions. For the low outcome uncertainty environment in Figure 2.5i, we replace the $70\%, 30\%$ reward probability with $90\%, 10\%$.

2.3.9 Regret and contextual uncertainty.

Let $\theta_t \in \mathbb{R}^A$ be the probability of getting a reward for each action at trial t. We define regret at trial T, R_T, as the expected differences in rewards between the retrospectively optimal action and the action taken,

$$R_T = \sum_{t=1}^{T} (\theta_t)_{a_t^*} - (\theta_t)_{a_t},$$

where a_t^* is the retrospectively optimal action.

To decode the contextual uncertainty, U, at Figure 2.4g and Figure 2.S7b, we consider the following nonlinear transformation of the MD activities x^{md}:

$$U = 2 - \frac{2}{1 + e^{-|x_1^{\text{md}} - x_2^{\text{md}}|}}.$$

Notice that when two MD populations have the same activity, the uncertainty is 1 and when they have a large difference in activity, the uncertainty is close to 0.

2.3.10 CogLink modeling for human probabilistic reversal task

Model. This section details the CogLink model, a slightly modified version of the original CogLink model (REF) tailored to the study's scope. The model consists of collections of rate neurons governed by differential equations. The core of the model includes the dlPFC-MDl circuit for contextual inferences, two OFC circuit copies to encode contextual action values, and the MDm circuit to update these values in the OFC. The dlPFC-MDl connections learn the environmental generative model through Hebbian learning, and the dlPFC-MDl circuit accumulates the learned likelihood to infer the current context. MDl then activates different OFC ensembles through interneuron pathways. MDm encodes the reward prediction error from the last trial and projects to the OFC to update PFC-OFC connections for learning contextual action values. For notation, let $S = 2$ be the number of stimuli, $A = 2$ be the number of actions and $C = 2$ be the number of contexts.

At trial t, the dlPFC consists of two populations, the first population, $x^{\text{pfc}_1} \in \mathbb{R}^S$, encodes the current cue,

$$x_s^{\text{pfc}_1} = \begin{cases} 1, \text{if } s = s_t \\ 0, \text{otherwise} \end{cases}, \tag{2.3.16}$$

while the second population, $x^{\text{pfc}_2} \in \mathbb{R}^{S\times A\times 2}$, encodes the sensorimotor-outcome associations at the last trial:

$$x^{\text{pfc}_2}_{s,a,r} = \begin{cases} 1, \text{if } s = s_{t-1}, a = a_{t-1}, r = r_{t-1} \\ 0, \text{otherwise} \end{cases}. \tag{2.3.17}$$

The MDl activities, denoted as $x^{\text{mdl}} \in \mathbb{R}^C$, evolve according to the equation:

$$\tau^{\text{eff}} \frac{dx^{\text{mdl}}}{dt} = -x^{\text{mdl}} + f_{\text{mdl}}(W^{\text{mdl}} x^{\text{mdl}}) + I^{\text{pfc/mdl}}. \tag{2.3.18}$$

Here, $\tau^{\text{eff}} = 5$ represents the effective time constant for accumulation dynamics, $W^{\text{mdl}} \in \mathbb{R}^{C\times C}$ denotes recurrent synaptic weights, given by

$$W^{\text{mdl}} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1, \end{bmatrix}.$$

The nonlinearity function, $f_{\text{md}} : \mathbb{R} \to \mathbb{R}$, is defined as

$$f_{\text{mdl}}(x) = \begin{cases} x - 1 & \text{if } -1 \le x \le 1 \\ 0 & \text{if } x > 1 \\ -2 & \text{otherwise} \end{cases},$$

and the PFC-MDl inputs, denoted as $I^{\text{pfc/mdl}} \in \mathbb{R}^C$, are given by

$$\forall c \in [C],\ I^{\text{pfc/mdl}}_c = f_{\text{pfc}}(W^{\text{pfc/mdl}}_c \cdot x^{\text{pfc}_1}). \tag{2.3.19}$$

Here, $W^{\text{pfc/mdl}}_c \in \mathbb{R}^{S\times A\times 2}$ represents PFC-MD connections projecting to MD neurons tuned to context c, and $\cdot$ signifies the tensor inner product. Additionally, $f_{\text{pfc}}(x) = [2.7 + \log(x)]_+$.

We update the PFC-MDl connections through Hebbian learning as follows:

$$\forall c \in [C], s \in [S], a \in [A], r \in [2],\ \Delta W^{\text{pfc/mdl}}_{c,s,a,r} = \eta_c \left(f_{\text{hebb}}(x^{\text{mdl}}_c) x^{\text{pfc}_1}_{s,a,r} - f_{\text{hebb}}(x^{\text{mdl}}_c) W^{\text{pfc/mdl}}_{s,c,a,r} \right) \tag{2.3.20}$$

Here, $W^{\text{pfc/mdl}}_{c,s,a,r}$ represents the synpaes $f_{\text{hebb}} : \mathbb{R} \to \mathbb{R}$ denotes the sigmoidal nonlinearity function,

$$f_{\text{hebb}}(x) = \left[\frac{1 - e^{4-4x}}{1 + e^{4-4x}} \right]_+$$

The learning rate is determined by $\forall c \in [C], a \in [A], \eta_c \in \mathbb{R}^2$, given by $\eta_c = \max\left\{0.1, \frac{f_{\text{hebb}}(x^{\text{md}}_c)}{6+N_c}\right\}$, where N_c represents a rolling sum of $f_{\text{hebb}}(x^{\text{md}}_c)$, and is updated as $N_c \leftarrow N_c + f_{\text{hebb}}(x^{\text{mdl}}_c)$.

MD neurons then modulate the downstream OFC ensembles via interneuron-mediated pathways. Specifically, the interneuron activities are defined as

$$\forall c \in [C],\ \tau_{\text{vip}} x^{\text{vip}}_c = -x^{\text{vip}}_c + x^{\text{mdl}}_c \tag{2.3.21}$$

and

$$\forall c \in [C],\ \tau_{\text{pv}} x^{\text{pv}}_c = -x^{\text{pv}}_c + x^{\text{mdl}}_{\bar{c}} \tag{2.3.22}$$

Here, the interneuron membrane time constant is $\tau_{\text{vip}} = \tau_{\text{pv}} = 0.1$, and $\bar{c}$ represents the context different from c. These activities modulate the OFC ensembles, $x^{\text{ofc}} \in \mathbb{R}^{C\times A}$, as follows:

$$\forall c \in [C], s \in [S], a \in [A],\ \tau^{\text{ofc}} \frac{dx^{\text{ofc}}_{c,a}}{dt} = -x^{\text{ofc}}_{c,a} + f(x^{\text{pfc}_2}_s) \circ f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c) W^{\text{pfc/ofc}}_{c,s,a}. \quad (2.3.23)$$

where

$$f_{\text{in}}(x) = \left[\frac{1 - e^{-2x}}{1 + e^{-2x}}\right]_+$$

and $W^{\text{pfc/ofc}} \in \mathbb{R}^{C\times S\times A}$ denotes PFC-OFC synapses.

These interneuron-mediated pathways also modulate plasticity. Specifically, given the activity of MDm, $x^{\text{mdm}} \in \mathbb{R}^C$,

$$\forall c \in [C], x^{\text{mdm}}_c = r_{t-1} - W^{\text{pfc/ofc}}_{c,s_{t-1},a_{t-1}}, \quad (2.3.24)$$

the interneurons-mediate pathways change the PFC-OFC plasticity as the follows:

$$\forall c \in [C], \Delta W^{\text{pfc/ofc}}_{c,s_{t-1}a_{t-1}} = \eta_{c,s_{t-1},a_{t-1}} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c) x^{\text{mdm}}. \quad (2.3.25)$$

where $\eta_{c,s,a} = \left\{0.2, \frac{1}{4+N_{c,s,a}}\right\}$ and $N_{c,s,a}$ is the rolling sum of $\mathbf{1}_{s=s_t,a=a_t} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c)$, with $N_{a,c} \leftarrow N_{a,c} + \mathbf{1}_{s=s_t,a=a_t} f_{\text{in}}(x^{\text{vip}}_c - x^{\text{pv}}_c)$.

To do action selection, we aggregate OFC action values onto motor neurons, $x^{\text{mct}} \in \mathbb{R}^A$, as the follows:

$$\forall a \in [A],\ I^{\text{mct}}_a = \frac{1}{3} \sum_{c\in[C]} x^{\text{ofc}}_{c,a} \quad (2.3.26)$$

and

$$\forall a \in [A],\ \tau^{\text{mct}} \frac{dx^{\text{mct}}_a}{dt} = -x^{\text{mct}}_a + g\left(\sum_{i=1}^{A} W^{\text{mct}}_{ai} x^{\text{mct}}_i\right) + I^{\text{mct}}_a. \quad (2.3.27)$$

Here, the membrane time constant $\tau^{mct} = 1$. The recurrent synaptic weights, $W^{\text{mct}} \in \mathbb{R}^{A\times A}$, are defined as:

$$W^{\text{mct}} = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ -1 & 1 & \cdots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \cdots & 1 \end{bmatrix}.$$

The nonlinearity function $g : \mathbb{R} \to \mathbb{R}$ is defined as:

$$g(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } 1 \geq x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The action a is chosen as a_t if either x^{mct}_a reaches the threshold $= 0.8$ within 5 seconds after the trial starts or chosen stochastically from a softmax distribution, $a_t \sim \text{softmax}(25x^{\text{mct}})$.

The simulation is conducted by discretizing the differential equation using $dt = 0.005$.

Model-based versus model-free splitting. To plot Figure 2.9c, we construct a model-free and a model-based algorithm and calculate the log posterior odd ratio of the two models given the actions of CogLink model. For the model-free model, we have the action value estimates $V \in \mathbb{R}^{S\times A}$ and we update the estimates as follows:

$$\Delta V_{s_t,a_t} = \eta_{s_t,a_t}(r_t - V_{s_t,a_t}). \tag{2.3.28}$$

Here, $\eta_{s,a} = \max\left\{0.2, \frac{1}{4+N_{s,a}}\right\}$, where $N_{s,a}$ represents a rolling sum of $\mathbf{1}_{s=s_t,a=a_t}$, and is updated as $N_{s,a} \leftarrow N_{s,a} + \mathbf{1}_{s=s_t,a=a_t}$. The action distribution at trial t is parametrized by $\text{softmax}(25V_{s_t,-})$.

For the model-based model, we consider a fixed hidden Markov model (HMM) with emit probability specified by the task parameter and the transition probability specified by the mean probability of switching, that is 1/25. One can then infer the most likely context state c_t at trial t in such HMM and we specify the action distribution at trial t to be $\text{softmax}\left(25\sum_{c\in[C]} c_{t,c}V_{c,s_t,-}\right)$. Here, $V_{c,s,a}$ denote the probability of receiving reward at context 1 given cue s and action a.

After we observe in Figure 2.9c that the MDl activity at other context x_2^{mdl} is highly correlated with the log posterior odd ratio of model-based versus model-free model, we categorize all the blocks with positive $z-$score of x_2^{mdl} as model-based block ($n = 308$) while blocks with negative $z-$score of x_2^{mdl} as model-free blocks ($n = 192$).

Causal effective connectivity analysis. Similar to the DCM analysis in human data, we choose the data 10 trials before the switch as steady state condition ($n = 500$), 10 trials after the switch in the model-based blocks as model-based condition ($n = 308$) and 10 trials after the switch in the model-free blocks as model-free condition ($n = 192$). Assume the underlying effective dynamic is $\dot{x} = Wx$ and the data is $\hat{x}$. Then we find the effective connectivity as $W_{\text{eff}} = \arg\min \int_0^T |Wx(t) - \hat{x}(t)|dt + |W|_2^2$. By discretizing the integral, this can be approximated as ridge regression. Since each brain area has multiple neurons with various activity patterns, we cannot sum up the all possible connections to represent the effective connectivity. To calculate the effective connectivity from one region, x_{pre}, to another, x_{post}, in Figure 2.1e, we calculate the projection of effective inputs onto postsynaptic activity's direction as follows

$$W_{eff}x_{pre} \cdot x_{post}/|x_{post}|. \tag{2.3.29}$$

For the cortical connectivity, we sum up effective connectivity of both OFC to dlPFC and dlPFC to OFC and for the thalamocortical connectivity, we sum up effective connectivity from both MDl and MDm to OFC and dlPFC.

2.3.11 Statistic test.

Data were first tested for normality using the Shapiro-Wilk test. All data presented in this paper are non-normally distributed; therefore, all statistical tests were conducted using nonparametric statistics. For all comparison of two groups, we used two way rank sum test. For comparison of more than two groups, we used Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test. All permutation tests are done using 10^6 resamples.

2.3.12 Data and code availability.

- Behavioral and neural activity data of the models have been deposited at FigShare and are publicly available as of the date of publication. DOI is 10.6084/m9.figshare.26065372
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOI is 10.5281/zenodo.13152289.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

2.4 Experimental Method for Human Probabilistic Reversal Task

2.4.1 Dataset for human probabilistic reversal task

Human participants. Forty participants (22 females, mean age ± SD: 24.5 ± 3.3 years) were recruited. All participants were right-handed and had normal or corrected to normal vision. The participants with a history of psychiatric or neurological disorders or who were taking regular medication were excluded. The study was approved by the local ethics committee of the Ruhr-University Bochum. All participants provided written informed consent prior to participation. Four participants were excluded due to technical issues with the fMRI scans. Thirty-six participants successfully completed the task during fMRI scanning. Participants who achieved less than 60% correct responses and/or lacked a clear learning and reversal effect in their performance during the fMRI experiment were excluded from further analysis. Based on these criteria, data from four participants were excluded. Consequently, the final analysis included data from 32 participants (16 females; mean age ± SD: 24.5 ± 3.5 years).

Tactile stimuli. The tactile stimuli were generated and delivered to the index fingertip of the right (dominant) hand using an MRI-compatible Braille device (Metec, Stuttgart, Germany). The Braille device was controlled using the Presentation software (version 20.1, Neurobehavioral Systems, Berkeley, CA, USA) by TCP-IP commands. The device consisted of eight plastic pins, aligned in two series of four pins (pin diameter 1.2 mm, rounded top, inter-pin spacing 2.45 mm) (Figure 2.8a). Eight alternative tactile stimulation patterns were used which always consisted of four raised and four lowered pins. Participants underwent a tactile detection test prior to task training and fMRI to ensure that all tactile stimulation patterns were correctly perceived. They were asked to report the pattern received until they accurately distinguished all patterns 100%.

Experiment design. Experimental design. We employed a probabilistic reversal learning Go/NoGo task as described recently [166]. In each block, two tactile patterns were randomly selected from the eight alternative patterns. 70% of trials with one tactile pattern were assigned to 'Go', and 70% of trials with the alternative tactile pattern were assigned to

'NoGo' (Figure 2.8b). By trial and error, participants had to learn which of the two available responses ('Go' and 'NoGo') had the higher reward probability for each of the two tactile patterns. In each individual block, the association between tactile stimuli and responses was reversed at a random trial between trail 20 to trial 25, requiring participants to adjust their behavior to gain reward (Figure 2.8b). Participants were informed of the probabilistic nature of the association and the existence of a rule switch in each block, but they were not given information about the levels of probability or the specific timing of the reversal.

In each trial, participants first received one out of two tactile stimulation patterns for 500ms on the index fingertip of the right (dominant) hand. A red fixation cross was simultaneously presented via fMRI-compatible LCD-goggles (Visuastim Digital, Resonance Technology Inc., Northridge, CA, USA). After the tactile cue, the red fixation cross turned green, instructing the participants to press the button (LumiTouch keypads, Photon Control Inc., Burnaby, BC, Canada) with the index finger of the left hand ('Go'), or refrain from pressing the button ('NoGo'). Participants had to press the button within 1000ms if action was needed. After the interval of 500-1500ms, the outcomes were presented for 500ms to indicate whether the action was correct or wrong. Trials were presented with randomized intertrial interval (ITI) ranging between 1500 and 3000ms in 100ms steps.

The task was organized in blocks of 45 trials, and consisted of 3 runs, each included four blocks. A novel pair of tactile patterns were used on each new block, which were presented to the participants at the beginning of each block. Before the fMRI scanning, each participant completed a short practice block with 90% probability to ensure she/he was able to follow the instructions. In total, the fMRI experiment consisted of 540 trials, which we split into three runs, each lasting about 16 min, resulting in a total scanning time of about 50 mins.

Within each block of the rule reversal task, the associations between stimuli and responses were reversed at a random trial dividing the block into two phases: (1) the initial learning phase, in which the participants learned the stimulus-response association for each stimulus, and (2) the reversal phase, in which they had to reverse their choice preference to gain reward. To investigate the dynamic changes along the learning process, we aligned the reversal phase using the reversal point and averaged the proportion of correct strategy across blocks. One sample t-test was leveraged to compare the difference of proportion of correct strategy between Switch and Steady State with a significance threshold of $p < 0.05$.

2.4.2 Model-free and model-based splitting and analysis for human data

Decision-making behaviors can be governed by two RL computational models: "model-free" and "model-based" RL strategies. These models share the same core algorithm but differ in complexity and the extent of task knowledge incorporated. Computationally, the "model-free" model assumes minimal information of the task design and maintains an expected probability of reward for each pairing of cue and action. This expected probability is updated per trial based on the reward prediction error (RPE) which is the difference between actual reward and expected reward. Specifically, the model considers a matrix of the value/preference of pairings between cues and actions (Q-matrix, with q_{ij} as value of action j given cue i, with

$i, j = 1$ or 2). On each trial, given cue i, the probability of action j is:

$$P_j = \frac{e^{\tau q_{ij}}}{e^{\tau q_{i1}} + e^{\tau q_{i2}}}.$$

Given the chosen action j, the Q-matrix is updated with a Q-learning rule: $q_{ij} = q_{ij} + \alpha \text{RPE}$, where α is the learning rate. A basic assumption is held as task knowledge: the Q-matrix is assumed to have anti-correlated entries (i.e. subjects are assumed to be aware that, for instance, $q_{11} + q_{12} = 1$). In other words, $q_{kl} = q_{kl} + (-1)^{k-i+l-j \mod 2} \rho\alpha \text{RPE}$ for the other 3 entries k, l, with ρ as the degree of anticorelation. In this model, τ, α, and ρ are the fit parameters. The "model-based" model extends from the "model-free" model with the additional knowledge that there are two (loosely defined) sets of stimulus-response association which the task jumps to and from (rule reversal) – one where cue 1 prefers Go (and cue 2 prefers NoGo) and one where cue 1 prefers NoGo (and cue 2 prefers Go). As such, the model has a confidence-based module which estimates the probability of rule reversal based on errors in recent trials. The method is similar to that in detailed in the previous study [167]. Briefly, a belief of rule reversal is computed as the posterior probability of rule reversal given the trial history, divided by that of no rule reversal. The belief is reformulated as a gaussian distribution, and outputs a state prediction error signal (SPE) as the portion of the distribution above a threshold (arbitrarily set as 1). This module consists of 2 fit parameters (standard deviation of the gaussian, and a perseverance factor augmenting the gaussian mean), resulting in 5 fit parameters of the "model-based" model (together with τ, α, and ρ).

Based on the computational "model-free" (MF) and "model-based" (MB) models, we divided the blocks into two categories: those better explained by the MF model and those better explained by the MB model. This categorization was performed by calculating the behavioral distance (% correct strategy) between the observed data and the fitted MF model, then subtracting the distance between the observed data and the fitted MB model. Blocks with lower values were classified as MF, while the remaining blocks were classified as MB. Additionally, the RPEs were computed from the MF model applied to the MF blocks, and the SPEs were derived from the MB model applied to the MB blocks. RPEs and SPEs play distinct roles in the learning processes of MF and MB systems, respectively. SPEs are critical for MB learning, refining the internal model of the environment by identifying discrepancies between predicted and actual state transitions. This process enables iterative updates and improvements to the model. RPEs are essential for MF learning, where action values are updated based on the difference between expected and actual rewards. Unlike SPEs, this process relies on direct experience and trial-and-error without constructing an explicit model of the environmen. To dissociate the neural regions associated with these prediction error signals, the RPEs and SPEs were incorporated into generalized linear models (GLMs) of fMRI data as parametric modulators.

To evaluate behavioral performance differences between the MF and MB blocks, we applied a linear mixed-effect model (LMM) to account for the repeated measures within the same participants. This approach accommodates both fixed and random effects, enabling more precise estimation of task differences while accounting for individual variability. The LMM was implemented using the fitlme function in MATLAB, with behavioral performance variables as the response variable and block categories (MF and MB) as the fixed effect. This

method ensures robust analysis of the complex data structure and provides reliable inferences regarding differences in behavioral performance between MF and MB strategies.

2.4.3 Human fMRI data acquisition

The fMRI data were collected on a Philips Achieva 3.0 T X-series scanner using a 32-channel head coil. A multi-band echo-planar imaging (EPI) sequence with a multi-band acceleration factor of 2 was used for functional scans. Thirty-eight transaxial slices parallel to the anterior-posterior commissure (AC-PC) covering the whole brain were acquired with a voxel size of $2 \times 2 \times 3$ mm^3, TR $= 2,200$ ms, TE $= 24$ ms, flip angle $= 90$, field of view $= 224$ mm, and no interslice gap. For each participant, high-resolution T1-weighted structural images were also acquired, with 176 transversally oriented slices covering the whole brain, to correct for geometric distortions and perform co-registration with the EPIs (isotropic T1 TFE sequence: voxel size: $1 \times 1 \times 1$ mm^3, field of view 240×176 mm^2).

2.4.4 FMRI data preprocessing and GLMs

Preprocessing. For each run, we acquired a total of 453 EPI volumes. To allow for T1-equilibration, five dummy scans preceded data acquisition in each run, which were removed before further processing. Each participant's EPI volumes were preprocessed and analyzed with the Statistical Parametric Mapping software SPM12 (Wellcome Department of Imaging Neuroscience, University College London, UK; http://www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB R2017b (MathWorks Inc). For preprocessing, images were first applied to slice time correction using sinc interpolation to the middle slice. Then, the T1w image was normalized to the Montreal Neurological Institute (MNI) reference space using the unified segmentation approach. Subsequently, the resulting transformation was applied to the individual EPI volumes to transform the images into standard MNI space and resample them into $2 \times 2 \times 2$ mm^3 voxel space. Spatial smoothing with a 6-mm FWHM Gaussian kernel was applied to the fMRI images only for univariate general linear modeling but not for RSA analyses. Data were high pass filtered at 1/128 Hz to remove low-frequency signal drifts. For each participant, the preprocessed fMRI data was analyzed in an event-related manner in the following GLMs.

General Linear Models (GLMs). The first univariate GLM which was used to analyze the univariate BOLD effect elicited by the rule reversals included two main regressors of interest per block, which accounted for the 10 trials in the Steady State before the reversals and the 10 trials of the Switch after the reversals. The onset was time-locked to the outcome presentation of each trial. For each of these two main regressors, the values of a trial-by-trial variables derived from the model-free (i.e., reward prediction errors, RPEs) and model-based (i.e., state prediction error, SPE) RL was independently defined as the parametric modulator. In the univariate GLMs, four additional regressors of no interest accounting for the unsigned trials before the reversal (the trials not belong to Steady State, time-locked to outcome), the unsigned trials after the reversal (the trials not belong to Switch, time-locked to outcome), presentation of the stimuli (all trials collapsed to a single regressor, time-locked to the onset of cue presentation) and invalid trials (i.e., late responses time-locked to outcome) were

included. To account for motion-related artifacts during the task, six head-motion parameters estimated during the realignment procedure were also defined as regressors of no interest in the univariate GLM. All regressors were convolved with the canonical hemodynamic response function in an event-related fashion.

Another univariate GLM was used to assess the functional connectivity of model-free and model-based MD with the PFC using PPI. Different with the first univariate GLM, the Switch regressor was divided into two regressors: one represented the model-free, and the other the model-based. The Steady State regressor and additional regressors of no interest were identical with the first univariate GLM.

Two multivariate GLMs were also included to assess the representational similarity between different types of trials during Switch using RSA. The multivariate GLMs were consisted of the unsmoothed fMRI data. In the first multivariate GLM, different with the first univariate GLM, the Switch regressor was divided into 4 regressors accounted for the four types of trials (cue1-'Go', cue2-'NoGo', cue1-'NoGo', cue2-'Go') during the Switch phase of the task. The onset of events within these regressors was locked to the onset of the outcome in each trial. The Steady State regressor and additional regressors of no interest were identical with the first univariate GLM. To assess the difference of neural representation between the model-free and model-based trials during Switch, another multivariate GLM was created from the first multivariate GLM. In this multivariate GLM, the Switch trials were categorized into model-free or model-based regressors based on whether they are from model-free or model-based blocks, with each further divided into four regressors accounted for the four types of trials (cue1-'Go', cue2-'NoGo', cue1-'NoGo', cue2-'Go'). This resulted in eight regressors for the Switch trials in total in the second multivariate GLM. The Steady State regressor and additional regressors of no interest were identical with the first multivariate GLM.

2.4.5 Representational similarity analysis (RSA)

To test how the multi-voxel response pattern in MD, dmPFC and CN represents the decision strategy after the reversals, we performed a representational similarity analysis (RSA) by assessing the representational similarity between different types of trials during Switch based on the first multivariate GLM. Multi-voxel measures of neural activity are quantitatively related to each other and to computational theory by comparing representational dissimilarity matrices (RDMs).

Construction of model RDMs. During the Switch, the trials can be categorized into four types (cue1-'Go', cue2-'NoGo', cue1-'NoGo', cue2-'Go'). The participants stuck to the old strategy used before the reversals (Keep) in trials with the associations of cue1-'Go' and cue2-'NoGo', while they updated their strategy to the new rule (Change) in trials with the alternative two associations (cue1-'NoGo', cue2-'Go'). Based on the predicted correlation distance for these four types of trials, two strategy model RDMs were constructed to investigate whether the multi-voxel spatial patterns of activity in dmPFC, dlPFC and OFC at the time of outcome presentation are sensitive either for the Keep or Change. The Keep Strategy representation model assumes that the activity in the respective brain region shows a greater representational similarity between trials where cue1-'Go' and cue2-'NoGo', whereas

the Change Strategy would show stronger similarities between trials where cue1-'NoGo' and cue2-'Go'.

Construction of ROI RDMs. RSA analysis was performed based on the second multivariate GLM to assess the difference of neural representations between the model-free and model-based trials during Switch for the PFC regions (dmPFC, dlPFC and OFC). The ROI of dmPFC was derived from the contrast of Switch > Steady State, while the ROIs of dlPFC and OFC were identified through the results of the PPI analysis (see below). These analyses revealed the functional connectivity with model-free and model-based MDs respectively using the threshold of $p < 0.001$, uncorrected. Different with the first RSA analysis based on all the Switch trials, the second RSA analysis was performed based on the model-free and model-based Switch trials separately.

Statistical analyses. For the RSA analysis, the differences in correlation coefficients between Change and Keep conditions (Keep-Change) were compared between the model-free and model-based RDMs. A paired two-sample t-test was conducted with a significance threshold of $p < 0.05$.

2.4.6 Dynamic causal modeling (DCM)

Time series extraction and Specification of DCMs. Based on the first univariate GLM results (Switch > Steady State), we performed the first DCM analysis by selecting those brain regions that significantly responded to the reversals (i.e., MD, dmPFC, and CN in the right hemisphere) to investigate the effective connectivity under different model assumptions. Subject-specific time series were extracted from the nearest local maximum within a sphere with a radius of 12 mm centered on the group maximum. The first Eigenvariate was then extracted across all voxels surviving $p = 0.05$ uncorrected within a 6 mm sphere centered on the individual peak voxel. The resulting BOLD time series were adjusted for effects of no interest (e.g., invalid trials, and movement parameters). Five participants had to be excluded from DCM analyses because we could not extract valid activity from one or more of the three ROIs.

The DCMs are specified in terms of fixed (endogenous) connections between brain areas and condition-dependent changes in their connection strength (i.e., modulatory or bilinear effects). We applied DCM, based on the second univariate GLM, for the inference of model parameters, namely whether a specific connection is more likely to exert an excitatory or an inhibitory effect on its target region in a given model. In the DCM, we focused on the strategy-related MD subdivisions (medial and lateral), along with their cortical counterparts (OFC and dlPFC, respectively). The constructed DCM networks consisted of reciprocal endogenous connections among all regions, except the two MDs because of the lack of recurrent connectivity in the thalamus, with the inputs given to both MDs. To better separate the signals between lateral and medial MD, we here used a mask from the AAL3 atlas (https://www.gin.cnrs.fr/en/tools/aal/) to extract time series from the right lateral and medial MD ROI. The masks of OFC and dorsal PFC were from the PPI analysis (Figure 2.8i).

Bayesian parameter averaging. The DCM, encompassing lateral MD, medial MD, dorsal PFC and OFC, was applied for the inference on the model parameters during the steady-state period, predominately model-free RL switch period, and predominately model-based RL switch period. The parameters of the given model were then summarized by Bayesian parameter averaging (BPA), which computes a joint posterior density for the entire group by combining the individual posterior densities. A posterior probability criterion of 90% was considered to reflect significant effective connectivity. The posterior means of cortico-cortical (connections between OFC and dPFC) and thalamo-cortical (outputs from two MDs) connections in the network were summed up. The bootstrap resampling (1000 iterations) was applied to assess the reliability of the posterior means of BPA.

2.4.7 Psychophysiological Interaction (PPI) analysis

To assess the functional connectivity of model-free and model-based MDs with PFC, we performed two PPI analyses, one using the model-free MD ($x = -2, y = -22, z = 10$) and the other using model-based MD ($x = 4, y = -20, z = 12$) as the seed region. The first Eigenvariate was calculated across all voxels within a 6-mm sphere centered on the peak MNI coordinates of MDs derived from the modulatory effect of model-free and model-based variables. The resulting BOLD time series were adjusted for effects of no interest (e.g., invalid trials and movement parameters) and deconvolved to generate time series required for constructing first-level GLMs for the PPIs. The PPI GLM at the single-subject level contained five regressors: two regressors representing model-free and model-based Switch, and two PPI regressors representing the interactions between the physiological variable (i.e., time series of the seed region) with model-free and model-based Switch. The last regressor represented the physiological variable.

We examined the connectivity of MD with PFC in the model-free and model-based condition separately. To this end, first-level contrast images were created using the PPI regressor of the interaction between the physiological variable and trials of model-free Switch, or the interaction between the physiological variable and trials of model-based Switch. Next, the first-level contrast images were applied to the group-level one-sample t-test. We performed a small volume correction (SVC) by restricting the search volume to the PFC mask with a threshold at FWE-corrected peak-level of $p < 0.05$. The PFC mask was defined with the Automated Anatomical Labelling (AAL) atlas.

2.5 Figures

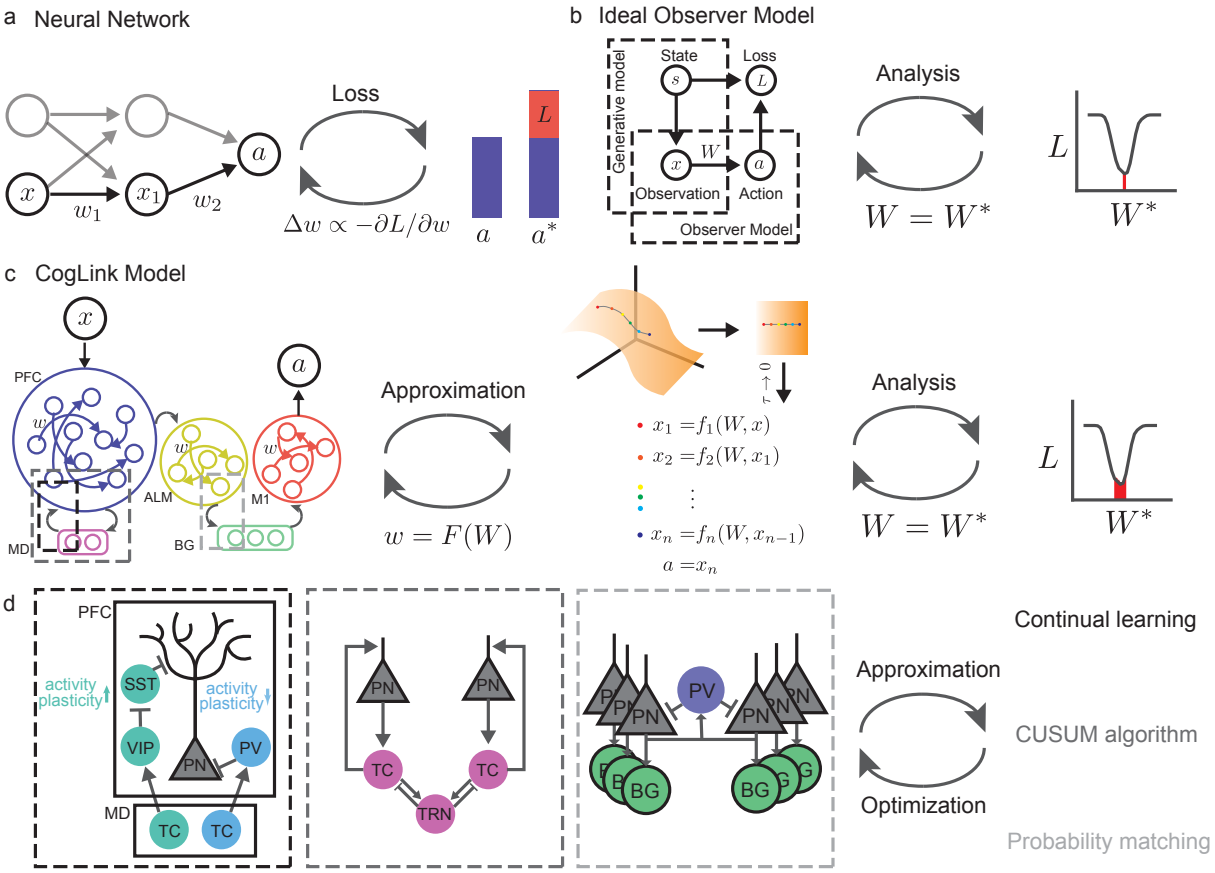

Figure 2.1: **CogLink bridges neural mechanisms and algorithms to optimize parameters for achieving complex cognitive tasks. a.** A task-optimized neural network updates its parameters using gradient descent to minimize a predefined loss function. **b.** The ideal observer model employs a generative model of the environment and chooses actions that minimize the loss over the posterior distribution of the state. This approach uses Bayesian inference to evaluate the posterior and inform optimal decision-making. **c.** The CogLink model integrates algorithmic approximations with neural dynamics, enabling interpretable computation. By constraining neural activity to a low-dimensional subspace, CogLink uses dimensionality reduction and separation of timescales to approximate neural trajectories as structured algorithms. These approximations allow the optimization of algorithmic parameters to inform the corresponding neural parameters in the CogLink model. Unlike exact solutions, which are computationally intractable, CogLink achieves asymptotic optimality through mathematical analysis, making the optimization process feasible. **d.** CogLink incorporates biological realism by modeling specific brain systems, including connectivity patterns, cell types, and learning rules. This biologically grounded design enables the identification of computational roles for each mechanism in hierarchical decision-making.

