Title

Revisiting The Actor-Critic System in Striatum during Decision Making

ISBN

9798297601062

Author

Qu, Albert Jiaxu

Publication Date

2025-08-01

Peer reviewed|Thesis/dissertation

Revisiting The Actor-Critic System in Striatum during Decision Making

By

Albert Jiaxu Qu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Behavioral and Systems Neuroscience, Psychology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Linda Wilbrecht, Chair
Professor Giles Hooker
Professor Anne Collins
Professor Michael Yartsev

Summer 2025

Revisiting The Actor-Critic System in Striatum during Decision Making

Abstract

Revisiting The Actor-Critic System in Striatum during Decision Making

By

Albert Jiaxu Qu

Doctor of Philosophy in Behavioral and Systems Neuroscience, Psychology

and the Designated Emphasis in

Computational and Genomic Biology

University of California, Berkeley

Professor Linda Wilbrecht, Chair

The actor-critic framework has provided a foundational model of reinforcement learning in biological systems. In one of the most popular version of the model, the dorsal striatum is thought to function as the "actor" implementing action selection, while the ventral striatum functions as the "critic" for credit assignment. While this model is highly successful, experimental evidence increasingly suggests that the computational and neural mechanisms underlying decision making may contain additional nuance or even extend beyond this classic framework. In this thesis, I present experimental and computational work that serves to refine our understanding of both the critic and the actor, based on recordings made in the striatal circuits of mice during decision making.

In the first study, I focused on the critic. Here, I investigated whether dopamine release in the nucleus accumbens reflects Bayesian inference computations beyond simple temporal difference learning. Using fiber photometry to measure dopamine signals during a probabilistic switching task, I compared predictions from Bayesian inference models against standard reinforcement learning models. Model comparison revealed that both mouse behavior and nucleus accumbens dopamine signals were better explained by models incorporating Bayesian inference for hidden state estimation, suggesting that the "critic" system performs sophisticated probabilistic computations rather than simple iterative value updates.

In the second study, I focused on the actor. Here, I examined how the dorsal striatum implements active choice rejection, a process that challenges traditional models of action selection. Using bilateral recordings from direct and indirect pathway neurons during a serial decision-making task (Restaurant Row), I discovered that active rejection decisions involve distinct opponency patterns compared to active accepting decisions. Remarkably,

rejection choices showed significant ipsilateral suppression, while acceptance choices showed classic contralateral activation. Optogenetic manipulations confirmed the causal role of this asymmetric activity pattern, with unilateral suppression of ipsilateral direct pathway neurons specifically enhancing rejection of marginally valuable offers.

To conclude, I will discuss how these findings might be used to revise and update the actor-critic framework. I propose that the neural model of the "critic" component be readily expanded to include the possibility of Bayesian inference mechanisms that enable rapid adaptation to environmental volatility through belief state representations. Then, I suggest that the neural model of the "actor" system can also be improved by including hemispheric opponency, while also accounting for hemispheric specialization in cost and benefit processing, when lateralized choice is being implemented. Together, these results demonstrate that biological decision making involves computational sophistication and circuit complexity that extends well beyond traditional reinforcement learning models, providing new insights into the neural basis of adaptive behavior and offering refined frameworks for understanding decision-making disorders.

To my wife Alexandra, my parents, and my friends

For your never-ceasing support of my PhD and research, and your sacrifices made along the line; for standing behind me throughout every hardship and joyful moments.

Contents

List of Figures

List of Tables

Acknowledgments

I thank Prof. Linda Wilbrecht, Prof. Giles Hooker, Prof. Anne Collins, Prof. Michael Yartsev, Prof. Samuel Gershman, Prof. Andrew MacAskill, Prof. Scott Linderman, Prof. Nuria Vendrell Llopis, Prof. Jose Carmena, Prof. William Fithian for their mentorship and guidance. I thank Dr. Lung-Hao Tai, Dr. Wan Chen Lin, Prof. Thomas Elston, Dr. Jacki Essig, Eric Hu, Jingjing Li, Lexi Zhou, Allison Nieto, Dr. Karyna Mischanchuk, and Christopher Hall, Dr. Juliana Chase, Dr. Madeline Klinger, Samantha Jackson, Cameron Jordan, Yang Zhang, Yilan Lin for many helpful discussions and collaboration, from whom my research has significantly benefited from. I also thank my friends Alexander, Emir, Fiona, Sandra, Danny, Alejandro, Katharine, Ambrose, Elisa, Austin, Jane for their moral support throughout my PhD years.

Chapter 1

Introduction

Reinforcement Learning (RL) has long served as an influential framework for understanding decision making in biological systems [1, 2]. RL can be instantiated in a number of ways, but one particularly influential and biologically plausible RL framework is the Actor-Critic model [3]. In the classic Actor-Critic model, the brain forms a summary representation of the environment, or **state**. Upon receiving feedback from the environment, the critic computes state value associations and evaluates reward prediction errors (RPE) to update the internal model, whereas the actor model uses the feedback signals from the environment to adjust policies, through mechanisms like RPE-scaled weight updates or policy gradient (a computational algorithm that updates policy towards the direction of increasing value) [3].

In the 1990s, neurophysiologists made the seminal discovery that midbrain dopamine neurons (DANs) increase their firing rate when outcome exceeds expectation and decrease their activities when expected rewards are omitted, essentially mimicking temporal difference errors [4]. This provided critical evidence supporting RL as a candidate mechanism for decision making in the living brain. The dopaminergic RPE signals are thought to update synaptic weights in their target release sites, including the striatum and frontal cortex, thereby reinforcing actions instrumental for obtaining the reward. Moreover, neurophysiological, anatomical and behavioral work over the years have consistently established striatal subregions in the basal ganglia as likely candidate hubs for the actor and critic systems [1,5,6], where ventral striatum serves as a critic component evaluating RPE, and dorsal striatum serves as an actor component responsible for action selection. This basic actor-critic of the basal ganglia elegantly explains how organisms learn from trial and error and aligns well with a wide range of experimental observation in reward learning across rodents, primates and human data [5, 7, 8].

Although successful, the classic actor-critic framework may still benefit from updates. Classic models for the critic systems like temporal difference (TD) learning assume that the organism reliably captures the state information of the environment. However, natural environments are rarely fully observable or static. Animals must deal with partial observability of "true" or underlying state of their environment and must infer them from ambiguous cues and exploration [9]. Also, real-world learning involves not only maximizing reward but

also avoiding cost, which requires integrating positive and negative value signals (possibly at the level of the striatum). Moreover, actor-critic RL models provide an algorithm layer description, but to fully capture the biological implementation level complexity we need to extend this framework to account for complexity of real neural circuits.

Here in this PhD thesis, I will discuss work that helps update and extend the critic framework in Chapter 2, the actor framework in Chapter 3, and a selection of statistical methods that made these discoveries possible in Chapter 4. But first, I will more thoroughly define the formal implementation of the actor-critic model and what is known about the neural architecture of the basal ganglia and its role in decision making.

1.1 The Actor-Critic Model

The actor-critic architecture represents one of the most influential reinforcement learning frameworks in both machine learning and computational neuroscience, providing a powerful model for understanding how biological systems learn to make optimal decisions through experience. First formalized by [10], this framework decomposes the reinforcement learning problem into two interacting components: an "actor" that selects actions and a "critic" that evaluates those actions.

Intuitively, the critic estimates the value function $V_\pi(s)$ or action value function $q_\pi(s, a)$, which represents the expected future reward from a given state (with given action in context of action value). It can potentially use temporal difference learning to compute a prediction error—the difference between actual and expected rewards—which serves as a teaching signal to update these value functions. The actor maintains a policy function that maps states to actions, and updates this policy based on the critic's evaluation using the policy gradient. This separation of policy and value estimation offers significant advantages, allowing for more stable learning in continuous action spaces and facilitating the learning of stochastic policies [3, 11].

In its computational formulation, the actor-critic model stems from the rich literature of Markov Decision Process (MDP) and Reinforcement Learning (RL) [3]. In a classic MDP framework, we assume that the environment surrounding an agent at any given time point can be summarized by a mathematical abstraction, or the "state" variable (s). At each state, an agent could take a list of possible of actions $a \in \mathcal{A}$, where $\mathcal{A}$ denotes this action set. After taking an action, the agent may receive a reward r, as well as transitioning to a new state s'. One example that illustrates this well is the time when I was learning to drive as a teenager. At each time point, I observed the road conditions (other vehicles, pedestrians, obstacles, etc.), my location, traffic signs, and many other vehicle variables like speed and steering wheel positions. All these sources of information can be encapsulated into a state variable s. The action set $\mathcal{A}$ may include pressing the gas pedal, pressing the brake, using turn signals, or steering the wheel leftwards or rightwards. At one point, when I accelerated and turned to change lanes, my state changed–my speed increased, lane position changed, and the relative

position of other vehicles may change as a result as well. During this transition, my coach told me "good job", which was rewarding to me as a student driver.

Another important aspect that can be seen from the above example is that the state transitions can be stochastic, or non-deterministic. For instance, as I drive towards an intersection, the green light could suddenly turn yellow, or a vehicle could be merging lanes in front of me. Each of these individual state variable changes may have their own statistical distribution. In the MDP framework, we make one convenient assumption that these transitions only depend on the current state we are in as well as the actions taken. This is largely suitable for a lot of applications, given enough variables are captured in the state representation. In the driving example, my vehicle position and speed, as well as other state variables, largely transition smoothly in an autoregressive fashion from the previous state. To incorporate this into the MDP framework, we use $P(s'|s,a)$ to represent the transition probability from state s to s' after taking action a. Reinforcement Learning models then naturally describe how an agent learns the optimal action policy (a mapping from each state to actions taken) by maximizing the reward r received from the system.

In this framework, canonically, we assume that each state has value $V_\pi(s)$ under an agent's action policy $\pi(a|s)$, where

$$\begin{aligned} V_\pi(s) &= \sum_a \pi(a|s) \sum_{s'} [P(s'|s,a)(r + \gamma V_\pi(s'))] \\ &= \sum_a \pi(a|s) Q(s,a) \end{aligned} \tag{1.1}$$

We can interpret the state value as a weighted value under the action policy π by considering the action probabilities and the value of each action $Q(s,a)$, Since $Q(s,a)$ can be viewed as an approximated value of taking action a at state s, in practice, we could estimate it using the random samples collected during iterations of navigating the learning environment via reward prediction errors $\delta_t = r - Q_t(s,a)$:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha(r - Q_t(s,a)) \tag{1.2}$$

Here α denotes the learning rate for the agent's learning process. Q-learning can be viewed as the simplest form of the critic learning, whereas the brain could adopt a more complex form of learning to approximate this Q-value function through environmental feedback.

Action selection under the Q-learning framework usually involves a simple transformation from Q values to choice probabilities via a softmax function (which reduces to the sigmoid function for binary decisions): $\pi(a_t = a|s) = \frac{e^{\beta Q(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\beta Q(s,a')}}$, where β is a hyperparameter controlling the exploitation-exploration tradeoff. The actor framework then expands on this to model arbitrary functional form in an action policy, and directly optimizes its policy function parameters to maximize rewards and perform action selection. Its computational framework naturally comes from policy gradient calculation. Assume that an agent operates using policy $\pi(a|s)$, we can readily compute value of a state S_t as $V_\pi(S_t) = \sum_a \pi(a|s) q_\pi(s,a)$ as above, where now we use $q_\pi(s,a)$ to distinctly denote action value under policy π. Now

if we take the gradient of the value function at the initial state S_0, assuming that the policy function is parameterized by θ, we get the policy gradient $J(\theta) = \nabla V_{\pi_\theta}(S_t)$. After some derivation, we can find that for a stochastic sample:

$$\begin{aligned}\nabla J(\theta) &\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\theta) \\ &= E_\pi[\sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\theta)] \\ &= E_\pi[\sum_a \pi(a|S_t,\theta) q_\pi(S_t,a) \nabla \log(\pi(a|S_t,\theta))] \end{aligned} \tag{1.3}$$

$$\text{Given: } \mu(s) = \frac{\eta_s}{\sum_{s'} \eta_{s'}},\ \eta_s = P_\pi(s_0 \to s) \tag{1.4}$$

Taking stochastic sample we have

$$J(\theta) \simeq q_\pi(S_t, A_t) \nabla \pi(A_t|S_t,\theta) \tag{1.5}$$

Note that here $q_\pi(S_t, A_t)$ refers to the action value Q-value function learned by the critic component. This allows the actor to utilize the critic information effectively to update its policy.