2.6 Supplemental Figures

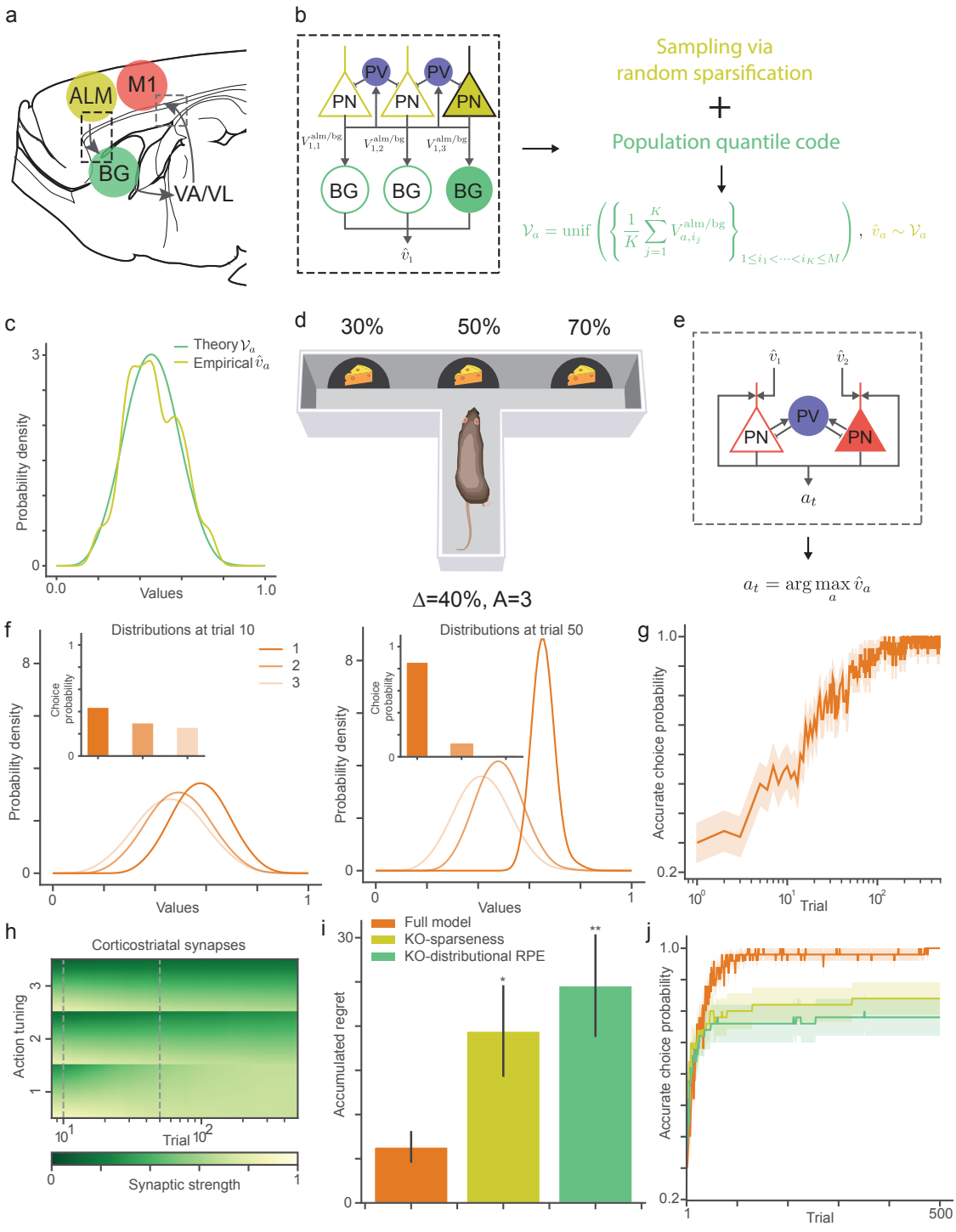

Figure 2.2: **Mechanistic details of basic CogLink support associative uncertainty encoding and balance exploration and exploitation.**

Figure 2.2: **Mechanistic details of basic CogLink support associative uncertainty encoding and balance exploration and exploitation. a.** Schematic of the basic CogLink network architecture. **b.** Premotor corticostriatal-like circuit implementing sampling from a distribution. In the BG-like circuit, a quantile population code encodes associative uncertainty as a distribution of action-value beliefs. Coupled with premotor random sparsification dynamics, the circuit samples from the distribution to extract uncertainty information for downstream motor processing. **c.** Comparison between the probability density function (p.d.f.) of theoretical and empirical distributions encoded in corticostriatal synapses ($n = 100$). The empirical distribution is derived from circuit simulations in **b**. **d.** Schematic of the A-AFC task in a stationary environment. The task is parameterized by the number of alternatives A and the expected difference Δ between the most and least rewarding options. **e.** Schematic of the motor cortex-like downstream action selection circuit. The circuit receives sampled inputs from the BG and selects the action with the highest sampled action value. **f.** P.d.f. plots of action-value beliefs and corresponding choice probabilities ($n = 100$) at trials 10 and 50. At trial 10, large overlaps between distributions promote exploration, while at trial 50, minimal overlaps lead to exploitation. **g.** Summary plot of accurate choice probability over trials (mean $\pm$ s.e.m., $n = 50$). **h.** Corticostriatal synaptic strengths summarized as a heatmap over trials (mean, $n = 50$). Each row represents a synapse, with 100 rows per block. **i.** Accumulated regret for the full model, KO-sparseness, and KO-distributional RPE variants (mean $\pm$ s.e.m., $n = 50$, $^{**}P < 10^{-2}$, $^{*}P < 0.05$; permutation test on the mean difference). **j.** Accurate choice probability for the full model, KO-sparseness, and KO-distributional RPE variants (mean $\pm$ s.e.m., $n = 50$).

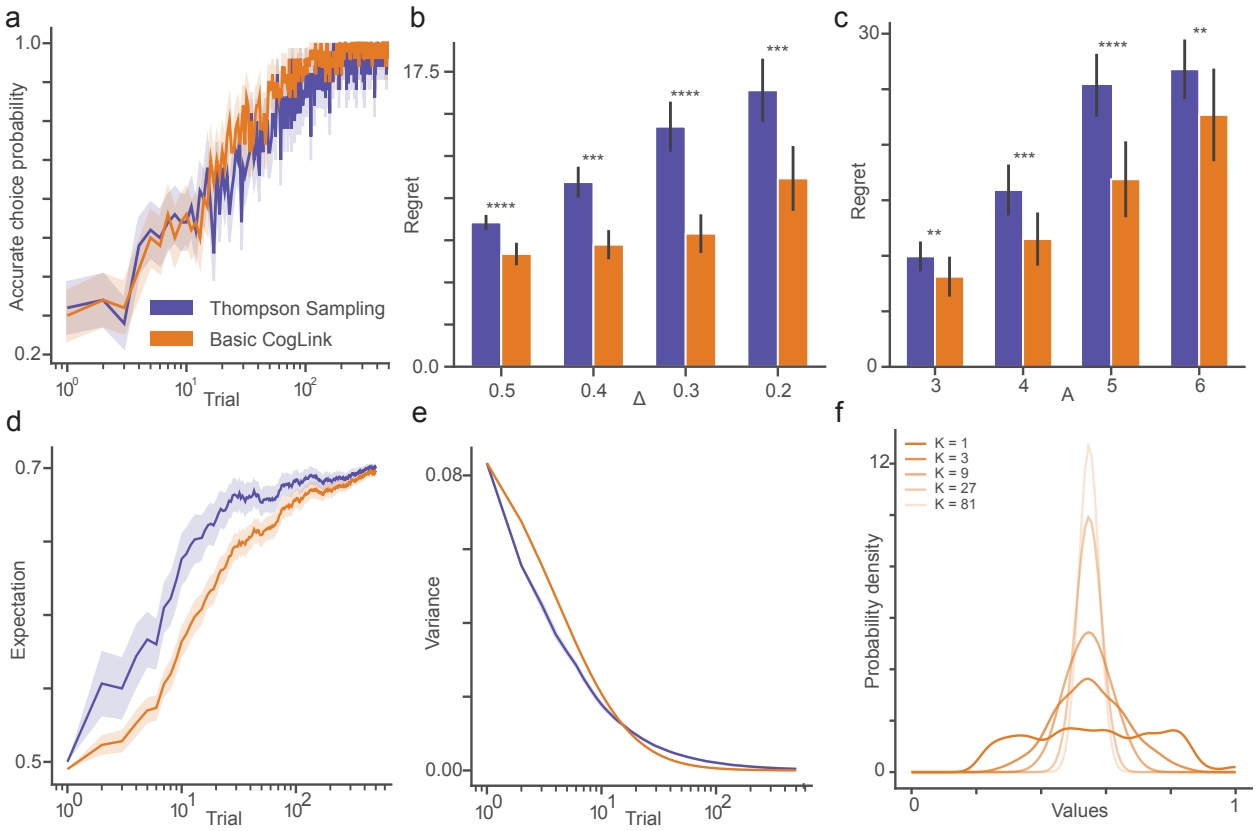

Figure 2.3: **Comparison between CogLink and Thompson Sampling.** We evaluate our model across various A-AFC tasks for 50 sessions of 500 trials. **a**. Summarized plot (mean±s.e.m., $n = 50$) for the accurate choice probability of Thompson Sampling (TS) and basic CogLink in AFC task with $\Delta = 0.4, A = 3$. CogLink achieves faster convergence and higher long-term accuracy compared to TS. **b**. Summarized plot (mean±s.e.m., $n = 50$) for regret across different Δ. CogLink consistently outperforms TS across all tested Δ values (****$P = 3.47 \times 10^{-5}$, ***$P = 1.25 \times 10^{-4}$, ****$P = 3.37 \times 10^{-7}$, *$P = 1.32 \times 10^{-4}$; two-way rank sum test). **c**. Summarized plot (mean±s.e.m., $n = 50$) for regret across different numbers of alternatives A. CogLink consistently outperforms TS across all tested A values ($P = 7.25 \times 10^{-3}$, ***$P = 1.04 \times 10^{-4}$, ****$P = 5.49 \times 10^{-7}$, **$P = 3.5 \times 10^{-3}$; two-way rank sum test). **d.** Summarized plot (mean±s.e.m., $n = 50$) for the expectation of the distribution of value beliefs over trials. Both CogLink and TS converge to similar expectations, and their trajectories are closely aligned throughout the trials, demonstrating comparable accuracy and adaptation in value estimation. **e.** Summarized plot (mean±s.e.m., $n = 50$) for the variance of the distribution of value beliefs, showing that both methods exhibit similar rates of uncertainty reduction over time. **f.** Empirical p.d.f. plot (mean±s.e.m., $n = 50$) for the distribution of value beliefs under varying premotor cortex sparsity K. Higher K values produce narrower distributions, emphasizing exploitation, while lower K values promote exploration.

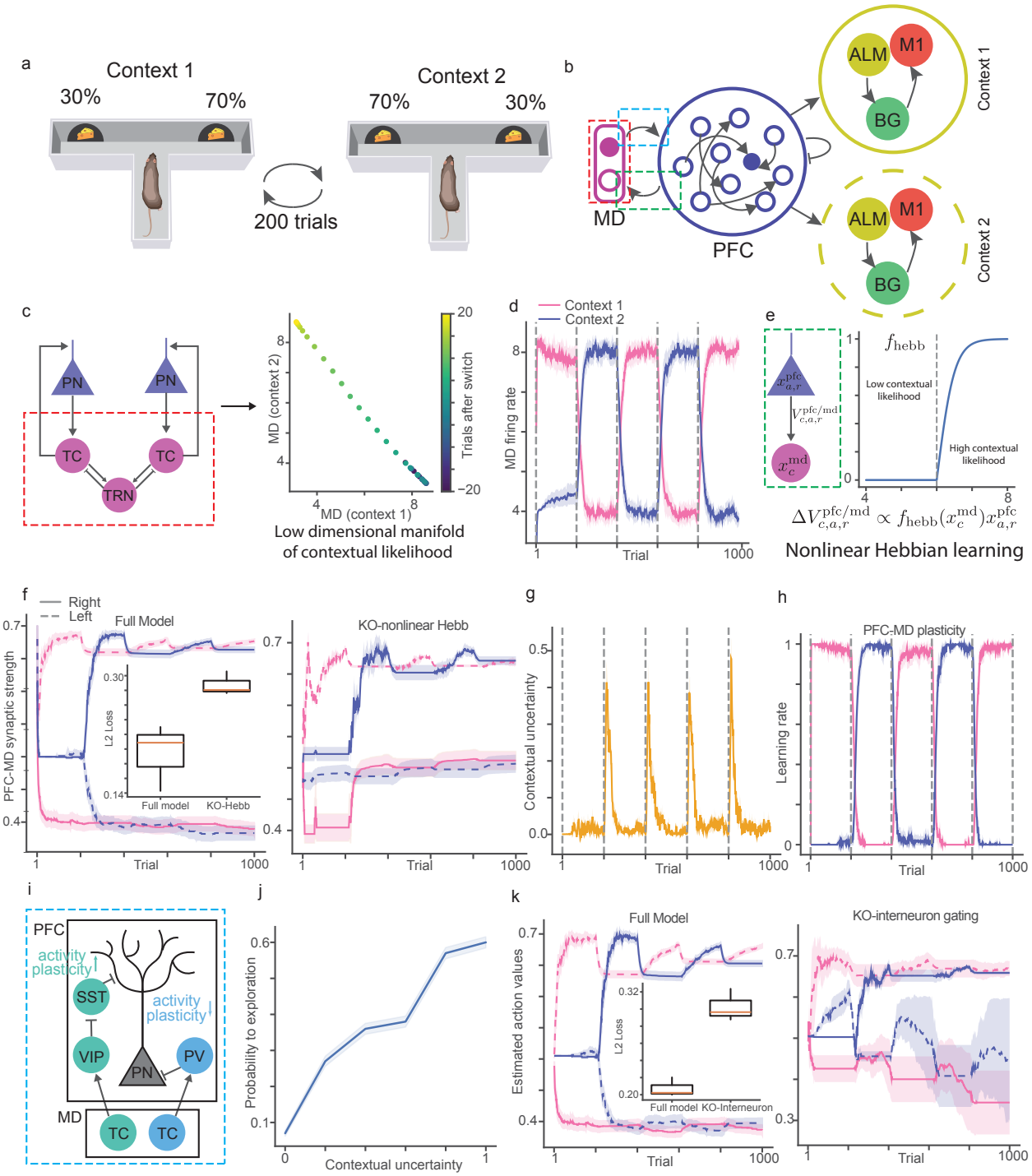

Figure 2.4: **Mechanistic details of augmented CogLink facilitate the encoding of contextual uncertainty and drive flexible switching behaviors.** Caption continues at the next page.

Figure 2.4: **Mechanistic details of augmented CogLink facilitate the encoding of contextual uncertainty and drive flexible switching behaviors. a.** Schematic of the probabilistic reversal task, where the context alternates every 200 trials. **b.** Schematic of the network architecture in the augmented CogLink, where the PFC-MD circuit infers the current context and activates downstream premotor circuits accordingly. **c.** Illustration of the PFC-MD circuit encoding contextual likelihood. PFC-MD connectivity constrains MD activity to reside on a low-dimensional manifold, representing the likelihood of contexts. **d.** Summarized plot (mean±s.e.m., $n = 50$) of MD activity across trials in the probabilistic reversal task, with dashed lines indicating context switches. **e.** Illustration of nonlinear Hebbian learning in PFC-MD synapses, which allows learning only when contextual likelihood is high. **f.** Summarized plot (mean±s.e.m., $n = 50$) of PFC-MD synaptic strengths encoding the contextual generative model for rewards. The left panel shows results from the full model, while the right panel represents the KO-nonlinear Hebb variant. The inset displays a boxplot ($n = 50$) of the L2 distance between the PFC-MD synaptic strengths and the true generative model. **g.** Summarized plot (mean±s.e.m., $n = 50$) of estimated contextual uncertainty across trials. Uncertainty peaks immediately after context switches, reflecting increased ambiguity about the context following a switch. **h.** Summarized plot (mean±s.e.m., $n = 50$) of learning rates for PFC-MD plasticity, showing modulation by contextual uncertainty. **i.** Diagram of interneuron-mediated thalamocortical projections. PV-mediated pathways suppress cortical activity and plasticity, while VIP-mediated pathways amplify them. **j.** Summarized plot (mean±s.e.m., $n = 50$) of exploration probability as a function of contextual uncertainty, where 0.5 indicates chance level (maximum exploration). **k.** Summarized plot (mean±s.e.m., $n = 50$) of corticostriatal synaptic strengths encoding contextual action values. The left panel shows results from the full model, and the right panel depicts the KO-interneuron gating variant. The inset provides a boxplot ($n = 50$) of the L2 distance between corticostriatal synaptic strengths and the true action values.

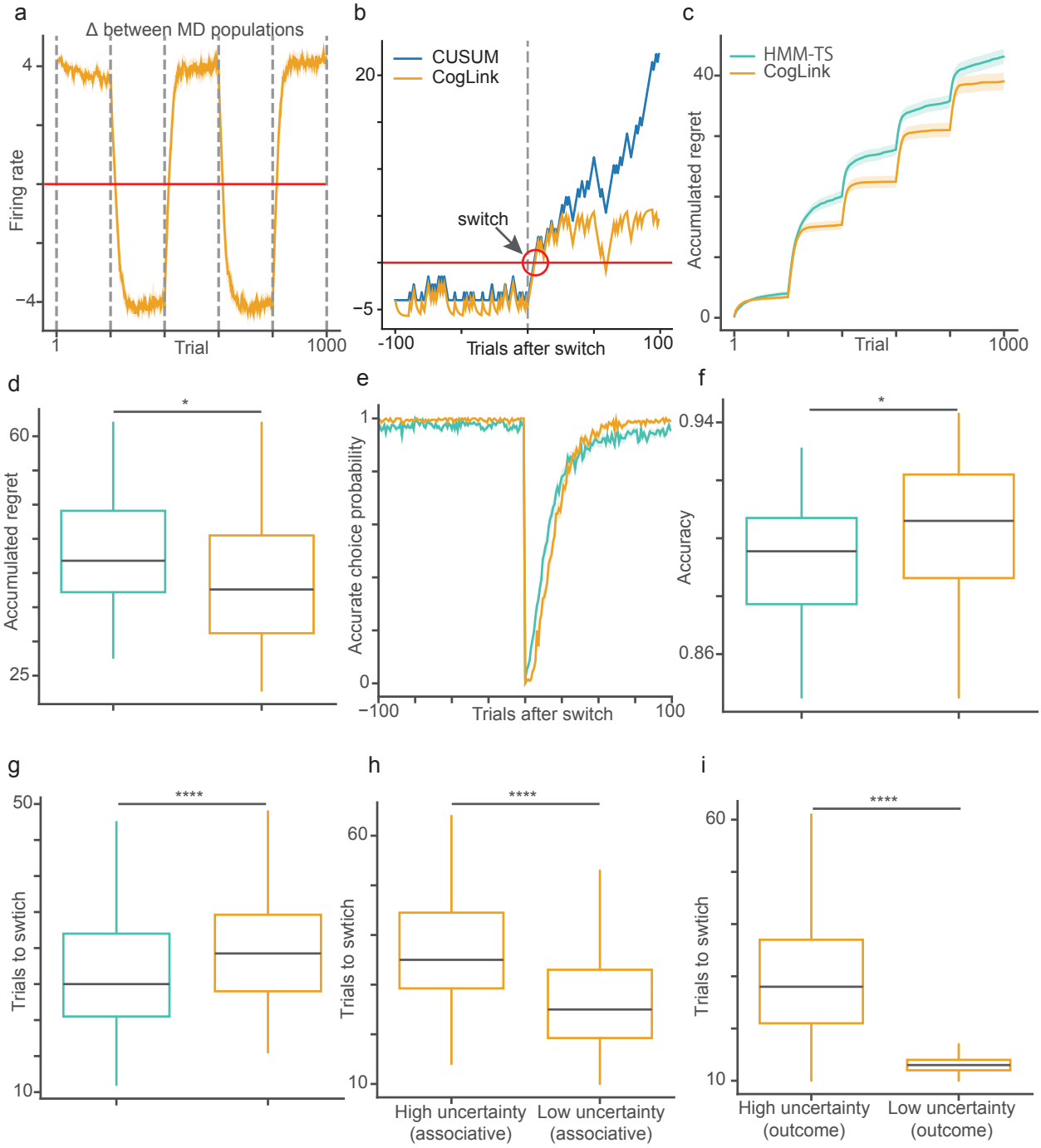

Figure 2.5: **The augmented CogLink achieves flexible decision-making and continual learning by managing hierarchical uncertainty.** Caption continues at the next page.

Figure 2.5: **The augmented CogLink achieves flexible decision-making and continual learning by managing hierarchical uncertainty. a.** Summarized plot (mean±s.e.m., $n = 50$) of the firing rate difference between two MD populations tuned to distinct contexts during a probability reversal task. **b.** Example trajectory comparing evidence accumulation in CogLink's MD circuit and the CUSUM algorithm during a context switch. **c.** Summarized plot (mean±s.e.m., $n = 50$) of accumulated regret over 1000 trials in probability reversal tasks, comparing CogLink to HMM-TS. **d.** Box plot ($n = 50$) of accumulated regret over 1000 trials in probability reversal tasks ($P < 0.05$; two-sided rank sum test). **e.** Summarized plot (mean±s.e.m., $n = 200$) showing recovery dynamics of accurate choice probability following a context switch in the probability reversal task. **f.** Box plot ($n = 50$) of overall accuracy over 1000 trials in probability reversal tasks ($P < 0.05$; two-sided rank sum test). **g.** Box plot ($n = 200$) of context-switching time (number of trials to reach 80% accuracy) in probability reversal tasks ($P < 10^{-4}$; two-sided rank sum test). **h.** Box plot comparing switching time under high ($n = 50$) and low ($n = 150$) associative uncertainty conditions in probability reversal tasks ($P < 10^{-4}$; two-sided rank sum test). **i.** Box plot ($n = 200$) comparing switching time under high and low outcome uncertainty conditions in probability reversal tasks ($^{****}P < 10^{-4}$; two-sided rank sum test).

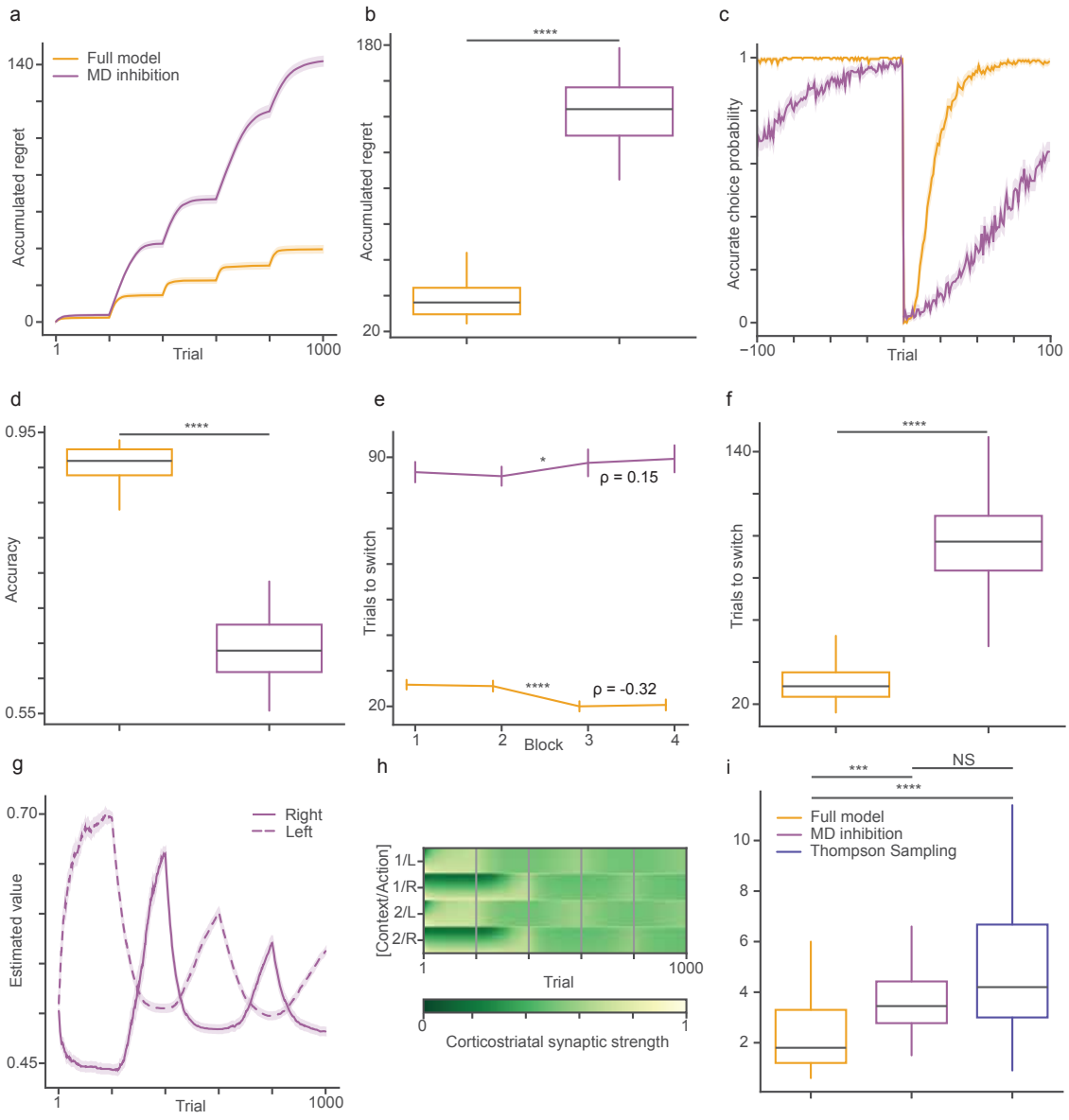

Figure 2.6: **Mediodorsal thalamus in CogLink is necessary for decision-making in a changing environment but not required in a stationary environment.** Captions continue at the next page.

Figure 2.6: **Mediodorsal thalamus in CogLink is necessary for decision-making in a changing environment but not required in a stationary environment.** We evaluate our model and MD inhibition model in the probabilistic reversal task for 50 sessions. **a**. Summarized plot (mean±s.e.m., $n = 50$) for average accumulated regret. **b**. Boxplot ($n = 50$) for the final regret ($^{****}P = 7.01 * 10^{-18}$; two-way rank sum test). **c**. Summarized plot (mean±s.e.m., $n = 200$, 4 switches for 50 trials) for accurate choice probability after a switch. **d**. Boxplot ($n = 50$) for the accuracy ($^{****}P = 7.01 * 10^{-18}$; two-way rank sum test). **e**. Summarized plot (mean±s.e.m., $n = 50$) for numbers of trials to switch ($^{*}P = 2.74 * 10^{-2}$, $^{****}P = 4.00 * 10^{-5}$; permutation test on Spearman's rank correlation coefficient ρ.) **f**. Boxplot ($n = 200$, 4 switches for 50 trials) for the switching time ($^{****}P = 6.50 * 10^{-60}$; two-way rank sum test). **g**. Summarized plot (mean±s.e.m., $n = 50$) for average estimated action values encoded in corticostriatal synapses. **h**. Summarized plot (mean) for average corticostriatal synaptic strengths. **i**. Boxplot ($n = 50$) for the regret in a stationary AFC task with $\Delta = 0.3, A = 2$ ($^{****}P < 10^{-4}$, $^{***}P < 10^{-3}$ NS $P > 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test).

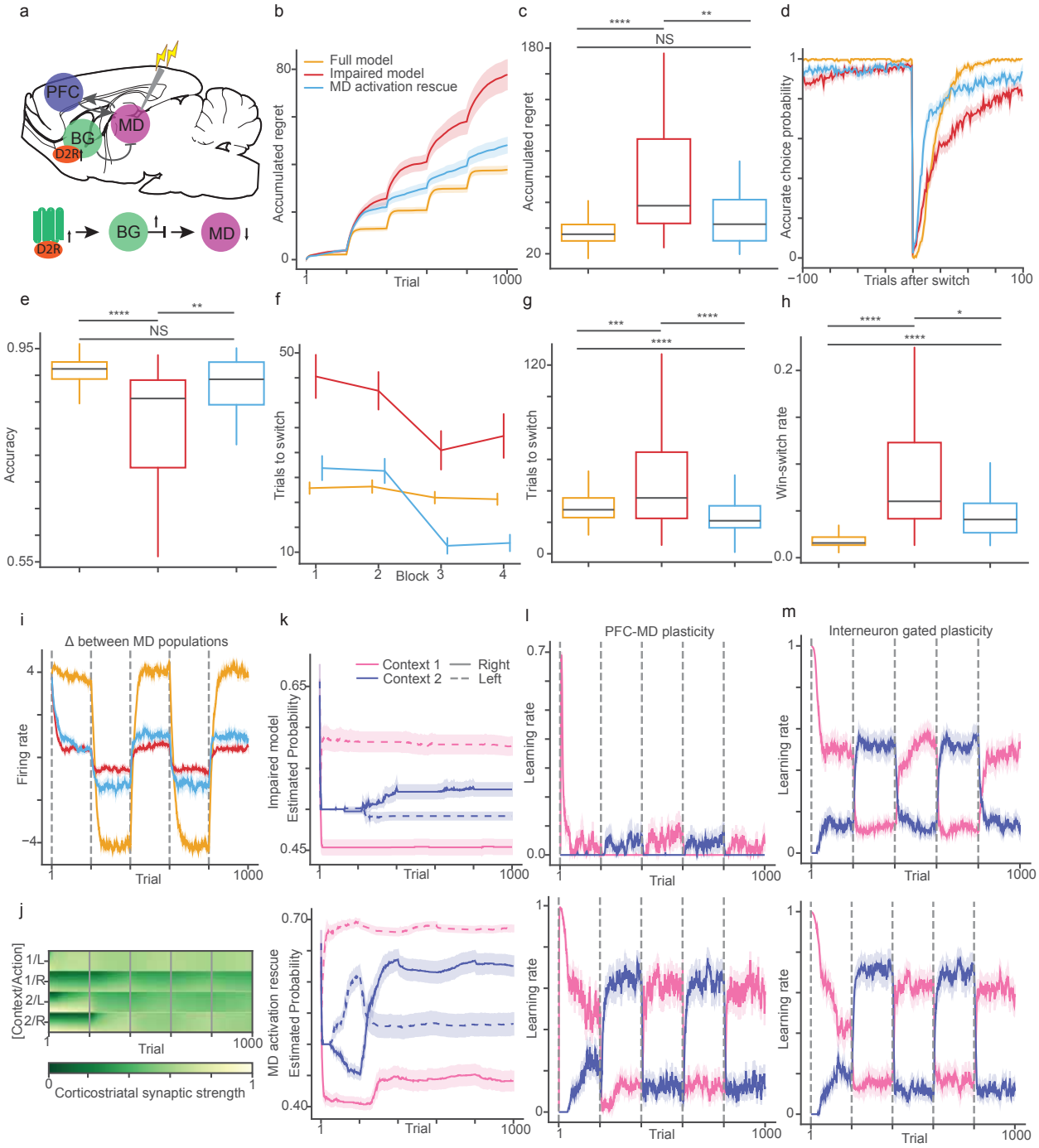

Figure 2.7: **Excess dopamine induces Schizophrenia-like signatures in CogLink and MD activation rescues the behaviors.** Caption continues at the next page.

Figure 2.7: **Excess dopamine on MD induced Schizophrenia-like signatures in CogLink and MD activation rescues the behaviors. a**. A schematic of our impaired model and MD activation rescue model. We posit that hyperactivation of striatal D2R leads to stronger inhibitory BG output which in terms reduce MD's excitability through shunting inhibition. We inject a current in MD to rescue the impaired model. **b**. Summarized plot (mean±s.e.m., $n = 50$) for average accumulated regret. **c**. Boxplot ($n = 50$) for the final regret ($^{****}P < 10^{-4}$, $^{**}P < 10^{-2}$, NS $P > 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **d**. Summarized plot (mean±s.e.m., $n = 200$, 4 switches from 50 trials) for average accurate choice probability after a switch. **e**. Boxplot ($n = 50$) for the accuracy ($^{****}P < 10^{-4}$, $^{**}P < 10^{-2}$, NS $P > 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). f. Summarized plot (mean±s.e.m., $n = 50$) for number of trials to switch. g. Boxplot ($n = 200$, 4 switches from 50 trials) for the switching time ($^{****}P < 10^{-4}$, $^{***}P < 10^{-3}$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **h**. Boxplot ($n = 50$) for the win-switch rate ($^{****}P < 10^{-4}$, $^{*}P < 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **i**. Summarized plot (mean±s.e.m., $n = 50$) for average activity difference in MD populations. **j**. Summarized plot (mean) for average corticostriatal synaptic strength for the impaired model. **k, l, m**. The top row represents data from the impaired model while the bottom row represents data from the rescue model. **k**. Summarized plot (mean±s.e.m., $n = 50$) for average estimated probability to receive a reward. **l**. Summarized plot (mean±s.e.m., $n = 50$) for average learning rate of PFC-MD plasticity. **m**. Summarized plot (mean±s.e.m., $n = 50$) for average learning rate of interneuron gated plasticity.

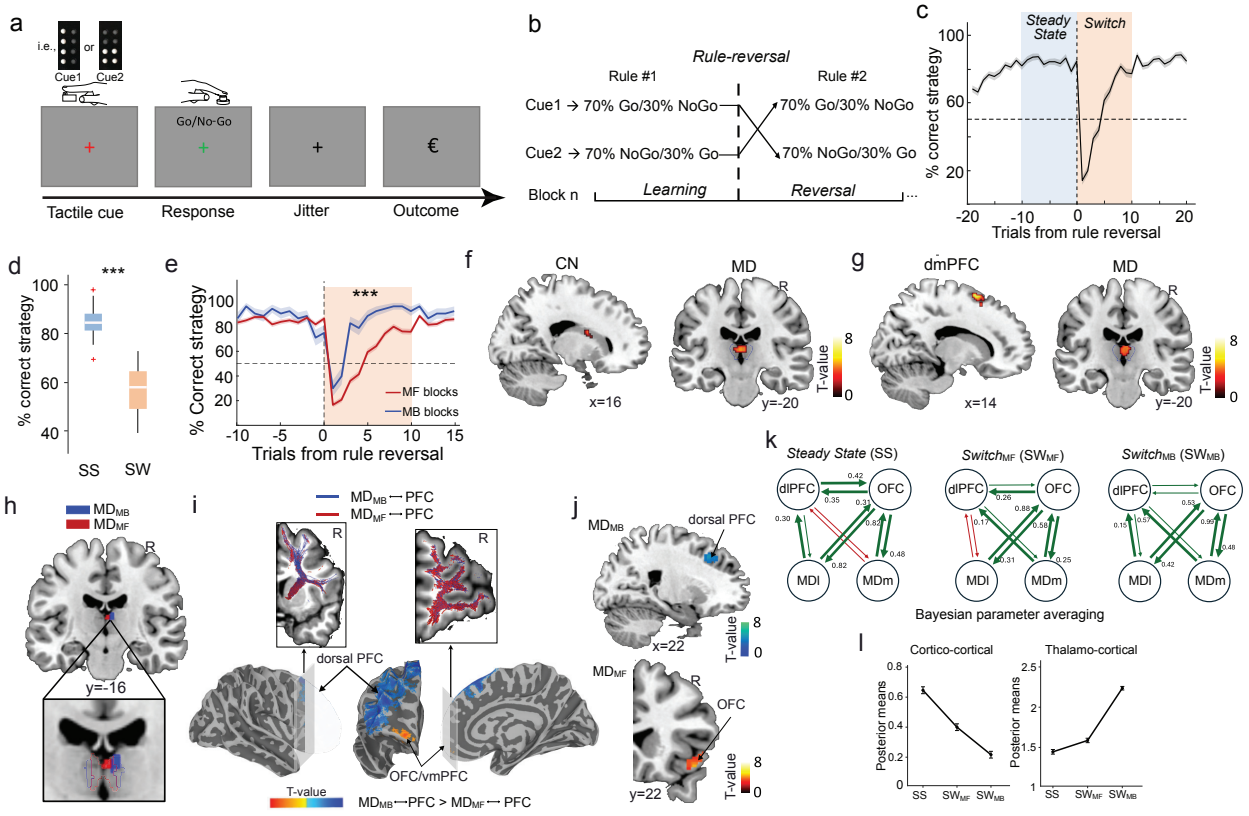

Figure 2.8: **Probability reversal task in humans and representation of reinforcement strategies in prefrontal-striatal network** Caption continues at the next page.

Figure 2.8: **Probability reversal task in humans and representation of reinforcement strategies in prefrontal-striatal network a.** Task schematic of the human rule reversal task. **b.** Illustration of a learning block. In each block, 70% of trials for one tactile pattern were assigned to 'Go,' while 70% of trials for the alternative pattern were assigned to 'NoGo.' The stimulus-response association was reversed at a random trial between the 20th and 25th trial. **c.** Group-averaged proportion of correct strategies across trials. The vertical dashed line indicates the rule reversal point, while the horizontal dashed line represents chance level. Shaded areas highlight Steady State (blue) and Switch (orange) phases, with shaded error bars representing SEM. **d.** Proportion of correct strategies was significantly higher during Steady State than Switch ($t_{(31)} = 13.20, p < 0.001$). **e.** Proportion of correct strategies in blocks dominated by model-free or model-based RL dynamics. The vertical dashed line represents the rule reversal, and the horizontal dashed line marks the chance level. **f.** Both the activity of CN ($x = 16$, $y = 2$, $z = 20$, $t(31) = 4.17$, FWE small-volume correction , $p = 0.027$) and MD ($x = -2$, $y = -22$, $z = 10$, $t(31) = 4.55$, FWE small-volume correction, $p = 0.011$) were significantly correlated with RPE during Switch. T-maps are displayed at $p < 0.001$, uncorrected, for display purpose only. **g**. The activity of both dmPFC ($x = 14$, $y = 12$, $z = 64$, $t(31) = 6.39$, FWE small-volume correction , $p < 0.001$) and MD ($x = 4$, $y = -20$, $z = 12$, $t(31) = 4.05$, FWE small-volume correction , $p = 0.036$) were significantly correlated with SPE in Switch. The outline in g and f indicates the locations of MD. **h.** The more medial model-free and the more lateral model-based related MD (MD_{MF} and MD_{MB}). The outline indicates the locations of medial and lateral MDs derived from the AAL3 atlas. **i.** Fiber density contrast between white-matter connections of MD_{MB}-PFC and MD_{MF}-PFC. The contrast map indicates that the model-based MD has preferential white-matter connection with dorsal PFC while model-free MD shows preferential white-matter connection with vmPFC/OFC. Tractography analysis is conducted on a separate sample of 113 participants. Contrast map is thresholded at $p <= 0.005$ at voxel level and cluster size > 40. Cold color: MD_{MB}-PFC $>$ MD_{MF}-PFC; Warm color: MD_{MB}-PFC $<$ MD_{MF}-PFC. Streamline visualization of the tractography between the MD ROIs and PFC in two representative brain slices is also shown. Blue: MD_{MB}-PFC; Red: MD_{MF}-PFC. **j.** The psychophysiological interaction (PPI) shows the MB-related MD connections with dorsal PFC and MF-related MD connections with the OFC. FWE small-volume correction was applied across the whole PFC ($p < 0.05$). T-maps are displayed at $p < 0.001$, uncorrected, for display purpose only. **k.** Dynamic causal modeling (DCM) analysis of human fMRI data. Bayesian parameter averaging was applied to estimate model parameters in a network comprising lateral MD, medial MD, OFC, and dlPFC with reciprocal endogenous connectivity among all regions (except between the two MDs) for Steady State (SS), model-free-related Switch trials (SW_{MF}), and model-based-related Switch trials (SW_{MB}). **l.** Posterior means of cortico-cortical (OFC to dlPFC) and thalamo-cortical (MD to PFC) connectivity across SS, SW_{MF}, and SW_{MB}. Error bars represent SEM derived from bootstrap resampling. MD: mediodorsal thalamus, MDm: medial MD, MDl: lateral MD, OFC: orbitofrontal cortex, dlPFC: dorsal PFC, MF: model-free, MB: model-based.