The biological plausibility of actor-critic models has made them particularly influential in neuroscience. The architecture maps naturally onto the organization of cortico-basal ganglia circuits, with the ventral striatum (particularly the nucleus accumbens) functioning as the critic and the dorsal striatum as the actor [1, 12]. This mapping is supported by a wealth of evidence: midbrain dopamine neurons signal reward prediction errors consistent with temporal difference (TD) learning [4, 13]; ventral striatal neurons evaluate outcomes and perform credit assignment integrating information from dopaminergic neurons [14, 15]; and dorsal striatal neurons implement action selection based on learned values [16–18].

Moreover, the actor-critic framework, or reinforcement learning framework in general, has made tremendous contribution to the machine learning field, from robotics to natural language processing. For instance, methods derived from classic actor-critic methods, like Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) [19, 20] has fueled advancement in Deep RL research and led to successful applications in Atari games, and robotic control. Moreover, in recent years, advanced Large Language Model (LLM) agents like GPTs greatly benefited from actor-critic models like PPO and Group Relative Policy Optimization (GRPO) in generating responses aligned with human preferences [21, 22]. Specifically, these models computed "advantages", which are generalized forms of value functions that measures how certain actions or trajectories provide significant improvement in rewards. And then they optimize the policy network subjected to constraints that limit significant perturbations from existing policy trajectories.

Through continuous refinement and integration with experimental findings, the actor-critic framework has evolved from a computational algorithm to a powerful conceptual model that bridges artificial intelligence and neuroscience. It greatly facilitates advancement in machine learning networks and provides a unifying perspective on how diverse neural circuits

coordinate to enable adaptive decision making in complex, dynamic environments, cementing its position as a cornerstone of computational approaches to understanding the biological basis of learning and decision making. Based on its current utility, improvements in the actor-critic framework could be valuable to both neurobiology and machine learning.

1.2 Neural Architecture of the Basal Ganglia and Its Role in Decision Making

Organization of the Basal Ganglia Circuitry

The basal ganglia form a group of subcortical nuclei that coordinate with cortical, thalamic and midbrain regions to support a variety of functions including action selection, motor control, reinforcement learning, and habit formation (Fig 1.1) [23, 24]. This network is organized into parallel circuits that connect cortex, basal ganglia, thalamus and midbrain output regions like deep superior colliculus (SC) and mesencephalic locomotor area (MLR), creating a circuit architecture that supports the integration of cognitive, motor, and limbic information for adaptive decision-making [25–27].

The canonical structure of each cortico-basal ganglia circuit begins with excitatory glutamatergic projections from cortical regions to the striatum, the primary input nucleus of the basal ganglia. The striatum can be divided into functional territories: the dorsolateral striatum (DLS) primarily receives input from sensorimotor cortices, the dorsomedial striatum (DMS) from associative cortices, and the ventral striatum (including nucleus accumbens) from limbic regions [28, 29].

From the striatum, information flows through the basal ganglia via two major output nuclei: the internal segment of the globus pallidus (GPi) and the substantia nigra pars reticulata (SNr) (Fig 1.1). These structures provide tonic inhibitory input to the thalamus, deep colliculus, and mesencephalic locomotor region (MLR) effectively gating basal ganglia outputs [27, 30–32]. The thalamus, in turn, sends excitatory projections back to the cortex, completing the loop. This arrangement creates a system where basal ganglia output can either facilitate or suppress output through disinhibition or maintenance of inhibition, respectively [30].

As mentioned above, this architecture is repeated in multiple parallel circuits that help to maintain partial segregation of computations while also possibly allowing for integration across functional domains. Tracing studies have identified at least five major functional circuits in primates defined by their cortical input origin: motor, oculomotor, dorsolateral prefrontal, lateral orbitofrontal, and anterior cingulate loops [33,34]. These loops process distinct information but as they converge at specific points, they may allowing for coordination across cognitive, motivational, and motor aspects of behavior [26].

Detailed anatomical studies have revealed additional complexity in this circuitry [35], including “hyper-direct” projections from the cortex to the subthalamic nucleus (STN) [36],

thalamostriatal projections that provide feedback within the loop [37], and arkypallidal inputs that project up from the to GP influences the striatum [38]. These findings have expanded our understanding of how information flows through the basal ganglia beyond the classical model.

The functional significance of this architecture lies in its ability to perform action selection through competitive dynamics. As proposed by [39], the basal ganglia can be viewed as a central selection mechanism that resolves conflicts between competing behavioral options. By integrating diverse inputs and controlling multi-targeted outputs, this circuit can effectively "decide" which actions to facilitate and which to suppress based on current goals and context [39–41].

Direct and Indirect Pathways: Parallel Streams for Action Selection

Within the cortico-basal ganglia circuitry, information processing is further organized into at least two major pathways distinguished by their connectivity, receptor expression, and functional roles: the direct (or "Go") pathway and the indirect (or "No-Go") pathway [32, 42, 43].

The direct pathway originates from striatal spiny projection neurons (SPNs) that predominantly express D1-type dopamine receptors (D1-SPNs). These neurons project directly to the basal ganglia output nuclei (GPi/SNr) and form inhibitory GABAergic synapses. Activation of the direct pathway thus inhibits the tonically active GPi/SNr neurons, resulting in disinhibition of thalamic targets and outputs through the MLR or colliculus, or via facilitation of cortical activity [27, 30, 44]. Consistent with this circuit arrangement, selective activation of D1-SPNs facilitates movement initiation and reinforces recently performed actions ([17, 27, 42]).

The indirect pathway, in contrast, arises from striatal SPNs that predominantly express D2-type dopamine receptors (D2-SPNs). These neurons project to the external segment of the globus pallidus (GPe), which in turn inhibits the subthalamic nucleus (STN) (Fig 1.1). The STN provides excitatory input to the GPi/SNr, completing the indirect pathway. Activation of D2-SPNs thus inhibits GPe, which disinhibits STN, leading to increased excitation of GPi/SNr blocking disinhibition at basal ganglia output targets [30, 45]. Selective activation of D2-SPNs suppresses movement [42, 46] and can drive avoidance behaviors [47].

Dopamine differentially modulates these pathways through distinct effects on D1 and D2 receptors. D1 receptor activation is thought to facilitate direct pathway activity through increased excitability and synaptic potentiation downstream of PKA activation, while D2 receptor activation is thought to suppress indirect pathway activity through supression of PKA activation [30, 48, 49]. This arrangement creates a system where dopamine release can modulate the balance between the two pathways to facilitate or suppress specific actions [50, 51].

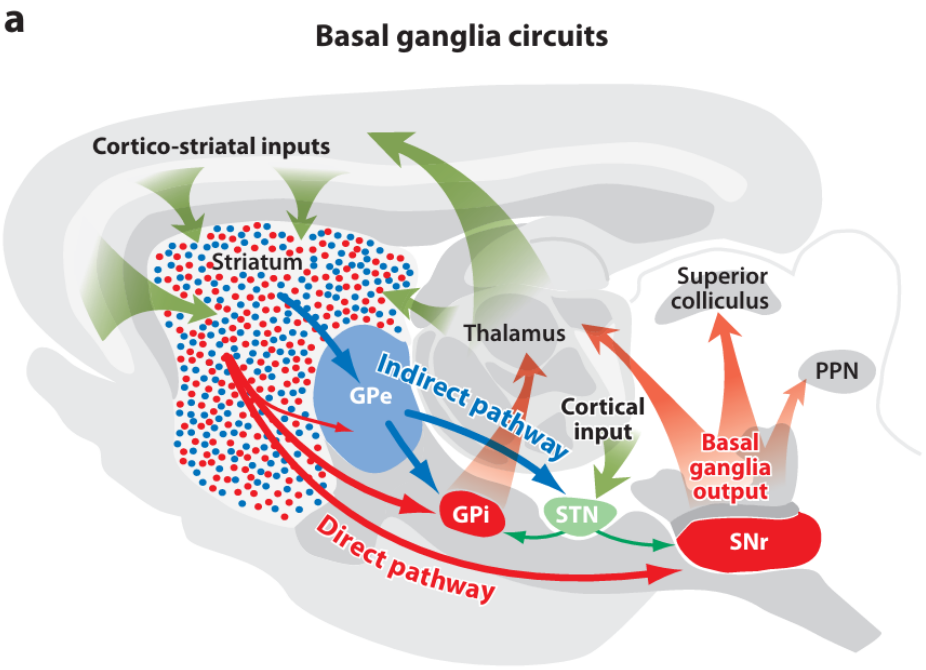

Figure 1.1: This schematic showcases the biological architecture of the direct and indirect pathways in the cortico-basal ganglia-thalamic loop, with illustrations of dopamine connections. (Adapted from Gerfen and Surmier 2011 [30])

While this classical model of direct and indirect pathways has provided a valuable framework, recent research has revealed additional complexity. Single-cell recording studies have shown that both D1-SPNs and D2-SPNs can be co-activated during movement initiation [52, 53], challenging the simple "Go/No-Go" dichotomy. Furthermore, anatomical and functional studies have identified substantial collateralization in D1-SPN projections and D2-SPNs [54], as well as a more intergrated architecture in ventral striatum [55, 56].

Beyond their role in movement control, the direct and indirect pathways contribute to reinforcement learning and decision making. The direct pathway promotes actions associated with positive outcomes, while the indirect pathway suppresses actions associated with negative outcomes, creating a dual-operator system for value-based selection [50, 51]. This functional organization is supported by the differential effects of dopamine on synaptic plasticity in each pathway: dopamine facilitates long-term potentiation in D1-SPNs and long-term depression in D2-SPNs [48, 49] likely through intracellular signaling downstream of PKA [48, 49].

Longstanding and recent work have also identified hemispheric specialization in pathway function. Bilateral recordings during decision tasks show that both pathways are activated in the hemisphere contralateral to movement, but with different temporal profiles [52]. The traditional model of competition between direct and indirect pathways is often only discussed

within a hemisphere, but given contralateral/ipsilateral opponency between hemispheres during orienting behaviors [39,46,57], it must also be expanded to incorporate bihemispheric coordination [58,59]. My thesis work to be discussed in detail below will expand opponency further by revealing that active accept and reject choices involve differential patterns of hemispheric opponency (see Chapter 3).

The direct and indirect pathways are found to be intermixed through all of the dorsal striatum but the larger dorsal striatal region itself can be subdivided into medial and lateral subregions with different function. In the dorsomedial striatum (DMS), the two pathways are known to regulate goal-directed actions [60,61], while in the dorsolateral striatum (DLS), the two pathways are thought to control habitual motor responses. In the ventral striatum, the D1-SPNs and D2-SPNs are not called the direct and indirect pathway because they are not as well segregated in terms of gene expression and projection target [56, 62], Here the two cell types are thought to contribute to appetitive and aversive learning, with the D1 expressing SPNs supporting approach behaviors and/or reinforcement and the D2 expressing SPNs contributing to avoidance [47] but also under some contexts positive reinforcement as well [63].

This organizational complexity enables the direct and indirect pathways (and ventral striatal D1-SPNs and D2-SPNs) to support a range of functions beyond simple action facilitation and suppression. Their coordinated activity, modulated by dopamine and shaped by experience-dependent plasticity, allows for adaptive selection among competing options based on expected value, contextual factors, and internal goals. Understanding how these pathways interact across hemispheres and functional territories is essential for developing comprehensive models of basal ganglia function in decision making.

1.3 Bayesian Inference As a Mechanism in Credit Assignment

One critical problem in decision making for human and animals is to integrate noisy sensory and past outcome information, infer the underlying rules of the environment, and generate actions to achieve their desired outcome. Simple perceptual decision making paradigms are informative in how sensory information is processed to generate decisions, but do not provide us with insights into how animals use past outcome experience in volatile environments where reward structures are constantly changing despite similar sensory evidence. For instance, when customers walk into their favorite restaurant, they often see a similar set of dish names and pictures. Occasionally there may be personnel changes in the chef team or changes in produce supplies that may change the food quality. These underlying changes in the kitchen team are unbeknownst to the customers, and are described here formally as the hidden or latent state of the data generating process, with the data being the food quality of the dishes. After a few observations, the brain needs to integrate these sources of information to determine whether the bad food quality was simply an unlucky occurrence or a fundamental

change in food quality due to personnel or management change. In general, from animals surviving in nature, to humans making life decisions, the ability to perform latent state inference and generate decisions that adapt to these hidden state changes are critical. A natural question arises: does the brain indeed carry out latent state inference computations? And what are the brain areas and mechanisms involved that integrate information from past outcomes and utilize them to generate decisions?

Bayesian statistics provides a powerful framework for latent state inference by integrating new observations into calculating a posterior belief distribution of latent states following Bayes' rule [64, 65]. Previous work has shown the feasibility for neurons to generate a probabilistic code for Bayesian inference [66]. Moreover, signatures of neural activity during perceptual decision making in areas like lateral intraparietal cortex (LIP) are found to be consistent with models using Bayesian inference computations [67, 68]. Behavioral modeling for human performing tasks that involve changing underlying reward structures suggested the plausibility of the brain using Bayesian inference to integrate outcome history information for future decision making [8].

Recent modeling work has suggested that Recurrent Neural Networks trained with a TD learning framework can iteratively integrate choice outcomes to produce belief-like representations [69]. A growing set of studies have begun to describe how cortical population codes could carry out Bayesian inference computations [66, 67, 70] and that hippocampus neurons could represent the contextual cognitive map and belief during decision making [71–73]. Consequently, it is possible to envision a theoretical neural model where integrations of distinct neural clusters could carry out the Bayes' rule propagation and update neural population codes encoding beliefs. After each outcome revelation, Bayesian inference and belief codes calculated and updated in the hippocampal-prefrontal circuits [73, 74] could be propagated via projections to the striatum and ventral tegmental area (VTA). Dopamine error feedback can then either act via striatal pathways or directly through limbic and mesocortical dopaminergic pathways to cause changes in cortical or hippocampal dynamics, leading to further belief updating. Interestingly, recent experimental observations on dopaminergic circuit functions during decision making have found evidence to support this model. [75] demonstrated in an inference task that substantia nigra pars compacta (SNc) dopamine neuron activities qualitatively reflect their ability to encode both experienced target values, as well as inferred values on unseen targets. [76] showed that manipulation of dopamine neurons in monkeys altered reversal behaviors that require inferences. [9, 77] carefully designed distinct Pavlovian paradigms and showed that VTA dopamine activity in mice is consistent with RPE predictions from a Bayesian inference based reinforcement learning model (BRL). Our work in Chapter 2 [78] then elucidated how dopamine release in mouse nucleus accumbens during instrumental tasks closely matched the RPE prediction of a hybrid model combining Bayesian inference and reinforcement learning, rather than other complex iterative RL processes alone. Moreover, recent work have also shown that the hidden state inference requires coordination between basal ganglia and other brain areas like hippocampus and frontal cortex to enable flexible adaptive behaviors [71, 79]. Future work could further distinguish between the dorsal and ventral striatal computations by comparing BRL latent and

RPE predictions with regional SPN activity and dopamine responses.

Understanding if the brain performs hidden state inference can provide answers about feasible neural mechanisms the brain may use to solve credit assignment problems. In Chapter 2, I will discuss this problem in further detail in a study where we recorded dopamine in nucleus accumbens and investigated if Bayesian inference models can effectively explain the behavioral and neural variance in mice as they make decisions. Specifically, we investigated how mice adapt to changes in the hidden reward structure in an instrumental task. We found evidence that models using Bayesian inference provided a better account of both behavioral data and dopamine signals in the nucleus accumbens (NAc) for mice compared to standard Reinforcement Learning (RL) models, even RL models that included sophisticated features such as counterfactuals, forgetting, and dynamic learning rate update. Moreover, we discovered that it is biologically plausible that mice employed a hybrid computational process that combined RL and Bayesian inference. Under this hypothesis, Bayesian inference is used to compute beliefs across hidden states, each of which contains its own action value map. Dopamine signals reward prediction errors that update these action values conditioned on these belief states. This work establishes a role for Bayesian inference in neural computation and behavior in a value-based decision making task, and helps to confirm a new understanding of the role of dopamine in learning.

1.4 Distinct Circuit Mechanisms in Action Selection

As described in section 1.2, the elegant anatomical structure of dorsal striatum featuring opposing pathways with distinct cell types, as well as rich connections to broad sensory, motor, and high-order regions, allows itself to play a critical role in decision making, especially value-based decision making [31]. The traditional actor-critic model characterizes the dorsal striatum as implementing action selection through the direct (facilitating actions) and indirect (suppressing actions) pathways [46, 47, 62]. This anatomical organization supports the "actor" role by translating value representations into appropriate behavioral responses.

However, when we consider continuous ongoing activity and the competition between two hemispheres, we realize this simple model may require more nuances. First, it has been observed that direct pathway and indirect pathway neurons are often co-activated during action initiation [52, 80]. This implies that the two pathways must operate in a multi-dimensional fashion instead of simply regulating single actions, since they otherwise would cancel each other out. It has been suggested that one action is selected while other competing actions are suppressed. In a study of the dorsolateral striatum (DLS), [46] found that when animals were covertly inhibiting themselves from making leftward and then rightward actions over time, iSPN activity was higher in the hemisphere contralateral to the currently latent/salient, but inhibited action. This cleverly revealed hemispheric opponency in the suppressive control of leftward or rightward orienting behaviors. Despite a clear role for iSPNs in action suppression in [46] we cannot assume that iSPNs also play a role when suppression of a choice involves generating an alternative action rather than simply withholding motor action. [81]

found that chemogenetic inhibition of DMS iSPNs did not disrupt learned suppression of a salient foraging choice, but DMS dSPN inhibition and DMS iSPN excitation did disrupt this learning [81]. This result led us to question the mapping of the Go/No Go-direct/indirect heuristic in striatal action models. One might automatically expect that if animals are learning to generate actions to avoid costs it should be controlled positive activity in the iSPNs due to their opponent role in the circuit and expression of dopamine D2 receptors (D2Rs). However, it also makes sense that in situations where costs should lead to active rejection actions like running and turning, not holding still, that these computations should instead require dSPN activation which can generate the necessary running or left/right orienting actions [17, 27].

To understand how the dorsal striatum actor contributes to active rejection, in Chapter 3, we investigated the role of dSPNs and iSPNs in the dorsomedial striatum in a serial decision-making task ("Restaurant Row"). This paradigm allowed us to isolate rejection behavior from action selection confounds present in traditional Go/No-go and two-alternative forced-choice tasks.

Thus, in the following chapters, I will explain updates to the critic model in Chapter 2 and updates to the actor model in Chapter 3. In Chapter 4, I will discuss the key computational methodologies that I used in my experimental analyses, which can be leveraged in a general way for studying decision making in the brain. The final chapter will then conclude this thesis and connect the chapters more broadly to the fast growing field of computational neuroscience.

Chapter 2

Dopamine Release in Nucleus Accumbens Signals Reward Prediction Errors Consistent with Bayesian Inference Value Computations

[1]Dopamine release in the nucleus accumbens has been hypothesized to signal the difference between observed and predicted reward, known as reward prediction error, suggesting a biological implementation for reinforcement learning. Rigorous tests of this hypothesis require assumptions about how the brain maps sensory signals to reward predictions, yet this mapping is still poorly understood. In particular, the mapping is non-trivial when sensory signals provide ambiguous information about the hidden state of the environment. Previous work using classical conditioning tasks has suggested that reward predictions are generated conditional on probabilistic beliefs about the hidden state, such that dopamine implicitly reflects these beliefs. Here we test this hypothesis in the context of an instrumental task (a two-armed bandit), where the hidden state switches stochastically. We measured choice behavior and recorded dLight signals that reflect dopamine release in the nucleus accumbens core. Model comparison among a wide set of cognitive models based on the behavioral data favored models that used Bayesian updating of probabilistic beliefs. These same models also quantitatively matched mesolimbic dLight measurements better than non-Bayesian alternatives. We conclude that probabilistic belief computation contributes to instrumental task performance in mice and is reflected in mesolimbic dopamine signaling.

[1]This work is published at PLoS computational biology as [78]

2.1 Background

Reinforcement learning (RL) algorithms hypothesize that, in value-based decision making tasks, animals maintain an action value map and update it using reward prediction errors (RPE), the difference between the observed and predicted reward associated with the chosen action [3]. The growing family of RL models proved to be remarkably successful in explaining trial-by-trial changes in behavior [82–84] and the responses of midbrain dopamine neurons in rodents and primates [4, 85, 86]. However, these standard RL (SRL) models may not adequately account for behavioral and neural data when immediate sensory observations alone are insufficient to guide optimal behavior, for example, when the underlying reward distribution changes over time due to the existence of a hidden state [75, 87, 88]. To solve such tasks, evidence suggests that the brain may represent a "belief state" (a probability distribution over hidden states), updated by Bayesian inference [9,68,76,77,88,89], or possibly learned implicitly in an end-to-end fashion [69]. Additionally, updates to these belief state-dependent action values may be encoded by dopamine neurons [9, 68, 76, 77, 89].

The significance of Bayesian inference becomes apparent when studying behavior in a two-armed bandit task (2ABT, Fig 2.1A), a serial reversal learning task where rewards for correct choices are delivered on a probabilistic schedule, and the correct port switches between two available options across blocks of trials within a single session. The fact that rewards are delivered on a probabilistic schedule generates ambiguity as to which port is currently the rewarded port following an unrewarded trial. Both mice and human subjects trained in this task are capable of maintaining stable behavior within a block, and then rapidly adapting to reward contingency changes following a block switch [8, 17, 90].

The simplest versions of standard RL models do not capture choice stability within a block and rapid switching when the rewarded port changes [8, 90]. Bayesian inference may offer better models of behavior and neural computation in tasks like the 2ABT due to their ability to capture stable within-block choice and rapid between-block switching [76]. A series of studies suggested that dopamine signaling may be better explained by models accounting for belief states [9, 68, 75, 77, 89]. However, the additional explanatory power of Bayesian models with respect to RL models is currently ambiguous, given that a growing family of complex RL models, geared with more parameters and nuanced functions, are capable of generating adaptive behaviors [90, 91]. A 2022 study of human behavior in the 2ABT found that a Bayesian model and an RL model variant equipped with counterfactual value updates both provided complementary explanations of behavior, but neither was decisively superior [8]. A 2022 study of mouse behavior in the 2ABT suggested a purely Bayesian account of reversal learning can overestimate mouse switch probability and fail to account for choice stickiness [90].

In 2024 two further studies were published in mice that compared Bayesian and RL models ability to explain both behavioral and dLight dopamine release data from the nucleus accumbens [71, 88]. Both found that behavior and imaging data were consistent with Bayesian inference. However, in both studies actions and outcomes were spread out over

multiple task steps or multiple seconds, time spans that engage areas such as hippocampus and PFC [71, 88, 92].

There is evidence that when actions and reinforcements follow each other in more fast paced tasks, that the role of striatum is enhanced [17, 93]. So, we cannot assume that computation in a fast paced 2ABT is comparable to a slower paced 2ABT even when both have latent block structure [71]. Therefore, this study provides an additional test of the robustness of Bayesian models in a faster-paced (therefore possibly more striatal based) assay of decision making.

Another limitation of the existing literature is that many studies compared only a few model variants in isolation, without including a larger growing set of models. Recently, Blanco-pozo et al. compared Bayesian models against model-free, model-based, and hybrid RL models in a two-step task [88]. Here we included a broader set of complex RL models commonly used in cognitive modeling to further elucidate how specific mechanisms, like asymmetric learning rate, counterfactual learning, dynamic learning rate, forgetting, and Bayesian inference, each differentially contribute to explaining mouse behavior and dopamine release dynamics in the nucleus accumbens in a simpler operant 2ABT task. In addition to comparing "pure" Bayesian and RL models, we can also examine hybrid models where Bayesian and RL processes are combined [9, 94, 95]. In these belief state RL hybrid models (BRL), Bayesian inference can be used to compute belief states, over which RL processes operate to learn policies appropriate for the current belief state.

Given these ambiguities in the past literature, our goal was to examine if Bayesian models or Bayesian-RL hybrid models could outperform sophisticated variants of RL models to explain behavior and outcome-related dopamine signals in our simple fast-paced 2ABT task with a hidden state. Mesolimbic dopamine neurons have been widely associated with RPE predictions. As one of their major projection destinations, the nucleus accumbens (NAc) is thought to play a critical role in credit assignment [7, 96–98]. Therefore, dopamine release in NAc is of particular scientific interest. To this end, we trained mice in a value-based 2ABT task while recording dLight signals in the NAc using fiber photometry to measure dopamine transients [99]. These data allowed us to test the hypotheses regarding the cognitive process that mice employ in the 2ABT at the behavioral level and to test whether these computations are reflected by mesolimbic dopamine dynamics in the NAc.

2.2 Results

Mice rapidly adapt to block switches in the 2ABT

To study the flexible updating of goal-directed behaviors under probabilistic conditions, we trained mice on a 2ABT task with block switches. Two weeks prior to training, mice were injected bilaterally with AAV dLight in the NAc core and implanted with an optic fiber and ferrule above the injection site.

In the 2ABT, mice encountered three ports equipped with infrared sensors. Animals were water restricted and trained to nose-poke into the central port to initiate a trial and then move to a left or a right port to obtain a water reward. Water rewards were available on either the left or right side, depending on the block. The correct choice was rewarded with water 75% of the time whereas no rewards were available at the other port. After they received a random number of rewards (uniformly sampled from 7 to 23 for the "Full Task" condition, Fig 2.1A-B), the correct port switched without any discriminative cue. In order to achieve consistent water rewards across the whole session, mice needed to readily update their choice using outcome feedback.