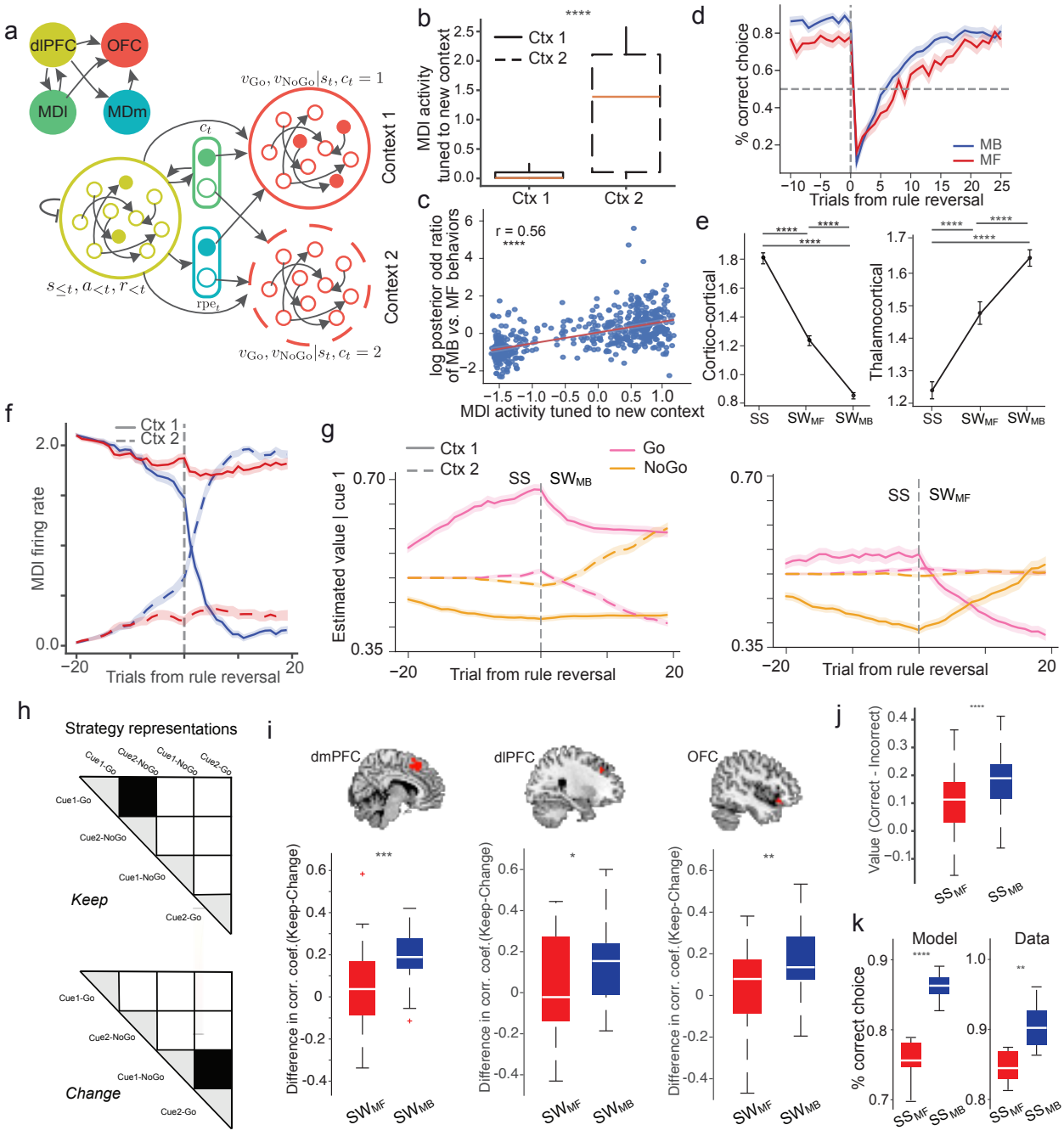

Figure 2.9: **CogLink model reveals the roles of thalamocortical connections for reinforcement learning during strategy switching and underlying circuit computations.** Caption continues at the next page.

Figure 2.9: **CogLink model reveals the roles of thalamocortical connections for reinforcement learning during strategy switching and underlying circuit computations. a.** A schematic of the brain areas and circuits modeled in our CogLink model. dlPFC encodes past sensorimotor-outcome associations as well as the current stimulus; MDl encodes the inferred context; MDm encodes the contextual reward prediction errors; OFC encodes the contextual action values given the inferred context and stimulus. We evaluate our CogLink model in the probabilistic reversal task for 500 blocks. **b.** The contextual encoding for MDl neurons tuned to the new context. **c.** A point plot ($n = 500$) of z-score of MDl activity tuned to the new context against the z-score of the log posterior odd ratio of MB versus MF behaviors (****$p = 2.00 \times 10^{-5}$; permutation test on Pearson's correlation coefficient $r = 0.56$). **d.** Summarized plot (mean±s.e.m.) for average accurate choice probability at MB ($n = 308$) and MF blocks ($n = 192$) from the model. **e.** Summarized plot (mean±s.e.m., $n = 500$) for the effective connectivity of the model at different conditions. The left column shows the effective intercortical connectivity (dlPFC to OFC and OFC to dlPFC) (****$p < 10^{-4}$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). The right column shows the effective thalamocortical connectivity (from both lateral and medial parts of MD to both dlPFC and OFC) (****$p < 10^{-4}$; Bonferroni corrected Kruskal-Wallis test with post hoc Dunn's test). **f.** Summarized plot (mean±s.e.m.) of MDl firing rate encoding the contextual signals in MF and MB blocks. **g.** Summarized plot (mean±s.e.m.) of estimated action values after presenting with cue 1 in MB and MF blocks for OFC neurons. **h.** The RDMs of the two models (i.e., representations of the Keep Strategy and Change Strategy based on the predicted correlation distance for the four types of trials (cue1→'Go', cue2→'NoGo', cue1→'NoGo' and cue2→'Go'). Black elements indicate similarity, white elements indicate dissimilarity between the response patterns of different types of trials. **i.** The RSA results separated for the MF and MB Switch trials (SW_{MF} and SW_{MB}) for the PFC regions (dmPFC, dlPFC and OFC) in humans. The differences of the representational dissimilarity between Keep Strategy and Change Strategy model in **h.** were calculated. Less differences were found for the SW_{MF} compared to SW_{MB} (***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$). **j.** A box plot of value difference between correct strategy and wrong strategy before reversal for model-based and model-free blocks (****$p < 10^{-4}$; two-sided rank sum test). **k.** The different behaviors between MB and MF blocks before the reversals for both the human data and the model (****$p < 10^{-4}$, **$p < 0.01$). MD: mediodorsal thalamus, MDm: medial MD, MDl: lateral MD, OFC: orbitofrontal cortex, dlPFC: dorsolateral PFC, MF: model-free, MB: model-based.

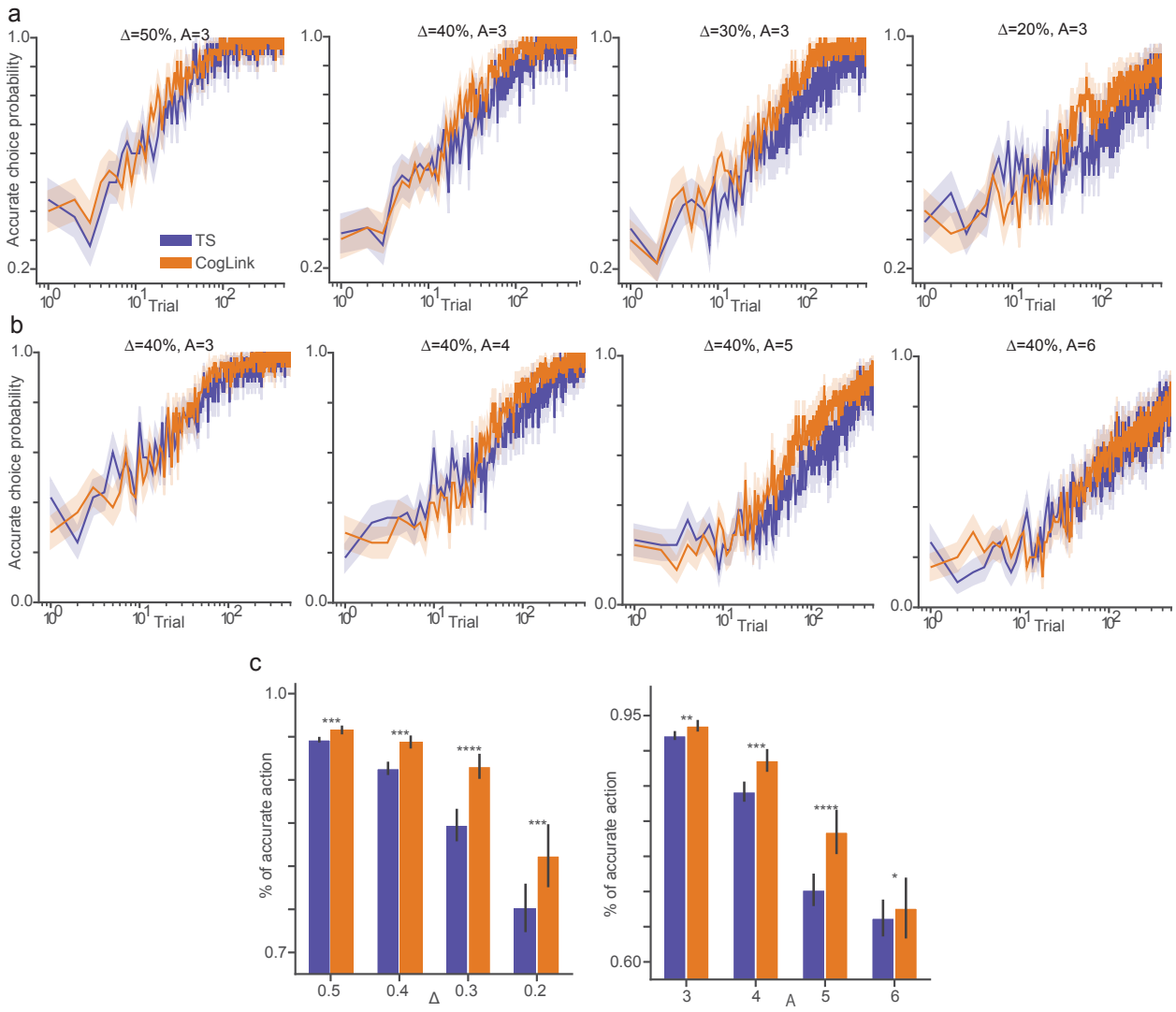

Figure 2.S1: **Comparison of basic CogLink model and Thompson Sampling (TS) in stationary environments. a, b.** Summarized plots (mean±s.e.m., $n = 50$) showing average accurate choice probability across stationary environments with varying reward gaps (Δ) and numbers of alternatives (A). In **a**, reward gaps decrease from $\Delta = 50\%$ to $\Delta = 20\%$ ($A = 3$), while in **b**, the number of alternatives increases from $A = 3$ to $A = 6$ ($\Delta = 40\%$). CogLink demonstrates faster convergence and higher long-term accuracy compared to TS across all conditions. **c.** Summarized plots (mean±s.e.m., $n = 50$) for average accuracy as a function of Δ (left) and A (right). CogLink outperforms TS significantly in all tested environments ($^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 10^{-3}$, $^{****}P < 10^{-4}$; two-way rank sum test). These results highlight CogLink's superior ability to balance exploration and exploitation under varying task difficulties.

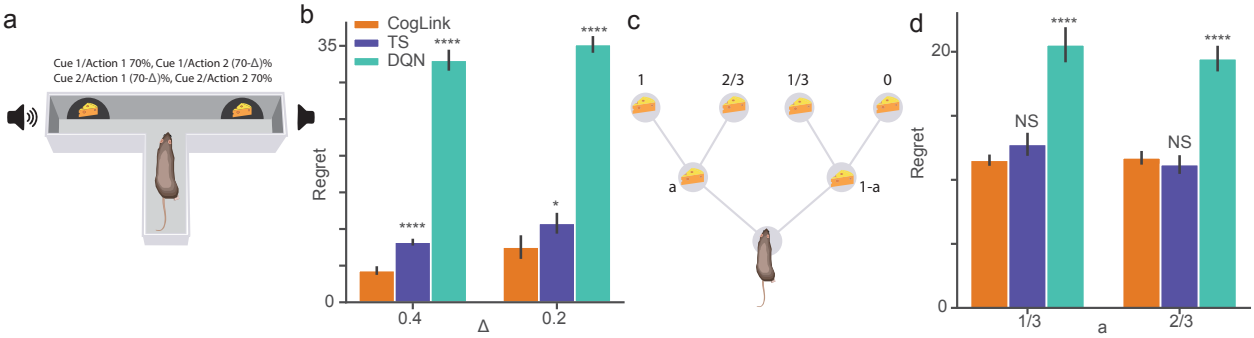

Figure 2.S2: **Comparison of basic CogLink, Thompson sampling (TS) and deep Q network (DQN) in various stationary environments. a.** Schematic of cued 2−AFC task parametrized by Δ, the gap between expected reward of two actions. **b.** Summarized plot (mean±s.e.m., $n = 50$) for regret in cued 2−AFC tasks across different Δ ($^{****}P < 10^{-4}$, $^{*}P < 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **c.** A schematic of a binary tree maze task with various probabilities to receive rewards at each node. **d.** Summarized plot (mean±s.e.m., $n = 50$) for regret in binary tree maze across different a ($^{****}P < 10^{-4}$, NS $P > 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test).

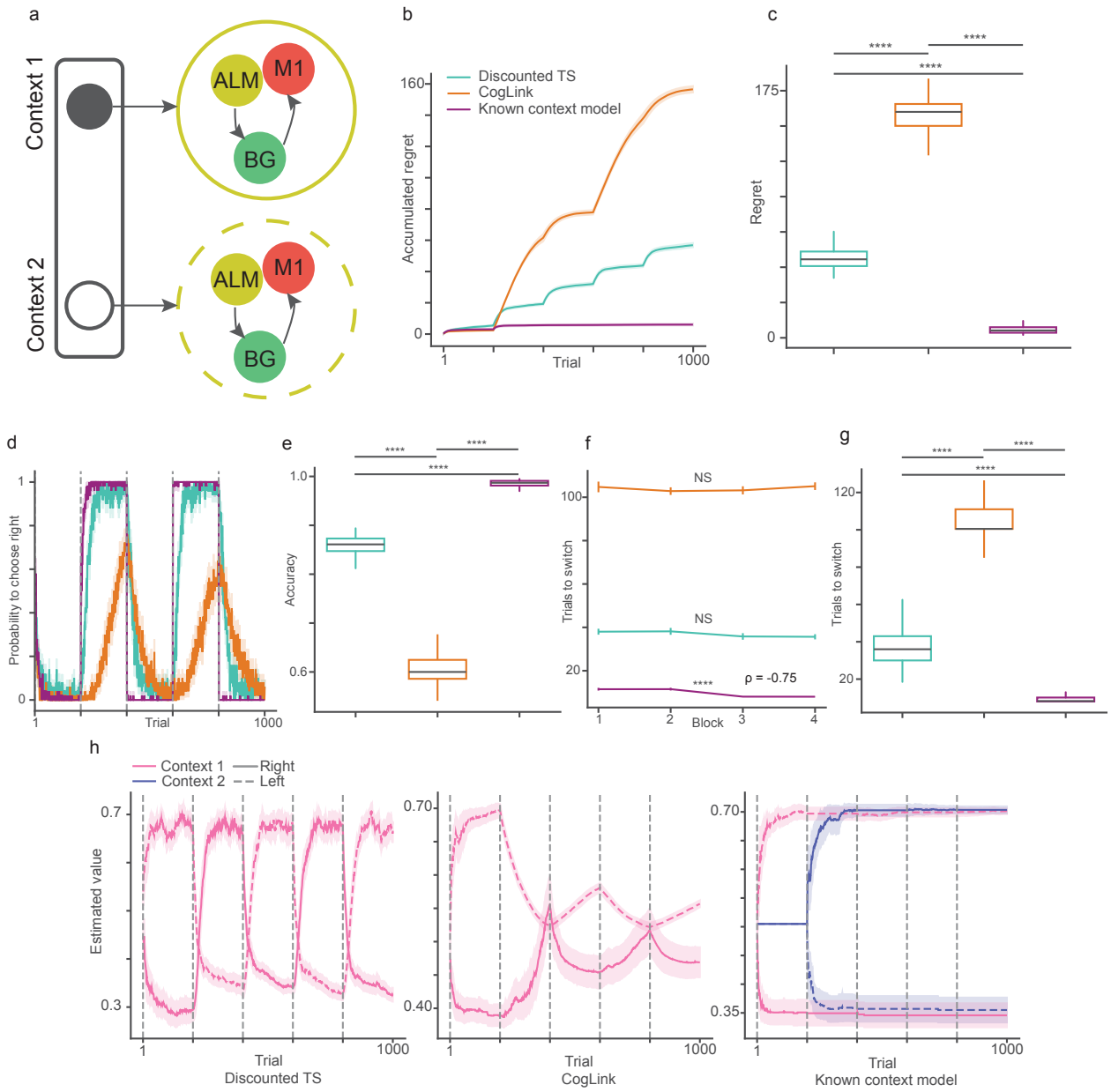

Figure 2.S3: **A contextual inference capacity is crucial for flexible decision-making in a dynamic environment.** We evaluate CogLink model, discounted TS and known context model in a probabilistic reversal task for 50 sessions of 1000 trials with probabilistic reversal every 200 trials. **a**. A schematic of a known context model. **b**. Summarized plot (mean±s.e.m., $n = 50$) for average accumulated regret. **c**. Boxplot ($n = 50$) for the final regret (****$P < 10^{-4}$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **d**. Summarized plot (mean±s.e.m., $n = 50$) for average probability to choose right. **e**. Boxplot ($n = 50$) for the accuracy (****$P < 10^{-4}$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). f. Summarized plot (mean±s.e.m., $n = 50$) for the number of trials to switch. (NS $P > 0.05$, ****$P < 10^{-4}$; permutation test on Spearman's rank correlation coefficient ρ.) g. Boxplot ($n = 200$, 4 switches from 50 trials) for the switching time (****$P < 10^{-4}$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test). **h.** Summarized plot (mean±s.e.m., $n = 50$) for average estimated action values.

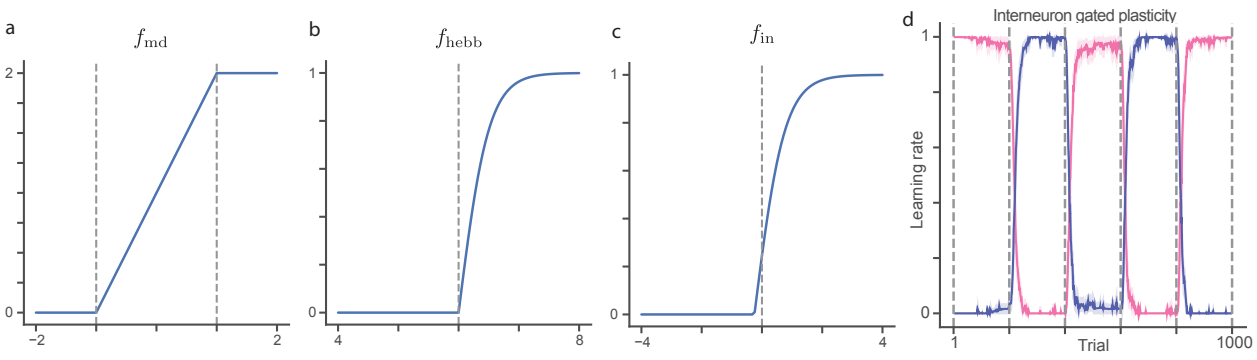

Figure 2.S4: **Nonlinearity used in CogLink and its effect on interneuron gated plasticity. a.** Nonlinearity for the MD frequency-current curve. **b.** Nonlinearity for the PFC-MD Hebbian plasticity. **c.** Nonlinearity for the VIP/PV cortical interneuron gated plasticity. **d.** Summarized plot (mean±s.e.m., $n = 50$) for normalized learning rate of interneuron gated plasticity in a probability reversal task.

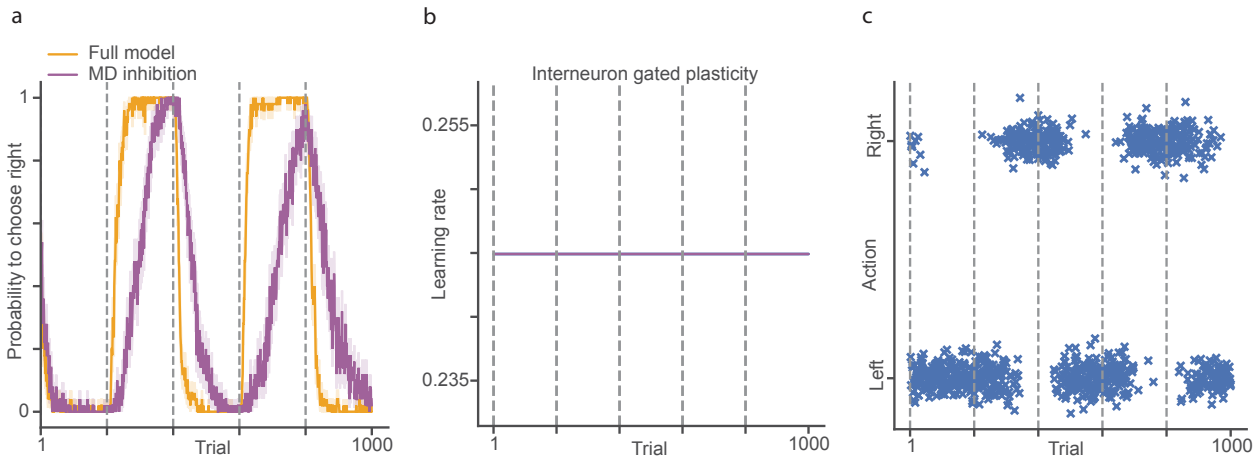

Figure 2.S5: **Behaviors and learning rates of MD inhibition model in a probability reversal task. a**. Summarized plot (mean±s.e.m., $n = 50$) for average probability to choose right. **b**. Summarized plot (mean±s.e.m., $n = 50$) for average learning rates for interneuron gated plasticity. Notice because MD is inhibited, the learning rate stays constant. **c**. Sample behavior of MD inhibition model in a probability reversal task.

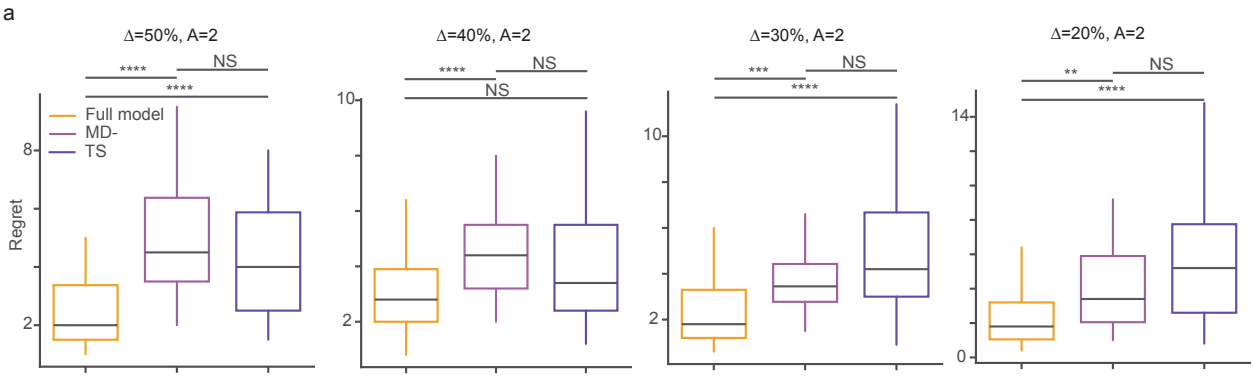

Figure 2.S6: **Comparison of thalamocortical model, MD inhibition model and Thompson sampling in various stationary AFC tasks.** We evaluate our models across various 2-AFC tasks for 50 sessions of 500 trials. **a**. Boxplot ($n = 50$) for the regret across different Δ ($^{****}P < 10^{-4}$, $^{***}P < 10^{-3}$, $^{**}P < 10^{-2}$, NS $P > 0.05$; Bonferroni-corrected Kruskal-Wallis test with post hoc Dunn's test).

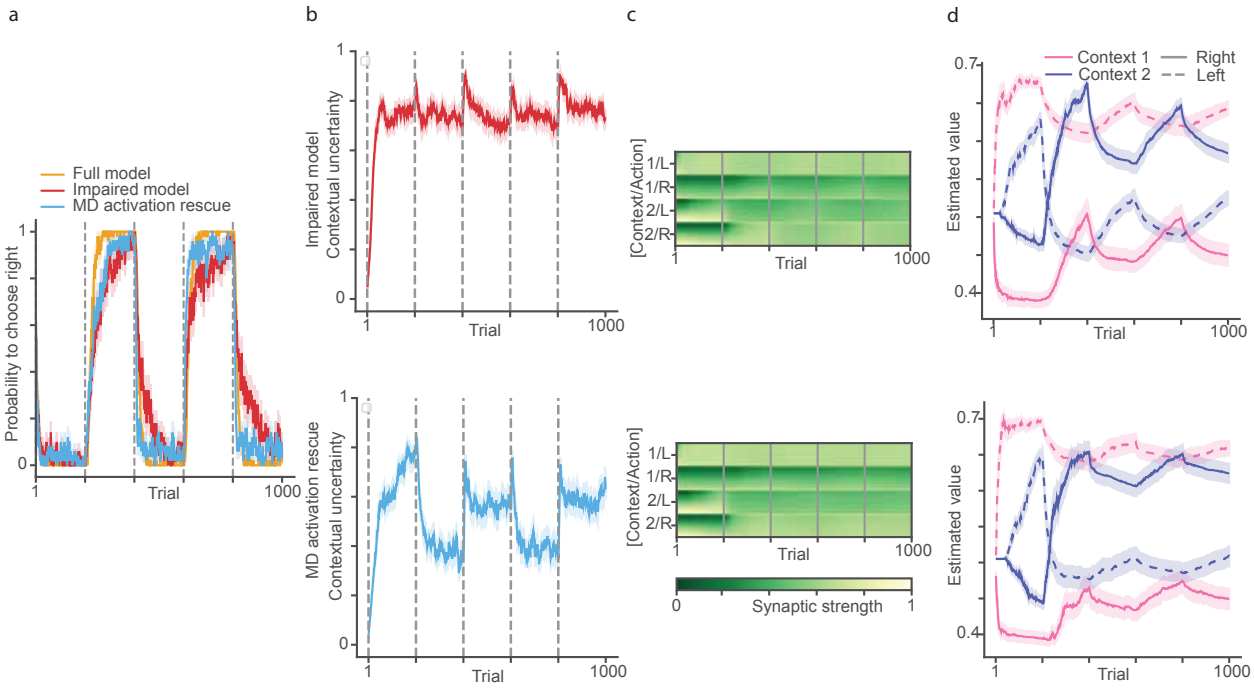

Figure 2.S7: **Comparison of the full thalamocortical model, impaired model and MD activation rescue model in a probabilistic reversal task.** **a**. Summarized plot (mean±s.e.m., $n = 50$) for average probability to choose right. **b-d**. The top row represents the impaired model and the bottom row represents the rescue model. **b**. Summarized plot (mean±s.e.m., $n = 50$) for average decoded contextual uncertainty values encoded in corticostriatal synapses **c**. Summarized plot (mean, $n = 50$) for average corticostriatal strength. **d**. Summarized plot (mean±s.e.m., $n = 50$) for average estimated action values encoded in corticostriatal synapses.

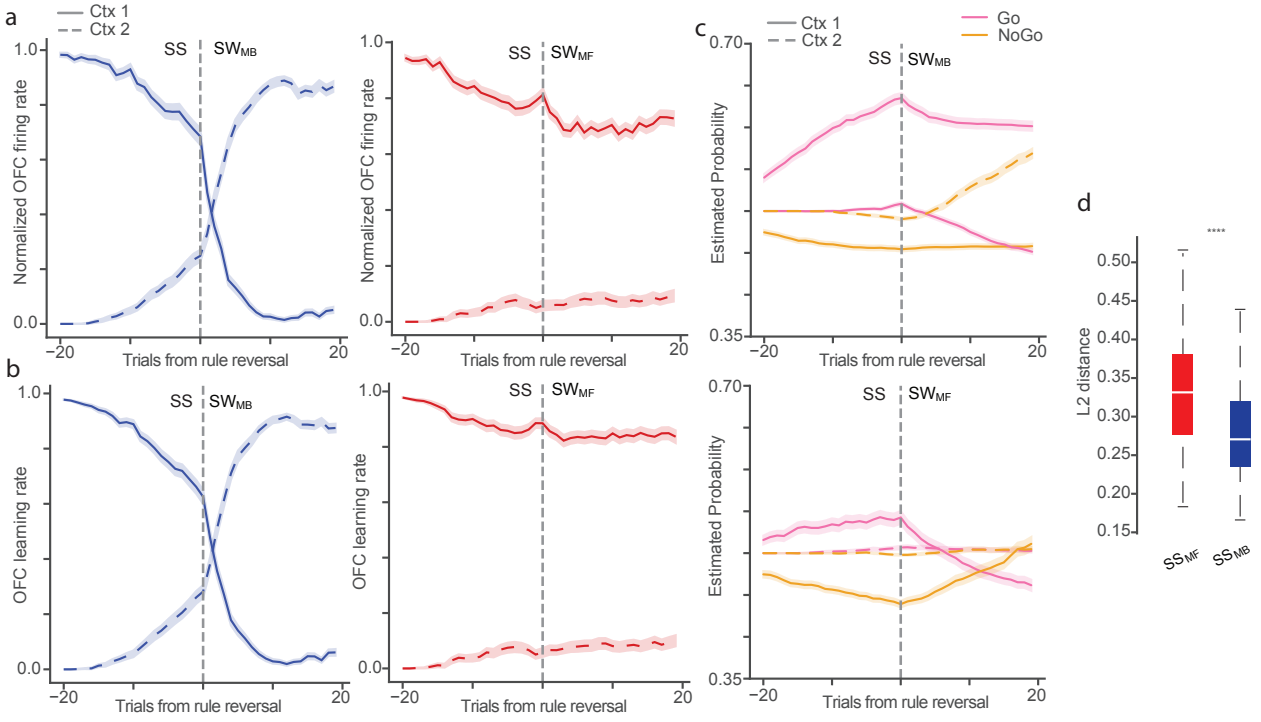

Figure 2.S8: **Normalized OFC firing rate and learning rate for CogLink. a.** Summarized plot (Mean±SEM) of normalized OFC firing rate at model-based ($n = 308$) and model-free ($n = 192$) blocks. **b.** Summarized plot (Mean±SEM) of normalized OFC learning rate at model-based and model-free blocks. **c.** Summarized plot (Mean±SEM) of generative model of receiving reward after presenting with cue 1 at model-based and model-free blocks. **d.** A box plot of L2 distance between estimated generative model before switch and true generative model for model-based and model-free blocks (****$p < 10^{-4}$; two-sided rank sum test)

References

[1] J. Bill, S. J. Gershman, and J. Drugowitsch. "Visual motion perception as online hierarchical inference". In: *Nat Commun* 13.1 (Dec. 2022), p. 7403.

[2] T. Rohe, A. C. Ehlis, and U. Noppeney. "The neural dynamics of hierarchical Bayesian causal inference in multisensory perception". In: *Nat Commun* 10.1 (Apr. 2019), p. 1907.

[3] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction". In: *Science* 331.6022 (2011), pp. 1279–1285. eprint: https://www.science.org/doi/pdf/10.1126/science.1192788.

[4] C. D. Mathys, E. I. Lomakina, J. Daunizeau, S. Iglesias, K. H. Brodersen, K. J. Friston, and K. E. Stephan. "Uncertainty in perception and the Hierarchical Gaussian Filter". In: *Front Hum Neurosci* 8 (2014), p. 825.

[5] D. C. Knill and A. Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *Trends Neurosci* 27.12 (Dec. 2004), pp. 712–719.

[6] K. P. Kording and D. M. Wolpert. "Bayesian integration in sensorimotor learning". In: *Nature* 427.6971 (Jan. 2004), pp. 244–247.

[7] D. J. Nott, C. Drovandi, and D. T. Frazier. "Bayesian Inference for Misspecified Generative Models". In: *Annual Review of Statistics and Its Application* 11.Volume 11, 2024 (2024), pp. 179–202.

[8] D. Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489.

[9] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proc Natl Acad Sci U S A* 111.23 (June 2014), pp. 8619–8624.

[10] I. E. Monosov. "How Outcome Uncertainty Mediates Attention, Learning, and Decision-Making". In: *Trends Neurosci* 43.10 (Oct. 2020), pp. 795–809.

[11] D. C. Knill and R. Whitman, eds. *Perception as Bayesian Inference*. Cambridge University Press, 1996.

[12] A. A. Stocker and E. P. Simoncelli. "Noise characteristics and prior expectations in human visual speed perception". In: *Nature Neuroscience* 9.4 (Mar. 2006), pp. 578–585.

[13] B. A. Wang, M. B. Wang, N. H. Lam, L. Mengxing, S. Li, R. D. Wimmer, P. M. Paz-Alonso, M. M. Halassa, and B. Pleger. "Thalamic regulation of reinforcement learning strategies across prefrontal-striatal networks". In: *Nature Communications* 16.1 (Oct. 2025).

[14] T. Zhou et al. "Enhancement of mediodorsal thalamus rescues aberrant belief dynamics in a mouse model with schizophrenia-associated mutation". In: *bioRxiv* (2024). eprint: https://www.biorxiv.org/content/early/2024/02/14/2024.01.08.574745.full.pdf.

[15] Y. Niv. "Reinforcement learning in the brain". In: *Journal of Mathematical Psychology* 53.3 (June 2009), pp. 139–154.

[16] A. Soltani and X. J. Wang. "A biophysically based neural model of matching law behavior: melioration by stochastic synapses". In: *J Neurosci* 26.14 (Apr. 2006), pp. 3731–3744.

[17] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. "Bayesian inference with probabilistic population codes". In: *Nat Neurosci* 9.11 (Nov. 2006), pp. 1432–1438.

[18] P. Hoyer and A. Hyvärinen. "Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002.

[19] R. P. Rao. "Bayesian computation in recurrent neural circuits". In: *Neural Comput* 16.1 (Jan. 2004), pp. 1–38.

[20] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick. "A distributional code for value in dopamine-based reinforcement learning". In: *Nature* 577.7792 (2020), pp. 671–675.

[21] J. D. Roitman and M. N. Shadlen. "Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task". In: *J Neurosci* 22.21 (Nov. 2002), pp. 9475–9489.

[22] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[23] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. "Gambling in a rigged casino: The adversarial multi-armed bandit problem". In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. 1995, pp. 322–331.

[24] N. Korda, E. Kaufmann, and R. Munos. "Thompson Sampling for 1-Dimensional Exponential Family Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc., 2013.

[25] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (Feb. 2015), pp. 529–533.

[26] C. S. Chen, D. Mueller, E. Knep, R. B. Ebitz, and N. M. Grissom. "Dopamine and Norepinephrine Differentially Mediate the Exploration–Exploitation Tradeoff". In: *The Journal of Neuroscience* 44.44 (Aug. 2024), e1194232024.

[27] K. Doya. "Metalearning and neuromodulation". In: *Neural Networks* 15.4–6 (June 2002), pp. 495–506.

[28] K. Sakai and R. E. Passingham. "Prefrontal interactions reflect future task operations". In: *Nat Neurosci* 6.1 (Jan. 2003), pp. 75–81.

[29] E. K. Miller and J. D. Cohen. "An integrative theory of prefrontal cortex function". In: *Annu Rev Neurosci* 24 (2001), pp. 167–202.

[30] M. Wolff and M. M. Halassa. "The mediodorsal thalamus in executive control". In: *Neuron* 112.6 (Mar. 2024), pp. 893–908.

[31] M. B. Wang and M. M. Halassa. "Thalamocortical contribution to flexible learning in neural systems". In: *Network Neuroscience* 6.4 (2022), pp. 980–997.

[32] D. N. Scott, A. Mukherjee, M. R. Nassar, and M. M. Halassa. "Thalamocortical architectures for flexible cognition and efficient learning". In: *Trends in Cognitive Sciences* 28.8 (Aug. 2024), pp. 739–756.

[33] M. Nakajima and M. M. Halassa. "Thalamic control of functional cortical connectivity". In: *Current Opinion in Neurobiology* 44 (June 2017), pp. 127–131.

[34] M. M. Halassa and S. Kastner. "Thalamic functions in distributed cognitive control". In: *Nature Neuroscience* 20.12 (Nov. 2017), pp. 1669–1679.

[35] M. M. Halassa and S. M. Sherman. "Thalamocortical Circuit Motifs: A General Framework". In: *Neuron* 103.5 (Sept. 2019), pp. 762–770.

[36] L. I. Schmitt, R. D. Wimmer, M. Nakajima, M. Happ, S. Mofakham, and M. M. Halassa. "Thalamic amplification of cortical connectivity sustains attentional control". In: *Nature* 545.7653 (May 2017), pp. 219–223.

[37] R. V. Rikhye, A. Gilra, and M. M. Halassa. "Thalamic regulation of switching between cortical representations enables cognitive flexibility". In: *Nat Neurosci* 21.12 (Dec. 2018), pp. 1753–1763.

[38] A. Mukherjee, N. H. Lam, R. D. Wimmer, and M. M. Halassa. "Thalamic circuits for independent control of prefrontal signal and noise". In: *Nature* 600.7887 (Dec. 2021), pp. 100–104.

[39] N. H. Lam, A. Mukherjee, R. D. Wimmer, M. R. Nassar, Z. S. Chen, and M. M. Halassa. "Prefrontal transthalamic uncertainty processing drives flexible switching". In: *Nature* 637.8044 (Nov. 2024), pp. 127–136.

[40] X. Chen, E. Sorenson, and K. Hwang. "Thalamocortical contributions to working memory processes during the n-back task". In: *Neurobiol Learn Mem* 197 (Jan. 2023), p. 107701.

[41] M. Canto-Bustos, F. K. Friason, C. Bassi, and A. M. Oswald. "Disinhibitory Circuitry Gates Associative Synaptic Plasticity in Olfactory Cortex". In: *J Neurosci* 42.14 (Apr. 2022), pp. 2942–2950.

[42] L. E. Williams and A. Holtmaat. "Higher-Order Thalamocortical Inputs Gate Synaptic Long-Term Potentiation via Disinhibition". In: *Neuron* 101.1 (Jan. 2019), pp. 91–102.

[43] G. V. Moustakides. "Optimal Stopping Times for Detecting Changes in Distributions". In: *The Annals of Statistics* 14.4 (1986), pp. 1379–1387.

[44] G. Lorden. "Procedures for Reacting to a Change in Distribution". In: *The Annals of Mathematical Statistics* 42.6 (1971), pp. 1897–1908.

[45] S. Chakraborty, N. Kolling, M. E. Walton, and A. S. Mitchell. "Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments". In: *Elife* 5 (May 2016).

[46] F. Alcaraz, F. Naneix, E. Desfosses, A. R. Marchand, M. Wolff, and E. Coutureau. "Dissociable effects of anterior and mediodorsal thalamic lesions on spatial goal-directed behavior". In: *Brain Struct Funct* 221.1 (Jan. 2016), pp. 79–89.

[47] K. Hwang, J. Bruss, D. Tranel, and A. D. Boes. "Network Localization of Executive Function Deficits in Patients with Focal Thalamic Lesions". In: *J Cogn Neurosci* 32.12 (Dec. 2020), pp. 2303–2319.

[48] S. C. Baker, A. B. Konova, N. D. Daw, and G. Horga. "A distinct inferential mechanism for delusions in schizophrenia". In: *Brain* 142.6 (June 2019), pp. 1797–1812.

[49] J. M. Sheffield, P. Suthaharan, P. Leptourgos, and P. R. Corlett. "Belief Updating and Paranoia in Individuals With Schizophrenia". In: *Biol Psychiatry Cogn Neurosci Neuroimaging* 7.11 (Nov. 2022), pp. 1149–1157.

[50] R. A. Adams, G. Napier, J. P. Roiser, C. Mathys, and J. Gilleen. "Attractor-like Dynamics in Belief Updating in Schizophrenia". In: *J Neurosci* 38.44 (Oct. 2018), pp. 9471–9485.

[51] M. R. Nassar, J. A. Waltz, M. A. Albrecht, J. M. Gold, and M. J. Frank. "All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs". In: *Brain* 144.3 (Apr. 2021), pp. 1013–1029.

[52] P. R. Corlett and P. Fletcher. "Modelling delusions as temporally-evolving beliefs". In: *Cogn Neuropsychiatry* 26.4 (July 2021), pp. 231–241.

[53] P. Corlett, J. Taylor, X.-J. Wang, P. Fletcher, and J. Krystal. "Toward a neurobiology of delusions". In: *Progress in Neurobiology* 92.3 (2010), pp. 345–369.

[54] A. S. Huang, R. D. Wimmer, N. H. Lam, B. A. Wang, S. Suresh, M. J. Roeske, B. Pleger, M. M. Halassa, and N. D. Woodward. "A prefrontal thalamocortical readout for conflict-related executive dysfunction in schizophrenia". In: *Cell Reports Medicine* 5.11 (Nov. 2024), p. 101802.

[55] A. Anticevic and M. M. Halassa. "The thalamus in psychosis spectrum disorder". In: *Frontiers in Neuroscience* 17 (Apr. 2023).

[56] A. Mukherjee and M. M. Halassa. "The Associative Thalamus: A Switchboard for Cortical Operations and a Promising Target for Schizophrenia". In: *The Neuroscientist* 30.1 (Aug. 2022), pp. 132–147.

[57] R. Paz and J. M. nez-Amaya. "The mediodorsal thalamic nucleus and schizophrenia". In: *J Psychiatry Neurosci* 33.6 (Nov. 2008), pp. 489–498.

[58] A. Anticevic, G. Yang, A. Savic, J. D. Murray, M. W. Cole, G. Repovs, G. D. Pearlson, and D. C. Glahn. "Mediodorsal and visual thalamic connectivity differ in schizophrenia and bipolar disorder with and without psychosis history". In: *Schizophr Bull* 40.6 (Nov. 2014), pp. 1227–1243.

[59] W. Byne, M. S. Buchsbaum, E. Kemether, E. A. Hazlett, A. Shinwari, V. Mitropoulou, and L. J. Siever. "Magnetic resonance imaging of the thalamic mediodorsal nucleus and pulvinar in schizophrenia and schizotypal personality disorder". In: *Arch Gen Psychiatry* 58.2 (Feb. 2001), pp. 133–140.

[60] N. D. Woodward, H. Karbasforoushan, and S. Heckers. "Thalamocortical dysconnectivity in schizophrenia". In: *Am J Psychiatry* 169.10 (Oct. 2012), pp. 1092–1099.

[61] E. Pomarol-Clotet, E. J. guez, R. Salvador, S. ó, J. J. Gomar, F. Vila, J. Ortiz-Gil, Y. Iturria-Medina, A. Capdevila, and P. J. McKenna. "Medial prefrontal cortex pathology in schizophrenia as revealed by convergent findings from multimodal imaging". In: *Mol Psychiatry* 15.8 (Aug. 2010), pp. 823–830.

[62] P. Seeman and T. Lee. "Antipsychotic drugs: direct correlation between clinical potency and presynaptic action on dopamine neurons". In: *Science* 188.4194 (June 1975), pp. 1217–1219.

[63] I. Creese, D. R. Burt, and S. H. Snyder. "Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs". In: *Science* 192.4238 (Apr. 1976), pp. 481–483.

[64] H. Y. Meltzer, S. Matsubara, and J. C. Lee. "Classification of typical and atypical antipsychotic drugs on the basis of dopamine D-1, D-2 and serotonin2 pKi values". In: *J Pharmacol Exp Ther* 251.1 (Oct. 1989), pp. 238–246.

[65] D. F. Wong et al. "Positron emission tomography reveals elevated D2 dopamine receptors in drug-naive schizophrenics". In: *Science* 234.4783 (Dec. 1986), pp. 1558–1563.

[66] A. Abi-Dargham et al. "Increased baseline occupancy of D2 receptors by dopamine in schizophrenia". In: *Proc Natl Acad Sci U S A* 97.14 (July 2000), pp. 8104–8109.

[67] M. Cazorla, M. Shegda, B. Ramesh, N. L. Harrison, and C. Kellendonk. "Striatal D2 receptors regulate dendritic morphology of medium spiny neurons via Kir2 channels". In: *J Neurosci* 32.7 (Feb. 2012), pp. 2398–2409.

[68] J. A. Waltz. "The neural underpinnings of cognitive flexibility and their disruption in psychotic illness". In: *Neuroscience* 345 (Mar. 2017), pp. 203–217.

[69] L. Deserno, R. Boehme, C. Mathys, T. Katthagen, J. Kaminski, K. E. Stephan, A. Heinz, and F. Schlagenhauf. "Volatility Estimates Increase Choice Switching and Relate to Prefrontal Activity in Schizophrenia". In: *Biol Psychiatry Cogn Neurosci Neuroimaging* 5.2 (Feb. 2020), pp. 173–183.

[70] N. D. Daw, Y. Niv, and P. Dayan. "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature Neuroscience* 8.12 (Nov. 2005), pp. 1704–1711.

[71] P. Dayan and Y. Niv. "Reinforcement learning: The Good, The Bad and The Ugly". In: *Current Opinion in Neurobiology* 18.2 (Apr. 2008), pp. 185–196.

[72] P. Dayan. "Goal-directed control and its antipodes". In: *Neural Networks* 22.3 (Apr. 2009), pp. 213–219.

[73] P. Smittenaar, T. H. FitzGerald, V. Romei, N. D. Wright, and R. J. Dolan. "Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans". In: *Neuron* 80.4 (Nov. 2013), pp. 914–919.

[74] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6 (Mar. 2011), pp. 1204–1215.

[75] S. J. Gershman, A. B. Markman, and A. R. Otto. "Retrospective revaluation in sequential decision making: A tale of two systems." In: *Journal of Experimental Psychology: General* 143.1 (2014), pp. 182–194.

[76] A. R. Otto, S. J. Gershman, A. B. Markman, and N. D. Daw. "The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive". In: *Psychological Science* 24.5 (Apr. 2013), pp. 751–761.

[77] B. Averbeck and J. P. O'Doherty. "Reinforcement-learning in fronto-striatal circuits". In: *Neuropsychopharmacology* 47.1 (Aug. 2021), pp. 147–162.

[78] H. H. Yin, B. J. Knowlton, and B. W. Balleine. "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning". In: *European Journal of Neuroscience* 19.1 (Jan. 2004), pp. 181–189.

[79] K. Wunderlich, P. Smittenaar, and R. J. Dolan. "Dopamine Enhances Model-Based over Model-Free Choice Behavior". In: *Neuron* 75.3 (Aug. 2012), pp. 418–424.

[80] B. W. Balleine and J. P. O'Doherty. "Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action". In: *Neuropsychopharmacology* 35.1 (Sept. 2009), pp. 48–69.

[81] W. H. Alexander and J. W. Brown. "Medial prefrontal cortex as an action-outcome predictor". In: *Nature Neuroscience* 14.10 (Sept. 2011), pp. 1338–1344.

[82] S. de Wit, P. R. Corlett, M. R. Aitken, A. Dickinson, and P. C. Fletcher. "Differential Engagement of the Ventromedial Prefrontal Cortex by Goal-Directed and Habitual Behavior toward Food Pictures in Humans". In: *The Journal of Neuroscience* 29.36 (Sept. 2009), pp. 11330–11338.

[83] S. E. Akkermans et al. "Frontostriatal functional connectivity correlates with repetitive behaviour across autism spectrum disorder and obsessive–compulsive disorder". In: *Psychological Medicine* 49.13 (Oct. 2018), pp. 2247–2255.

[84] R. V. Rikhye, R. D. Wimmer, and M. M. Halassa. "Toward an Integrative Theory of Thalamic Function". In: *Annual Review of Neuroscience* 41.1 (July 2018), pp. 163–183.

[85] J. M. Shine, L. D. Lewis, D. D. Garrett, and K. Hwang. "The impact of the human thalamus on brain-wide information processing". In: *Nature Reviews Neuroscience* 24.7 (May 2023), pp. 416–430.

[86] J. Q. Kosciessa, U. Lindenberger, and D. D. Garrett. "Thalamocortical excitability modulation guides human perception under uncertainty". In: *Nature Communications* 12.1 (Apr. 2021).

[87] A. Hummos, B. A. Wang, S. Drammis, M. M. Halassa, and B. Pleger. "Thalamic regulation of frontal interactions in human cognitive flexibility". In: *PLOS Computational Biology* 18.9 (Sept. 2022). Ed. by S. Palminteri, e1010500.

[88] K. Hwang, M. A. Bertolero, W. B. Liu, and M. D'Esposito. "The Human Thalamus Is an Integrative Hub for Functional Brain Networks". In: *The Journal of Neuroscience* 37.23 (Apr. 2017), pp. 5594–5607.

[89] N. D. Daw. "Are we of two minds?" In: *Nature Neuroscience* 21.11 (Oct. 2018), pp. 1497–1499.

[90] R. J. Dolan and P. Dayan. "Goals and Habits in the Brain". In: *Neuron* 80.2 (Oct. 2013), pp. 312–325.

[91] N. N. Foster et al. "The mouse cortico–basal ganglia–thalamic network". In: *Nature* 598.7879 (Oct. 2021), pp. 188–194.

[92] G. E. Alexander, M. R. DeLong, and P. L. Strick. "Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex". In: *Annual Review of Neuroscience* 9.1 (Mar. 1986), pp. 357–381.

[93] J. Cox and I. B. Witten. "Striatal circuits for reward learning and decision-making". In: *Nat Rev Neurosci* 20.8 (Aug. 2019), pp. 482–494.

[94] C. C. Petersen and S. Crochet. "Synaptic computation and sensory processing in neocortical layer 2/3". In: *Neuron* 78.1 (Apr. 2013), pp. 28–48.

[95] A. L. Barth and J. F. Poulet. "Experimental evidence for sparse firing in the neocortex". In: *Trends Neurosci* 35.6 (June 2012), pp. 345–355.

[96] J. N. Kerr, C. P. de Kock, D. S. Greenberg, R. M. Bruno, B. Sakmann, and F. Helmchen. "Spatial organization of neuronal population responses in layer 2/3 of rat barrel cortex". In: *J Neurosci* 27.48 (Nov. 2007), pp. 13316–13328.

[97] S. Kato et al. "Action Selection and Flexible Switching Controlled by the Intralaminar Thalamic Neurons". In: *Cell Reports* 22.9 (Feb. 2018), pp. 2370–2382.

[98] T. Minamimoto, Y. Hori, and M. Kimura. "Roles of the thalamic CM–PF complex—Basal ganglia circuit in externally driven rebias of action". In: *Brain Research Bulletin* 78.2–3 (Feb. 2009), pp. 75–79.

[99] R. D. Wimmer, L. I. Schmitt, T. J. Davidson, M. Nakajima, K. Deisseroth, and M. M. Halassa. "Thalamic control of sensory selection in divided attention". In: *Nature* 526.7575 (Oct. 2015), pp. 705–709.

[100] D. M. Wolpert and M. Kawato. "Multiple paired forward and inverse models for motor control". In: *Neural Netw* 11.7-8 (Oct. 1998), pp. 1317–1329.

[101] J. B. Heald, M. Lengyel, and D. M. Wolpert. "Contextual inference underlies the learning of sensorimotor repertoires". In: *Nature* 600.7889 (Dec. 2021), pp. 489–493.

[102] J.-H. Kim, K. Daie, and N. Li. "A combinatorial neural code for long-term motor memory". In: *Nature* 637.8046 (Nov. 2024), pp. 663–672.

[103] E. G. Jones. "The thalamic matrix and thalamocortical synchrony". In: *Trends Neurosci* 24.10 (Oct. 2001), pp. 595–601.

[104] S. M. Sherman and R. W. Guillery. *Exploring the Thalamus and Its Role in Cortical Function, Second Edition*. English. Hardcover. The MIT Press, Dec. 2005.

[105] M. Tanaka. "Cognitive signals in the primate motor thalamus predict saccade timing". In: *J Neurosci* 27.44 (Oct. 2007), pp. 12109–12118.

[106] Y. B. Saalmann and S. Kastner. "The cognitive thalamus". In: *Front Syst Neurosci* 9 (2015), p. 39.

[107] H. Zhou, R. J. Schafer, and R. Desimone. "Pulvinar-Cortex Interactions in Vision and Attention". In: *Neuron* 89.1 (Jan. 2016), pp. 209–220.

[108] S. S. Bolkan, J. M. Stujenske, S. Parnaudeau, T. J. Spellman, C. Rauffenbart, A. I. Abbas, A. Z. Harris, J. A. Gordon, and C. Kellendonk. "Thalamic projections sustain prefrontal activity during working memory maintenance". In: *Nat Neurosci* 20.7 (July 2017), pp. 987–996.

[109] Z. V. Guo, H. K. Inagaki, K. Daie, S. Druckmann, C. R. Gerfen, and K. Svoboda. "Maintenance of persistent activity in a frontal thalamocortical loop". In: *Nature* 545.7653 (2017), pp. 181–186.

[110] W. Guo, A. R. Clause, A. Barth-Maron, and D. B. Polley. "A Corticothalamic Circuit for Dynamic Switching between Feature Detection and Discrimination". In: *Neuron* 95.1 (July 2017), pp. 180–194.

[111] A. Mukherjee, N. Bajwa, N. H. Lam, C. Porrero, F. Clasca, and M. M. Halassa. "Variation of connectivity across exemplar sensory and associative thalamocortical loops in the mouse". In: *Elife* 9 (Oct. 2020).

[112] Y. Wang and Q.-Q. Sun. "A prefrontal motor circuit initiates persistent movement". In: *Nature Communications* 15.1 (June 2024).

[113] S. J. Gershman. "Deconstructing the human algorithms for exploration". In: *Cognition* 173 (Apr. 2018), pp. 34–42.

[114] J. D. Cohen, S. M. McClure, and A. J. Yu. "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration". In: *Philos Trans R Soc Lond B Biol Sci* 362.1481 (May 2007), pp. 933–942.

[115] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. "Humans use directed and random exploration to solve the explore-exploit dilemma". In: *J Exp Psychol Gen* 143.6 (Dec. 2014), pp. 2074–2081.

[116] W. J. Ma and M. Jazayeri. "Neural coding of uncertainty and probability". In: *Annu Rev Neurosci* 37 (2014), pp. 205–220.

[117] E. Y. Walker, S. Pohl, R. N. Denison, D. L. Barack, J. Lee, N. Block, W. J. Ma, and F. Meyniel. "Studying the neural representations of uncertainty". In: *Nat Neurosci* 26.11 (Nov. 2023), pp. 1857–1867.

[118] K. Akiti, I. Tsutsui-Kimura, Y. Xie, A. Mathis, J. E. Markowitz, R. Anyoha, S. R. Datta, M. W. Mathis, N. Uchida, and M. Watabe-Uchida. "Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction". In: *Neuron* 110.22 (Nov. 2022), pp. 3789–3804.

[119] M. O'Neill and W. Schultz. "Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value". In: *Neuron* 68.4 (Nov. 2010), pp. 789–800.

[120] P. Masset, T. Ott, A. Lak, J. Hirokawa, and A. Kepecs. "Behavior- and Modality-General Representation of Confidence in Orbitofrontal Cortex". In: *Cell* 182.1 (July 2020), pp. 112–126.

[121] G. Orban, P. Berkes, J. Fiser, and M. Lengyel. "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex". In: *Neuron* 92.2 (Oct. 2016), pp. 530–543.

[122] E. Y. Walker, R. J. Cotton, W. J. Ma, and A. S. Tolias. "A neural basis of probabilistic computation in visual cortex". In: *Nat Neurosci* 23.1 (Jan. 2020), pp. 122–129.

[123] L. S. Geurts, J. R. H. Cooke, R. S. van Bergen, and J. F. M. Jehee. "Subjective confidence reflects representation of Bayesian probability in cortex". In: *Nat Hum Behav* 6.2 (Feb. 2022), pp. 294–305.

[124] R. Echeveste, L. Aitchison, G. Hennequin, and M. Lengyel. "Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference". In: *Nat Neurosci* 23.9 (Sept. 2020), pp. 1138–1149.

[125] S. Deneve. "Bayesian spiking neurons I: inference". In: *Neural Comput* 20.1 (Jan. 2008), pp. 91–117.

[126] M. A. van der Meer, A. Johnson, N. C. Schmitzer-Torbert, and A. D. Redish. "Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task". In: *Neuron* 67.1 (July 2010), pp. 25–32.

[127] B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, and N. D. Daw. "Model-based choices involve prospective neural activity". In: *Nat Neurosci* 18.5 (May 2015), pp. 767–772.

[128] J. Scher, N. Daw, P. Dayan, and J. P. O'Doherty. "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4 (May 2010), pp. 585–595.

[129] S. W. Lee, S. Shimojo, and J. P. O'Doherty. "Neural computations underlying arbitration between model-based and model-free learning". In: *Neuron* 81.3 (Feb. 2014), pp. 687–699.

[130] W. Schultz, P. Dayan, and P. R. Montague. "A neural substrate of prediction and reward". In: *Science* 275.5306 (Mar. 1997), pp. 1593–1599.

[131] T. Akam, I. Rodrigues-Vaz, I. Marcelo, X. Zhang, M. Pereira, R. F. Oliveira, P. Dayan, and R. M. Costa. "The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection". In: *Neuron* 109.1 (Jan. 2021), pp. 149–163.

[132] P. P. Witkowski, S. A. Park, and E. D. Boorman. "Neural mechanisms of credit assignment for inferred relationships in a structured world". In: *Neuron* 110.16 (Aug. 2022), pp. 2680–2690.

[133] N. W. Schuck, M. B. Cai, R. C. Wilson, and Y. Niv. "Human Orbitofrontal Cortex Represents a Cognitive Map of State Space". In: *Neuron* 91.6 (Sept. 2016), pp. 1402–1412.

[134] A. Soltani and E. Koechlin. "Computational models of adaptive behavior and prefrontal cortex". In: *Neuropsychopharmacology* 47.1 (Jan. 2022), pp. 58–71.

[135] E. G. Jones, ed. *The Thalamus*. Springer US, 1985.

[136] M. Nakajima, L. I. Schmitt, and M. M. Halassa. "Prefrontal Cortex Regulates Sensory Filtering through a Basal Ganglia-to-Thalamus Pathway". In: *Neuron* 103.3 (Aug. 2019), pp. 445–458.

[137] J. M. Phillips, N. A. Kambi, and Y. B. Saalmann. "A Subcortical Pathway for Rapid, Goal-Driven, Attentional Filtering". In: *Trends Neurosci* 39.2 (Feb. 2016), pp. 49–51.

[138] P. R. Montague, P. Dayan, and T. J. Sejnowski. "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". In: *J Neurosci* 16.5 (Mar. 1996), pp. 1936–1947.

[139] H. M. Bayer and P. W. Glimcher. "Midbrain dopamine neurons encode a quantitative reward prediction error signal". In: *Neuron* 47.1 (July 2005), pp. 129–141.

[140] N. S. Bamford, R. M. Wightman, and D. Sulzer. "Dopamine's Effects on Corticostriatal Synapses during Reward-Based Behaviors". In: *Neuron* 97.3 (Feb. 2018), pp. 494–510.

[141] J. C. R. Whittington and R. Bogacz. "Theories of Error Back-Propagation in the Brain". In: *Trends Cogn Sci* 23.3 (Mar. 2019), pp. 235–250.

[142] M. Minsky. "Steps toward Artificial Intelligence". In: *Proceedings of the IRE* 49.1 (1961), pp. 8–30.

[143] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (Oct. 1986), pp. 533–536.

[144] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton. "Backpropagation and the brain". In: *Nat Rev Neurosci* 21.6 (June 2020), pp. 335–346.

[145] M. McCloskey and N. J. Cohen. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". In: ed. by G. H. Bower. Vol. 24. Psychology of Learning and Motivation. Academic Press, 1989, pp. 109–165.

[146] R. M. French. "Catastrophic forgetting in connectionist networks". In: *Trends Cogn Sci* 3.4 (Apr. 1999), pp. 128–135.

[147] D. Kumaran, D. Hassabis, and J. L. McClelland. "What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated". In: *Trends Cogn Sci* 20.7 (July 2016), pp. 512–534.

[148] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan. "Measuring Catastrophic Forgetting in Neural Networks". In: *AAAI Conference on Artificial Intelligence*. 2018.

[149] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. "Continual lifelong learning with neural networks: A review". In: *Neural Netw* 113 (May 2019), pp. 54–71.

[150] I. R. Fiete and H. S. Seung. "Gradient Learning in Spiking Neural Networks by Dynamic Perturbation of Conductances". In: *Phys. Rev. Lett.* 97 (4 July 2006), p. 048104.

[151] M. Schiess, R. Urbanczik, and W. Senn. "Somato-dendritic Synaptic Plasticity and Error-backpropagation in Active Dendrites". In: *PLOS Computational Biology* 12.2 (Feb. 2016), pp. 1–18.

[152] L. Kusmierz, T. Isomura, and T. Toyoizumi. "Learning with three factors: modulating Hebbian plasticity with errors". In: *Curr Opin Neurobiol* 46 (Oct. 2017), pp. 170–177.

[153] B. A. Richards and T. P. Lillicrap. "Dendritic solutions to the credit assignment problem". In: *Curr Opin Neurobiol* 54 (Feb. 2019), pp. 28–36.

[154] J. Sacramento, R. Ponte Costa, Y. Bengio, and W. Senn. "Dendritic cortical microcircuits approximate the backpropagation algorithm". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018, pp. 8735–8746.

[155] J. Kornfeld, M. Januszewski, P. Schubert, V. Jain, W. Denk, and M. Fee. "An anatomical substrate of credit assignment in reinforcement learning". In: *bioRxiv* (2020). eprint: https://www.biorxiv.org/content/early/2020/02/19/2020.02.18.954354.full.pdf.

[156] Y. H. Liu, S. Smith, S. Mihalas, E. Shea-Brown, and U. Sümbül. "A solution to temporal credit assignment using cell-type-specific modulatory signals". In: *bioRxiv* (2020). eprint: https://www.biorxiv.org/content/early/2020/11/23/2020.11.22.393504.full.pdf.

[157] R. C. O'Reilly. "Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm". In: *Neural Computation* 8.5 (1996), pp. 895–938.

[158] P. R. Roelfsema and A. van Ooyen. "Attention-gated reinforcement learning of internal representations for classification". In: *Neural Comput* 17.10 (Oct. 2005), pp. 2176–2214.

[159] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. "Random synaptic feedback weights support error backpropagation for deep learning". In: *Nat Commun* 7 (Nov. 2016), p. 13276.

[160] P. R. Roelfsema and A. Holtmaat. "Control of synaptic plasticity in deep cortical networks". In: *Nat Rev Neurosci* 19.3 (Feb. 2018), pp. 166–180.

[161] P. V. Gejman, A. R. Sanders, and J. Duan. "The role of genetics in the etiology of schizophrenia". In: *Psychiatr Clin North Am* 33.1 (Mar. 2010), pp. 35–66.

[162] D. Levenstein et al. "On the Role of Theory and Modeling in Neuroscience". In: *J Neurosci* 43.7 (Feb. 2023), pp. 1074–1088.

[163] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2001.

[164] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. "Context-dependent computation by recurrent dynamics in prefrontal cortex". In: *Nature* 503.7474 (Nov. 2013), pp. 78–84.

[165] E. Majani, R. Erlanson, and Y. Abu-Mostafa. “On the K-Winners-Take-All Network”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 1. Morgan-Kaufmann, 1988.

[166] B. A. Wang, M. Veismann, A. Banerjee, and B. Pleger. “Human orbitofrontal cortex signals decision outcomes to sensory cortex during behavioral adaptations”. In: *Nature Communications* 14.1 (June 2023).

[167] M. Sarafyazd and M. Jazayeri. “Hierarchical reasoning by neural circuits in the frontal cortex”. In: *Science* 364.6441 (May 2019).

Chapter 3

Neural architectures for flexible hierarchical planning

3.1 Introduction

Goal-directed behavior in spatial and abstract domains relies on hierarchical planning: long sequences are organized into intermediate objectives that define progress, focus attention, and structure control. Daily routines make this vivid. Traveling to the airport proceeds through subgoals—reach the subway, make the transfer, clear security, arrive at the gate. A laboratory immunostaining workflow advances through modules—fixation, blocking, primary antibody, secondary antibody, imaging. This organization supports both navigation and multi-step decision making by elevating planning from moment-to-moment actions to temporally abstract units. In line with classic ideas about chunking and the organization of action [1–3], contemporary behavioral and neural evidence shows that people and animals exploit hierarchical structure. Human fMRI reveals state abstractions and subgoal-related prediction errors during multistep tasks, and computational work formalizes how hierarchical planning accelerates search and generalization across tasks [4–15].

Hierarchical model-based reinforcement learning (H-MBRL) provides a formal scaffold for this organization by representing an internal model as an abstract graph of macro-states (subgoals) connected by temporally extended actions (options) [16–18]. A high-level planner selects a short sequence of options to reach the goal, while low-level controllers execute within each option. Learning progressively updates the graph topology, refining options, and the values of subgoals, enabling efficient search, reuse of subroutines, and rapid transfer. This algorithmic decomposition also yields clear neurobiological predictions: if planning operates over subgoals and options, distinct circuit elements should encode macro-states, subgoal values, and options, and their interaction should reorganize over learning. Converging evidence aligns with this view. Hippocampus is necessary for model-based planning and generates prospective sequences via replay that can support internal simulation and consolidation [19–21]; orbitofrontal cortex (OFC) carries a cognitive map of task states and subgoal values [10, 22]; anterior cingulate cortex (ACC) predicts future states and contributes to model-based action selection [23]. Basal ganglia provide reinforcement-learning signals.

However, many circuit models of navigation/planning are non-hierarchical and rely on

"follow-the-bump" value ramp dynamics from the goal; when the reward is sparse, over long horizons, attenuation and neural noise flatten distal gradients, yielding unreliable choices [24–33]. This gap between H-MBRL's hierarchical computations and their circuit implementation motivates our approach.

In this chapter, we propose a multi-area circuit that both discovers subgoals and plans hierarchically over them. Because hippocampal replay compresses transition structure and trajectories into neural timescales, we introduce a normative objective that maximizes the sum of within-cluster principal eigenvalues estimated from replay. We derive a plasticity rule that optimizes this objective and show in simulations that it learns topology-, goal-, and context-aware hierarchies across diverse environments. An orbitofrontal–anterior cingulate (OFC–ACC) circuit then performs hierarchical planning on the learned hierarchy: OFC computes subgoal values via hierarchical successor representations, while ACC performs prospective, path-level winner-take-all to select option sequences maximizing accumulated reward. Critically, the selected high-level option injects a subgoal value signal into lower-level OFC populations, establishing proximal waypoints and mitigating gradient attenuation in flat models. In the next chapter, we build on these discrete-state results and extend them to navigation over continuous manifolds.

Together, these results provide a normative, biologically grounded account of how hippocampal replay drives hierarchical subgoal discovery and how OFC–ACC interactions implement hierarchical planning over learned subgoals, offering a plausible neural realization of H-MBRL.

3.2 Results

We model the environment as a directed graph $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$, where $\mathcal{S}^0$ is the set of vertices and $\mathcal{E}^0 \subseteq \mathcal{S}^0 \times \mathcal{S}^0$ is the set of directed edges. For $v \in \mathcal{S}^0$, define the outgoing neighborhood

$$\mathcal{N}(v) \;=\; \{\, w \in \mathcal{S}^0 : (v, w) \in \mathcal{E}^0 \,\}.$$

To define action dynamics, let $\mathcal{A}$ be a finite action set and let $l : \mathcal{E}^0 \to \mathcal{A}$ be a labeling function such that, for all $v \in \mathcal{S}^0$, the restriction $l|_{\{(v,w)\in\mathcal{E}^0\}}$ is injective. For $a \in \mathcal{A}$ and $v \in \mathcal{S}^0$, define $f_a(v) = w$ if $(v, w) \in \mathcal{E}^0$ and $l(v, w) = a$.

Given a start vertex v_0 and a goal vertex g, the optimal navigation problem is to find a sequence $a_1, \ldots, a_T$ minimizing T such that

$$f_{a_T} \circ \cdots \circ f_{a_1}(v_0) \;=\; g.$$

3.2.1 Why flat value-ramp navigation fails

A common strategy is to construct a bump-like value signal around the goal (e.g., via a successor representation) that decays with distance [24–33],

$$V_g(v) \;=\; \gamma^{\mathrm{dist}(v,g)}, \tag{3.2.1}$$

and choose greedily among neighbors,

$$a \;=\; \arg\max_{a\in\mathcal{A}} V_g\big(f_a(v)\big). \tag{3.2.2}$$

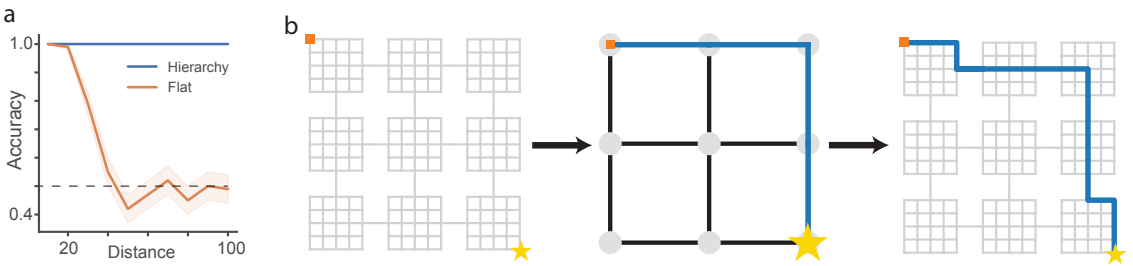

Figure 3.1: **Flat versus hierarchical navigation. a** Accuracy on a 1D line as a function of goal distance (mean ± s.e.m.; $n = 100$ trials). The dashed line denotes chance (0.5). Hierarchical planning maintains near-perfect accuracy across distances, whereas flat value-ramp navigation degrades with distance. **b** Schematic of hierarchical planning in a 3×3 "rooms" maze. The agent first plans over a coarse graph of rooms to obtain a high-level route (middle), then refines the plan at the fine scale to execute the path (right). Hierarchy shortens the effective planning horizon by replacing the distant goal with proximal subgoals.

While standard in reinforcement learning, this scheme faces a neural implementation drawback. Let $d = \mathrm{dist}(v, g)$. The value gap between a geodesic step and the best off-geodesic competitor satisfies

$$\Delta \;\le\; \gamma^{d-1}(1-\gamma), \tag{3.2.3}$$

so both the decision SNR (Δ/σ under channel noise σ) and the one-step TD error decay exponentially with distance, degrading decision reliability for distant goals. Introducing hierarchy reduces the effective horizon to the next subgoal, restoring decision SNR and TD magnitudes and enabling robust navigation (Figure 3.1a).

3.2.2 Hierarchical navigation

We formalize hierarchical navigation by recursively clustering the state space and planning over the induced quotient graphs. Let $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$ be a directed graph. For each level $l = 1, \dots, L$, recursively partition the level-$(l-1)$ states $\mathcal{S}^{l-1}$ into clusters $\mathcal{C}^{l-1} = \{\mathcal{C}_1^{l-1}, \dots, \mathcal{C}_{k_l}^{l-1}\}$ and these clusters define the vertex set $\mathcal{S}^l = \{1, \dots, k_l\}$ of the level-l quotient graph $G^l = (\mathcal{S}^l, \mathcal{E}^l)$. We place an edge $(i, j) \in \mathcal{E}^l$ whenever a boundary edge connects some $u \in \mathcal{C}_i^{l-1}$ to some $v \in \mathcal{C}_j^{l-1}$ in G^{l-1}. We define the cluster map $[\cdot]_l : \mathcal{S}^{l-1} \to \mathcal{S}^l$ by $[u]_l = i$ iff $u \in \mathcal{C}_i^{l-1}$, with inverse membership $i_l : \mathcal{S}^l \to 2^{\mathcal{S}^{l-1}}$, $i_l(i) = \mathcal{C}_i^{l-1}$. The composition $\pi_l = [\cdot]_l \circ \cdots \circ [\cdot]_1$ maps base states to their level-l representatives, so a base goal $g^0 \in \mathcal{S}^0$ lifts as $g^l = \pi_l(g^0)$.