As soon as the full task started (see Methods for details), we recorded unilaterally from the NAc in mice performing the 2ABT for a total of 14 daily sessions, alternating the hemisphere and allowing one non-recording session every third day (Fig 2.1B). Just after pre-training, mice chose the high reward probability port 72.9% (95% CI: [0.669, 0.767]) of the time on average. Over the course of 14 training sessions, performance steadily increased (Fig 2.1F, linear regression slope coefficient 0.0026, t statistic: 2.743, p=0.008, CI: [0.001, 0.005]). In the first 7 sessions, mice took 2.49 trials on average (95% CI: [2.39, 2.59]) to switch to the correct port after a block switch. In the 8-14th session this reduced to 2.36 trials on average (95% CI: [2.28, 2.45]). Mice switched and committed to the correct port for 3 or more trials (average 3.12 trials, 95% CI: [3.00, 3.26]) following the first 7 sessions. These data were comparable to mice performing the task without a fiber implant or cable [17].

To identify what task features the mice might consider before switching ports, we analyzed the stay/switch behaviors after two trials of outcome history. To simplify the comparison, we specifically chose the trials after the mice consecutively stayed in the same port for two trials or more. We encoded the outcome history at t-2, t-1 as RR, RU, UR, RU, (23607, 5719, 9428, 8520 trials each) with R representing rewarded trials and U representing unrewarded trials. For instance, a trial with "RU" stay history means that a mouse encountered a reward at trial t-2 and then was unrewarded at trial t-1, all at the same port. We found a significant reduction in the probability of repeating the previous choice after one unrewarded outcome (RR vs. RU, Fisher's exact test: 6.47, $p \leq$ 1e-4, Cohen's d: 0.6242, 95% CI: [0.6, 0.65]). Moreover, mice reduced their stay rate at the chosen port after two consecutive unrewarded outcomes compared to one unrewarded outcome (RU vs. UU, Fisher's exact test: 3.98, $p \leq$ 1e-4, Cohen's d: 0.6715, 95% CI: [0.64, 0.7]; Fig 2.1E).

More formally, we conducted comprehensive logistic regression analyses to describe how outcome histories influenced behaviors across individual mice, with a particular focus on capturing the decaying impact of historical reward and choice information. By implementing two distinct regression approaches (Fig 4.1, [90, 101]), we were able to examine how mice integrate recent (1 trial back) and more distant trial histories into their decision-making process. The first model explored the general decay of choice outcome weights across trial histories, revealing consistent patterns of how past experiences progressively influence current choices. The second model specifically evaluated whether recent rewards can "block" or reduce the effect of more distant outcome histories. This nuanced analysis allows us to parse the complex cognitive mechanisms underlying mouse behavioral decision-making, providing

insight into how mice integrate sequential information to guide their choices. By fitting these models to individual animal data, we can characterize the subtle variations in decision-making strategies across different subjects with higher precision, thus offering a more robust and mechanistic understanding of the experimental results.

Cognitive model fitting confirms that mouse behavior data favors Bayesian models or complex RL models compared to simple RL models

To investigate the nature of computations that mice use to rapidly adapt their choices following block switches, we compared a purely Bayesian model (BIfp, see methods) adopted from [8], a hybrid BRL model (BRLfwr, see methods) inspired by [9], and several RL models (Fig 2.2A): a model with asymmetric learning rates for positive and negative RPEs (RL4p, dubbed the simple standard RL model), a model that additionally used counterfactual updates for unchosen option values (RLCF), the recursively formulated logistic regression (RFLR) model developed by [90], an RL model that simulates forgetting via decaying the Q value of unchosen options (RLFQ3p), a foundational dynamic learning model developed by Pearce and Hall that adapts learning rate with outcome uncertainties [102], and another recent dynamic learning variant that adjusts negative learning rate adaptively based on expected and unexpected uncertainty (RL_meta) [91]. The model space under selection consists of a rich set of cognitive mechanisms that have been shown in previous work to perform well in explaining mouse and human behaviors in the 2ABT [8, 90, 102], including Bayesian inference, dynamic learning, forgetting, counterfactual, stickiness, and asymmetrical learning (Table 2.1). We carefully chose this model space in order to identify which cognitive mechanism alone can best explain mouse behavior and dopamine release in NAc. Due to the intricate connection between these cognitive models, as well as their mathematical equivalence under certain restricted conditions [90], a definitive classification or grouping can be rather difficult. However, for the convenience and scope of this paper, we focused on evaluating the effectiveness of Bayesian inference in explaining empirical data against other more complex RL models, and opted to illustrate this with the conceptual diagram in Fig 2.2A. For instance, though one can show its connection to Hidden Markov Models (HMM) under restricted conditions, the RFLR model was grouped with other RL models due to its mathematical equivalence to the forgetting Q-learning model [90, 103]. Modeling details can be found in the Methods and descriptions shown in Table 2.1.

All models listed compute choice values (or implicit values via reward probabilities in the case of BIfp), map these to choice probabilities, and then update their values (and beliefs in the case of the Bayesian models) after receiving reward feedback. For models that do not explicitly utilize RPE for model updating, like BIfp, we can calculate "pseudo-RPEs" by taking the difference between the observed and expected reward. Critically, the BRL and RL models differ in how they perform value computation. The RL models update choice values stored in a look-up table, while the BRL models computed values as a sum of choice

Table 2.1: **O**verview of different cognitive models.

model name	description	parameters	distinct mechanisms
RL4p	Q learning	α^+, α^-, β, ϕ (stickiness)	asymmetric learning, perseverance
RLCF	Q learning with counterfactual updates [8]	α^+, α^-, β, ϕ	counterfactual, asymmetric learning, perseverance
RFLR	Recursive Formulation Logistic Regression [90]	α, ϕ, τ (decay rate)	perseverance, forgetting, connections to HMM [90]
BRLfwr	Belief state RL model with reward weight updates and fixed initial weight	β, ϕ, q (latent switch rate), α	perseverance, Bayesian inference
BIfp	Bayesian Inference Model [104]	β, ϕ, q	perseverance, Bayesian inference
RLFQ3p	Reinforcement Learning with Forgetting ζ (3 parameter)	β, α^+, α^- ($\zeta = 1 - \frac{\alpha^+ + \alpha^-}{2}$)	asymmetric learning, forgetting
RL_meta	RL with meta learning [91]	α^+, α^-, β, ϕ, ζ, α_ν, ψ	dynamic learning, asymmetric learning, forgetting, perseverance
PearceHall	Pearce-Hall dynamic learning model [102]	α_+, α^-, ϕ, α_ν, ζ	perseverance, asymmetric learning, forgetting, perseverance

values in each belief state weighted by posterior belief probabilities. The updating process is schematized in Fig 2.2D-E.

The RL4p model and its variants have been extensively used in previous studies (e.g., [105–109]), which have provided evidence that asymmetric learning rates and choice perseveration ("stickiness") are often helpful in capturing animal behaviors. RL4p serves as a baseline against which all more complex models should be compared. Despite its past empirical success, a critical limitation of the RL4p model is that it fails to capture the observation that animals appear to update values for unchosen options (counterfactual updating; [8, 110–112]). For example, in reversal learning tasks like the 2ABT, observing an unexpected reward omission dramatically increases the likelihood of a switch, accompanied by neural responses that anticipate the new reward contingencies [75, 113]. Counterfactual

updating is usually formalized by updating values for unchosen actions in the opposite direction from the values of chosen actions. This qualitatively mimics the behavior of Bayesian models [104].

We fitted all models to the choice data of all 14 sessions for 5 mice (Fig 2.1F) in the 2ABT using maximum likelihood estimation. To qualitatively analyze similarities among this rich set of cognitive models, we simulated behaviors using parameter ranges fit to the mouse behavioral data. Then we fit each of the 8 main model variants to each of the simulated behavioral data sets and tested for model identifiability based on the frequency with which the ground-truth model was chosen by the model selection criterion, the Akaike Information Criterion (AIC). This revealed relatively poor identifiability between the Bayesian model and the hybrid BRL, as well as poor identifiability between the complex RL models (RL_meta, RLFQ3p, RFLR). These results indicate that the 2ABT cannot be used to discriminate within each family of model on the basis of behavioral data, but can discriminate across families. Furthermore, the 2ABT is adequate for rejecting the standard RL model (RL4p, Fig 2.2B,C).

With RL4p as a baseline, we found that both BRLfwr and BIfp significantly improved the AIC measure compared to the RL4p model (BIfp: $\Delta\text{AIC} = -536.57 \pm 59.48$, BRLfwr: $\Delta\text{AIC} = -530.59 \pm 61.86$). The RLCF also explained the mouse behavior better than the RL4p model (RLCF: $\Delta\text{AIC} = -361.92 \pm 70.52$), with a higher relative AIC on average than the BIfp model ($\Delta\text{AIC} = 174.64 \pm 47.83$) (Fig 2.2C). Additionally, other complex RL models also outperformed the RL4p model ($\Delta\text{AIC} = -652.53 \pm 49.26$, -677.84 ± 65.69, -656.44 ± 62.16, -646.45 ± 53.62 for RFLR, RL_meta, PearceHall, and RLFQ3p respectively, Fig 2.2C). When we evaluated the qualitative difference for model fitting across subjects, we noted that due to the limitation of sample size, the lowest attainable uncorrected p-value against a one-sided alternative using non-parametric pairwise tests (like permutation test) is 0.03125, which can readily lose power when multiple test corrections are used. Furthermore, it is reasonable to assume that relative AIC values across subjects follow a symmetric and normal-adjacent distribution. Therefore, we used parametric t-test and controlled family-wise error rate (FWER) via Holm's method with Bonferroni adjustment. We found that both Bayesian models like BRLfwr (p corrected: 0.018) and BIfp (p corrected: 0.015), and complex RL models (p corrected for relative AIC for RFLR: 0.004, RLFQ3p: 0.006, for RL_meta: 0.009, PearceHall: 0.009) provided better accounts of mouse behavior in the 2ABT than the standard RL model. However, there was not sufficient discriminative power to identify whether Bayesian models or any member of the complex RL models was superior based on the simple quantitative account of all behavioral data as a whole (corrected p value for BRLfwr vs RL_meta: 0.39, BRLfwr vs PearceHall: 0.679, BRLfwr vs RFLR: 0.679, BRLfwr vs RLFQ3p: 0.679, though BRLfwr performs significantly better than RLCF: p corrected: 0.020).

Bayesian models and complex RL models qualitatively recover mouse behaviors around trials feature volatile outcome changes

To further understand the mechanism by which the Bayesian model, hybrid BRL, or the complex RL models explained mouse behavior, we compared qualitative signatures of the BIfp, BRLfwr, and RL models with those of mouse data. One hallmark behavioral signature of mouse adaptation to block switches is improved choice accuracy as the number of trials after a block switch increases (Fig 2.1D). Due to the model similarities found in the previous model identification analysis, (Fig 2.2B,C), we focused on a restricted subset of model space: BRLfwr, RLCF, RL4p, RFLR and PearceHall. We omitted BIfp, RLFQ3p, and RL_meta from the analysis due to their qualitative similarity to BRLfwr, RFLR, and PearceHall respectively. Accordingly, we first compared the rate at which choice accuracy improved after block switches for simulated model behaviors against mouse data. Consistent with our hypothesis, we observed that BRLfwr and RLCF predicted faster adaptation to block switches compared to RL4p, resembling the adaptation rate of mouse switching behavior (Fig 2.2F). When we focused on single-trial choice-outcome trajectories, we qualitatively observed that the Bayesian model and complex RL models updated their switch probability differently (or faster in most cases) from the RL4p model after unrewarded observations or near block switches (Fig 2.2G). From this, one further hypothesis naturally arises: the Bayesian models and complex RL models outperform the RL4p model because of their ability to detect choice-outcome sequences that are suggestive of a block switch or volatile choice-outcome contingencies.

To test this hypothesis, it is imperative to obtain a more granular understanding of how different trial outcome histories drive a mouse to switch ports, and to determine if BRLfwr can successfully predict switch probabilities under different outcome histories, especially around block switches. Following [90], we categorized all trials based on their three past trial outcome histories, using capital A/B to denote rewarded outcomes and lower-case a/b for unrewarded outcomes. Due to the symmetrical task structure, we denoted the trial at t-3 as A/a regardless of its spatial location, and trials at t-2 or t-1 as A/a if the mouse chose the same port, or B/b if the mouse chose a different port, in reference to trial t-3 (Fig 2.3A). For instance, if a mouse was unrewarded at port 1 at t-3, rewarded at port 2 at t-2, unrewarded again at port 2 at t-1, we would describe the trial outcome history as aBb. We then calculated the switch probability as the probability to choose a different port at trial t compared to trial t-1.