Intuitively, the hierarchy groups nearby places into a "room," rooms into a "building," and buildings into a "neighborhood". Planning proceeds at the coarsest level that still advances toward the goal and then refines that choice down the hierarchy, so the agent always aims for a proximal waypoint rather than the distant target itself.

This construction mitigates distance-dependent value flattening. If each cluster and the top-level graph have diameter

$$O\Big(\tfrac{\log\left((1-\gamma)/\varepsilon\right)}{\log(1/\gamma)}\Big),$$

with ε the effective neural-noise scale, then within each level the value gap between a geodesic

step and the best off-geodesic competitor is $\Omega(\varepsilon)$, preserving decision signal-to-noise and stabilizing one-step temporal-difference (TD) signals at subgoal horizons. Intuitively, by shortening the horizon to the next "room," the value advantage of the correct move stays above the noise floor.

We implement a greedy hierarchical policy. Let v^0 be the current base state and write $\mathcal{N}_l(\cdot)$ for the outgoing neighborhood in G^l. At level l, define the subgoal-conditioned value

$$V_u^l(v) \;=\; \gamma^{\mathrm{dist}_{G^l}(v,u)},$$

the discounted reachability on G^l. At the top level,

$$s^L \;=\; \arg \max_{w \in \mathcal{N}_L\left(\pi_L(v^0)\right)} V_{g^L}^L(w),\; u^{L-1} = i_L(s^L) \tag{3.2.4}$$

which selects the neighboring region that most increases discounted reachability of the lifted goal. The choice is then refined down the hierarchy: for $l = L-1, \ldots, 0$,

$$s^l \;=\; \arg \max_{w \in \mathcal{N}_l\left(\pi_l(v^0)\right)} V_{u^l}^l(w),\; u^{l-1} = i_l(s^l) \tag{3.2.5}$$

so each level picks the next region that best approaches the previously chosen waypoint. Finally, select a primitive action $a \in \mathcal{A}$ with $f_a(v^0) = s^0$ and execute it. In effect, the agent chooses "neighborhood," then "building," then "room," and only then the specific "step," keeping the value advantage robust to noise while also reducing planning cost (Figure 3.1b).

In the remainder of the chapter, we detail how a multi-area circuit implements this scheme. Our hippocampal model recursively clusters states to learn a hierarchical world model via hippocampal replay; orbitofrontal cortex (OFC) model learns the subgoal-conditioned values $V_{u^l}^l$; and anterior cingulate cortex (ACC) model plans prospectively and selects subgoals u^l that maximize these values.

3.2.3 Hippocampal replay–driven hierarchical learning

The hierarchical planning scheme in the previous section specifies how an available hierarchy can be exploited to shorten the effective planning horizon. What it does not yet explain is where such a hierarchy comes from, nor how animals acquire hierarchies that appear well aligned with task structure. Not all hierarchies are created equal. From past theoretical work, good hierarchies accelerate learning and support transfer to new tasks [1, 6, 7]. From animal and human experimental work, hierarchies must also reflect environmental geometry, capturing bottlenecks and community-like organization [6, 7, 34]; they must be modulated by task demands, with higher resolution around task-relevant or goal-proximal regions [7, 35, 36]; and they can be organized around salient shared cues, such as route color in a metro map [5]. The question, therefore, is how a neural circuit could learn a hierarchy that simultaneously encodes geometry, task-dependent emphasis, and common cue-based grouping.

The hippocampus is a brain region that helps form memories for places and events; during rest and sleep, it rapidly "replays" short sequences of recent events. We argue that hippocampal replay supplies exactly the statistics needed for hierarchical learning. By compressing experience into neural timescales, replay presents transition structure at

a timescale appropriate to engage synaptic plasticity [37, 38]. Replay is also shaped by current goals and reward history, causing trajectories to oversample task-relevant parts of the environment [39, 40], consistent with goal-related over-representations in hippocampus and entorhinal cortex [35, 36]. Thus, replay provides both a geometric map of which states connect and a task-weighted emphasis on the regions that matter most.

Formally, we model the hippocampus as a generative process over short trajectories (pathlets):

$$s = (s_1, \cdots, s_n) \sim \mathcal{D}_{\text{hipp}}, \tag{3.2.6}$$

and we feed these pathlets to the hierarchical learning circuit through an exponentially weighted filter

$$\phi(s) = \sum_{i=1}^{n} \gamma^{n-i} e_{s_i}. \tag{3.2.7}$$

Here, γ is the discount factor and e_{s_i} is the one-hot encoding of the state s_i. To see why these pathlets capture environmental geometry, notice the following. Conditioning on the endpoint $s_n = v$, the expected feature

$$\mathbb{E}[\phi(s) \mid s_n = v] = \psi(v) \tag{3.2.8}$$

acts as a discounted random-walk kernel centered at v. In the limit $n \to \infty$ when $\mathcal{D}_{\text{hipp}}$ is uniform random walk, this kernel satisfies

$$\psi(v)^\top \psi(u) = \gamma^{\text{dist}(u,v)} N_{u,v}(1 + O(\gamma)), \tag{3.2.9}$$

where $\text{dist}(u, v)$ is the graph distance and $N_{u,v}$ is the number of distinct shortest paths between u and v. Replay-derived similarities therefore decay with distance and increase with the multiplicity of short routes, capturing both geometry and connectivity of the environment. By modulating the $\mathcal{D}_{\text{hipp}}$ to be goal-directed, such kernel encodes both environmental geometry and task-relevant information.

A natural way to obtain clusters from such features is k-means clustering [41]. However, k-means has two limitations in this setting. First, it prefers roughly isotropic clusters and so ignores elongated structure with coherent direction; this makes it poorly suited when we want to group potentially far-away states along external attributes such as the same metro line grouping observed in Balaguer et. al. [5]. Second, it is not obvious how to implement such clustering in a biologically plausible neural circuit.

To overcome these issues, we maximize the sum of within-cluster principal eigenvalues rather than the within-cluster similarity. Intuitively, we look for clusters whose replay-induced covariance is as close to rank 1 as possible: clusters in which pathlets all share a dominant direction. Concretely, we optimize

$$\max_{\{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^{K} \lambda_1\Big(\mathbb{E}[xx^\top \mathbf{1}_{x \in \mathcal{C}_k}]\Big), \tag{3.2.10}$$

where x denotes replay features (e.g. $x = \phi(s)$), $\mathbf{1}_{x \in \mathcal{C}_k}$ selects cluster k, and $\lambda_1(\cdot)$ is the top eigenvalue. Under this objective, a k-means community that actually contains two directions

(for example, two intersecting metro routes) will be split, whereas a set of states that all lie on the same colored line but span a long distance will still be grouped together. This makes the learned hierarchy not only sensitive to geometry and task-driven replay biases but also to common salient cues, and produces clusters that are easier to interpret in behavioral terms.

In the Methods section, we show that this eigenvalue-maximization problem can be reduced to a dual optimization problem that is implementable by a neural circuit.

Theorem 3.4.9. *The clustering obtained by*

$$\arg \max_{\{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^K \lambda_1\big(\mathbb{E}[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}]\big)$$

is equivalent to

$$\arg \min_{y,\,\|y\|_0=1,\,W} \max_B \mathbb{E}\big[-\operatorname{Tr}(x^\top W y)\big] + \frac{1}{2}\|W\|_F^2 + \mathbb{E}\Big[\sum_{i=1}^K b_i(y_i^2-1)\Big],$$

where $y \in \mathbb{R}^K$ is a one-sparse code whose non-zero index indicates the cluster assignment, W is a weight matrix, and $B = \operatorname{diag}(b_1, \ldots, b_K)$ enforces unit-norm constraints.

This dual form leads directly to a local circuit, obtained by mirror descent on the space of one-sparse codes:

$$\tau_n \frac{dr}{dt} = Wx - Br, \qquad r^* = B^{-1}Wx, \qquad y = \mathrm{WTA}(r^*), \tag{3.2.11}$$

with Hebbian and homeostatic plasticity

$$\tau_w \frac{dW}{dt} = yx^\top - W, \qquad \tau_w \frac{db_i}{dt} = y_i^2 - 1. \tag{3.2.12}$$

The one-sparse code y implements clustering: for each input x, a single unit wins, $i^\star = \arg\max_i r_i^*$, and its index is the assigned cluster. Hebbian plasticity in W then aligns the winning unit with the dominant replay pattern of its assigned inputs, yielding the principal direction within that cluster. The resulting clusters define the vertices of a quotient graph. By stacking such modules, and feeding cluster codes from one level as inputs to the next, we obtain a replay-driven, hierarchical learning circuit that constructs a hierarchical graph $G^1, \ldots, G^L$ from an environmental graph $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$.

To verify that the circuit indeed captures environmental geometry, task demands and common salient cues, we simulated it on these benchmark environments. On the Schapiro et al. graph [34], it correctly detected the three clusters and learned clean cluster-specific receptive fields (Figure 3.2a). On the four-room environment [6], it recovered the macro room structure (Figure 3.2b). On the Tower of Hanoi, whose game tree is explicitly hierarchical, the model captured multi-scale triangular structure (Figure 3.2c). On the metro map [5], our variance-aware, eigenvalue-based clustering recovered the ground-truth routes, while k-means did not (Figure 3.2d). Finally, in a grid world with either random-walk replay or goal-directed replay toward the bottom-right corner, the learned clusters showed much higher resolution around the goal only in the goal-directed case, matching hippocampal over-representation near rewarded locations [35, 36] (Figure 3.2e).

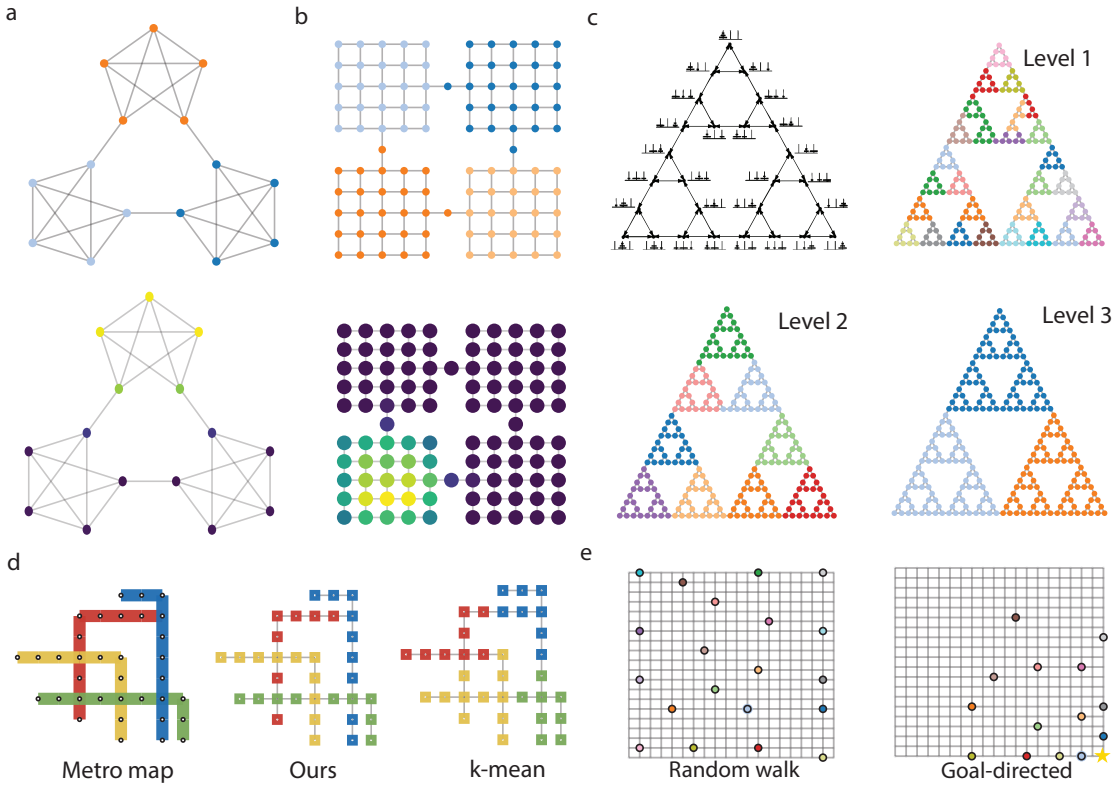

Figure 3.2: **Replay-driven hierarchical learning captures geometry, common cue and task-dependent clusters.** We applied the circuit to environments used in the literature. **a** Schapiro et al. graph [34]: the circuit recovers the three latent communities (top) and cluster units develop receptive fields selective for each community (bottom). **b** Four-room graph from Solway et al. [6]: the circuit learns room-level structure. **c** Tower-of-Hanoi game tree [6]: stacking the circuit recovers the multi-scale triangular hierarchy. **d** Metro map from Balaguer et al. [5]: augmenting states with route-color features allows the circuit to group states by line, whereas k-means fails. **e** Grid world under two replay regimes: random-walk replay produces nearly uniform tiling, whereas goal-directed replay toward the bottom-right corner yields higher cluster density around the goal, consistent with goal-related over-representation in CA1 [35, 36].

3.2.4 OFC-ACC interaction implements hierarchical planning

Once the hierarchy has been learned from hippocampal replay, we can apply the hierarchical navigation policy in Equation 3.2.5 directly on the inferred multiscale graph. What remains is to identify cortical circuitry capable of instantiating the two key ingredients of this policy: (i) representing subgoal-conditioned values at each level of the hierarchy and (ii) prospectively plan trajectories with the highest subgoal-conditioned values.

We propose that the OFC–ACC circuit is a natural candidate. On the OFC side, human fMRI work has reported subgoal-related value signals in OFC [10], suggesting that OFC is well positioned to represent value functions that are conditional on an intermediate subgoal. This aligns with our requirement that, for each level l, the circuit maintains a hierarchical

subgoal-conditioned value function $V^l_{u^l}$ for subgoal u^l. On the ACC/mPFC side, recordings have revealed prospective sequences during model-based or sequential choice, consistent with a role in planning multiple steps ahead rather than simply encoding the current action [23, 42]. Thus ACC is a plausible substrate for the prospective, path-level planning that selects among competing subgoal-directed trajectories.

Combining these two computations yields a circuit-level implementation of hierarchical planning. Given a hierarchy $G^1, \dots, G^L = (\mathcal{S}^1, \mathcal{E}^1), \dots, (\mathcal{S}^L, \mathcal{E}^L)$ learned from replay over the base graph $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$, and given a current level-l subgoal $u^l \in i_{l+1}(\mathcal{S}^{l+1})$, OFC supplies the level-l subgoal-conditioned value function $V^l_{u^l}$. ACC then performs a finite-horizon prospective planning, choosing a path $p = (x_1, \dots, x_H)$ on G^l that maximizes

$$\sum_{\delta=1}^{H-1} R^l_\delta(x_\delta),$$

where $R^l_\delta(x) = \mathbf{1}[x \in u^l]$ for $1 \le \delta \le H-1$ and $R^l_H(x) = V^l_{u^l}(x)$. In words, the ACC planner selects the trajectory with the highest OFC's subgoal-conditioned value. The selected trajectory supplies the next lower-level subgoal as

$$u^{l-1} = i_l(x_2),$$

i.e. the next state along the chosen level-l path becomes the target for the level-$(l-1)$ OFC, which in turn provides values to the level-$(l-1)$ ACC planner. Recursively applying this OFC-ACC interaction realizes the full top-down hierarchical plan. In simulations on a nine-room maze, this circuit first identifies a room-level route and then successively instantiates intermediate subgoals to complete navigation (Figure 3.3).

In the following sections, we detail how the ACC circuit can implement the prospective planning over subgoal-conditioned OFC's value function, and how the OFC circuit can compute and update the family of value functions $\{V^l_u\}$ needed to support this process.

3.2.5 ACC computation of prospective planning via path-selection attractors

To implement the prospective, path-level component of hierarchical planning, we propose that ACC realizes a path-selection attractor that settles onto the multi-step trajectory most compatible with the current subgoal. The key idea is that for each planning depth, ACC represents a distribution over states that is jointly constrained by (i) the subgoal-conditioned value signal supplied by OFC and (ii) the requirement that adjacent planning layers form a legal path on the current hierarchical graph. With these two constraints in place, the circuit effectively computes the marginals of a Gibbs distribution over path values, thereby placing probability mass on trajectories with higher subgoal-conditioned value.

For each hierarchy level l, the ACC circuit maintains, at planning layer $\delta = 1, \dots, H$, an activity vector

$$r^l_\delta \in \Delta^{N^l} = \{r \in \mathbb{R}^{N^l}_{\ge 0} : \mathbf{1}^\top r = 1\},$$

where $N^l = |\mathcal{S}^l|$ and Δ^{N^l} is the probability simplex over level-l states. We also define the log-activity $z^l_\delta = \log r^l_\delta$. From OFC, the circuit receives the hierarchical, subgoal-conditioned value

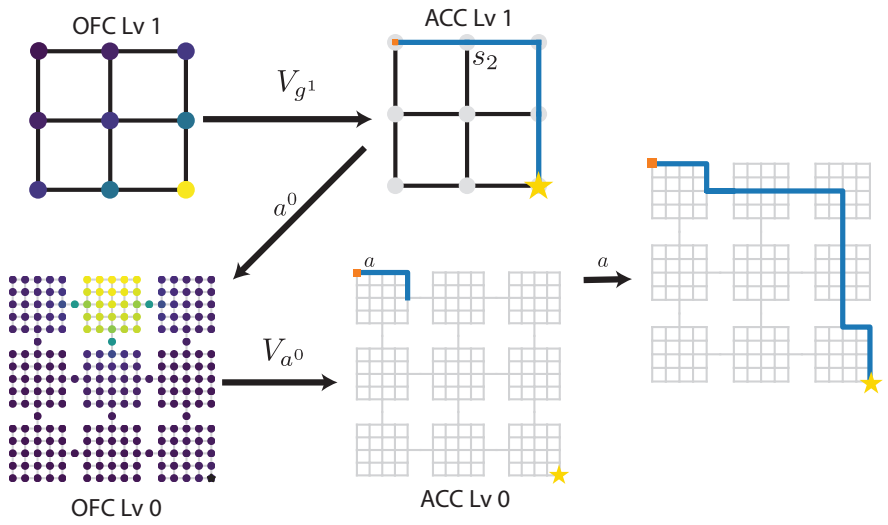

Figure 3.3: **Hierarchical navigation of a nine-rooms maze through OFC–ACC interaction.** OFC first computes subgoal-conditioned value functions at the coarsest level. ACC then selects the high-level trajectory with the highest accumulated value and passes the next waypoint as a subgoal to the lower-level OFC circuit. The lower-level OFC recomputes values around this proximal subgoal, enabling the lower-level ACC to plan at the primitive-action scale and complete the hierarchical navigation.

functions $V^1_{u^1}, \ldots, V^L_{u^L}$. At level l, the goal of the ACC circuit is to perform a finite-horizon prospective planning step, choosing a path $p = (x_1, \ldots, x_H)$ on G^l that maximizes

$$\sum_{\delta=1}^{H-1} R^l_\delta(x_\delta),$$

where $R^l_\delta(x) = \mathbf{1}[x \in u^l]$ for $1 \le \delta \le H-1$ and $R^l_H(x) = V^l_{u^l}(x)$.

We introduce a log-normalized reward,

$$\widehat{R}^l_\delta \;=\; \beta R^l_\delta - \log \sum_i \exp(\beta R^l_{\delta,i}),$$

with inverse temperature $\beta > 0$, so that reward enters additively in log-space. Let $A^l \in \{0,1\}^{N^l \times N^l}$ be the adjacency matrix of the level-l graph, with $A^l_{ij} = 1$ iff $(i,j) \in \mathcal{E}^l$. The ACC dynamics is

$$\tau_n \frac{dz^l_{\delta,i}}{dt} = -z^l_{\delta,i} + \widehat{R}^l_{\delta,i} + \log\Big(\sum_j A^l_{ji}\, r^l_{\delta-1,j}\Big) + \log\Big(\sum_k A^l_{ik}\, r^l_{\delta+1,k}\Big), \qquad r^l_\delta \propto e^{z^l_\delta}, \quad (3.2.13)$$

together with per-layer normalization $z^l_\delta \leftarrow z^l_\delta - \log \sum_i e^{z^l_{\delta,i}}$ to ensure $r^l_\delta \in \Delta^{N^l}$.

The two log terms have a direct interpretation. The "incoming" term

$$\log\Big(\sum_j A^l_{ji}\, r^l_{\delta-1,j}\Big)$$

forces the distribution at layer δ to be compatible with the distribution at layer $\delta-1$ (there must exist an edge into i from a state on which the previous layer places mass), while the "outgoing" term

$$\log\Big(\sum_k A^l_{ik}\, r^l_{\delta+1,k}\Big)$$

forces compatibility with the next layer. The reward term $\widehat{R}^l_{\delta,i}$ biases the attractor toward states preferred by OFC at that time step.

At a fixed point of Equation 3.2.13, we obtain

$$r^{l,*}_{\delta,i} = e^{\widehat{R}^l_{\delta,i}}\Big(\sum_j A^l_{ji}\, r^{l,*}_{\delta-1,j}\Big)\Big(\sum_k A^l_{ik}\, r^{l,*}_{\delta+1,k}\Big). \tag{3.2.14}$$

Let

$$\Pi^l \;=\; \Big\{\pi=(x_1,\dots,x_H)\in(\mathcal{S}^l)^H:\; A^l_{x_\delta,x_{\delta+1}}=1\;\;\forall\,\delta\in[H-1]\Big\}$$

be the set of legal level-l paths of horizon H, and define the path score

$$S^l_\beta(\pi) \;=\; \sum_{\delta=1}^{H}\beta\, R^l_\delta(x_\delta),$$

with associated Gibbs distribution

$$\mathbb{P}^l_\beta(\pi) \;=\; \frac{\exp\{S^l_\beta(\pi)\}}{Z^l_\beta}, \qquad Z^l_\beta=\sum_{\pi\in\Pi^l}\exp\{S^l_\beta(\pi)\}. \tag{3.2.15}$$

We can then prove the following in the Methods.

Theorem 3.4.18. *Consider the dynamics in Equation 3.2.13 on a level-l graph of horizon H and legal path set Π^l. Then the fixed point $r^{l,*}_\delta$ equals the layerwise node marginal of the Gibbs distribution Equation 3.2.15,*

$$r^{l,*}_{\delta,i} \;=\; \mathbb{P}^l_\beta(x_\delta=i),$$

and Equation 3.2.13 converges to this fixed point.

Thus the ACC attractor is computing, in parallel across layers, the marginals of a path distribution that is already weighted by subgoal-conditioned value. If we let

$$\pi^\star=\arg\max_{\pi\in\Pi^l}\sum_{\delta=1}^{H}R^l_\delta(x_\delta)$$

be the optimal path, then as $\beta\to\infty$ the Gibbs distribution $\mathbb{P}^l_\beta$ concentrates on $\pi^\star$, and the fixed point concentrates on the layerwise supports of $\pi^\star$:

$$r^{l,*}_\delta(\beta) \;\xrightarrow[\beta\to\infty]{}\; e_{x^\star_\delta}.$$

When the maximizer is unique, the ACC circuit therefore recovers the optimal path in a distributed form. After convergence, ACC simply reads out the second state on this path and projects the next lower-level subgoal

$$u^{l-1}=i_l(x_2),$$

to level-$(l-1)$ OFC, which in turn drives level-$(l-1)$ ACC in the next planning cycle.

3.2.6 OFC computation of hierarchical successor representations

Because subgoals can change on the fly as the ACC planner refines the trajectory, OFC needs a mechanism that can re-evaluate subgoal values rapidly without relearning the underlying transition structure. The successor representation (SR) provides exactly this flexibility: it factorizes value into (i) a predictive map of future state occupancies and (ii) a goal- or subgoal-specific reward vector.

The (flat) SR under policy π is the resolvent

$$M_\pi \;\equiv\; (I - \gamma P_\pi)^{-1} \;=\; \sum_{t=0}^{\infty} \gamma^t P_\pi^t, \tag{3.2.16}$$

whose (s, s') entry is the expected discounted future occupancy of s' when starting from s and following π. If we define a subgoal-conditioned reward

$$r_u(s) \;=\; \mathbf{1}[s \in u], \qquad u \subseteq \mathcal{S}^0,$$

then subgoal values are simply

$$V_u^\pi \;=\; M_\pi\, r_u. \tag{3.2.17}$$

Once M_π has been learned, changing the subgoal from u to u' reduces to changing r_u to $r_{u'}$ from ACC and multiplying; the predictive part M_π is reused. This makes SR a natural computation for OFC, which is hypothesized to flexibly compute subgoal values.

To extend SR to the hierarchical setting above, we adopt an options view in which decisions are made only at decision epochs, i.e. when the agent enters a new macrostate and the ACC planner selects the next subgoal for the lower level. Fix a hierarchy and let

$$(c_1, \tau_1), (c_2, \tau_2), \ldots$$

denote the sequence of macrostates visited at level l together with the durations τ_k spent executing the option that carries the agent from c_k to c_{k+1}, where $c_k \in \mathcal{S}^l$. For a fixed high-level policy π, we define the policy-induced jump kernel

$$P_\pi(i, j) \;=\; \Pr\big(c_{k+1} = j \,\big|\, c_k = i,\, \pi\big), \tag{3.2.18}$$

and a per-epoch discount

$$\Gamma_\pi(i, j) \;:=\; \mathbb{E}[\gamma^{\tau_k} \,|\, c_k = i,\, c_{k+1} = j,\, \pi]\,, \tag{3.2.19}$$

which is the expected discount factor accumulated while executing the option from i to j. In parallel with the flat case, we can define subgoal-conditioned rewards over macrostates, $r_u(i) = \mathbf{1}[i \in u]$.

By exact analogy with (3.2.16), the hierarchical successor representation is the resolvent of the discounted jump operator:

$$M_\pi \;:=\; \big(I - \Gamma_\pi \cdot P_\pi\big)^{-1}, \qquad V_u \;=\; M_\pi\, r_u. \tag{3.2.20}$$

This matrix predicts, for each macrostate, the discounted occupancy of all future macrostates under the current high-level policy, and it correctly accounts for the fact that some options

last longer than others (via Γ_π). As in the flat case, changing the subgoal at that level only requires changing r_u.

This computation admits a straightforward recurrent implementation. Consider

$$\tau_n \frac{dx}{dt} = -x + Wx + r_u, \tag{3.2.21}$$

with $x \in \mathbb{R}^{|\mathcal{S}^l|}$ and W a synaptic weight matrix. If

$$W = \Gamma_\pi \cdot P_\pi,$$

then the fixed point satisfies

$$x^* = Wx^* + r_u \quad \Longrightarrow \quad x^* = (I - \Gamma_\pi \cdot P_\pi)^{-1} r_u = M_\pi r_u = V_u,$$

so the network converges to the desired subgoal-conditioned value.

It remains to learn W. Each unit maintains an *entry-gated* eligibility trace

$$e_i \;\leftarrow\; \gamma e_i + \mathbf{1}[c_t = i],$$

which captures how long the current option has persisted in state i (and so encodes the option duration into the effective discount). A simple Hebbian-like update,

$$\Delta W_{ij} \;\propto\; e_j \, \mathbf{1}[c_{t+1} = i] - W_{ij}, \tag{3.2.22}$$

pushes W toward the empirical discounted transition statistics, i.e. toward $\Gamma_\pi(i,j)P_\pi(i,j)$. In expectation, the stationary point of this rule is exactly $W = \Gamma_\pi \cdot P_\pi$, yielding the hierarchical SR that OFC needs to supply subgoal-conditioned values to ACC in real time.

3.3 Discussion

We propose a neural account of flexible hierarchical planning that integrates three levels of description: (i) a circuit mechanism in which hippocampal replay optimizes a variance-aware normative clustering objective to build multiscale structure; (ii) an ACC module that realizes hierarchical path-selection attractors, maximizing subgoal value and emitting the next subgoal for the lower level; and (iii) an OFC module that computes hierarchical subgoal values via successor representations. Together, these components orchestrate hierarchical planning and yield a biologically plausible instantiation of hierarchical model-based reinforcement learning, bridging the gap between algorithmic theory and circuit implementation.

3.3.1 Behavioral and neural evidence of hierarchical planning

Humans and animals behave in ways that reveal nested subgoals and option-like routines rather than flat stimulus–response chains. Classic theory anticipated this by arguing that complex action is organized into hierarchically structured "chunks" [1, 2]. Contemporary behavioral studies make the case quantitatively: people infer community structure and bottlenecks that support efficient hierarchical plans [6, 7, 34], and their choices are better

explained by hierarchically structured reinforcement learning than by flat alternatives [3]. In realistic planning tasks, such as navigating a virtual subway, participants choose routes consistent with planning over a coarse, bottleneck-level map before refining details and their response time correlates with hierarchical distance rather than flat graph distance, exhibiting hallmarks of hierarchical control in behavior [5, 7]. Program-generation experiments further show that human plans are composed of reusable subroutines, indicating explicit compositional structure [8]. Parallel work in rodents echoes these principles: mice discover and exploit subgoal waypoints to stitch efficient routes through complex spaces, revealing an internal decomposition into intermediate objectives [9].

Neural data align with this behavioral picture and localize complementary roles across circuits. In multistep human tasks, Blood-Oxygen-Level-Dependent (BOLD) signals dissociate subgoal-related from goal-related prediction errors, providing a neural signature of hierarchical reinforcement learning [4]; in virtual subway navigation, frontoparietal and prefrontal activity tracks hierarchical planning progress [5]. Orbitofrontal cortex represents and evaluates subgoals during model-based behavior, linking valuation to hierarchical structure [10]. Medial frontal cortex encodes distributed representations of task progress and prospective state predictions, consistent with a controller that monitors advancement through an abstract action graph [11, 12]. Systems-level frameworks situate these computations along a rostro–caudal gradient of prefrontal abstraction, offering an anatomical scaffold for hierarchical control [13, 14, 43]. Effective connectivity analyses during hierarchical choice support directed interactions among these regions that are consistent with top–down, multilevel planning [15]. Together, these findings indicate that neural representations and inter-area dynamics are organized to support hierarchical planning rather than flat action selection.

3.3.2 Biological plausibility and implication of our hierarchical planning circuit

Our hierarchical planning circuit incorporates distinct circuit mechanisms and neural representations in a multi-area neural circuit consists of hippocampus, ACC and OFC. Below, I will detail the biological plausibility and implication of our models.

Hippocampal replay compresses behavior into neural timescales that are well matched to synaptic plasticity, offering a natural teaching signal for structure learning. During rest and sleep, hippocampal ensembles rapidly "replay" recent trajectories at a rate far faster than experienced time, bringing multi-step transition structure into a temporal window that can effectively drive plasticity [37, 38]. Replay content is not neutral: it is shaped by goals and recent reward history, biasing samples toward task-relevant routes and thereby emphasizing precisely those transitions that should be consolidated for future planning [39, 40]. On the other hand, hippocampus place codes also exhibit multiscale structure that aligns with task geometry and relevance. CA1 place cells have small place fields that grow progressively larger toward ventral hippocampus [44] while they can also over-represent goal-proximal regions and salient waypoints, effectively increasing spatial resolution where decision demands are high [35, 36]. In our framework, this compressed, goal-biased sampling supplies the statistics needed to learn task-relevant hierarchical structures shown in the hippocampal place cells. Specifically, replay-driven features induce a similarity kernel that decays with graph distance

and scales with the multiplicity of short paths, encouraging clustering along corridors, rooms, and routes that matter for efficient navigation and subgoal discovery.

Prospective planning signals in medial frontal cortex support the role we assign to ACC/mPFC as a path-selection attractor. During sequential, model-based choice, rodent and primate ACC/mPFC exhibit forward-looking activity patterns that preview upcoming states or action sequences, rather than merely reflecting current motor output [23, 42]. Complementary human work implicates midcingulate cortex in tracking progress through multistep tasks and representing control-relevant variables at an abstract level, consistent with a circuit poised to evaluate partial plans and enforce path consistency across steps [11, 12]. These data align with our proposal that ACC computes layerwise marginals over legal trajectories biased by subgoal-conditioned value, settling onto the multi-step plan that best advances the current waypoint.

Orbitofrontal cortex (OFC) provides a natural substrate for subgoal values. In a virtual subway task that requires planning over a hierarchical network, vmPFC/OFC encodes path value in both lower-level goal distance and higher-level number of switching routes, consistent with hierarchical planning [5]. More recently, human fMRI has revealed explicit subgoal representations and valuation in OFC during a multi-step task that necessitates discovering and pursuing intermediate goals [10]. Because subgoals change rapidly, either upon attainment or following replanning, the brain must re-evaluate their value without rebuilding the entire model. The successor representation (SR) affords exactly this flexibility by factorizing value into a predictive occupancy map and a reward vector, so that subgoal values can be recomputed by updating only the reward on the subgoal set [45, 46]. Extending SR to semi-Markov, option-level transitions yields a hierarchical SR over macrostates, which is precisely the computation we ascribe to OFC for supplying hierarchical subgoal-conditioned values to ACC's prospective search [16].

The model yields several testable predictions spanning hippocampus, ACC/mPFC, and OFC. First, hierarchical structure should be experience-dependent: during initial exploration, replay statistics should progressively sharpen cluster boundaries and increase representational resolution near task-relevant regions, with CA1/CA3 ensembles reorganizing from broadly overlapping fields into community-like structure around bottlenecks; behavior should shift from distance-sensitive, noisy "flat" choices to robust, subgoal-based routes with shorter planning latencies and fewer dead-ends. Disrupting NMDA-receptor–dependent plasticity during exploration should prevent this reorganization: replay remains diffuse, cluster selectivity fails to emerge, and animals revert to flat value-ramp behavior that works at short distances but breaks down over long horizons and detours.

Second, ACC/mPFC should be necessary for using subgoals to prospectively plan. Transient silencing during planning abolishes multi-step prospective sequences and reduces inter-layer path consistency; downstream, OFC loses subgoal-conditioned value signals while retaining coarse, flat goal-distance signals, yielding behavior characteristic of flat-value navigation—impaired over long distances yet largely preserved for short-range routes.

Third, OFC should be necessary for extending planning horizons via hierarchical successor representations. Transient inactivation during choice would leave ACC's prospective sequences intact but deprive the system of the bootstrapped subgoal values needed to plan beyond a fixed depth; behavior should therefore exhibit a hard cap on effective planning horizon: animals can initiate a plan but fail to carry it past the planning depth. Neurally, OFC units

should no longer express subgoal-conditioned value or SR-like predictive occupancy, whereas ACC activity should still reflect path marginals that respect legality constraints but need not converge to the optimal sequence. Critically, performance should recover selectively when external subgoal cues or instructed waypoints are close enough to substitute for the missing OFC values, rescuing long-horizon navigation without reinstating OFC computations.

3.3.3 Related works on subgoal discovery circuits

A rich body of work in computer science has developed algorithms for subgoal discovery, spanning structural/graph-theoretic methods that pick bottlenecks or bridges in a state-transition graph [47, 48], clustering/representation methods that use spectral (e.q. successor-representation, graph laplacian) or learned representation to define abstract regions and their centroids as subgoals [49–51], trajectory- or demonstration-based segmentation that turns recurring milestones into options/skills [52–54], and information-theoretic/end-to-end discovery that optimizes for diverse, predictive, or intrinsically valuable intermediate states [55–57]. Interestingly, these algorithmic strategies resonate with emerging evidence from systems neuroscience: the hippocampus has been modeled as a predictive map that encodes successor representations, in which bottleneck states naturally emerge in the eigenvectors—providing a circuit-plausible substrate for structural and representation-based approaches [46]. Rodent behavior likewise shows that animals spontaneously identify and exploit structural bottlenecks (e.g., obstacle edges) as effective subgoals during escape [9, 58]. Human neuroimaging further reveals explicit subgoal (and option) representations and valuation in frontal cortex during hierarchical behavior [4, 10]. Our work builds on these convergences by proposing a neural circuit with plasticity mechanisms that learns hierarchical subgoals from hippocampal-replay trajectories, capturing environmental geometry in the spirit of structural methods while connecting to a normative clustering objective for optimization-driven abstraction. Specifically, our normative objective differs fundamentally from k-means: by maximizing the sum of within-cluster principal eigenvalues, it preferentially captures clusters with a dominant directional mode, yielding semantically meaningful groupings (e.g., distinct metro lines) that k-means typically blurs.

3.3.4 Related works on hierarchical planning circuits

Prior circuit models of navigation planning rely on spreading activation, wavefront propagation, or bump-attractor dynamics in successor-like coordinates [24–33]. Although effective, these approaches are largely non-hierarchical and require goal signals to propagate over long distances; under attenuation and neural noise, distal goals induce shallow gradients and unreliable action selection. We address this limitation with a hierarchical, goal-conditioned, model-based RL framework in which a high-level controller selects abstract subgoals and a lower-level controller realizes them with primitive actions [59–61]. In our architecture, the high-level ACC population receives subgoal-conditioned values from OFC and selects the next waypoint for lower-level OFC–ACC populations, continually redirecting control toward proximal targets rather than distant goals and thereby improving decision signal-to-noise. Functionally, the scheme factorizes hierarchical planning into complementary components: OFC learns model-free subgoal-value functions via successor representation,

while ACC performs model-based prospective search as a finite-horizon rollout of depth h on the hierarchical graph, using OFC's subgoal values as terminal costs. Algorithmically, this brief look-ahead directly tightens value targets: because the Bellman operator is a γ-contraction, planning h steps ahead cuts the residual value error by roughly $O(\gamma^h)$, making choices more reliable early in learning [62, 63].

3.3.5 Possible extensions and future work

We focused on deterministic navigation, but the framework naturally extends to stochastic Markov decision processes. Environmental uncertainty and sampling dynamics can be represented with distributional population codes, with Winner-Take-All (WTA) readout selecting discrete outcomes [64, 65]. A parallel extension targets cognitive domains in which latent variables—such as context—must be inferred online. Hippocampus has been proposed to derive positional-like codes from latent sequence learning [66, 67]; applying our replay-driven hierarchical learner to the inferred latent-state graph recasts cognitive decision making as navigation on a "cognitive map" of abstract states. Because individual replay bouts are typically confined within a context, higher-level clusters should acquire context selectivity, yielding latent context abstractions.

A second direction is to modularize state representations around objects and their attributes. Real-world tasks are naturally object-centric: actions alter object properties, and structure factorizes across entities. Object-oriented MDPs (OO-MDPs) formalize this view by representing states as sets of typed objects with attributes and by specifying action schemas that update object fields [68]. Embedding OO-MDP structure into our circuit could compress representations, improve sample efficiency, and enhance transfer to novel compositions.

Finally, many sequential decisions unfold in continuous spaces. While our present instantiation is discrete, the same principles, replay-driven abstraction, subgoal-conditioned valuation, and prospective path selection, can be translated to continuous domains. In the next chapter, we outline an approach that adapts these components to continuous state and action spaces.

3.4 Methods

3.4.1 Hippocampal replay model

Given an environmental graph $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$, we model hippocampus as a generative process over short trajectories:

$$s = (s_1, \cdots, s_n) \sim \mathcal{D}_{\text{hipp}}, \tag{3.4.1}$$

and we feed these pathlets to the hierarchical learning circuit through an exponentially weighted filter

$$\phi(s) = \sum_{i=1}^{n} \gamma^{n-i} e_{s_i}. \tag{3.4.2}$$

In our simulation $n = 10$ and $\gamma = 0.9$. For the uniform condition, we generate random walk uniformly in random. For the goal-directed condition, at each time step, with 30% of chance

the trajectory produces a random walk and 70% of chance choose a neighbor that follows the shortest path toward the goal. For replay in a hierarchy, each level is generated with the same parameter setting. Replay data can also be augmented with a feature. Let $f(s)$ to be the feature associate with the state s. Then given a replay $(s_1, \cdots, s_n)$, the feature-augmented replay input is defined as the exponentially weighted filter over the concatenation of the one-hot state encoding and the augmented features

$$\phi(s) = \sum_{i=1}^{n} \gamma^{n-i} [e_{s_i}, f(s_i)]. \tag{3.4.3}$$

Given an environmental graph $G^0 = (\mathcal{S}^0, \mathcal{E}^0)$, we model hippocampus as a generator of short "pathlets"

$$s = (s_1, \ldots, s_n) \sim \mathcal{D}_{\text{hipp}}, \tag{3.4.4}$$

which are converted into inputs for the hierarchy learner via an exponentially weighted trace

$$\phi(s) \;=\; \sum_{i=1}^{n} \gamma^{n-i}\, e_{s_i}, \tag{3.4.5}$$

where e_{s_i} is the one-hot basis vector for state s_i. In all simulation, we use $n = 10$ and $\gamma = 0.9$.

We consider two replay regimes. In the uniform condition, s is an unbiased random walk on G^0. In the goal-directed condition, the walk follows a mixture policy: at each step, with probability 0.3 it takes an unbiased random step; with probability 0.7 it chooses uniformly among neighbors that lie on a shortest path to the goal (ties broken uniformly). For hierarchical replay, each level is generated with the same parameters on its quotient graph.

Replay can also carry auxiliary features. Let $f(s)$ denote a feature vector associated with state s. Given a pathlet $(s_1, \ldots, s_n)$, the feature-augmented input concatenates state and feature channels before discounting:

$$\phi(s) \;=\; \sum_{i=1}^{n} \gamma^{n-i}\, [\, e_{s_i},\; f(s_i)\,], \tag{3.4.6}$$

which preserves discounted occupancy structure while exposing salient features (e.g., route color) to the clustering circuit.

3.4.2 Diverse environments for hierarchical learning

We evaluated the replay-driven hierarchy learner across canonical benchmarks. For Figure 3.2a, we used the community-structured graph from [34]. For Figure 3.2b, we used the four-rooms environment from [6]. For Figure 3.2c, we used the Tower-of-Hanoi game tree with three poles and five disks following [69]. For Figure 3.2d, we used the metro graph from the virtual subway task in [5], augmenting replay inputs with route-color features. For Figure 3.2e, we used a 20×20 grid world. For Figure 3.3, we generalized the four-rooms layout to a nine-rooms maze to assess multiscale planning. Unless stated otherwise, simulations used the uniform replay condition; in the grid world we additionally tested a goal-directed replay regime, and in the metro map we employed feature-augmented replay with route colors.

3.4.3 Hippocampal hierarchical learning model

Remember the hierarchical learning circuit is defined as the follows:

$$\tau_n \frac{dr}{dt} = Wx - Br, \qquad r^* = B^{-1}Wx, \qquad y = \mathrm{WTA}(r^*), \tag{3.4.7}$$

with $B = \mathrm{diag}(b_1, \ldots, b_K)$ and Hebbian/homeostatic plasticity

$$\tau_w \frac{dW}{dt} = yx^\top - W, \qquad \tau_w \frac{db_i}{dt} = y_i^2 - 1. \tag{3.4.8}$$

In simulation, each replay feature $\phi(s)$ serves as an input x. We perform a single Euler update per replay,

$$\Delta W = \eta\,(yx^\top - W), \qquad \Delta b_i = \eta\,(y_i^2 - 1),$$

with learning rate $\eta = 10^{-3}$.

For each hierarchical level we sample 10^4 replays from and train the circuit using the above updates. Cluster assignments are then read out by the winner index $i^\star = \arg\max_i r_i^*$. The induced partition $\{\mathcal{C}_i^{l-1}\}_{i=1}^K$ defines the vertex set of the level-l quotient graph; directed edges are placed whenever any boundary edge connects two member states in the level-$(l-1)$ graph. The procedure is applied recursively: once G^l is constructed, fresh replays at level l are generated and fed to a new module to obtain G^{l+1}.

For Figure 3.2, cluster assignment uses one-hot state inputs $x = e_s$ for the Schapiro community graph, the rooms environments and the Tower of Hanoi graph. For the virtual subway task, x is the concatenation of state one-hot and the route-color feature. Prototypical responses for cluster i are visualized as the normalized weight row W_i/b_i, which approximates the principal direction learned by the unit for that cluster.

3.4.4 Variance-aware clustering via a neural circuit

In this section, we prove Theorem 3.4.9, which converts the normative objective

$$\max_{\{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^K \lambda_1\Big(\mathbb{E}\big[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}\big]\Big)$$

, maximizing the sum of within-cluster principal eigenvalues, into a dual optimization that admits a neurally implementable circuit. The proof proceeds by (i) passing to a sample-average approximation (ii) expressing the spectral objective via its variational (Rayleigh–Ritz) form, (iii) encoding cluster assignments with a one-sparse code y that enforces disjoint supports, (iv) introducing Lagrange multipliers to impose unit-norm constraints, and (v) completing the square to obtain a saddle problem in (W, B) whose first-order stationarity yields the Hebbian/homeostatic plasticity used by our circuit. This establishes the equivalence between the spectral clustering objective and the neural dual, providing the theoretical basis for the learning rule derived in the main text.

Theorem 3.4.9. *The clustering obtained by*

$$\arg\max_{\{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^K \lambda_1\big(\mathbb{E}[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}]\big)$$

is equivalent to

$$\arg\min_{y,\,\|y\|_0=1,\,W} \max_B \mathbb{E}\big[-\operatorname{Tr}(x^\top W y)\big] + \frac{1}{2}\|W\|_F^2 + \mathbb{E}\Big[\sum_{i=1}^K b_i(y_i^2-1)\Big],$$

where $y \in \mathbb{R}^K$ is a one-sparse code whose non-zero index indicates the cluster assignment, W is a weight matrix, and $B = \operatorname{diag}(b_1,\ldots,b_K)$ enforces unit-norm constraints.

Proof. Let $x \in \mathbb{R}^d$ denote replay features and let $X = [x_1,\ldots,x_T] \in \mathbb{R}^{d\times T}$ be T i.i.d. samples from the replay distribution. For a cluster $\mathcal{C}_k \subset \{1,\ldots,T\}$, the empirical second moment is

$$\frac{1}{T} X_{\mathcal{C}_k} X_{\mathcal{C}_k}^\top \xrightarrow[T\to\infty]{} \mathbb{E}[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}],$$

so maximizing the sum of within-cluster top eigenvalues

$$\max_{\{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^K \lambda_1\big(\mathbb{E}[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}]\big)$$

is equivalent, in the large-sample limit, to

$$\max_{\{\mathcal{C}_k\}} \sum_{k=1}^K \lambda_1\big(\mathbb{E}[xx^\top \mathbf{1}_{x\in\mathcal{C}_k}]\big) = \lim_{T\to\infty} \max_{\{\mathcal{C}_k\}} \sum_{k=1}^K \frac{1}{T}\lambda_1\big(X_{\mathcal{C}_k} X_{\mathcal{C}_k}^\top\big) \tag{3.4.10}$$

$$= \lim_{T\to\infty} \max_{\{\mathcal{C}_k\}} \sum_{k=1}^K \frac{1}{T}\lambda_1\big(X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k}\big), \tag{3.4.11}$$

where we used that $X_{\mathcal{C}_k} X_{\mathcal{C}_k}^\top$ and $X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k}$ have the same nonzero eigenvalues.

For each k, write the top eigenvalue as a Rayleigh quotient:

$$\lambda_1\big(X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k}\big) = \max_{\|v_k\|=1} v_k^\top \big(X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k}\big) v_k.$$

Hence

$$\lim_{T\to\infty} \max_{\{\mathcal{C}_k\}} \sum_{k=1}^K \frac{1}{T}\lambda_1\big(X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k}\big) = \lim_{T\to\infty} \max_{\{\mathcal{C}_k\},\,\{v_k:\|v_k\|=1\}} \frac{1}{T}\sum_{k=1}^K v_k^\top X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k} v_k. \tag{3.4.12}$$

Now pad each v_k to a T-dimensional vector $\tilde{v}_k \in \mathbb{R}^T$ with support restricted to the indices in $\mathcal{C}_k$, i.e. $\operatorname{supp}(\tilde{v}_k) \subset \mathcal{C}_k$ and $\|\tilde{v}_k\| = 1$. Then

$$v_k^\top X_{\mathcal{C}_k}^\top X_{\mathcal{C}_k} v_k = \tilde{v}_k^\top X^\top X \tilde{v}_k,$$

and the objective becomes

$$\lim_{T\to\infty} \max_{\{\mathcal{C}_k\},\, \{\tilde{v}_k : \|\tilde{v}_k\|=1,\, \mathrm{supp}(\tilde{v}_k)\subset\mathcal{C}_k\}} \frac{1}{T}\sum_{k=1}^{K} \tilde{v}_k^\top X^\top X \tilde{v}_k. \tag{3.4.13}$$

Construct a matrix $Y \in \mathbb{R}^{K\times T}$ whose kth row is $\sqrt{T}\,\tilde{v}_k^\top$. Because each $\tilde{v}_k$ is supported on a disjoint cluster $\mathcal{C}_k$, each *column* of Y is one-sparse: every sample x_t is assigned to exactly one cluster. Moreover, each row of Y has norm $\sqrt{T}$, so

$$YY^\top = \mathrm{diag}(\|\sqrt{T}\,\tilde{v}_1\|^2, \ldots, \|\sqrt{T}\,\tilde{v}_K\|^2) = TI_K.$$

With this Y,

$$\sum_{k=1}^{K} \tilde{v}_k^\top X^\top X \tilde{v}_k = \frac{1}{T}\,\mathrm{Tr}(X^\top XY^\top Y),$$

and the objective can be written as

$$\lim_{T\to\infty} \max_{Y:\, YY^\top = TI} \frac{1}{T^2}\,\mathrm{Tr}(X^\top XY^\top Y). \tag{3.4.14}$$

Introduce Lagrange multipliers $B = \mathrm{diag}(b_1, \ldots, b_K)$ to enforce $YY^\top = TI$, and write

$$\lim_{T\to\infty} \min_{Y} \max_{B} -\frac{1}{T^2}\,\mathrm{Tr}(X^\top XY^\top Y) + \mathrm{Tr}\Big(B\big(\tfrac{1}{T}YY^\top - I\big)\Big). \tag{3.4.15}$$

We now introducing an auxiliary weight matrix W by using complete-the-square arguments:

$$\lim_{T\to\infty} \min_{Y,W} \max_{B} -\frac{1}{T}\,\mathrm{Tr}(X^\top W^\top Y) + \frac{1}{2}\|W\|_F^2 + \mathrm{Tr}\Big(B\big(\tfrac{1}{T}YY^\top - I\big)\Big). \tag{3.4.16}$$

Passing to the limit and taking expectation over x recovers the stochastic form

$$\min_{y,W} \max_{B}\ \mathbb{E}\big[-\mathrm{Tr}(x^\top Wy)\big] + \frac{1}{2}\|W\|_F^2 + \mathbb{E}\Big[\sum_{i=1}^{K} b_i(y_i^2 - 1)\Big], \tag{3.4.17}$$

where $y \in \mathbb{R}^K$ is now a one-sparse code indicating the cluster assignment of sample x. This is exactly the objective in the theorem statement. □

3.4.5 ACC circuits compute node marginals of a Gibbs distribution

In this section we prove Theorem 3.4.18, i.e. that the proposed ACC dynamics computes, at convergence, the layerwise node marginals of the Gibbs distribution over legal paths, and that the dynamics is globally convergent.

Theorem 3.4.18. *Consider the dynamics in Equation 3.2.13 on a level-l graph of horizon H and legal path set Π^l. Then the fixed point $r_\delta^{l,*}$ equals the layerwise node marginal of the Gibbs distribution Equation 3.2.15,*

$$r_{\delta,i}^{l,*} \;=\; \mathbb{P}_\beta^l(x_\delta = i),$$

and Equation 3.2.13 converges to this fixed point.

We begin with a simple convex-analytic identity that we will use to show convexity of the ACC energy.

Lemma 3.4.19. *For every $u > 0$,*

$$-\log u \;=\; \sup_{v>0}\bigl(-uv + 1 - \log v\bigr). \tag{3.4.20}$$

Proof. Consider $f(u) = -\log u$ on $u > 0$. Its convex conjugate is

$$f^*(v) \;=\; \sup_{u>0}\bigl(uv - f(u)\bigr) = \sup_{u>0}\bigl(uv + \log u\bigr).$$

Differentiating in u gives $v + 1/u = 0$, so the maximizer is $u^* = -1/v$, which is feasible only when $v < 0$. Plugging back,

$$f^*(v) = \begin{cases} -1 - \log(-v), & v < 0, \\ -\infty, & v \geq 0. \end{cases}$$

By Fenchel–Young equality,

$$-\log u \;=\; \sup_{w<0}\bigl(uw - f^*(w)\bigr) = \sup_{w<0}\bigl(uw + 1 + \log(-w)\bigr).$$

Reparametrize $v = -w > 0$ to obtain

$$-\log u \;=\; \sup_{v>0}\bigl(-uv + 1 - \log v\bigr),$$

as claimed. □

Next we show that the ACC dynamics admits a strictly convex Lyapunov energy function and thus converges to a unique minimizer.

Lemma 3.4.21. *Define, for a fixed hierarchy level l, the energy*

$$\mathcal{F}^l(r) \;=\; \sum_{\delta=1}^{H}\sum_{i=1}^{N^l} r_{\delta,i}\left(\log r_{\delta,i} - \widehat{R}_{\delta,i}\right) \;-\; \sum_{\delta=1}^{H-1}\Bigl(r_\delta^\top \log\bigl(A^l r_{\delta+1}\bigr) + r_{\delta+1}^\top \log\bigl((A^l)^\top r_\delta\bigr)\Bigr), \tag{3.4.22}$$

where $\log(\cdot)$ acts elementwise and $r_\delta \in \Delta^{N^l}$. Then $\mathcal{F}^l$ is strictly convex over $\prod_{\delta=1}^{H}\Delta^{N^l}$ and hence admits a unique global minimizer.

Proof. The term $\sum_{\delta,i} r_{\delta,i}\log r_{\delta,i}$ is strictly convex on the simplex and the linear term $-\sum_{\delta,i} r_{\delta,i}\widehat{R}_{\delta,i}$ is convex. It therefore suffices to show that

$$(a, b) \;\mapsto\; -a^\top \log(A^l b)$$

is convex for $a, b \in \Delta^{N^l}$. By Lemma 3.4.19,

$$-a_i \log(A^l b)_i = \sup_{v_i>0}\bigl(-v_i(A^l b)_i + a_i - a_i \log v_i\bigr),$$

and summing over i gives

$$-a^\top \log(A^l b) = \sup_{v>0}\big(-v^\top A^l b + a^\top \mathbf{1} - a^\top \log v\big).$$

A supremum of affine functions in (a, b) is convex, so $-a^\top \log(A^l b)$ is convex. The same argument applies to the reverse term $-b^\top \log((A^l)^\top a)$. Adding these terms to a strictly convex entropy term preserves strict convexity on the product of simplices, so $\mathcal{F}^l$ has a unique global minimizer. □

We now show that $\mathcal{F}^l$ is a Lyapunov function for the ACC dynamics.

Lemma 3.4.23. *The energy $\mathcal{F}^l$ in (3.4.22) is a Lyapunov function for the dynamics in Equation 3.2.13.*

Proof. Recall the ACC dynamics

$$\tau_n \frac{dz^l_{\delta,i}}{dt} = -z^l_{\delta,i} + \widehat{R}^l_{\delta,i} + \log\Big(\sum_j A^l_{ji} r^l_{\delta-1,j}\Big) + \log\Big(\sum_k A^l_{ik} r^l_{\delta+1,k}\Big), \qquad r^l_{\delta,i} \propto e^{z^l_{\delta,i}},$$

so that $\dot{r}^l_{\delta,i} = \frac{1}{\tau_n}(-z^l_{\delta,i} - \rho^l_{\delta,i})\, r^l_{\delta,i}$, where

$$\rho^l_{\delta,i} \;=\; -\log\big(A^l r_{\delta+1}\big)_i \;-\; \log\big((A^l)^\top r_{\delta-1}\big)_i.$$

A direct differentiation of (3.4.22) gives

$$\frac{\partial \mathcal{F}^l}{\partial r_{\delta,i}} \;=\; 1 + z^l_{\delta,i} + \rho^l_{\delta,i}.$$

Therefore

$$\begin{aligned}\frac{d}{dt}\mathcal{F}^l(r(t)) &= \sum_{\delta,i} \frac{\partial \mathcal{F}^l}{\partial r_{\delta,i}}\, \dot{r}_{\delta,i} = \sum_{\delta,i}(1 + z^l_{\delta,i} + \rho^l_{\delta,i})\dot{r}_{\delta,i} \\ &= \sum_{\delta,i}(z^l_{\delta,i} + \rho^l_{\delta,i})\dot{r}_{\delta,i} = -\frac{1}{\tau_n}\sum_{\delta,i}(z^l_{\delta,i} + \rho^l_{\delta,i})^2\, r^l_{\delta,i} \;\le\; 0.\end{aligned}$$

The third equality holds because $\sum_{\delta,i} \dot{r}_{\delta,i} = 0$. Thus $\mathcal{F}^l$ is nonincreasing along trajectories. By Lemma 3.4.21, $\mathcal{F}^l$ has a unique minimizer, hence $\mathcal{F}^l$ is a Lyapunov function. □

We can now prove the main result.

Proof of Theorem 3.4.18. Define singleton potentials

$$\phi^l_\delta(i) \;=\; \exp\big[\beta R^l_{\delta,i}\big]$$

and edge potentials

$$\psi^l_\delta(i,j) \;=\; A^l_{i,j} \in \{0,1\},$$

for $\delta = 1, \ldots, H-1$. These define a Markov random field over paths $(x_1, \ldots, x_H)$ on the level-l graph whose joint density is exactly

$$\mathbb{P}^l_\beta(\pi) \;=\; \frac{1}{Z^l_\beta} \exp\Big\{\sum_{\delta=1}^{H} \beta R^l_\delta(x_\delta)\Big\},$$

i.e. the Gibbs distribution in the theorem statement. The fixed-point equation of the ACC dynamics, Equation 3.2.14,

$$r^{l,*}_{\delta,i} = e^{\widehat{R}^l_{\delta,i}} \Big(\sum_j A^l_{ji} r^{l,*}_{\delta-1,j}\Big)\Big(\sum_k A^l_{ik} r^{l,*}_{\delta+1,k}\Big),$$

is precisely the stationarity condition of belief–propagation messages on this pairwise graphical model: each layer's belief is proportional to its local potential times the messages received from its neighbors. On a tree-structured chain of length H, belief propagation is exact and its fixed point equals the true node marginals of the Gibbs distribution.

What remains is to show that the ACC dynamics converges to this fixed point. Lemma 3.4.21 shows that the energy $\mathcal{F}^l$ is strictly convex and has a unique global minimum; Lemma 3.4.23 shows that $\mathcal{F}^l$ is a Lyapunov function for the dynamics in Equation 3.2.13. Hence every trajectory converges to the unique minimizer of $\mathcal{F}^l$, and this minimizer satisfies the fixed-point equation above. Therefore $r^{l,*}_{\delta,i} = \mathbb{P}^l_\beta(x_\delta = i)$, as claimed. □

References

[1] K. S. Lashley. "The Problem of Serial Order in Behavior". In: *Cerebral Mechanisms in Behavior: The Hixon Symposium*. Ed. by L. A. Jeffress. New York: John Wiley & Sons, 1951, pp. 112–136.

[2] R. Cooper and T. Shallice. "Contention scheduling and the control of routine activities". In: *Cognitive Neuropsychology* 17.4 (June 2000), pp. 297–338.

[3] M. K. Eckstein and A. G. E. Collins. "Computational evidence for hierarchically structured reinforcement learning in humans". In: *Proceedings of the National Academy of Sciences* 117.47 (Nov. 2020), pp. 29381–29389.

[4] J. J. Ribas-Fernandes, A. Solway, C. Diuk, J. T. McGuire, A. G. Barto, Y. Niv, and M. M. Botvinick. "A Neural Signature of Hierarchical Reinforcement Learning". In: *Neuron* 71.2 (July 2011), pp. 370–379.

[5] J. Balaguer, H. Spiers, D. Hassabis, and C. Summerfield. "Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network". In: *Neuron* 90.4 (May 2016), pp. 893–903.

[6] A. Solway, C. Diuk, N. Córdova, D. Yee, A. G. Barto, Y. Niv, and M. M. Botvinick. "Optimal Behavioral Hierarchy". In: *PLoS Computational Biology* 10.8 (Aug. 2014). Ed. by O. Sporns, e1003779.

[7] M. S. Tomov, S. Yagati, A. Kumar, W. Yang, and S. J. Gershman. "Discovery of hierarchical representations for efficient planning". In: *PLOS Computational Biology* 16.4 (Apr. 2020). Ed. by D. Pascucci, e1007594.

[8] C. G. Correa, S. Sanborn, M. K. Ho, F. Callaway, N. D. Daw, and T. L. Griffiths. "Exploring the hierarchical structure of human plans via program generation". In: *Cognition* 255 (Feb. 2025), p. 105990.

[9] P. Shamash, S. Lee, A. M. Saxe, and T. Branco. "Mice identify subgoal locations through an action-driven mapping process". In: *Neuron* 111.12 (June 2023), 1966–1978.e8.

[10] C. D. Grossman, V. Man, and J. P. O'Doherty. "The representation and valuation of subgoals in the human brain during model-based hierarchical behavior". In: *bioRxiv* (2025). eprint: https://www.biorxiv.org/content/early/2025/03/25/2025.03.24.645084.full.pdf.

[11] T. R. Colin, I. Ikink, and C. B. Holroyd. "Distributed Representations for Cognitive Control in Frontal Medial Cortex". In: *Journal of Cognitive Neuroscience* 37.5 (2025), pp. 941–969.

[12] C. B. Holroyd, J. J. F. Ribas-Fernandes, D. Shahnazian, M. Silvetti, and T. Verguts. "Human midcingulate cortex encodes distributed representations of task progress". In: *Proceedings of the National Academy of Sciences* 115.25 (June 2018), pp. 6398–6403.

[13] M. M. Botvinick, Y. Niv, and A. G. Barto. "Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective". In: *Cognition* 113.3 (Dec. 2009), pp. 262–280.

[14] M. M. Botvinick. "Hierarchical models of behavior and prefrontal function". In: *Trends in Cognitive Sciences* 12.5 (May 2008), pp. 201–208.

[15] Q. Liang, J. Li, S. Zheng, J. Liao, and R. Huang. "Dynamic Causal Modelling of Hierarchical Planning". In: *NeuroImage* 258 (Sept. 2022), p. 119384.

[16] R. S. Sutton, D. Precup, and S. Singh. "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". In: *Artificial Intelligence* 112.1–2 (Aug. 1999), pp. 181–211.

[17] R. Parr and S. Russell. "Reinforcement Learning with Hierarchies of Machines". In: *Advances in Neural Information Processing Systems*. Ed. by M. Jordan, M. Kearns, and S. Solla. Vol. 10. MIT Press, 1997.

[18] A. G. Barto and S. Mahadevan. "Recent Advances in Hierarchical Reinforcement Learning". In: *Discrete Event Dynamic Systems* 13.4 (Oct. 2003), pp. 341–379.

[19] K. J. Miller, M. M. Botvinick, and C. D. Brody. "Dorsal hippocampus contributes to model-based planning". In: *Nature Neuroscience* 20.9 (July 2017), pp. 1269–1276.

[20] B. E. Pfeiffer and D. J. Foster. "Hippocampal place-cell sequences depict future paths to remembered goals". In: *Nature* 497.7447 (Apr. 2013), pp. 74–79.

[21] J. D. Shin, W. Tang, and S. P. Jadhav. "Dynamics of Awake Hippocampal-Prefrontal Replay for Spatial Learning and Memory-Guided Decision Making". In: *Neuron* 104.6 (Dec. 2019), 1110–1125.e7.

[22] N. W. Schuck, M. B. Cai, R. C. Wilson, and Y. Niv. "Human Orbitofrontal Cortex Represents a Cognitive Map of State Space". In: *Neuron* 91.6 (Sept. 2016), pp. 1402–1412.

[23] T. Akam, I. Rodrigues-Vaz, I. Marcelo, X. Zhang, M. Pereira, R. F. Oliveira, P. Dayan, and R. M. Costa. "The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection". In: *Neuron* 109.1 (Jan. 2021), pp. 149–163.

[24] T. Zhang, M. Rosenberg, Z. Jing, P. Perona, and M. Meister. "Endotaxis: A neuromorphic algorithm for mapping, goal-learning, navigation, and patrolling". In: *eLife* 12 (Feb. 2024).

[25] N. A. Schmajuk and A. D. Thieme. "Purposive behavior and cognitive mapping: a neural network model". In: *Biological Cybernetics* 67.2 (June 1992), pp. 165–174.

[26] H. Voicu and N. Schmajuk. “Exploration, Navigation and Cognitive Mapping”. In: *Adaptive Behavior* 8.3–4 (June 2000), pp. 207–223.

[27] A. V. Samsonovich and G. A. Ascoli. “A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval”. In: *Learning & Memory* 12.2 (Mar. 2005), pp. 193–208.

[28] P. Gaussier, A. Revel, J. P. Banquet, and V. Babeau. “From view cells and place cells to cognitive map learning: processing stages of the hippocampal system”. In: *Biological Cybernetics* 86.1 (Jan. 2002), pp. 15–28.

[29] B. Schölkopf and H. A. Mallot. “View-Based Cognitive Mapping and Path Planning”. In: *Adaptive Behavior* 3.3 (Jan. 1995), pp. 311–348.

[30] O. Trullier and J.-A. Meyer. “Animat navigation using a cognitive graph”. In: *Biological Cybernetics* 83.3 (Aug. 2000), pp. 271–285.

[31] L.-E. Martinet, D. Sheynikhovich, K. Benchenane, and A. Arleo. “Spatial Learning and Action Planning in a Prefrontal Cortical Network Model”. In: *PLoS Computational Biology* 7.5 (May 2011). Ed. by O. Sporns, e1002045.

[32] F. Ponulak and J. J. Hopfield. “Rapid, parallel path planning by propagating wavefronts of spiking neural activity”. In: *Frontiers in Computational Neuroscience* 7 (2013).

[33] D. S. Corneil and W. Gerstner. “Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015.

[34] A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, and M. M. Botvinick. “Neural representations of events arise from temporal community structure”. In: *Nature Neuroscience* 16.4 (Feb. 2013), pp. 486–492.

[35] C. N. Boccara, M. Nardin, F. Stella, J. O’Neill, and J. Csicsvari. “The entorhinal cognitive map is attracted to goals”. In: *Science* 363.6434 (Mar. 2019), pp. 1443–1447.

[36] S. A. Hollup, S. Molden, J. G. Donnett, M.-B. Moser, and E. I. Moser. “Accumulation of Hippocampal Place Fields at the Goal Location in an Annular Watermaze Task”. In: *The Journal of Neuroscience* 21.5 (Mar. 2001), pp. 1635–1644.

[37] K. Louie and M. A. Wilson. “Temporally Structured Replay of Awake Hippocampal Ensemble Activity during Rapid Eye Movement Sleep”. In: *Neuron* 29.1 (Jan. 2001), pp. 145–156.

[38] M. F. Carr, S. P. Jadhav, and L. M. Frank. “Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval”. In: *Nature Neuroscience* 14.2 (Jan. 2011), pp. 147–153.

[39] A. K. Gillespie, D. A. Astudillo Maya, E. L. Denovellis, D. F. Liu, D. B. Kastner, M. E. Coulter, D. K. Roumis, U. T. Eden, and L. M. Frank. “Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice”. In: *Neuron* 109.19 (Oct. 2021), 3149–3163.e6.

[40] H. Igata, Y. Ikegaya, and T. Sasaki. "Prioritized experience replays on a hippocampal predictive map for learning". In: *Proceedings of the National Academy of Sciences* 118.1 (Dec. 2020).

[41] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[42] K. Kaefer, M. Nardin, K. Blahna, and J. Csicsvari. "Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching". In: *Neuron* 106.1 (Apr. 2020), 154–165.e6.

[43] D. Badre. "Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes". In: *Trends in Cognitive Sciences* 12.5 (May 2008), pp. 193–200.

[44] K. B. Kjelstrup, T. Solstad, V. H. Brun, T. Hafting, S. Leutgeb, M. P. Witter, E. I. Moser, and M.-B. Moser. "Finite Scale of Spatial Representation in the Hippocampus". In: *Science* 321.5885 (July 2008), pp. 140–143.

[45] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. "The successor representation in human reinforcement learning". In: *Nature Human Behaviour* 1.9 (Aug. 2017), pp. 680–692.

[46] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. "The hippocampus as a predictive map". In: *Nature Neuroscience* 20.11 (Oct. 2017), pp. 1643–1653.

[47] A. McGovern and A. G. Barto. "Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 361–368.

[48] I. Menache, S. Mannor, and N. Shimkin. "Q-Cut—Dynamic Discovery of Sub-goals in Reinforcement Learning". In: *Machine Learning: ECML 2002*. Springer Berlin Heidelberg, 2002, pp. 295–306.

[49] M. Stolle and D. Precup. "Learning Options in Reinforcement Learning". In: *Abstraction, Reformulation, and Approximation*. Springer Berlin Heidelberg, 2002, pp. 212–223.

[50] A. Bar, R. Talmon, and R. Meir. "Option discovery in the absence of rewards with manifold analysis". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[51] M. C. Machado, C. Rosenbaum, X. Guo, M. Liu, G. Tesauro, and M. Campbell. "Eigenoption Discovery through the Deep Successor Representation". In: *International Conference on Learning Representations*. 2018.

[52] G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto. "Constructing Skill Trees for Reinforcement Learning Agents from Demonstration Trajectories". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010.

[53] S. Niekum, S. Osentoski, G. Konidaris, and A. G. Barto. "Learning and generalization of complex tasks from unstructured demonstrations". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012, pp. 5239–5246.

[54] D. Tang, X. Li, J. Gao, C. Wang, L. Li, and T. Jebara. "Subgoal Discovery for Hierarchical Dialogue Policy Learning". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[55] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. "Diversity is All You Need: Learning Skills without a Reward Function". In: *International Conference on Learning Representations*. 2019.

[56] S. Lee, S.-W. Lee, J. Choi, D.-H. Kwak, and B.-T. Zhang. "Micro-Objective Learning : Accelerating Deep Reinforcement Learning through the Discovery of Continuous Subgoals". In: *CoRR* abs/1703.03933 (2017).

[57] A. Mesbah, R. Hosseini, S. P. Shariatpanahi, and M. N. Ahmadabadi. "Subgoal Discovery Using a Free Energy Paradigm and State Aggregations". In: *CoRR* abs/2412.16687 (2024). arXiv: 2412.16687.

[58] P. Shamash, S. F. Olesen, P. Iordanidou, D. Campagner, N. Banerjee, and T. Branco. "Mice learn multi-step routes by memorizing subgoal locations". In: *Nature Neuroscience* 24.9 (July 2021), pp. 1270–1279.

[59] P. Dayan and G. E. Hinton. "Feudal Reinforcement Learning". In: *Advances in Neural Information Processing Systems 5, [NIPS Conference]*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 271–278.

[60] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. "FeUdal networks for hierarchical reinforcement learning". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 3540–3549.

[61] O. Nachum, S. Gu, H. Lee, and S. Levine. "Data-efficient hierarchical reinforcement learning". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 3307–3317.

[62] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, Apr. 1994.

[63] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2018.

[64] M. B. Wang, N. Lynch, and M. M. Halassa. "The neural basis for uncertainty processing in hierarchical decision making". In: *Nature Communications* 16.1 (Oct. 2025).

[65] B. A. Wang, M. B. Wang, N. H. Lam, L. Mengxing, S. Li, R. D. Wimmer, P. M. Paz-Alonso, M. M. Halassa, and B. Pleger. "Thalamic regulation of reinforcement learning strategies across prefrontal-striatal networks". In: *Nature Communications* 16.1 (Oct. 2025).

[66] D. George, R. V. Rikhye, N. Gothoskar, J. S. Guntupalli, A. Dedieu, and M. Lázaro-Gredilla. "Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps". In: *Nature Communications* 12.1 (Apr. 2021).

[67] R. V. Raju, J. S. Guntupalli, G. Zhou, C. Wendelken, M. Lázaro-Gredilla, and D. George. "Space is a latent sequence: A theory of the hippocampus". In: *Science Advances* 10.31 (Aug. 2024).

[68] C. Diuk, A. Cohen, and M. L. Littman. "An object-oriented representation for efficient reinforcement learning". In: *Proceedings of the 25th international conference on Machine learning - ICML '08*. ICML '08. ACM Press, 2008, pp. 240–247.

[69] F. Donnarumma, D. Maisto, and G. Pezzulo. "Problem Solving as Probabilistic Inference with Subgoaling: Explaining Human Successes and Pitfalls in the Tower of Hanoi". In: *PLOS Computational Biology* 12.4 (Apr. 2016). Ed. by O. Sporns, e1004864.