We then compared the average mouse switch probability after each outcome history against the mean predictions of the three different cognitive models (Fig 2.3B; sum of squared errors (SSE): BRLfwr: 0.1796, RL4p: 0.8270, RLCF: 0.2748, RFLR: 0.1537, RL_meta: 0.0787, PearceHall: 0.1084). Notably, we observed that the RL4p model overestimated the switch probability for outcome history contexts where mice encountered rewards in both ports within the past 3 trials (e.g., ABb, AaB, etc., Fig 2.3A). Since these outcome contexts occurred only around block switches, mice usually stayed at the newly rewarded ports (mean switch rate: 0.2143). BRLfwr showed a similarly low switch rate after these contexts

(0.1801), whereas RL4p model predicted a high average switch rate of 0.4328 (RLCF: 0.2248, RFLR: 0.2778).

For instance, after the outcome context of ABb, the RL4p model overestimated the probability of the mice switching back to port A (Fig 2.3B), because it was only able to increment the action value for port B, instead of capturing the underlying reward distribution change. A mouse using a learning mechanism like RL4p would mistakenly think that selecting port A was still as valuable as it was prior to this outcome sequence, given that no reward omissions were recently experienced at that port, leading to regressive errors back to port A. Interestingly, the RL_meta model was able to explain adaptive behaviors with even higher accuracy than BRLfwr due to its adaptive power via learning rate tuning when reward context changes rapidly. We furthered our model comparison using AIC (ΔAIC) relative to the RL4p standard RL model to examine mouse behaviors around block switch (defined as the first 5 trials after block switch). We found similar ΔAIC results for BRLfwr and dynamic RL models (BRLfwr: -163.6763, RL_meta: -163.9207, PearceHall: -158.9066) (Fig 2.3C). However, RL_meta was also the most complex model with 7 parameters, with poor parameter identifiability. One might therefore argue that that the Bayesian model BRLfwr was better because it exhibited good parameter recovery and was more parsimonious.

Predictions from BRLfwr and BIfp capture nucleus accumbens core dLight dynamics better than RL models

Next, we investigated whether Bayesian models or RL models provide qualitatively and quantitatively accurate predictions of dopamine release in the NAc triggered by action outcome feedback. Importantly, as noted above, BIfp does not inherently calculate an RPE term, since it uses Bayesian inference for model updates. Therefore, to allow the comparison between model predictions of BIfp and the dopamine signals, we calculated a pseudo-RPE for BIfp, the difference between observed and expected reward.

Our experimental mice were implanted with optical fibers bilaterally in NAc core to enable recording of dLight signals. Histology images were visually inspected after the experiment to verify the implant tip location and viral expression (Fig 2.4A). We analyzed the dopamine responses in NAc by aligning dopamine signals to the "outcome" event, the time point when water rewards were either delivered to the mice or were omitted. After mice received a water reward, we observed a dLight dopamine signal (Z(DA)) increase on average (estimate: 1.6345, CI: [1.613, 1.656], p$\leq$1e-5) in ports both contralateral and ipsilateral to the recording hemisphere, consistent with previous literature [86, 114, 115]. When reward was omitted at the peripheral port, we observed a reduction (OLS estimate: -1.5840, CI: [-1.604, -1.564], p$\leq$1e-5) in the dLight signal in NAc (Fig 2.4B-C).

To test the effect of animal choice switching, choice laterality (with respect to dopamine recording hemisphere), and reward or reward omissions on dopamine responses at outcome phase, we fitted a simple OLS model with all four factors. We observed a significant Z(DA) difference (OLS coefficient estimate for switch: 0.5030, CI: [0.454, 0.552], p $\leq$ 1e-5) in trial-

averaged dLight responses to the "outcome" event between switch trials and stay trials. Signals triggered by reward were larger on average on trials where mice switched to a new port, compared to when they stayed with their past choice (Fig 2.4BC). Negative signals observed after reward omissions in trials where mice stayed were larger when compared to when they had just switched (Fig 2.4BC). Our data were consistent with expectations that rewards are followed by increases in dopamine release in NAc (estimate: -1.6345, CI: [1.613, 1.656], $p \leq$ 1e-5) and unrewarded outcomes with a decrease (estimate: -1.5840, CI: [-1.604, -1.564], $p \leq$ 1e-5). There was no significant effect of hemisphere (laterality, contra or ipsi relative to reward port) on NAc dopamine release ($p = 0.523$) (Fig 2.4BC).

When we visualized single trial dopamine traces around "outcome" events, we found that the temporal span of the dopamine response qualitatively aligned with the amount of time mice stayed at the peripheral reward port (Fig 2.4E). Similarly, after unrewarded outcomes, we found that the duration of the decrease in dLight response qualitatively aligned with the time mice stayed in the peripheral reward port, reaching signal trough typically just after mice left the port. These results together suggest that when modeling dopamine responses, the time duration that mice stayed in the peripheral reward port after an "outcome" event (dubbed "port duration") needs to be taken into consideration (Fig 2.4D-E). For this reason, we included port duration as a covariate in our regression analyses.

We reasoned that if the internal reward prediction updates of the mice are represented by dopamine in the NAc and resemble that of BRLfwr or BIfp, then BRLfwr or BIfp RPE calculations should explain a larger amount of variance in the dopamine responses in NAc at the time of "outcome" events than RL models. To simplify the testing procedure, we obtained dopamine summary statistics by taking the peak (as the maximum value) of dopamine transients for rewarded trials and troughs (as the minimum value) for unrewarded trials (dubbed DA-PT) during the one second window after the outcome, marked in gray in Fig 2.4B. To control for behavioral confounding variables, we included session number, port duration, egocentric action, movement time, and center port poke duration as covariates in addition to predicted RPE values from BIfp, BRLfwr, RL4p, and RLCF (note that we omitted RFLR model since it does not have a straightforward RPE formulation), each in a disjoint regression covariate set. Using maximum likelihood estimation with featurized covariates (see Methods) on the five different RPE covariate sets, we found that both BIfp and BRLfwr models yielded larger relative log-likelihood (Δllk CV) than the RL4p baseline, evaluated on cross validation sets (70-30) for each animal (Fig 2.4F, Δllk CV: BRLfwr: 80.58, 95% CI: [60.84, 102.79], BIfp: 118.64, 95% CI: [65.71, 217.50]). Both Bayesian models also outperformed other complex RL models on this measure (Δllk CV: RLCF: -19.23, 95% CI: [-47.03, 4.64], RLFQ3p: 18.85, 95% CI: [4.98, 33.31], RL_meta: 22.54, 95% CI: [7.96, 33.90], PearceHall: 19.50, 95% CI: [6.04, 32.83]).

To illustrate the quantitative differences between model predictions, and to identify the aspects of the data that the BRL and BIfp models captured but the RL models did not, we investigated conditions where different models gave qualitatively different predictions. First, we looked at the changes in dLight responses to rewards as mice experienced a series of outcomes in consecutive stay trials from aaa to AAA (lowercase "a" indicates unrewarded

and uppercase "A" indicates rewarded). As mice encountered more rewards in their recent history, dLight dopamine responses to reward gradually decreased (consistent with decreased RPE) but did not completely flatten to no response (Fig 2.4G), partially explaining the quantitatively worse fit (Fig 2.4F). This pattern observed in our dLight data from the NAc core ('data' indicated by a black line) was not well captured by the RL models in later trials but was captured by the Bayesian model and the hybrid BRL model. To quantify, we trained a separate OLS model for each cognitive model to predict dopamine response using only simulated RPEs. We compared the model fitness across different cognitive models with cross validated log-likelihood (BRLfwr: -985.89, RL4p: -991.80, RLCF: -992.87, PearceHall: -993.81, RLFQ3p: -994.74, RL_meta: -995.02) (Fig 2.4H). Indeed, BRLfwr RPE predictions achieved the highest model fitness to dopamine data. The RL models updated action values on each trial as mice encountered consecutive rewards, estimating a recent average of choice outcomes. This feature may make RL models overly sensitive to recent outcome histories and liable to overestimating the action values, particularly around block switches. In our dataset, the best fitting RL4p model had an average α^+ of 0.85 (95% CI: [0.77, 0.94]), and average α^- of 0.73 (95% CI: [0.71, 0.75]). The large α (learning rate) values likely enabled RL4p to maintain the flexibility to adapt to block switches, but they also posed issues regarding biological feasibility. RL_meta and PearceHall model were able to mitigate this issue with a dynamic learning rate that adapts to outcome uncertainty. However, due to its iterative nature, as the amount of reward in the recent outcome history increased, it steadily decreased the RPE towards zero, which did not match what we observed in the dopamine signal.

Both BIfp and BRLfwr were able to explain the adaptation of dopamine transients after increasing consecutive rewards better than the RL models. These models accounted for possible block switches and maintained values of both a high reward context and a low reward context that are not associated with any given port (Fig 2.4H). Instead, these values were assigned to specific ports in proportion to the model's belief state on any given trial. We speculate that this information about task structure provided robustness in response to probabilistic reward omissions after selecting the more rewarding port. This also allowed the best fitting BRLfwr and BIfp model to predict a small but non-negative RPE response on rewarded stay trials (Fig 2.4H), suggesting a degree of uncertainty as to whether the block had switched and capturing a pattern observed in the NAc dLight signals in later consecutive trials (right side of Fig 2.4H). Furthermore, we observed that among trials after an immediate past trial reward, the dopamine response was highly consistent, independent of more distant outcome histories. Only BRLfwr captured this phenomenon qualitatively, in striking contrast to other complex RL models.

The Bayesian model and the hybrid BRL model were able to modulate expectations about both choices using inference based on one single choice outcome. This leads to the prediction that at the outcome phase of each choice, an increasing number of rewards gathered at the opposing port in past trials will reduce the RPE signal reflected in dopamine, especially trials where animals switched to a different port from the previous trial (animal switch trial). To investigate if this prediction was upheld in the dopamine data, we fitted both the mouse data and the model simulated data with linear mixed effects models (LMER, see Methods).

Specifically, we modeled the dopamine values, or the predicted RPEs for simulated data, as a linear weighted sum of current reward (Reward), rewards observed at the currently chosen port (trial t) over the past 4 trials (R_chosen) and past rewards observed at the opposite port over the past 4 trials (R_unchosen), conditioned on whether animal switched trials or animal stayed with their prior choice t-1 on the current trial t. On animal switch trials, both BIfp, BRLfwr, and RLCF predicted a positive effect of R_unchosen, but other RL models predicted little to no effect due to a forgetting effect or asymmetrical value updates. LMER fitting revealed a significantly negative effect of R_chosen on dopamine at outcome phase during animal switch trials (slope estimate: -0.128, 95% CI [-0.195, -0.062], p=0.02), and a significantly positive effect of R_unchosen on dopamine during animal switch trials (slope estimate: 0.126, 95% CI [0.035, 0.217], p=0.03, Fig 2.4I). On animal stay trials, the result was more complicated. All models predicted a negative relationship between R_chosen and RPE signal, matching the dopamine observation (slope estimate: -0.09, 95% CI: [-0.11, -0.07], p=0.01). However, low sample size of high R_unchosen value trials a noisy negative estimate of effect of R_unchosen, which none of the models generated consistent predictions (Fig 2.4J). This elevated level estimation errors for R_unchosen makes it a bad target for model arbitration, while the other three measures may be a more credible source of evidence. To sum up, the relationship between past reward history and dopamine values qualitatively match the prediction of the BRLfwr model for switch trials and the predictions of BIfp or RLCF model for stay trials. Taken together with the functional similarity of BIfp and BRLfwr (Fig 2.2B), we conclude that the dopamine data matched the RPE predictions of Bayesian models quantitatively and qualitatively, but we could not further discriminate between BIfp or BRLfwr.

2.3 Methods

Animal protocol

Mice (all C57 Bl/6 male, bred in-house aged 104-167 days) were housed on a 12 h reversed light-dark cycle (lights on at 22:00) with nesting material. Prior to surgery they were group housed with access to food and water ad libitum. We conducted all animal procedures, which follow the principles outlined by the NIH Guide for the Care and Use of Laboratory Animals, according to the protocol (AUP-2015-11-8145-3) approved by the University of California, Berkeley Institutional Animal Care and Use Committee (IACUC) and Office of Laboratory Animal Care (OLAC).

Surgery protocol

Mice were anesthetized with isoflurane gas for stereotaxic surgery. Meloxicam was given on the day of surgery and daily for 48h after. Coordinates for the Nucleus Accumbens Core were bregma coordinate: 1.20mm anterior, ± 1.2mm medial-lateral, -4.1mm ventral. AAV-CAG-dLight1.3b or AAV9-syn-dLight1.2b were injected using a Nanojet II (Drummond scientific).

Neurophotometrics NA 0.37 or 0.48 400μm optic fibers were placed approximately 100μm above the injection site. Dental cement was used to secure the implant to the skull. During a recovery period of 7-14 days, mice were singly-housed and fed ad libitum.

Histology and Imaging

Following behavior, animals were transcardially perfused with 4% paraformaldehyde (PFA) in an 0.1M phosphate buffer (PB) solution (pH = 7.4). Brains were collected and fixed overnight, followed by a transfer to 0.1M PB. To visualize striatal photometry fibers, perfused brains were sliced coronally at 50μm using a vibratome (VT1000S Leica Biosystems; Buffalo Grove, IL). Immunohistochemistry was performed to amplify the GFP signal (1:1000 chicken anti-GFP, Aves Labs, Inc.; GFP-1020 followed by 1:1000 goat anti-chicken AlexaFluor 488, Invitrogen by Thermo Fisher Scientific; A11039). Slides were mounted on slides with Fluoromount-G (Southern Biotech). Slices anterior and posterior to the fiber tract were imaged at 10X on an AxioScan Z.1 fluorescent microscope (CRL Molecular Imaging Center, UC Berkeley) to confirm targeting. Detailed resources are fully described in Table 2.

Probabilistic Switching 2ABT Task

We trained 5 male mice on a 2ABT behavioral task in which the location of the water reward was periodically switched between the two potential rewarded ports at random intervals. Trials were initiated by a nose poke in the center port and concluded after the mouse subsequently chose one of two reward ports, located on either side of the center port. Nose pokes were detected by a infrared photodiode. Once the initiating poke was sensed, lights that encompass the reward ports were triggered to cue the animals to the viability of a potential water reward. The mouse receives a 2 μl water reward or no reward depending on their correct or incorrect, respectively, port choice for each trial.

Only one peripheral port was rewarded at a time. The setting for the number and frequency of water rewards was altered during each phase of the task. Teaching animals the task, we started them in the "operant" phase, where water rewards were offered in the peripheral ports, and then moved to the "pokeseq" phase, where mice learned to poke the middle port before going to retrieve a peripheral port water reward. Successful learning during this phase was measured as the mouse getting more than 700 rewards in less than five hours. Some food pellets were placed into the arena while the mice were learning the task to motivate them. We omitted mice from the experiment if they didn't perform successfully on "pokeseq" after five days.

Once the mice learned to poke the center port to initiate the peripheral port water rewards, we no longer put food into the arenas and advance the mice onto the "learning switch" phase. During the three subphases of the learning switch, we teach the mice that only one port is rewarded at a frequency of 90%, 80%, and 78% for each subphase respectively. Additionally, we taught the mice that the rewarded port switches after obtaining a random number of cumulative rewards set within the ranges of 27-43, 17-33, and 12-28 for each

subphase respectively. Water rewards became less predictable and more variable with each subsequent phase.

Successful learning for the three learning switch phases were set at a score higher than 69.5% on both the left and right port within four, three, and two hours for each subphase respectively. If the mice seemed unmotivated, we placed some food into the arena. Once they completed the task with food, we removed the food and had them reach criterion for their appropriate subphase.

Fibers were first introduced to the mice after they complete the last learning switch subphase and are utilized for neural recording only after the mice successfully reach the learning switch criteria once again. During the "full task", used for fiber photometry recording, mice received a water reward with a 75% chance if they picked the correct port, but received nothing if they picked the wrong port for each trial. Once mice obtained a random number of cumulative rewards, sampled from 7 to 23 uniformly, from one port, the reward was switched to the other port. This structure minimized the possibility for the mice to predict the timing of the switch.

When participating in this task, the mice were water restricted but maintained ad libitum access to food. They were supplemented with additional water if they earned less than 500 μl water from the task. Mice were weighed daily, before and after the task, and were given additional water if needed, at least 30 minutes after training sessions, to ensure that they maintained around 85% of their original weight.

Fiber Photometry (FP)

We recorded from all 5 mice with fiber implants in nucleus accumbens core (NAc) using Neurophotometrics FP3001 system at 20Hz. dLight signals are measured in 470nm channel, while isosbestic control signal is measured in 415nm channel for artifact removal. During the learning switch phase, we start plugging in fiber implants without recording to get mice accustomed to moving and behaving with fibers. Starting from the first "full task" session, we alternate between a left hemisphere recording, right hemisphere recording, and no recording schedule. The no recording day was implemented to avoid signal bleaching. All FP recordings are synchronized to behavioral and video data via custom TTL systems and bonsai programs and saved for further processing. Recorded signals are preprocessed using custom python program to control for motion artifacts using 415nm reference channels [116]. Specifically, we first subtracted the channel-specific trends by approximating a smooth fit using airPLS algorithm across both channels. Then we perform a robust linear regression from reference channel to signal channel, and subtract the fitted 415nm artifact signal from 470nm channel recordings and obtain the processed dLight signals. All dLight signals are subsequently z-scored to account for extraneous variations between sessions.

Task Behavior Analysis

All final behavioral analysis was conducted with custom implemented python packages. Behavioral data were first collected using a custom implemented 2ABT control module in matlab, and then preprocessed into behavioral data frames with each row representing a separate trial with various columns containing information regarding task meta data, mice choice data, and task relevant variables. For trials that mice failed to initiate and make a choice, we included the data but noted the relevant variables as NaNs/null. For most behavioral analysis, outcome histories as well as covariates with null entries were dropped. All model plotting are generated via `seaborn` packages, and test statistics and various measures are generated using `statsmodels, pingouin and sklearn`.

We divided the tasks into initiation phase, execution phase, outcome phase and inter trial interval (ITI) phase. The initiation phase includes both "center in" and "center out" events when mice poke their nose into and out of the center port to initiate a trial. To execute a choice, mice leave the center port and then enter one of two peripheral ports to make a selection. When the outcome of their choice is revealed to the mice, they linger at the port for water reward or time-out until they leave the port during the "side out event".

We defined the following critical events:

- Center in (CI): when mice first poke into the center port to start a new trial.
- Center out (CO): when mice leave the center port to select ports.
- Side in (SI): when mice poke peripheral port.
- Outcome (O): when outcome is revealed, typically immediately after side in (latency on the order of system latency).
- Zeroth side out (SO0): first time when mice move the nose poke out of the port trigger IR beam break.
- First side out (SO1): last time of a consecutive set of beam breaks at the same side peripheral port. Zeroth side out and first side out can often be the same when the animal leaves the port only after drinking all the water.
- Last side out (SOf): last side out beam breaks before next center port poke to initiate a new trial.

We define the following critical timing measures:

- Movement time (`MVMT`) : CI - SOf (t-1), interval between final side port departure and new trial initiation.
- Inter-trial interval (`ITI`): CI - SO1(t-1), interval between the initial side port departure and new trial initiation. This quantity is often the same as MVMT above, as mice typically do not reenter side port before a new trial after leaving side port.

- Center duration (`center_dur`): CO-CI, duration that mice lingered at center port after their trial initiation.
- Port duration (`port_dur`): SO1-O, intervals between outcome and side port departure.
- Side out bout latency (`SO_lat`): SO1-SO0, which represents the gap between two nose poke bouts at side ports. They are typically close to 0, as mice generally proceeded to the next trial after finishing drinking at the side port. Since tnose pokes are measured with infrared beam breaks, critically this is also used to control for atypical trials with excessive beam breaks.

Neural Data Analysis

All analysis was carried out with a custom Python module for aligning, visualization, and data modeling. After baseline correction, we visually inspected both 415nm channel recording and 470nm channel recording, as well as the AUC-ROC scores between the rescaled baseline values against signal channel recordings. We picked sessions that meets the following conditions: AUC-ROC higher than 0.9, longer tail in distributions in 470nm recording compared to rescaled 415nm recordings suggesting clear fluorescence increases, and no sharp discontinuity in both channels' recordings. We use baseline-corrected and z-scored dLight signals for in-depth analysis, called Z(DA). For neural data alignment, we interpolated Z(DA) around event times and appended them to the trial level dataframe.

Aligned neural signals are appended to the trial level data frames as additional columns. Aligned Z(DA) is baselined by subtracting out its value at "outcome" time (we denote this procedure as de-base for simplicity). For trial average dopamine visualization, we first identified a list of relevant columns as grouping variables of interest, we then dropped any row with one or more NaN entries in the relevant columns. These contain trials that either missed neural data or behavioral data, happening mostly at the beginning or the end of the session, accounting for $\sim$ 0.5% of the trials. The trial averaged plots are then plotted using seaborn with the error bar being bootstrapped confidence intervals. When reporting summary statistics, we used bootstrapped confidence intervals to characterize mean estimates. For difference comparisons, we used suitable paired or unpaired tests and reported both the testing statistics as well as effect size, and the confidence intervals associated with it. For model RPE prediction fitting to RPE, we fitted the following regression: `outcome_DA_PT ~ rpe:rpe_pos + rpe:rpe_neg + port_dur:rpe_pos + port_dur:rpe_neg + ego_action + session_num + log__MVMT + log__center_dur`.

In the above formulation, `rpe:rpe_pos` and `rpe:rpe_neg` respectively encoded positive and negative RPE responses in dopamine. `ego_action` describes the effect of movement laterality on dLight signal. We included `port_dur, MVMT, center_dur` to control any potential movement-related components of dLight signals. To minimize the effects of intertrial events on neural signals, we selected sessions with `center_dur <= 0.8` (mice lingering less than 0.8 seconds at center port), `port_dur <= 6` (intervals shorter than 6s between outcome and side port departure), `MVMT <= 3` (intervals shorter than 3s between the final side port

departure and next trial initiation). We also filtered out trials with `SO_lat01 > 1`, where the trials had gaps longer than 6s between two nose poke bouts at side ports. These trials typically consist of trials where mice revisited the port multiple times, or simply mistriggered the infrared sensors used to measure nose pokes, biasing the task structure. After this cleaning, we were left with 20952 trials ($\sim 68\%$ of the dataset) across 34 sessions with valid dopamine recordings. With the processed data, we fit the above model to dopamine data, and observed consistent results across model metrics.

Linear Mixed Effects Modeling

Linear Mixed Effect Modeling was used to capture the effect of past rewards on the current chosen port, or the opposing port. We modeled using `rpy2` with the `lme4` package with the following formula:

```
DA ~ Reward:Switch|Subject + R_chosen:Switch|Subject + Switch|Subject

  + R_unchosen:Switch|Subject + 1|Subject + Reward:Stay|Subject
  + R_chosen:Stay|Subject + R_unchosen:Stay|Subject
```

For model simulated RPEs, we changed the output variable to be RPE predictions instead and ran regular regressions without subject level effects due to the uniform sample size in each simulated session.

To determine the optimal lags, we did cross validation and compared multiple models with distinct feature space to arrive at R+Sw model at lag 4 (Fig 2.5), which was enough to capture similar levels of variance compared to L0:R. We used 4 lags because it allows us to capture a relatively high amount of past outcome levels, with only the expense of 1% variance explained. Additionally, we fit the models to individual animals, and compared them to neural data and found varied but qualitatively consistent results across animals.

Quantification and statistical analysis

Statistical analysis was performed using Python with standard packages: scipy, statsmodels, pingouin. We included n=5 subjects with 14 sessions each. Statistical details of each analysis can be found in each figure legend, result section or correspondent method section. Error bars are 95% bootstrapped confidence intervals, and Holm-Bonferroni correction was applied when appropriate.

2.4 Discussion

Our results showed that mice were capable of rapidly adapting to block switches in an instrumental task with probabilistic and volatile contingencies, without sacrificing the robustness of their choice policy when encountering stochastic unrewarded outcomes after correct choices.

Both the pure Bayesian inference model and the belief state RL model were able to capture this balance and outperformed standard RL models in explaining mouse behavior (Fig 2.2C, Fig 2.3C). While behavioral data alone was not sufficient for us to arbitrate among Bayesian models, hybrid BRL, and complex RL models, models using Bayesian inferences (both BIfp and BRLfwr) for state estimation were able to provide a more accurate account of the dopamine release events in the NAc core following choice outcomes (Fig 2.4G-J). Critically, both Bayesian and hybrid model had explicit belief state representations of different block identities (Fig 2.2DE), which provided a mechanism for encoding the rapidly changing reward contingencies in the 2ABT task. This mechanism successfully explained the ability to efficiently switch after consecutive unrewarded outcomes while also disregarding occasional reward omissions within a block. The RL component in the BRL models allowed the model to use RPEs to update its mapping from choices to values conditioned on belief states. However, we did not find decisive evidence that BRL models explained the mouse data better than pure Bayesian models. The principal conceptual advantage of BRL models is that they make direct use of RPE signals, whereas pure Bayesian models do not. This highlights the importance of neural data in adjudicating between these models.