Chapter 4

Neural architectures for flexible navigation on manifolds

4.1 Introduction

Humans and animals often succeed in new or changing tasks with only limited prior experience, a feat made possible by building rich internal representations of the world. When an organism encounters a complex environment, whether a physical landscape or a conceptual problem space, it can draw on these representations, or "cognitive maps," to plan novel routes and achieve goals. For example, an animal might need to adapt its path when a familiar trail is blocked by a temporary barrier, leveraging prior knowledge of local landmarks to find an alternative route. Likewise, someone solving a multi-step puzzle can transition among partially known states, guided by an internal map of how different ideas connect. In both cases, the power to move beyond brute-force trial and error depends on learning and exploiting the underlying structure.

The last chapter considered how the brain might construct and traverse such maps when the state space is discrete. Here we focus on continuous geometry. Experimental evidence in rodent and Drosophila navigation suggests that neural manifolds play a central role: grid cells in rodents, for instance, encode spatial location in a toroidal manifold, while head direction cells in Drosophila capture orientation information in a compact, circular manifold. Yet it remains an open question how animals extend such representations to more arbitrary topologies, from winding mazes to abstract reasoning tasks where the "locations" are ideas rather than physical points in space.

We propose that the brain can approximate these neural manifolds by stitching together piecewise linear patches, known as simplicial complexes (SCs). Drawing on circuits inspired by rodent and Drosophila navigation systems, we show how each individual simplex can function as a stable attractor, enabling robust "local" navigation through closed-loop feedback. We then identify the bifurcation conditions under which multiple simplex attractors can be joined, forming larger simplicial complexes reminiscent of gluing triangular tiles to approximate curved surfaces. Finally, we show how a planning mechanism can navigate these glued-together simplexes by computing the shortest path on a hypergraph that captures the connectivity between them. This approach allows an agent to hop flexibly among attractors,

each representing different regions or concepts, and reach a specified goal without exhaustively exploring every possibility.

By unifying local control with global route planning, our model provides a plausible neural implementation for how cognitive maps can be both learned and deployed. In rodents, for instance, grid cells form continuous attractors that support vector-based navigation, while place cells form spatial fields and navigate through successor representation-like planning. In Drosophila, ring attractors help shift the fly from one "sector" of orientation space to the appropriate goal direction via planning-like circuitry. We reinterpret these biologically observed continuous manifolds as simplicial complexes, offering a computational framework to understand navigation in real neural circuits through the coordination of local control and global route planning.

Ultimately, the capacity to form and exploit such complex manifolds appears fundamental to flexible, goal-directed behavior across species. By integrating stable attractors for local control with a global planning system, our circuit model demonstrates how brains can navigate both physical and conceptual terrains, illuminating a deep connection between the geometry of neural representations and the adaptive behaviors they enable. This perspective opens promising avenues for understanding how animals construct high-dimensional maps that facilitate versatile problem-solving, even with limited direct experience.

4.2 Background

Animals possess the remarkable ability to form complex cognitive maps, enabling them to navigate physical environments and reason about abstract concepts. These cognitive maps are believed to be encoded in low-dimensional neural manifolds [1]. In rodents, grid cells in the medial entorhinal cortex (MEC) exhibit a hexagonal firing pattern that maps spatial locations, effectively creating an internal coordinate system for navigation. This activity is shown to arise from a toroidal continuous attractor, which maintains stable spatial representations and enables path integration by tracking movement in the absence of external cues [2]. Similarly, head direction cells provide a sense of orientation by firing when the animal's head is aligned in a specific direction. In Drosophila, these cells collectively form a ring attractor, a recurrent network architecture that ensures a stable representation of heading and allows for continuous integration of directional changes [3]. These continuous attractors play a fundamental role in constructing neural manifolds that represent physical space and support spatial navigation.

Beyond spatial navigation, neural manifolds are increasingly recognized as key structures for representing abstract cognitive variables. Studies have shown that the hippocampus encodes not only spatial maps but also non-spatial task-relevant variables, such as latent cognitive states, learned rules, and relational structures [4–6]. Moreover, spatial and abstract representations can be jointly embedded in a shared neural manifold, suggesting a general computation to create task-relevant low-dimensional manifolds [1]. Theoretical models suggest that these manifold structures provide an efficient framework for cognitive inference, supporting behaviors such as decision-making, reasoning, and memory-guided planning [7].

Despite these insights, the mechanisms by which the brain constructs and navigates arbitrary neural manifolds—whether for spatial orientation or abstract reasoning—remain incompletely understood. Understanding how these diverse cognitive maps are formed and

utilized is crucial for elucidating the neural basis of flexible behavior and complex thought processes in animals.

4.3 Results

4.3.1 Simplicial complexes and action dynamics

Representing arbitrary manifolds is challenging due to strong nonlinearities of manifolds. To combat the nonlinearity, we approximate a manifold by a simplicial complex (SC), a piecewise-linear space obtained by gluing together higher-dimensional generalizations of triangles called simplices. For example, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron (Figure 4.1a). Formally, an n-dimensional simplicial complex S is a finite collection Σ of simplices of dimension at most n such that (i) any face of a simplex in Σ is also in Σ, and (ii) the intersection of any two simplices in Σ is either empty or a common face.

In order to navigate on the SC, we also need to define how action dynamics evolve the state on SC. We first specify local action dynamics within each simplex. For a simplex $\sigma \in \Sigma$ with $\dim(\sigma) = d$, let $\Gamma^{(\sigma)} \in \mathbb{R}^{d\times d}$ denote a local action basis. The within-simplex dynamics are

$$\tau_{\mathrm{e}} \frac{dx}{dt} \;=\; \Gamma^{(\sigma)} a, \qquad x \in \sigma,\; a \in \mathbb{R}^d, \tag{4.3.1}$$

where x is the position, τ_{e} is an environmental time constant, and a is the local action coordinate in σ.

To extend Equation 4.3.1 to the whole complex, we must resolve boundaries where local coordinates change across adjacent simplices. We therefore track both a global action $a(t) \in \mathbb{R}^n$ (with only the first d entries used on a d-simplex) and an active-simplex process $A(t) \in \Sigma$,

$$\tau_{\mathrm{e}}\, \dot{x}(t) \;=\; \sum_{\sigma\in\Sigma} \mathbf{1}\big\{x(t) \in \sigma,\; A(t) = \sigma\big\}\; \Gamma^{(\sigma)}\, a_{1:\dim(\sigma)}(t), \tag{4.3.2}$$

which selects the appropriate local chart and action coordinates for the currently active simplex (Figure 4.1b). Transitions of $A(t)$ at shared faces implement the change of chart; in practice these can be governed by a planner or a policy that chooses the next simplex when $x(t)$ reaches a face.

Given a goal $x_g \in S$, the navigation objective is to find σ-selection $A(x, x_g)$ and local actions $a(x, x_g)$ so that the trajectory converges to the goal $x \to x_g$ with $x(t)$ evolving under Equation 4.3.2 (Figure 4.1c).

We tackle navigation in three steps. First, in Section 4.3.2 we represent a single simplex as a neural attractor and implement feedback control for within-simplex motion. Second, in Section 4.3.3, we derive conditions that allow multiple simplex attractors to be stably glued into larger complexes. Finally, in Section 4.3.4, we plan across the complex by treating simplices as nodes of a hypergraph and computing shortest paths at the macro scale before handing control back to local dynamics.

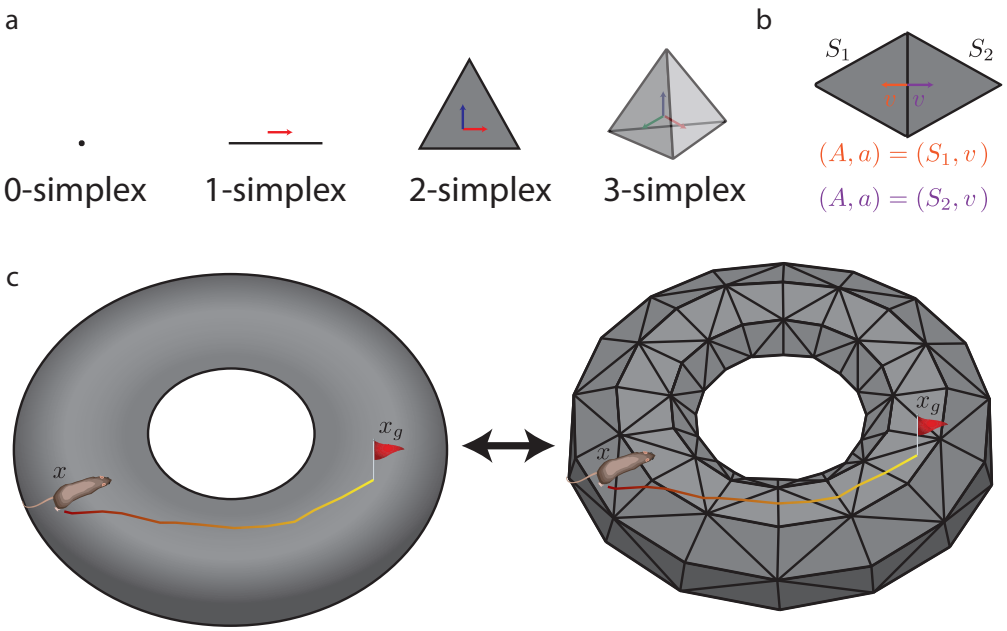

Figure 4.1: **Simplicial complex and action dynamics. a.** Canonical examples of simplices. **b.** Global motion on an SC selects a local chart $\sigma = A(t)$ and applies the corresponding $\Gamma^{(\sigma)}$ to the active coordinates of $a(t)$. **c.** Navigation on an arbitrary manifold approximated by an SC via sequential chart selection and local control.

4.3.2 Representing a simplex as an attractor

We begin with a single complex and its action dynamics embedded in a recurrent neural network. Let Δ_n be an n–simplex with action basis Γ, and consider a linear embedding $f : \Delta_n \to \mathbb{R}^m_{\geq 0}$,

$$f(x) = W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}, \tag{4.3.3}$$

chosen so that the boundary maps to the boundary of the nonnegative orthant, $f(\partial\Delta_n) \subseteq \partial\mathbb{R}^m_{\geq 0}$. Our goal in this section is to construct a circuit that (i) realizes $f(\Delta_n)$ as an attracting (center) manifold and (ii) simulates the prescribed action dynamics on this manifold.

The circuit comprises three populations. *State neurons* $y^{\mathrm{s}} \in \mathbb{R}^m_{\geq 0}$ receive the sensory drive $f(x)$ and implement the internal dynamics that mirror motion on the embedded simplex. *Goal neurons* $y^{\mathrm{g}} \in \mathbb{R}^m_{\geq 0}$ encode target inputs $f(x_{\mathrm{g}})$. *Motor neurons* $y^{\mathrm{m}+}, y^{\mathrm{m}-} \in \mathbb{R}^n_{\geq 0}$ read out a control signal $a = y^{\mathrm{m}+} - y^{\mathrm{m}-}$ from state and goal activity to act on the environment.

State dynamics follow rectified linear integration with recurrent and feedforward drive,

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + \left[W^{\mathrm{s}} y^{\mathrm{s}} + W^{\mathrm{m}+/\mathrm{s}} y^{\mathrm{m}+} + W^{\mathrm{m}-/\mathrm{s}} y^{\mathrm{m}-} + b^{\mathrm{s}} \right]_+, \tag{4.3.4}$$

where τ_{n} is the membrane time constant, $W^{\mathrm{s}} \in \mathbb{R}^{m \times m}$ are recurrent synapses, $W^{\mathrm{m}\pm/\mathrm{s}}$ project from motor to state populations, and b^{s} is a tonic bias. Goal neurons receive goal inputs $f(x_{\mathrm{g}})$,

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{g}}}{dt} = -y^{\mathrm{g}} + f(x_{\mathrm{g}}), \tag{4.3.5}$$

and motor activities are

$$y^{\mathrm{m}+} = \left[W^{\mathrm{s/m}+} y^{\mathrm{s}} + W^{\mathrm{g/m}+} y^{\mathrm{g}} \right]_+, \qquad y^{\mathrm{m}-} = \left[W^{\mathrm{s/m}-} y^{\mathrm{s}} + W^{\mathrm{g/m}-} y^{\mathrm{g}} \right]_+, \tag{4.3.6}$$

with $W^{\mathrm{s/m}\pm}, W^{\mathrm{g/m}\pm} \in \mathbb{R}^{n\times m}$.

In the Methods section, we first show that the embedded simplex is a center manifold for the state dynamics, and, when motor drive is absent, a stable attracting set whose flow matches the pushforward of the environmental action dynamics. The following Theorem 4.5.1 says that under mild alignment and rank conditions, choosing the recurrent and motor weights as above makes $f(\Delta_n)$ an invariant center manifold (normally attracting when there is no motor drive), on which the network's dynamics match exactly the environmental action dynamics pushed forward by f.

Theorem 4.5.1. *Let*

$$W^{\mathrm{s}} = -\left(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\right)(\hat{b}^{\mathrm{e/s}})^{\dagger} \; + \; W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger}, \qquad W^{\mathrm{m}+/\mathrm{s}} = \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}}\Gamma, \qquad W^{\mathrm{m}-/\mathrm{s}} = -\frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}}\Gamma,$$

and define $a = y^{\mathrm{m}+} - y^{\mathrm{m}-}$. *Here* $\hat{b}^{\mathrm{e/s}} = (I - W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger})\, b^{\mathrm{e/s}}$ *and* $(\cdot)^{\dagger}$ *denotes the pseudoinverse. Assume* $\langle \hat{b}^{\mathrm{e/s}}, b^{\mathrm{s}} \rangle \geq 0$ *and that* $W^{\mathrm{e/s}}$ *has full column rank. Then the affine set* $f(\Delta_n)$ *is a center manifold for the state dynamics. When* $y^{\mathrm{m}+} = y^{m-} = 0$, *the manifold* $f(\Delta_n)$ *is normally attracting. Moreover, the dynamics restricted to the center manifold coincides with the pushforward of the environmental action dynamics under* f.

Intuitively, this theorem states that the recurrent neural network tracks the environmental dynamics and are stable under perturbation. In simulations with two state neurons, the system exhibits a line attractor aligned with $f(\Delta_1)$, and a velocity unit traverses the attractor under motor-induced drive (Figure 4.2a). Even in the absence of sensory input, the circuit supports path integration and maintains an internal estimate of position (Figure 4.2b).

We next construct an inverse model that implements goal-directed control on the simplex. The following Theorem 4.5.15 says that by choosing the synaptic weight to motor neurons properly, the model can invert the pushforward dynamic to implement close-loop feedback control on the environmental state.

Theorem 4.5.15. *Let*

$$W^{\mathrm{s/m}+} = -(W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{g/m}+} = (W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{s/m}-} = (W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{g/m}-} = -(W^{\mathrm{e/s}}\Gamma)^{\dagger}$$

and all the conditions in Theorem 4.5.1 satisfy. Then the induced environmental dynamics are

$$\tau_{\mathrm{e}} \frac{dx}{dt} = x_{\mathrm{g}} - x,$$

and the state $x(t)$ *converges exponentially to* x_{g}.

Intuitively, this theorem shows that by combining the simulation of the state dynamics and inverting the environmental input to implement close-loop feedback control. Indeed, when we simulate it on a 3-simplex, this inverse mapping enables closed-loop navigation within a simplex and navigate toward the goal (Figure 4.2c).

Example 4.3.7 (A standard simplex attractor)**.** *Let*

$$b^{\mathrm{e/s}} = e_1, \qquad W^{\mathrm{e/s}} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \vdots \\ -1 & 0 & 0 & \dots & 1 & 0 \\ -1 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}. \tag{4.3.8}$$

Then the resulting recurrent weight is

$$W^{\mathrm{s}} = I - \mathbf{1}\mathbf{1}^{\top}, \tag{4.3.9}$$

which realizes a simplex attractor (Figure 4.2a). We call this particular embedding of simplex as a standard simplex embedding.

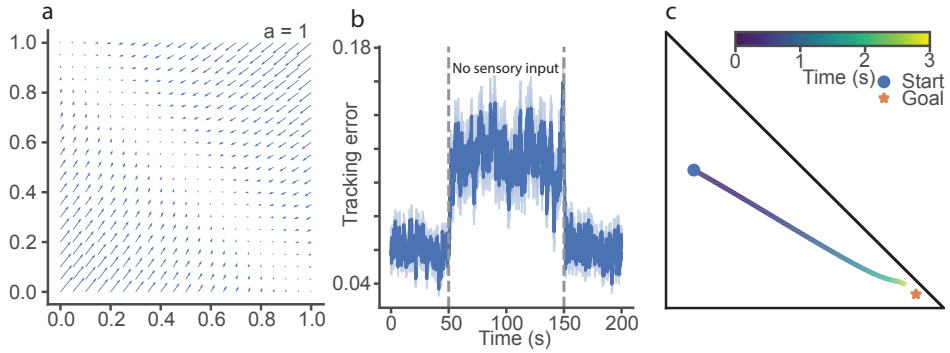

Figure 4.2: **Representing a simplex as a continuous attractor and enabling navigation via feedback control. a**, Phase portrait showing a line attractor embedded in a two-neuron system, with a velocity unit progressing along the attractor. **b**, Tracking error (mean ± s.e.m., $n = 50$) of the internally encoded position (Euclidean norm). Dashed lines mark onset and offset of sensory feedback. **c**, Example trajectory illustrating goal-directed navigation via the inverse model.

4.3.3 Stitching simplices to represent complex simplicial structure

Complex cognitive maps can be assembled by stitching multiple simplex attractors to form an arbitrary simplicial complex. In this section we derive conditions under which two simplex manifolds can be coupled so that their union is invariant and locally attracting under a single rectified-linear recurrent network.

Consider the recurrent network

$$\tau\dot{y} \;=\; -y + [Wy + b]_+, \qquad y \in \mathbb{R}^m_{\geq 0},$$

with neuron indices partitioned into two simplices σ_1, σ_2, their shared set $S := \sigma_1 \cap \sigma_2$, and exclusives $E_1 := \sigma_1 \setminus S$, $E_2 := \sigma_2 \setminus S$. After reordering indices as $[E_1\ S\ E_2]$ write

$$W \;=\; \begin{bmatrix} W_{11} & W_{1S} & W_{12} \\ W_{S1} & W_{SS} & W_{S2} \\ W_{21} & W_{2S} & W_{22} \end{bmatrix}, \qquad b \;=\; (b_{E_1}, b_S, b_{E_2}).$$

We assume each simplex is already realized as an attractor and that the representing weights coincide on the shared neurons. Concretely, define the block–sparse systems

$$W_1 = \begin{bmatrix} W_{11} & W_{1S} & 0 \\ W_{S1} & W_{SS} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad b_1 = (b_{E_1}, b_S, 0), \qquad W_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & W_{SS} & W_{S2} \\ 0 & W_{2S} & W_{22} \end{bmatrix}, \quad b_2 = (0, b_S, b_{E_2}).$$

Assume the W_1, b_1 dynamics realize a simplex attractor $\mathcal{M}_1 \subset \mathbb{R}^m_{\geq 0}$ supported on σ_1 (i.e. $y_{E_2} = 0$ on $\mathcal{M}_1$), and the W_2, b_2 dynamics realize a simplex attractor $\mathcal{M}_2$ supported on σ_2 (i.e. $y_{E_1} = 0$ on $\mathcal{M}_2$). Further assume the two manifolds coincide on the shared face,

$$\mathcal{M}_1 \cap \{y_{E_1} = 0\} = \mathcal{M}_2 \cap \{y_{E_2} = 0\}.$$

Since W_1, W_2 individually realize $\mathcal{M}_1, \mathcal{M}_2$ as simplex attractors, it remains to determine conditions on the cross-blocks W_{12}, W_{21} that make $\mathcal{M}_1 \cup \mathcal{M}_2$ an attracting set for the full network. By checking that all points on $\mathcal{M}_1 \cup \mathcal{M}_2$ are fixed points of the full dynamics, we obtain:

Lemma 4.3.10. *Suppose the cross-blocks (W_{12}, W_{21}) satisfy the simplex-level WTA inequalities*

$$W_{21}\, y_{E_1} + W_{2S}\, y_S + b_{E_2} \leq 0 \quad \textit{for all } y \in \mathcal{M}_1, \tag{C1}$$
$$W_{12}\, y_{E_2} + W_{1S}\, y_S + b_{E_1} \leq 0 \quad \textit{for all } y \in \mathcal{M}_2. \tag{C2}$$

Then the sets $\{y \in \mathcal{M}_1 : y_{E_2} = 0\}$ and $\{y \in \mathcal{M}_2 : y_{E_1} = 0\}$ are forward invariant under the full *dynamics with cross-blocks. Consequently, when the state lies on (or sufficiently near) $\mathcal{M}_1$, all neurons in E_2 remain identically zero (and symmetrically for $\mathcal{M}_2$ and E_1). Hence $\mathcal{M}_1 \cup \mathcal{M}_2$ is an invariant simplicial-complex attractor stitched with mutual exclusivity of the exclusive parts.*

Because each $\mathcal{M}_i$ is a simplex, the linear forms in (C1)–(C2) attain their maxima at vertices; it is therefore sufficient to verify the inequalities on the vertices of $\mathcal{M}_1$ and $\mathcal{M}_2$. In particular, if

$$W_{12} = a_{12}\, \mathbf{1}\mathbf{1}^\top, \qquad W_{21} = a_{21}\, \mathbf{1}\mathbf{1}^\top,$$

then choosing $a_{12}, a_{21} < 0$ with sufficiently large magnitude enforces (C1)–(C2), producing simplex-level winner–take–all coupling. As the pair (a_{12}, a_{21}) varies, fixed-point bifurcations delineate the stitching regime: intuitively, given two stable simplex attractors with coincident connections on the shared neurons, sufficiently strong mutual inhibition between exclusive neurons implements a simplex-level WTA that stitches the two attractors into a single invariant simplicial-complex manifold. Simulations confirm successful stitching when both are sufficiently negative (Figure 4.3a).

Using this principle, we constructed a toroidal attractor by stitching triangles. Persistent homology of the state-trajectory point cloud reveals one connected component (0th homology), two loops (1st), and one void (2nd), matching the torus topology (Figure 4.3b).

Example 4.3.11 (Stitching two line segments)**.** *Consider two 1-simplices joined at one vertex,* $\sigma_1 = \{1, 2\}$ *and* $\sigma_2 = \{2, 3\}$*. Let*

$$W = \begin{bmatrix} 0 & -1 & a_{12} \\ -1 & 0 & -1 \\ a_{21} & -1 & 0 \end{bmatrix}, \qquad b = (1, 1, 1).$$

Restricted to σ_1 *or* σ_2*, this recovers the standard embedding of Example 4.3.7. If* $a_{12}, a_{21} < -1$*, the vertex checks of (C1)–(C2) hold and the two segments stitch cleanly. If* $a_{12} = a_{21} = 0$*, the center manifold expands to the triangle. If only one of* a_{12}, a_{21} *is* < -1*, only the corresponding segment remains stable. For* $-1 < a_{12}, a_{21} < 0$*, the stable manifold collapses to the intersection (shared edge). If* $a_{12}, a_{21} > 0$*, the dynamics lose stability and diverge (Figure 4.3b).*

Example 4.3.12 (Embedding an arbitrary simplicial complex)**.** *Using the standard embedding, one can realize an arbitrary simplicial complex as a center manifold. Given an* n*-simplicial complex* Σ *with vertex set* $[n]$*, define* $W \in \mathbb{R}^{n \times n}$ *and* $b = \mathbf{1}$ *by*

$$W_{ii} = 0, \quad W_{ij} = W_{ji} = \begin{cases} -1 & \text{if } \exists\, \sigma \in \Sigma \text{ with } i, j \in \sigma, \\ -2 & \text{otherwise.} \end{cases}$$

Then the induced dynamics form a simplicial-complex attractor whose invariant faces match Σ*.*

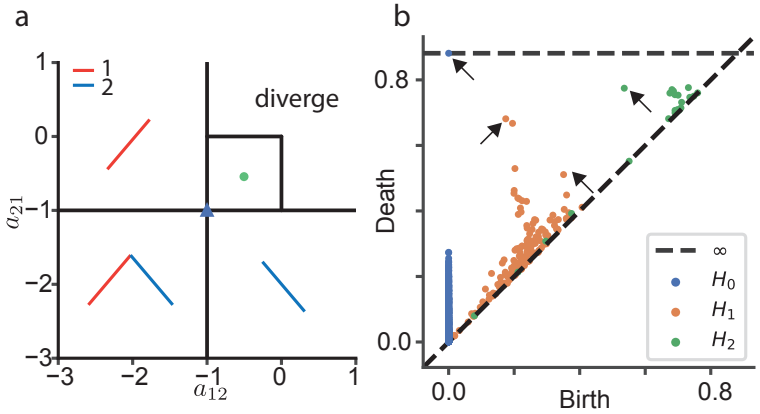

Figure 4.3: **Stitching simplices to form complex attractors. a**, Phase diagram of stationary points versus cross-coupling strengths (a_{21}, a_{12}). **b**, Persistence diagram of a toroidal attractor obtained by stitching triangles (one connected component, two 1D loops, one 2D void).

4.3.4 Planning to navigate across simplicial complex

Once we have represented the simplicial complex, the next step is to enable navigation on it. The previous vector-based feedback control in Section 4.3.2 is not applicable here, as it is not meaningful to perform vector calculus on vectors located in different simplices. To address this, we treat the simplicial complex as a hypergraph, where each vertex represents a

simplex and an edge exists if two simplices share a vertex (Figure 4.4a). Navigating across simplices then becomes equivalent to planning the shortest path on this hypergraph. Once the model reaches the goal simplex, the control circuit takes over to navigate the state to the goal within that simplex.

This component of the network is essentially the discrete planning module from chapter 3. The only difference is that the state input now arises from the simplicial complex, and the action becomes the direction from the center of the current simplex σ to the center of the intersection $\sigma \cap \sigma'$ with the next simplex (Methods, Section 4.5.3). The recurrent dynamics compute the shortest distance to the goal simplex via the OFC successor representation derived in the chapter 3 (Figure 4.4c). The ACC circuit then plans one step ahead through competition dynamics: ACC neurons receive inputs from both state and goal neurons and select the neighboring simplex that is closest to the goal (Figure 4.4d, e). The model successfully navigates across simplices and reaches the goal simplex. Once there, the control system activates to guide the model to the final goal within the simplex (Figure 4.4b; Methods, Section 4.5.3).

As in the discrete case in chapter 3, the planning circuit also benefits from hierarchy when the distance to the goal is large. We thus also include the same hierarchical planning mechanism over the induced hypergraph to achieve robust long-horizon navigation. We simulate hierarchical planning on a chain of 1-simplices. As shown in Figure 4.5a–d, the hierarchical graph structure establishes intermediate subgoals by projecting high-level next-step neurons as inputs to lower-level goal neurons. Leveraging this hierarchical structure enables navigation on graphs with arbitrarily large diameters, as shown in Chapter 3 (Figure 3.1a).

4.4 Discussion

We introduced a circuit framework for navigating complex cognitive maps by decomposing behavior into local control on piecewise-linear patches and global route planning over their connectivity. The central technical move is to approximate neural manifolds with simplicial complexes: individual simplexes serve as stable attractors that support robust, feedback-driven movement, while "gluing" conditions specify when multiple attractors can be stitched into a larger structure without sacrificing stability. Planning then reduces to computing shortest paths on a hypergraph whose nodes are simplexes and whose edges mark shared faces, allowing the system to hop efficiently among local attractors to reach distant goals. This unifies two strands of work, continuous attractors for local integration and graph-based planning for global routing, within a single dynamical architecture.

4.4.1 Connections to rodent and *Drosophila* navigation circuits

Our framework aligns with key organizational principles in mammalian and insect navigation systems. In rodents, the hippocampus is thought to form cognitive maps that jointly encode physical and abstract variables on a low-dimensional neural manifold [1], supporting flexible, goal-directed behavior. Our circuit offers an explicit implementation of such maps: simplicial complexes (SCs) are realized as continuous attractors navigated by combining feedback control with planning. The control module functions analogously to grid-cell circuits, sustaining

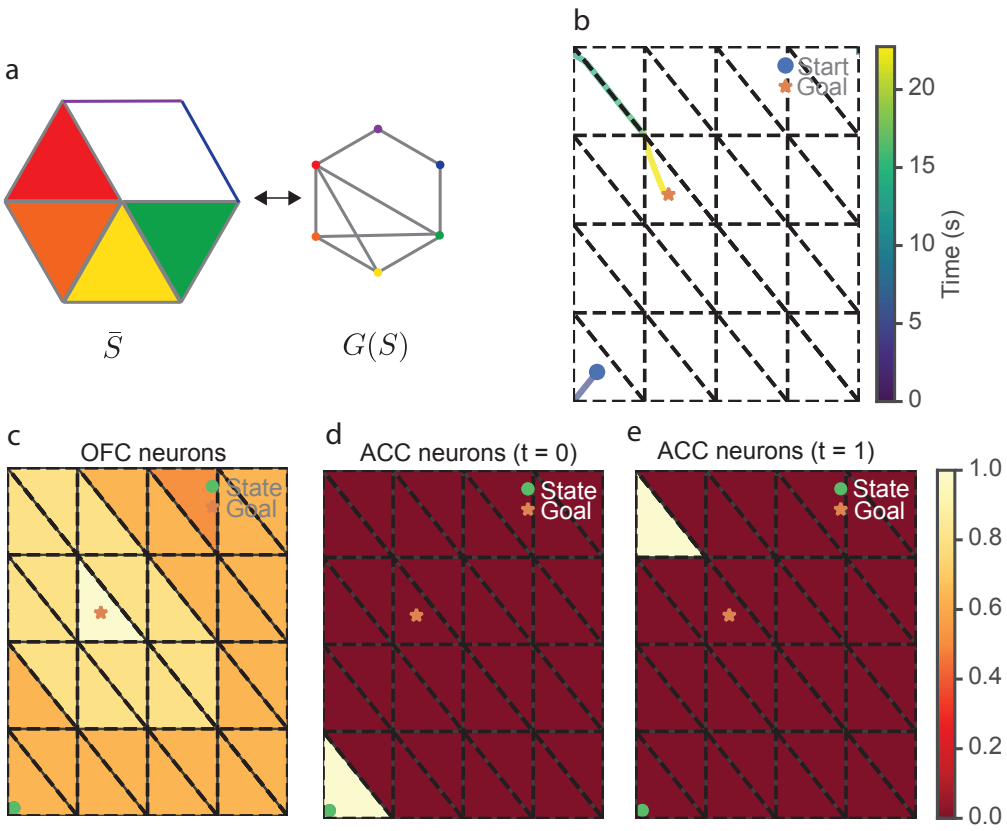

Figure 4.4: **Planning to navigate across a simplicial complex. a.** Example correspondence between a simplicial complex $\bar{S}$ and its hypergraph $G(S)$. **b.** Example trajectory on a torus, demonstrating coordination between planning (across simplices) and control (within a simplex). **c.** Spatial field of OFC neurons, where activity represents proximity to the goal. **d.** Spatial field of ACC neurons at $t = 0$, tuned to current spatial location (place-like responses). **e.** Spatial field of ACC neurons at $t = 1$, tuned to one-step-forward plans toward neighboring simplices.

continuous-attractor dynamics for path integration, whereas the planning module resembles place-cell computations, assigning spatial fields to individual units and selecting routes via a successor-representation-like computation.

This decomposition yields a computational hypothesis for interactions among hippocampal cell types during planning-based navigation. Our model reproduces neural signature parallel place cells [8], reward cells [9] and even trajectory-dependent activity reminiscent of splitter cells [10] observed in the hippocampus (Figure 4.4c-e). Moreover, the modularity of SCs affords reuse and recombination of submaps, offering a concrete account of hippocampal remapping as rearrangements of stitched simplices under context changes.

Recent advances from the *Drosophila* connectome suggest a more granular correspondence between circuit motifs and model components. In our construction, the canonical ring attractor can be interpreted as eight stitched simplices, each spanning 90° of heading space. The forward model within the control pathway corresponds to EPG and PENa populations that instantiate a ring manifold and update heading, respectively [3]. The inverse model

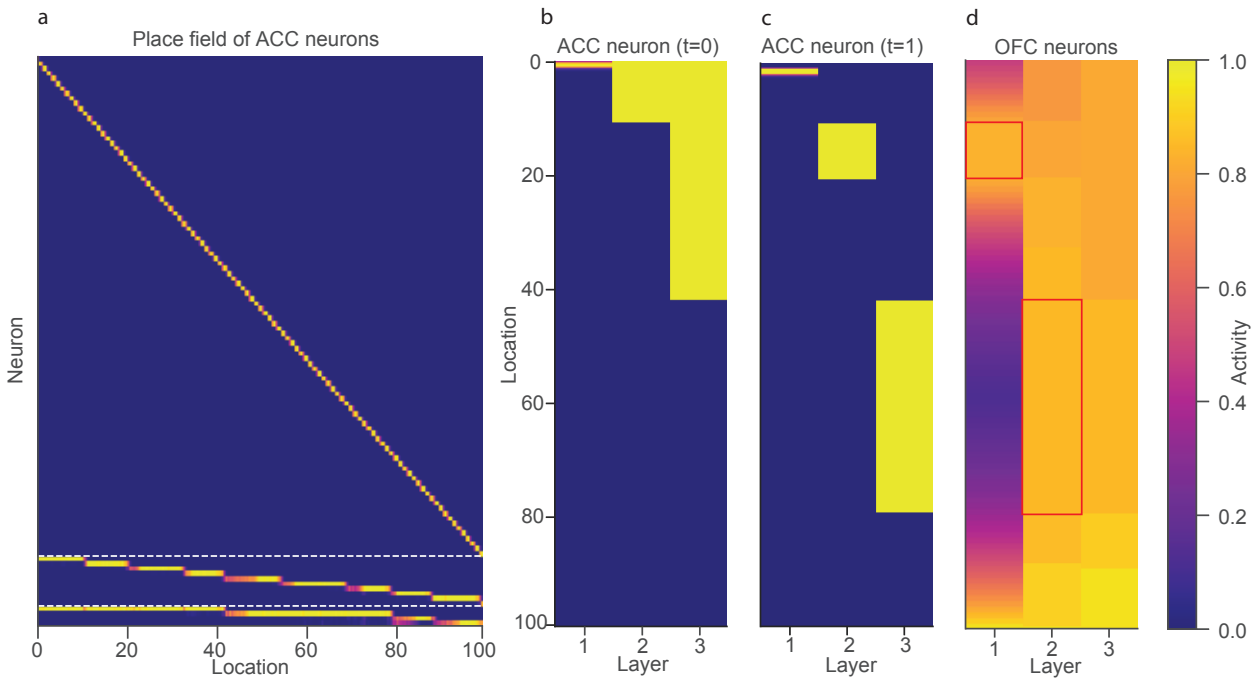

Figure 4.5: **Learning to navigate a hierarchical graph.** The network navigates on a line with start $v_0 = 1$ and goal $v_G = 100$. **a.** Hierarchical place field learned. **b.** Spatial heat map of hierarchical ACC neurons at $t = 0$ (defined as spatial field multiplied by neuronal activity). **c.** Spatial heat map of hierarchical ACC neurons at $t = 1$, tuned to one-step-forward plans toward neighboring simplices. **d.** Spatial heat map of hierarchical OFC neurons.

resembles the PFL3 pathway, which reads out current state (EPG) and goal signals (FC2) to drive steering outputs [11]. Notably, in PFL3 the state and goal inputs are offset by $\sim 90°$; thus PFL3 units are maximally engaged when current and goal headings differ by 90°, initiating turning.

A longstanding question is why PFL1 circuitry is required given the apparent overlap with PFL3. PFL1 similarly receives state input and, plausibly, goal input, with an offset of $\sim 135°$, yielding peak activation when the goal is far from the current heading [12]. Our framework suggests a division of labor: when the angular separation from the goal exceeds 90°, planning across simplices recruits PFL1 to mediate multi-step reorientation. Based on the connectome, we further hypothesize that FC1 neurons act as planning units encoding one-step forward moves between stitched simplices, providing a discrete bridge from the current heading simplex to the appropriate neighboring sector.

These correspondences do not imply strict homology, but they illustrate how a unified control-planning architecture can account for continuous-attractor dynamics, goal-directed readout, and discrete reorientation within a single mechanistic scheme spanning rodents and *Drosophila*. Targeted perturbations and electrophysiology of EPG/PENa, PFL1/PFL3, and FC1/FC2 during large-angle goals could directly test these predictions.

4.4.2 Representation of neural manifolds

Across species and tasks, population activity often lies on low-dimensional manifolds embedded in high-dimensional neural space. In navigation, rodent entorhinal–hippocampal circuits exhibit toroidal structure arising from grid-cell periodicity and ring-like organization in head-direction networks; place-cell populations form manifolds that combine physical location with abstract, task-dependent features. In insects, the central complex implements a ring attractor that stabilizes heading during navigation.

Theoretical accounts have explained many of these observations. Continuous-attractor models yield line, ring, and torus manifolds via structured recurrent connectivity with local excitation and inhibition. Yet most existing models target relatively simple geometries (e.g., rings or tori) and do not offer a general construction for complex manifolds such as those that jointly encode abstract task variables and physical space observed in hippocampal place cells.