Building on the behavioral modeling, we showed that the Bayesian and hybrid models do a good job predicting NAc dopamine activity at the time of outcomes (reward or reward omission). In particular, RPEs generated by the BRL model and pseudo-RPEs derived from the BI models qualitatively and quantitatively match the history-dependent pattern of dopamine activity (Fig 2.4F-J). This finding supports a growing literature showing that dopamine signaling of RPEs depends on a belief state (or some approximation of a belief state; see [69]) in partially observable tasks [9,68,76,77,89]. Furthermore, our results show that in the context of a probabilistic instrumental task, dopamine release in NAc compute RPE-like signals not just based on observable input, but also internally generated state information. This confirms an expanding literature showing that "model-free" RL computations in the midbrain and striatum take in higher-level inputs such as beliefs about latent state [9, 71, 75, 95], as well as information about reward outcomes and reward expectations [117–120].

The BRL model is closely related to structure learning models that learn multiple context-dependent policies. Such models often assume some form of Bayesian inference at the more abstract level, with RL supporting learning of specific policies [94, 121, 122]. Theoretical work, supported by experimental findings, has also shown that dynamic adaptation at the more abstract level of belief states/latent contexts/rules may also be supported by RL-like computations [104, 123, 124]. The current work cannot dissociate these possibilities.

There are several limitations to our work that can be addressed in future investigations. Here we adopted a 75-0 reward probability design from the legacy of prior lab literature. In this design, in any given block, an inferior port always yields no reward. Consequently, mice rarely switch to a new port and get rewarded after observing rewards in the previously chosen port. Previous work has noted that inference-based cognitive models are capable of decreasing the expectations about alternative options after observing a reward for a selected option [71, 75, 113]. Therefore, mice using inference would exhibit a higher RPE following a reward in a new port after also observing a reward in the old port, a reward-switch-

reward sequence. Yet in our dataset, there are only 94 trials initiating such sequences out of 30,707 trials in total, across multiple animal sessions, giving us limited power to address this prediction.

In our task design, mice can initiate new trials at their discretion after the outcome, with no instructed delays. Especially after unrewarded outcomes, mice leave the peripheral port shortly after, approaching the center port for a new trial. As a result, in trials when mice leave the peripheral port too soon, dopamine responses coincided temporally with the dopamine ramp associated with approaching center ports [125, 126]. To mitigate this, we only analyzed trials with specific port durations (see Methods), which limited our statistical power. More generally, the scaling effect of port duration on dopamine could reflect a multiplexed role of dopamine in vigor, action initiation, or other functional diversity of dopamine neurons [127–135]. Also, our work does not directly address the possibility of other alternative computational hypotheses about dopamine, like directly setting the adaptive learning rate [136] or signaling retrospective inferences about causal targets [137].

Another caveat of our experimental design was that it incorporated a multi-stage pre-training protocol before we started recording dopamine data in the "full task" condition. Mice started with blocks with 100% reward probability, which was brought down in stages to 90%, 80% and then 78%, before the "full task" at 75%. Since this pre-training took approximately 10 days on average, it may have been enough for mice to develop Bayesian inference strategies before dopamine recordings started. When we fit models to single session behavioral data during the "full task" phase, we did not observe any changes consistent with transitions from one cognitive strategy to another across days in the full task. As a result, our current dataset cannot address the question of whether, in the earliest days of pretraining, mice transitioned from initial reinforcement learning to a more sophisticated belief-state inference approach.

A final limitation of our study is the fact that that the RPE that correlates with dopamine signaling only plays a minimal role in predicting behavior, as evidenced by the fact that BRL models did not outperform pure Bayesian models with no RL component. We believe that better discrimination between these models can come from theory-guided experimental design, possibly using a more complex task that places a greater demand on learning policies within each belief state [138, 139]. Moreover, in future work, more careful examinations of distinct cognitive mechanisms may also allow us to develop new models to further illuminate the questions raised in our study.

2.5 Chapter Summary

In natural environments, animals and humans often experience ambiguous outcome feedback about the hidden reward structure of the environment. We emulated this in a two-armed bandit task with switching reward blocks, and showed that mice are capable of adapting rapidly to changes in hidden reward structures after observing specific outcome feedback sequences. Computational model fitting to behavioral data suggested that models performing

Bayesian updating of beliefs better explained mouse behavioral data than standard RL models. These Bayesian models also quantitatively and qualitatively matched the dopamine release in the nucleus accumbens core better than the non-Bayesian alternatives. Together, we conclude that probabilistic belief updates are critical to behavioral adaptation and RPE signaling in the mesolimbic dopaminergic system during instrumental learning in a partially observable environment.

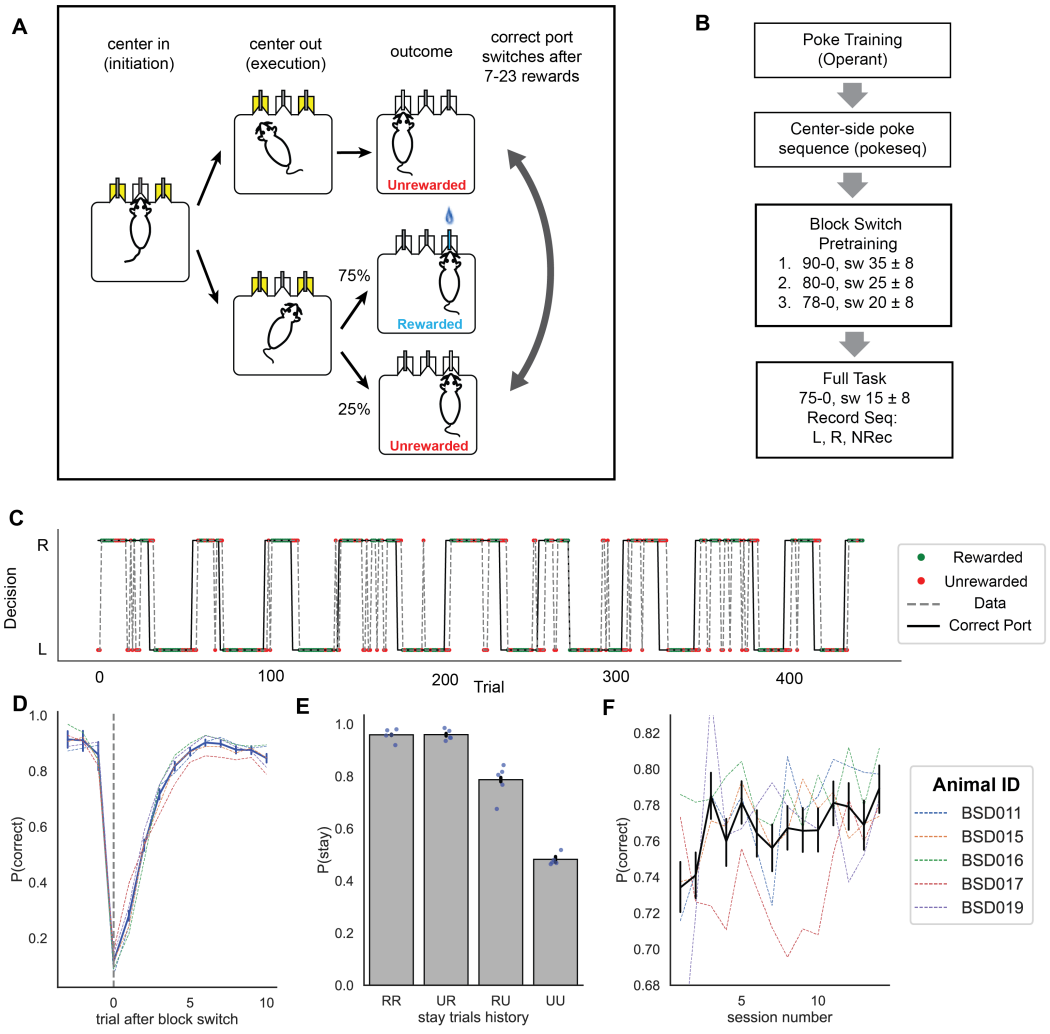
A
center in (initiation)
center out (execution)
outcome
correct port switches after 7-23 rewards
Unrewarded
75%
Rewarded
25%
Unrewarded
B
Poke Training (Operant)
Center-side poke sequence (pokeseq)
Block Switch Pretraining
1. 90-0, sw 35 ± 8
2. 80-0, sw 25 ± 8
3. 78-0, sw 20 ± 8
Full Task
75-0, sw 15 ± 8
Record Seq:
L, R, NRec
C
Decision
R
L
0
100
200
300
400
Trial
Rewarded
Unrewarded
Data
Correct Port
D
P(correct)
trial after block switch
E
P(stay)
RR
UR
RU
UU
stay trials history
F
P(correct)
session number
Animal ID
BSD011
BSD015
BSD016
BSD017
BSD019

Figure 2.1 *(preceding page)*: **M**ice adapt rapidly to block switches in a probabilistic reversal task. (A) Illustration of the two-armed bandit task, divided into initiation, execution, and outcome phases, similar to [17, 100]. In the illustrated trial, the right port is rewarded with 0.75 probability and the left port is unrewarded. After 7-23 rewarded trials, the correct port switches. (B) Training protocol. The recording phase took place in the "Full Task" phase. In the pretraining phases, the structure of the task was the same as the in the full task phase, except the reward contingencies and block lengths were different. Each contingency is labeled by numbers indicating the proportion of correct and incorrect choices that were rewarded. For example, "90-0" in the first pretraining phase indicates that 90% of correct choices were rewarded. The block length in each phase is indicated by its mean and range. For example, "sw 35 ± 8" in the first pretraining phase indicates that switches occurred after the animal earned between 27 and 43 rewards. During the 14 sessions of mouse behavior data collection, we recorded dLight signals using a "left hemisphere (L), right hemisphere (R), no neural recording pure behavior (NRec)" sequence. (C) Raw behavioral trajectory taken from the first half of a sample session. Black line indicates correct reward port locations while dashed gray line indicates actual mouse behavior. Green dots and red dots mark rewarded and unrewarded trials, respectively. (D) Probability of making a correct choice (i.e., choosing the high probability port) as a function of the number of trials around a block switch. The vertical dashed line shows trials at which rewarded block changes. Each colored dashed line plot shows behavioral performance for individual animals. (E) Probability of staying (repeating the last choice) after experiencing different outcome histories in the same port. RR: two consecutive rewards; UR: unrewarded outcome followed by rewarded; RU: rewarded outcome followed by unrewarded; UU: two consecutive unrewarded outcomes. (F) Performance across 14 sessions. Dashed lines show individual animal trajectories. Error bars show 95% bootstrapped confidence intervals.

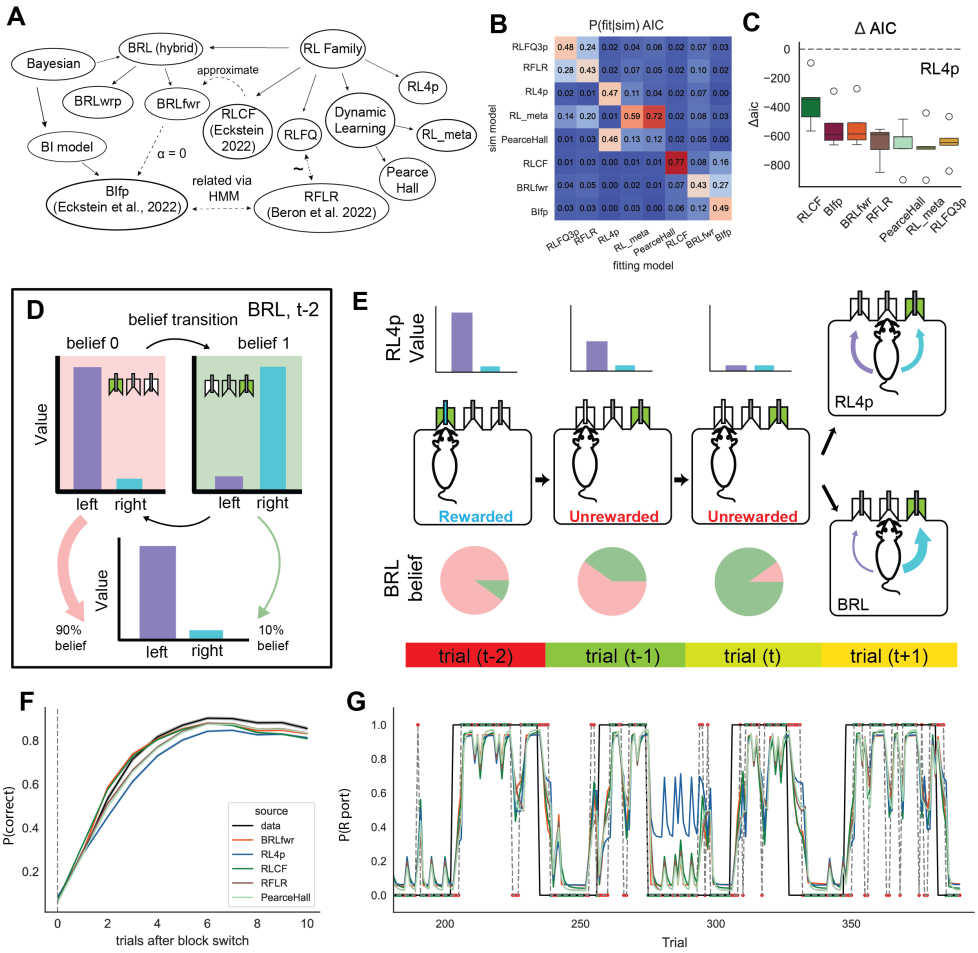

Figure 2.2 *(preceding page)*: **B**ayesian and reinforcement learning models. (A) Relationships between cognitive models (see Methods for more details). (B) Confusion matrix outlines results for model identification analysis. Each entry i, j represents the percentage of time that the column j fitting model best explained data generated by row i simulating model. The row orders are sorted via dendrogram based on model similarity (see Methods). (C) Model comparison using relative AIC compared to RL4p: ΔAIC = AIC(model) − AIC(RL4p), with lower values indicating better fit. (D) Illustration of value computation for BRL model family, which updates beliefs via Bayes' rule and then uses these beliefs to compute values. (E) Illustration using a four-trial sequence (similar to [100]) to show the differences between RL4p and BRL. Top: purple and cyan bars show the choice values conditioned on the belief state; Bottom: pie charts show the belief state for BRL; the animal's policy is selected as a function of the value within their belief states. (F) Behavior of different models compared to mouse data (black line). Trial 0 is when the program has switched the rewarded side in a block switch. (G) Example behavioral trajectory (probability of choosing the rightward port) predicted by different models. Mouse data are marked by a dashed line and block structure is marked by a solid line. Rewarded trials are marked as green dots and unrewarded trials are marked as red dots. Error bars show 95% bootstrapped confidence intervals.