Our framework provides a constructive and mechanistic route to represent and use neural manifolds for control and planning. Given an arbitrary manifold, we represent it by a *simplicial complex*, a piecewise-linear space assembled from simplexes. We then realize the target as a *center manifold* of a rectified recurrent circuit built from simple building blocks: each simplex is implemented as a continuous attractor with closed-form synaptic structure, and arbitrary topology is obtained by stitching simplexes along shared faces with analytically tunable cross-couplings. To support path integration and navigation on such manifolds, the model coordinates *local* feedback control with *global* planning: planning operates across stitched simplexes on the adjacency structure, while vector-based control operates within a simplex where directions are well defined.

4.4.3 Learning simplicial complex attractors

A central limitation of our proposal is that we have *designed* the weights analytically; how such weights could be *learned* from experience remains open. Here we outline a set of intuitions for how learning might proceed in a biologically plausible way.

At the level of a single simplex attractor, an affine chart, one can imagine an error-driven feedback rule that pushes the recurrent dynamics toward a desired hyperplane, for instance by adjusting synapses in proportion to a teaching signal comparing the network's current flow $f(x)$ to a target chart activity y^{s}. Such a rule can make the intended hyperplane an invariant set of the dynamics, but it also exposes a weakness that is easy to miss: if the training signal only specifies where the manifold should be and never explicitly labels *off*-manifold deviations, then the learned system may fail to contract in the *transverse* off-manifold directions. In other words, the network can learn to "stay on" the chart without learning to "return to" the chart when perturbed. To obtain robust attractor behavior, learning must therefore also estimate the geometry of the inputs and actively suppress variance orthogonal to the manifold. A natural candidate is an online subspace-learning process: Hebbian plasticity can amplify the dominant modes of variation, the manifold directions, while complementary anti-Hebbian or inhibitory mechanisms can identify and damp the residual components, the transverse directions. Functionally, this acts like learning a local PCA basis from experience and then implementing decay or inhibitory feedback along the transverse components so that off-manifold activity contracts. In parallel, forward and inverse controllers can be learned

with three-factor rules that reduce motor-prediction errors and goal errors, using modulatory signals that mark prediction mismatch or task success.

Practically, these ingredients suggest a curriculum. Early learning identifies candidate charts and their adjacencies from behavioral trajectories, using simple clustering augmented with topological signatures. With candidate charts in hand, local affine maps can be fit online and then stabilized until transverse variance reliably contracts. Finally, control and planning can be calibrated with sparse supervision and replay; the planning module can then be learned using the learning rule developed in Chapter 3.

The more difficult problem is learning the *embedding* of the simplicial complex itself. In this section we assumed a feedforward embedding that maps the boundary of the simplicial complex to the boundary of the positive orthant, which is what makes a ReLU-based representation of the resulting piecewise-linear structure so natural. Two questions follow: how can the system learn an embedding that satisfies this boundary-alignment condition, and among all such embeddings, what principle selects an optimal one? We speculate that a normative trade-off between metabolic cost and reliability could drive the solution. Because spiking is energetically expensive, embeddings that simulate the same dynamics with less total activity are favored; yet activity cannot be too small because neural noise degrades downstream readout. This suggests an optimum that balances energy efficiency against signal-to-noise ratio. Importantly, optimizing such a trade-off often yields sparse codes, and sparsity tends to align representational boundaries with coordinate axes, making orthant-aligned embeddings more likely. In this view, the boundary-alignment assumption is not an arbitrary assumption but an emergent consequence of learning under energy and noise constraints.

4.4.4 More biological representations for simplicial complex attractors

The representations used in this chapter are intentionally minimal, but they do not capture several aspects of known neurobiology. A first mismatch concerns our control-circuit representation. We drew an analogy between the simplex attractor used for control and grid-cell representations in the hippocampal-entorhinal system, but our current construction does not reproduce the hallmark hexagonal structure of grid-cell firing. Functionally, the present scheme behaves more like a collection of localized fields than like a periodic code. Yet grid cells are widely thought to provide an efficient representation of affine variables and to support vector-like computations [13] through their phase geometry. This suggests a more biologically grounded alternative: replace each simplex chart with a *set* of grid-cell modules whose combined activity encodes a higher-dimensional affine state. In this view, the "chart" is not a affine hyperplane implemented by a distributed periodic code whose linearizable coordinates can be read out locally. Moreover, the intrinsic path-integration dynamics of grid-cell circuits could implement much of the forward controller in our control loop, reducing the need to posit a separate analytic forward model. Under this interpretation, the control circuit in our architecture could plausibly be realized by grid-cell-like attractor dynamics rather than by the simplex attractors as written.

A second mismatch is geometric: we assumed each chart is affine and used rectifying nonlinearities that naturally produce piecewise-linear structure, whereas biological manifolds

are often smoothly curved. A straightforward extension is to relax the piecewise-linearity assumption. If the neural nonlinearity is closer to sigmoidal saturation than to a hard rectifier, then the same general construction can yield *nonlinear* invariant sets: instead of affine hyperplanes, the local attractors become curved hypersurfaces. One can then apply the same stitching logic as in the simplicial-complex case to assemble a global, curved neural manifold. This preserves the hierarchical intuition, local charts plus hypergraph transitions, while moving toward representations that more closely resemble the smooth geometry often inferred from neural population activity.

4.5 Methods

4.5.1 Representing a simplex as an attractor

In this section, we prove the following theorem, showing how we can embed a simplex as a simplex attractor.

Theorem 4.5.1. *Let*

$$W^{\mathrm{s}} = -\big(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\big)\,(\hat{b}^{\mathrm{e/s}})^{\dagger} \;+\; W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger}, \qquad W^{\mathrm{m+/s}} = \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}\, W^{\mathrm{e/s}}\Gamma, \qquad W^{\mathrm{m-/s}} = -\frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}\, W^{\mathrm{e/s}}\Gamma,$$

and define $a = y^{\mathrm{m}+} - y^{\mathrm{m}-}$. Here $\hat{b}^{\mathrm{e/s}} = (I - W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger})\, b^{\mathrm{e/s}}$ and $(\cdot)^{\dagger}$ denotes the pseudoinverse. Assume $\langle \hat{b}^{\mathrm{e/s}}, b^{\mathrm{s}} \rangle \geq 0$ and that $W^{\mathrm{e/s}}$ has full column rank. Then the affine set $f(\Delta_n)$ is a center manifold for the state dynamics. When $y^{\mathrm{m}+} = y^{m-} = 0$, the manifold $f(\Delta_n)$ is normally attracting. Moreover, the dynamics restricted to the center manifold coincides with the pushforward of the environmental action dynamics under f.

We break the proof into three steps: (i) show that the affine set

$$\mathcal{M} \;:=\; f(\Delta_n) \;=\; \{\, W^{\mathrm{e/s}}x + b^{\mathrm{e/s}} \;:\; x \in \Delta_n \,\}$$

is invariant for the state dynamics when $y^{\mathrm{m}} = 0$; (ii) show that directions transverse to $\mathcal{M}$ are contracting, so $\mathcal{M}$ is an attracting center manifold; and (iii) show that when the motor drive $a = y^{\mathrm{m}+} - y^{\mathrm{m}-}$ is present, the motion of y^{s} restricted to $\mathcal{M}$ is exactly the pushforward through f of the environmental/action dynamics on Δ_n.

Let's review the setup of the dynamics.

Setup and notation. Let

$$P \;:=\; W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger} \in \mathbb{R}^{m\times m}$$

be the orthogonal projector onto the column space $\mathrm{col}(W^{\mathrm{e/s}})$. Decompose the embedding bias $b^{\mathrm{e/s}}$ into the part inside the column space and the part orthogonal to it:

$$b^{\mathrm{e/s}} \;=\; Pb^{\mathrm{e/s}} + (I-P)b^{\mathrm{e/s}} \;=\; Pb^{\mathrm{e/s}} + \hat{b}^{\mathrm{e/s}},$$

where

$$\hat{b}^{\mathrm{e/s}} := (I-P)b^{\mathrm{e/s}} \in \ker(P).$$

By construction $\hat{b}^{\mathrm{e/s}}$ is orthogonal to $\mathrm{col}(W^{\mathrm{e/s}})$.

The state-neuron dynamics (with rectification active but in the orthant where all relevant arguments are positive) are

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + \left[\, W^{\mathrm{s}} y^{\mathrm{s}} + W^{\mathrm{m+/s}} y^{\mathrm{m+}} + W^{\mathrm{m-/s}} y^{\mathrm{m-}} + b^{\mathrm{s}} \,\right]_+. \tag{4.5.2}$$

Because $f(\Delta_n) \subset \mathbb{R}^m_{\geq 0}$, there is a neighborhood of $\mathcal{M}$ in which all arguments of $[\cdot]_+$ stay nonnegative; on that neighborhood $[\cdot]_+$ acts as the identity, and the dynamics reduce to the linear-affine form

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + W^{\mathrm{s}} y^{\mathrm{s}} + W^{\mathrm{m+/s}} y^{\mathrm{m+}} + W^{\mathrm{m-/s}} y^{\mathrm{m-}} + b^{\mathrm{s}}. \tag{4.5.3}$$

The theorem chooses

$$W^{\mathrm{s}} = -(\, b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}} \,)\, (\hat{b}^{\mathrm{e/s}})^{\dagger} + P, \tag{4.5.4}$$

$$W^{\mathrm{m+/s}} = \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}} \Gamma, \qquad W^{\mathrm{m-/s}} = -\frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}} \Gamma, \tag{4.5.5}$$

and defines the motor output

$$a \;:=\; y^{\mathrm{m+}} - y^{\mathrm{m-}} \in \mathbb{R}^n.$$

Now we are ready for the proof.

Proof. **Step 1: $\mathcal{M}$ is invariant and made of equilibria when $y^{\mathbf{m}} = 0$.** The affine set

$$\mathcal{M} \;:=\; f(\Delta_n) \;=\; \{\, W^{\mathrm{e/s}} x + b^{\mathrm{e/s}} \;:\; x \in \Delta_n \,\}$$

is invariant for the state dynamics when $y^{\mathrm{m}} = 0$.

Set $y^{\mathrm{m+}} = y^{\mathrm{m-}} = 0$. Then the state dynamics become

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + W^{\mathrm{s}} y^{\mathrm{s}} + b^{\mathrm{s}}. \tag{4.5.6}$$

Let $x \in \Delta_n$ be arbitrary and put

$$y^{\mathrm{s}} = f(x) = W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}.$$

We must show that y^{s} is an equilibrium of Equation 4.5.6.

Compute the right-hand side:

$$-y^{\mathrm{s}} + W^{\mathrm{s}} y^{\mathrm{s}} + b^{\mathrm{s}} = -\big(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}\big) + W^{\mathrm{s}} \big(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}\big) + b^{\mathrm{s}}.$$

Using the definition Equation 4.5.4 of W^{s},

$$W^{\mathrm{s}} z = -(\, b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}} \,)\, (\hat{b}^{\mathrm{e/s}})^{\dagger} z + Pz \quad \text{for any } z \in \mathbb{R}^m.$$

Apply this with $z = W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}$:

$$W^{\mathrm{s}}(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}) = -(\, b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}} \,)(\hat{b}^{\mathrm{e/s}})^{\dagger}(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}) + P(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}).$$

Because $\hat{b}^{\mathrm{e/s}} \in \ker(P)$, it is orthogonal to $\mathrm{col}(W^{\mathrm{e/s}})$ and hence

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}(W^{\mathrm{e/s}}x) = 0.$$

Also, since $b^{\mathrm{e/s}} = Pb^{\mathrm{e/s}} + \hat{b}^{\mathrm{e/s}}$ and $(\hat{b}^{\mathrm{e/s}})^{\dagger}\hat{b}^{\mathrm{e/s}} = 1$, we get

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}b^{\mathrm{e/s}} = 1.$$

Therefore

$$-(\, b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\,)(\hat{b}^{\mathrm{e/s}})^{\dagger}(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}) = -(\, b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\,) \cdot 1 = -b^{\mathrm{s}} + \hat{b}^{\mathrm{e/s}}.$$

Next,

$$P(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}) = PW^{\mathrm{e/s}}x + Pb^{\mathrm{e/s}} = W^{\mathrm{e/s}}x + Pb^{\mathrm{e/s}} = W^{\mathrm{e/s}}x + (b^{\mathrm{e/s}} - \hat{b}^{\mathrm{e/s}}),$$

since $b^{\mathrm{e/s}} = Pb^{\mathrm{e/s}} + \hat{b}^{\mathrm{e/s}}$. Putting these together,

$$\begin{aligned} W^{\mathrm{s}}(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}) &= \left(-b^{\mathrm{s}} + \hat{b}^{\mathrm{e/s}}\right) + \left(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}} - \hat{b}^{\mathrm{e/s}}\right) \\ &= W^{\mathrm{e/s}}x + b^{\mathrm{e/s}} - b^{\mathrm{s}}. \end{aligned}$$

Finally,

$$-y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + b^{\mathrm{s}} = -(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}) + \left(W^{\mathrm{e/s}}x + b^{\mathrm{e/s}} - b^{\mathrm{s}}\right) + b^{\mathrm{s}} = 0.$$

Thus every point

$$y^{\mathrm{s}} = f(x) = W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}$$

is an equilibrium of Equation 4.5.6, which proves that $\mathcal{M}$ is invariant when $y^{\mathrm{m}} = 0$.

Step 2 transverse stability. Keep the state dynamics with zero motor drive:

$$\tau_{\mathrm{n}}\frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + b^{\mathrm{s}}, \qquad W^{\mathrm{s}} = -\left(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\right)(\hat{b}^{\mathrm{e/s}})^{\dagger} + P, \tag{4.5.7}$$

where

$$P := W^{\mathrm{e/s}}(W^{\mathrm{e/s}})^{\dagger} \quad \text{and} \quad \hat{b}^{\mathrm{e/s}} := (I - P)b^{\mathrm{e/s}}.$$

Note that $\hat{b}^{\mathrm{e/s}} \in \ker(P)$ and

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}z = \frac{(\hat{b}^{\mathrm{e/s}})^{\top}z}{\|\hat{b}^{\mathrm{e/s}}\|^2} \quad \text{for any } z.$$

Define the *off-manifold error* as

$$e \;:=\; (I - P)\left(y^{\mathrm{s}} - b^{\mathrm{e/s}}\right). \tag{4.5.8}$$

Then $e = 0$ iff $y^{\mathrm{s}} \in b^{\mathrm{e/s}} + \mathrm{col}(W^{\mathrm{e/s}}) = f(\Delta_n)$, so e is exactly the transverse component.

Differentiate Equation 4.5.8 and use Equation 4.5.7:

$$\tau_{\mathrm{n}}\frac{de}{dt} = (I - P)\left(-y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + b^{\mathrm{s}}\right).$$

Insert $W^{\mathrm{s}} = -\big(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\big)(\hat{b}^{\mathrm{e/s}})^{\dagger} + P$:

$$\begin{aligned}\tau_{\mathrm{n}}\frac{de}{dt} &= (I-P)\big(-y^{\mathrm{s}} + Py^{\mathrm{s}} - (b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}})(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} + b^{\mathrm{s}}\big) \\ &= (I-P)(-y^{\mathrm{s}} + Py^{\mathrm{s}}) \;-\; (I-P)\big(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\big)(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} \;+\; (I-P)b^{\mathrm{s}}.\end{aligned}$$

Since $(I-P)(-y^{\mathrm{s}} + Py^{\mathrm{s}}) = -(I-P)y^{\mathrm{s}} = -e - (I-P)b^{\mathrm{e/s}} = -e - \hat{b}^{\mathrm{e/s}}$, we get

$$\tau_{\mathrm{n}}\frac{de}{dt} = -e - \hat{b}^{\mathrm{e/s}} - (I-P)\big(b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}\big)(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} + (I-P)b^{\mathrm{s}}.$$

Introduce the shorthand

$$\Delta b \;:=\; (I-P)\big(b^{\mathrm{s}} - b^{\mathrm{e/s}}\big) = (I-P)b^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}} \;\in\; \ker(P),$$

i.e. Δb is precisely the difference of the *off-manifold* parts of b^{s} and $b^{\mathrm{e/s}}$. Then the above becomes

$$\begin{aligned}\tau_{\mathrm{n}}\frac{de}{dt} &= -e - \hat{b}^{\mathrm{e/s}} - \big(\Delta b + \hat{b}^{\mathrm{e/s}}\big)(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} + \Delta b + \hat{b}^{\mathrm{e/s}} \\ &= -e + \Delta b - \Delta b(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} - \hat{b}^{\mathrm{e/s}}(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}}.\end{aligned}$$

Now note that $\hat{b}^{\mathrm{e/s}}(\hat{b}^{\mathrm{e/s}})^{\dagger}$ is the orthogonal projector onto $\mathrm{span}\{\hat{b}^{\mathrm{e/s}}\}$, so

$$\hat{b}^{\mathrm{e/s}}(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} = (\text{component of } y^{\mathrm{s}} \text{ along } \hat{b}^{\mathrm{e/s}}).$$

Moreover, by definition of e we can write any y^{s} uniquely as

$$y^{\mathrm{s}} = f(x) + z, \quad z \in \mathbb{R}^m, \quad \text{with} \quad e = (I-P)z.$$

Since $f(x) = W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}$ and $\hat{b}^{\mathrm{e/s}} \perp \mathrm{col}(W^{\mathrm{e/s}})$, we have

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}f(x) = (\hat{b}^{\mathrm{e/s}})^{\dagger}b^{\mathrm{e/s}} = 1,$$

because $b^{\mathrm{e/s}} = Pb^{\mathrm{e/s}} + \hat{b}^{\mathrm{e/s}}$ and $(\hat{b}^{\mathrm{e/s}})^{\dagger}\hat{b}^{\mathrm{e/s}} = 1$. Hence,

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}y^{\mathrm{s}} = (\hat{b}^{\mathrm{e/s}})^{\dagger}(f(x)+z) = 1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}z = 1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}e,$$

because $(\hat{b}^{\mathrm{e/s}})^{\dagger}Pz = 0$ and $e = (I-P)z$ lies in the same subspace as $\hat{b}^{\mathrm{e/s}}$. Substitute this into the error dynamics:

$$\begin{aligned}\tau_{\mathrm{n}}\frac{de}{dt} &= -e + \Delta b - \Delta b\big(1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}e\big) - \hat{b}^{\mathrm{e/s}}\big(1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}e\big) \\ &= -e + \Delta b - \Delta b - \Delta b(\hat{b}^{\mathrm{e/s}})^{\dagger}e - \hat{b}^{\mathrm{e/s}} - \hat{b}^{\mathrm{e/s}}(\hat{b}^{\mathrm{e/s}})^{\dagger}e \\ &= -e - \big(\Delta b + \hat{b}^{\mathrm{e/s}}\big)(\hat{b}^{\mathrm{e/s}})^{\dagger}e.\end{aligned}$$

Recall that $\Delta b + \hat{b}^{\mathrm{e/s}} = (I-P)b^{\mathrm{s}}$. Thus

$$\tau_{\mathrm{n}}\frac{de}{dt} = -e - (I-P)b^{\mathrm{s}}\,(\hat{b}^{\mathrm{e/s}})^{\dagger}e = -\big(I + (I-P)b^{\mathrm{s}}(\hat{b}^{\mathrm{e/s}})^{\dagger}\big)e. \tag{4.5.9}$$

Equation 4.5.9 is linear in e and has the form

$$\tau_{\mathrm{n}}\dot{e} = -(I + uv^{\top})e, \quad \text{with} \quad u = (I-P)b^{\mathrm{s}},\ v^{\top} = (\hat{b}^{\mathrm{e/s}})^{\dagger}.$$

A rank-1 update of the identity has eigenvalues

$$\lambda_1 = 1 + v^{\top}u, \qquad \lambda_2 = \cdots = \lambda_m = 1.$$

Therefore, the eigenvalues of the matrix $-\frac{1}{\tau_{\mathrm{n}}}(I + uv^{\top})$ governing e are

$$-\frac{1}{\tau_{\mathrm{n}}}(1 + v^{\top}u), \qquad -\frac{1}{\tau_{\mathrm{n}}} \quad \text{(with multiplicity } m-1\text{)}.$$

All transverse directions orthogonal to $\hat{b}^{\mathrm{e/s}}$ decay with rate $1/\tau_{\mathrm{n}}$. Along the single direction coupled to the rank-1 term, the decay rate is

$$\frac{1}{\tau_{\mathrm{n}}}(1 + v^{\top}u) = \frac{1}{\tau_{\mathrm{n}}}\Big(1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}(I-P)b^{\mathrm{s}}\Big).$$

But

$$(\hat{b}^{\mathrm{e/s}})^{\dagger}(I-P)b^{\mathrm{s}} = \frac{\big((I-P)b^{\mathrm{e/s}}\big)^{\top}(I-P)b^{\mathrm{s}}}{\|(I-P)b^{\mathrm{e/s}}\|^2} = \frac{\langle (I-P)b^{\mathrm{e/s}}, (I-P)b^{\mathrm{s}}\rangle}{\|(I-P)b^{\mathrm{e/s}}\|^2}.$$

Hence

$$1 + (\hat{b}^{\mathrm{e/s}})^{\dagger}(I-P)b^{\mathrm{s}} = 1 + \frac{\langle (I-P)b^{\mathrm{e/s}}, (I-P)b^{\mathrm{s}}\rangle}{\|(I-P)b^{\mathrm{e/s}}\|^2}.$$

Therefore, the transverse dynamics are exponentially stable *iff*

$$\|(I-P)b^{\mathrm{e/s}}\|^2 + \big\langle (I-P)b^{\mathrm{e/s}},\ (I-P)b^{\mathrm{s}}\big\rangle \ > \ 0. \tag{4.5.10}$$

By the assumption in the theorem, Equation 4.5.10 holds. Thus, all eigenvalues of the linearization in transverse directions are negative, so $f(\Delta_n)$ is an exponentially attracting manifold.

Step 3: Pushforward of the environmental/action dynamics. Now allow motor activity. Using Equation 4.5.3 and the special choice Equation 4.5.5 we get

$$\tau_{\mathrm{n}}\frac{dy^{\mathrm{s}}}{dt} = -y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}W^{\mathrm{e/s}}\Gamma y^{\mathrm{m}+} - \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}W^{\mathrm{e/s}}\Gamma y^{\mathrm{m}-} + b^{\mathrm{s}} \tag{4.5.11}$$

$$= -y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}W^{\mathrm{e/s}}\Gamma(y^{\mathrm{m}+} - y^{\mathrm{m}-}) + b^{\mathrm{s}} \tag{4.5.12}$$

$$= -y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}}W^{\mathrm{e/s}}\Gamma a + b^{\mathrm{s}}. \tag{4.5.13}$$

Now *restrict* to the manifold $\mathcal{M}$, i.e. assume

$$y^{\mathrm{s}} = f(x) = W^{\mathrm{e/s}}x + b^{\mathrm{e/s}} \quad \text{for some } x \in \Delta_n,$$

and stay in the positive orthant so that no rectification is triggered. As in Step 1, we already computed

$$-y^{\mathrm{s}} + W^{\mathrm{s}}y^{\mathrm{s}} + b^{\mathrm{s}} = 0 \quad \text{whenever } y^{\mathrm{s}} = W^{\mathrm{e/s}}x + b^{\mathrm{e/s}}.$$

Therefore Equation 4.5.13 simplifies on $\mathcal{M}$ to

$$\tau_{\mathrm{n}} \frac{dy^{\mathrm{s}}}{dt} = \frac{\tau_{\mathrm{n}}}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}} \Gamma a,$$

i.e.

$$\frac{dy^{\mathrm{s}}}{dt} = \frac{1}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}} \Gamma a. \tag{4.5.14}$$

But $y^{\mathrm{s}} = f(x) = W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}$, so

$$\frac{dy^{\mathrm{s}}}{dt} = \frac{d}{dt}\left(W^{\mathrm{e/s}} x + b^{\mathrm{e/s}}\right) = W^{\mathrm{e/s}} \frac{dx}{dt}.$$

Comparing with Equation 4.5.14, we obtain

$$W^{\mathrm{e/s}} \frac{dx}{dt} = \frac{1}{\tau_{\mathrm{e}}} W^{\mathrm{e/s}} \Gamma a.$$

Since $W^{\mathrm{e/s}}$ has full column rank, we can multiply both sides on the left by $(W^{\mathrm{e/s}})^{\dagger}$ to get

$$\frac{dx}{dt} = \frac{1}{\tau_{\mathrm{e}}} \Gamma a.$$

This is exactly the environmental/action dynamics one would posit on the simplex when the action basis is Γ and the timescale is τ_{e}:

$$\tau_{\mathrm{e}} \frac{dx}{dt} = \Gamma a.$$

Hence, on the manifold $\mathcal{M} = f(\Delta_n)$ the neural state y^{s} evolves exactly as the *pushforward* of the environmental state x evolves under the action dynamics; i.e.

$$\frac{d}{dt} f(x(t)) = Df(x(t)) \frac{dx}{dt} = W^{\mathrm{e/s}} \cdot \frac{1}{\tau_{\mathrm{e}}} \Gamma a,$$

which coincides with Equation 4.5.14. This proves the theorem. □

4.5.2 Our neural circuit forms an inverse model within a simplex via feedback control.

In this section, we will prove the following theorem

Theorem 4.5.15. *Let*

$$W^{\mathrm{s/m+}} = -(W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{g/m+}} = (W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{s/m-}} = (W^{\mathrm{e/s}}\Gamma)^{\dagger}, \quad W^{\mathrm{g/m-}} = -(W^{\mathrm{e/s}}\Gamma)^{\dagger}$$

and all the conditions in Theorem 4.5.1 satisfy. Then the induced environmental dynamics are

$$\tau_{\mathrm{e}} \frac{dx}{dt} = x_{\mathrm{g}} - x,$$

and the state $x(t)$ converges exponentially to x_{g}.

Proof. Notice $a = y^{m+} - y^{m-}$. Plug in the definition of y^{m+}, y^{m-}, we get

$$y^{m+} - y^{m-} = [W^{\mathrm{s/m+}} y^{\mathrm{s}} + W^{\mathrm{g/m+}} y^{\mathrm{g}}]_+ - [W^{\mathrm{s/m-}} y^{\mathrm{s}} + W^{\mathrm{g/m-}} y^{\mathrm{g}}]_+. \tag{4.5.16}$$

Plug in the definition of the weights, we get

$$[(W^{\mathrm{e/s}}\Gamma)^\dagger (y^{\mathrm{g}} - y^{\mathrm{s}})]_+ - [(W^{\mathrm{e/s}}\Gamma)^\dagger (y^{\mathrm{s}} - y^{\mathrm{g}})]_+ = (W^{e/s}\Gamma)^\dagger (y^{\mathrm{g}} - y^{\mathrm{s}}). \tag{4.5.17}$$

Now we can plug it in the envronmental dynamic

$$\tau_e \frac{dx}{dt} = \Gamma a. \tag{4.5.18}$$

We get

$$\tau_e \frac{dx}{dt} = \Gamma (W^{\mathrm{e/s}}\Gamma)^\dagger (y^{\mathrm{g}} - y^{\mathrm{s}}). \tag{4.5.19}$$

Since $\tau_{\mathrm{e}} \gg \tau_{\mathrm{n}}$, we can assume $y^{\mathrm{g}}, y^{\mathrm{s}}$ are at their fixed point. Notice at fixed point $y^{\mathrm{g}} - y^{\mathrm{s}} = f(x_{\mathrm{g}}) - f(x) = W^{\mathrm{e/s}}(x_g - x)$. We have

$$\tau_e \frac{dx}{dt} = \Gamma (W^{\mathrm{e/s}}\Gamma)^\dagger W^{\mathrm{e/s}} (x_g - x) = x_g - x \tag{4.5.20}$$

as desired. □

4.5.3 Coordination between feedback control on a simplicial complex and planning on the hypergraph

After representing a simplicial complex as an attractor, we project its activity to the hypergraph layer to enable global planning. Precisely, let Σ be the simplicial complex. For each simplex $\sigma \in \Sigma$, let its embedding be $f_\sigma(x) = W_\sigma^{\mathrm{e/s}} x + b_\sigma^{\mathrm{e/s}}$, and let I_σ denote the index set of neurons active on σ. The hypergraph state is $z^{\mathrm{S}} \in \mathbb{R}^{|\Sigma|}$. For each $\sigma \in \Sigma$, note that

$$f_\sigma(\sigma) = \{\, y : \ \langle y,\ \hat{b}_\sigma^{\mathrm{e/s}} \rangle = \|\hat{b}_\sigma^{\mathrm{e/s}}\|_2^2,\ y \geq 0 \,\},$$

and thus $(\hat{b}_\sigma^{\mathrm{e/s}})^\dagger y = 1$ for $y \in f_\sigma(\sigma)$. So we define the readout

$$\tau_{\mathrm{n}} \frac{dz_\sigma^{\mathrm{S}}}{dt} \ = \ - z_\sigma^{\mathrm{S}} + \Theta\big((\hat{b}_\sigma^{\mathrm{e/s}})^\dagger y_{I_\sigma}^{\mathrm{s}}\big), \qquad \Theta(x) = \mathbf{1}[x \geq 1]\ x. \tag{4.5.21}$$

In this scheme the hypergraph input is effectively 1 when y lies on the corresponding simplex and 0 otherwise.

We then apply the planning circuit from the previous chapter on the hypergraph. The only modification is the action output. Let $z^{\mathrm{N}} \in \mathbb{R}^{|\Sigma|}$ represent the selection of the next simplex. Suppose the current simplex is σ and σ' is selected as the next target; the controller steers toward the midpoint of the intersection $\sigma \cap \sigma'$. Concretely,

$$y^{\mathrm{M+}} = \left[\sum_\sigma \sum_{\sigma'} - (W^{\mathrm{e/s}}\Gamma)^\dagger M_\sigma\, z_\sigma^{\mathrm{S}} + \ (W^{\mathrm{e/s}}\Gamma)^\dagger M_{\sigma\sigma'}\, z_\sigma^{\mathrm{S}} z_{\sigma'}^{\mathrm{N}} \right]_+, \tag{4.5.22}$$

$$y^{\mathrm{M}-} = \left[\sum_{\sigma} \sum_{\sigma'} \; (W^{\mathrm{e/s}}\Gamma)^{\dagger} M_{\sigma} \, z_{\sigma}^{\mathrm{S}} - \; (W^{\mathrm{e/s}}\Gamma)^{\dagger} M_{\sigma\sigma'} \, z_{\sigma}^{\mathrm{S}} z_{\sigma'}^{\mathrm{N}} \right]_{+} . \tag{4.5.23}$$

Here, M_σ denotes the center of simplex σ and, in barycentric coordinates, $M_\sigma = [0, \frac{1}{d+1}, \ldots, \frac{1}{d+1}]$ for $d = \dim \sigma$; $M_{\sigma\sigma'}$ denotes the center of the face $\sigma \cap \sigma'$. The net motor command is

$$a \;=\; y^{\mathrm{M}+} - y^{\mathrm{M}-} + y^{\mathrm{m}+} - y^{\mathrm{m}-}.$$

When the model is already in the goal simplex (i.e. $\sigma = \sigma'$), the face-directed term cancels, $y^{\mathrm{M}+} - y^{\mathrm{M}-} = 0$, and the local feedback controller $y^{\mathrm{m}+} - y^{\mathrm{m}-}$ acts alone. Thus, in the combined circuit, the planning module drives transitions between simplices until the goal simplex is reached, at which point the intra-simplex controller takes over to reach the final goal within the simplex.

References

[1] E. H. Nieh, M. Schottdorf, N. W. Freeman, R. J. Low, S. Lewallen, S. A. Koay, L. Pinto, J. L. Gauthier, C. D. Brody, and D. W. Tank. "Geometry of abstract learned knowledge in the hippocampus". In: *Nature* 595.7865 (June 2021), pp. 80–84.

[2] R. J. Gardner, E. Hermansen, M. Pachitariu, Y. Burak, N. A. Baas, B. A. Dunn, M.-B. Moser, and E. I. Moser. "Toroidal topology of population activity in grid cells". In: *Nature* 602.7895 (Jan. 2022), pp. 123–128.

[3] J. D. Seelig and V. Jayaraman. "Neural dynamics for landmark orientation and angular path integration". In: *Nature* 521.7551 (May 2015), pp. 186–191.

[4] N. W. Schuck and Y. Niv. "Sequential replay of nonspatial task states in the human hippocampus". In: *Science* 364.6447 (June 2019).

[5] R. M. Tavares, A. Mendelsohn, Y. Grossman, C. H. Williams, M. Shapiro, Y. Trope, and D. Schiller. "A Map for Social Navigation in the Human Brain". In: *Neuron* 87.1 (July 2015), pp. 231–243.

[6] S. A. Park, D. S. Miller, H. Nili, C. Ranganath, and E. D. Boorman. "Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps". In: *Neuron* 107.6 (Sept. 2020), 1226–1238.e8.

[7] J. C. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. Behrens. "The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation". In: *Cell* 183.5 (Nov. 2020), 1249–1263.e23.

[8] B. E. Pfeiffer and D. J. Foster. "Hippocampal place-cell sequences depict future paths to remembered goals". In: *Nature* 497.7447 (Apr. 2013), pp. 74–79.

[9] J. L. Gauthier and D. W. Tank. "A Dedicated Population for Reward Coding in the Hippocampus". In: *Neuron* 99.1 (July 2018), 179–193.e7.

[10] E. R. Wood, P. A. Dudchenko, R. Robitsek, and H. Eichenbaum. "Hippocampal Neurons Encode Information about Different Types of Memory Episodes Occurring in the Same Location". In: *Neuron* 27.3 (Sept. 2000), pp. 623–633.

[11] P. Mussells Pires, L. Zhang, V. Parache, L. F. Abbott, and G. Maimon. "Converting an allocentric goal into an egocentric steering signal". In: *Nature* 626.8000 (Feb. 2024), pp. 808–818.

[12] B. K. Hulse et al. “A connectome of the Drosophila central complex reveals network motifs suitable for flexible navigation and context-dependent action selection”. In: *eLife* 10 (Oct. 2021).

[13] D. Bush, C. Barry, D. Manson, and N. Burgess. “Using Grid Cells for Navigation”. In: *Neuron* 87.3 (Aug. 2015), pp. 507–520.

Chapter 5

Concluding Remarks

In this thesis, we study how animals learn environmental structure and utilize it for planning and inference to generate flexible behavior. Whether we are planning a route through a familiar city or navigating an awkward social exchange, flexibility depends on more than stimulus-response associations. It requires planning over internal models that capture the structure of the environment and support generalization under changing demands. Throughout the thesis, we develop circuit-level accounts of flexible behavior spanning three regimes: decision making under latent contexts through cortico-thalamic interactions between PFC and MD, hierarchical navigation supported by hippocampal replay and downstream computations in OFC and ACC, and navigation on continuous manifolds that coordinate local control with global routing in a manner parallel to hippocampal navigation.

Across these settings, a shared principle emerges. Slow plasticity learns structure in a form that can be reused across episodes, while fast neural dynamics plan and infer over that structure to select actions under changing demands. This interplay between timescales is not simply a modeling convenience. It is a candidate organizing principle for biological cognition. Plasticity is mediated by slow biochemical processes and often depends on protein synthesis, making it well suited for consolidating internal models that persist over time. Neural dynamics, by contrast, provide rapid computation through fast electrical signaling, enabling flexible behavior on the timescale of decisions. This perspective also shifts the unit of explanation from single neurons to interacting circuits operating across timescales. The computations studied here arise from coordinated population dynamics shaped by slow plasticity.

Neuroscience is, in many ways, the study of emergent properties. Individual neurons are wired into specialized circuits with distinct computational roles, and cognition arises from their coordinated interactions. When parts of this hardware break down, the resulting circuit-level disruptions can manifest as psychiatric disease. A central challenge for the field is to connect mechanistic descriptions of neural implementation to algorithmic accounts of cognition, and to understand how perturbations at molecular, genetic, or circuit levels cascade into failures of cognitive function that appear as clinical symptoms.

A unifying path forward lies in tightening the loop between theory and experiment. The models in this thesis are intended as proofs of principle that mechanistic circuit models can be both biologically grounded and computationally explicit. By linking circuit elements to specific computations, these models make predictions about how cognitive variables should

be represented and computed, how these representations influence downstream behavior, and which failure modes should arise under targeted perturbations. They also provide a natural framework for computational psychiatry, since the same perturbations can be applied in models and in animals or patients to compare behavioral changes, neural dynamics, and breakdown signatures. In my future work, I plan to work closely with experimental collaborators and multimodal measurements, using theoretical tools to bridge circuit mechanisms, cognitive function, and symptom expression within a single coherent framework. More broadly, I hope this thesis contributes to a growing style of theoretical neuroscience in which mechanistic circuit models do not merely fit behavior, but serve as interpretable computational blueprints for how brains achieve cognition in a complex world.