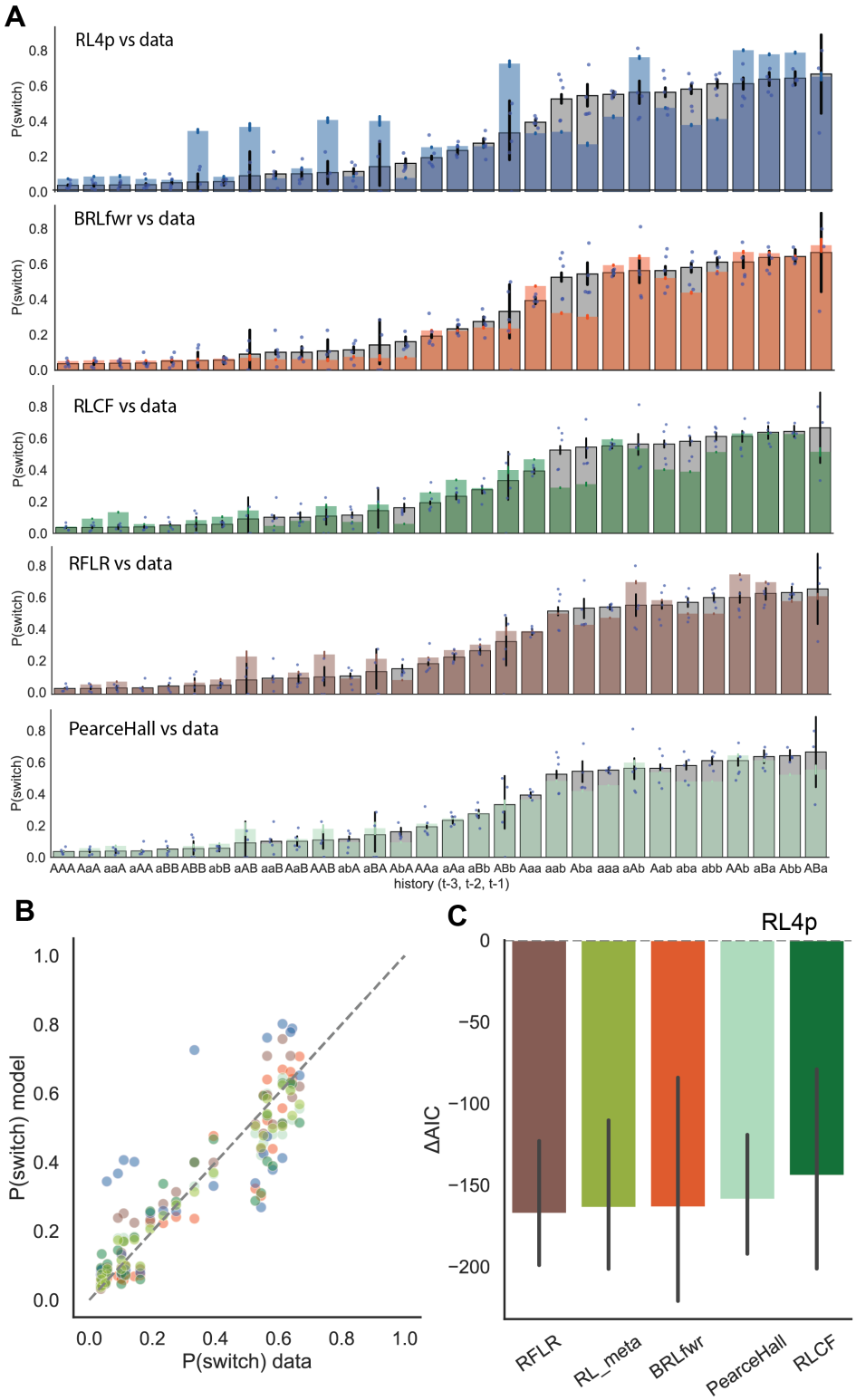
A
RL4p vs data
BRLfwr vs data
RLCF vs data
RFLR vs data
PearceHall vs data
P(switch)
0.0
0.2
0.4
0.6
0.8
AAA AaA aaA aAA aBB ABB abB aAB aaB AaB AAB abA aBA AbA AAa aAa aBb ABb Aaa aab Aba aaa aAb Aab aba abb AAb aBa Abb ABa
history (t-3, t-2, t-1)
B
P(switch) model
P(switch) data
0.0
0.2
0.4
0.6
0.8
1.0
C
RL4p
ΔAIC
0
−50
−100
−150
−200
RFLR
RL_meta
BRLfwr
PearceHall
RLCF

Figure 2.3 *(preceding page)*: **B**RL and complex RL models outperform standard RL by better explaining mouse behaviors around block switches. (A) Switch probability by different trial outcome histories described by action-outcome pairs three trials back. Gray bars showed mouse average probability of switching for each outcome history, deep blue dots represent individual mice. From top to bottom: mouse data overlaid with BRLfwr, RL4p, RLCF, RFLR, PearceHall model predictions of switch rate, respectively. (B) Switch probability predicted by different models scales with probability of mice switching port selections in different outcome contexts described. Colors represent different models, sharing the same legend as C (orange: BRLfwr, wine red: BIfp, dark green: RLCF, brown: RFLR, blue: RL4p, light cyan: PearceHall, dark yellow: RL_meta) (C) Relative AIC with respect to RL4p (dashed line at ΔAIC = 0) showing model fit to mouse data around block switch. Error bars show 95% bootstrapped confidence intervals.

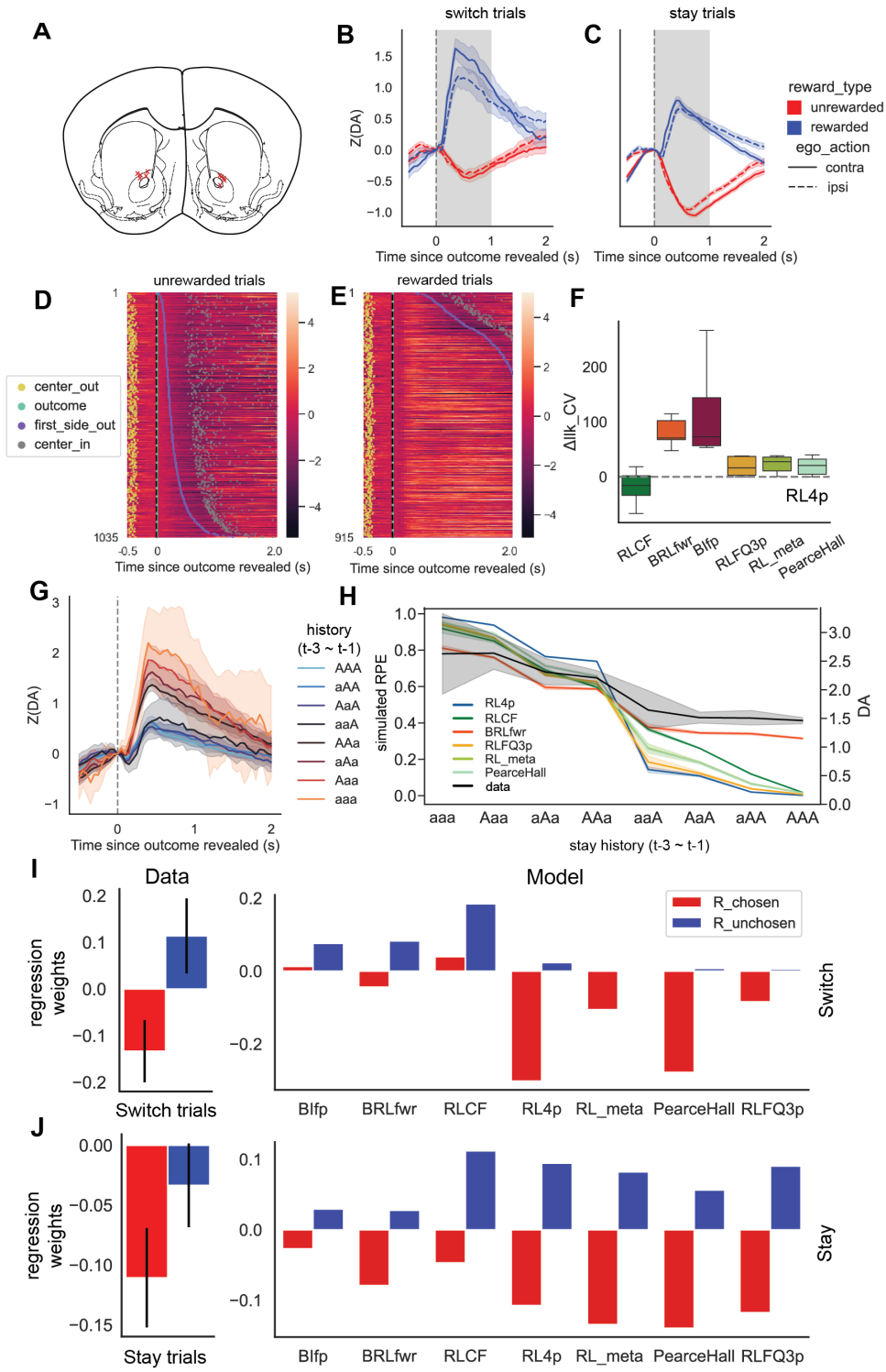
A
B
switch trials
C
stay trials
Z(DA)
Time since outcome revealed (s)
reward_type
unrewarded
rewarded
ego_action
contra
ipsi
D
unrewarded trials
E
rewarded trials
center_out
outcome
first_side_out
center_in
1035
915
F
ΔIlk_CV
RL4p
RLCF
BRLfwr
Blfp
RLFQ3p
RL_meta
PearceHall
G
history
(t-3 ~ t-1)
AAA
aAA
AaA
aaA
AAa
aAa
Aaa
aaa
H
simulated RPE
DA
RL4p
RLCF
BRLfwr
RLFQ3p
RL_meta
PearceHall
data
stay history (t-3 ~ t-1)
I
Data
Model
R_chosen
R_unchosen
regression weights
Switch trials
Switch
J
Stay trials
Stay

Figure 2.4 *(preceding page)*: **N**Ac dLight dopamine dynamics consistent with RPE predictions by models with Bayesian inference. (A) Implant fiber locations indicated on mouse brain atlas with red crosses, similar to [100]. (B-C) Trial average of NAc dLight signals (z-scored, as described in Methods) aligned to outcome events. Shaded area indicates the one second where the peak or trough is taken for neural regression. (B) shows switch trials and (C) shows stay trials. Rewarded trials are in blue and unrewarded trials are in red. Trials where mice picked the port contralateral to the recording hemisphere are plotted with solid lines; trials in which mice picked the ipsilateral port are plotted with dashed lines. (D-E) Example session single trial dLight responses plotted in heatmaps, trials sorted by the time mice spent in the reward port (see Methods for further details). (D) shows a heatmap for unrewarded trials, and (E) shows rewarded trials. Increase in dLight signal is indicated by brighter shades of red and decreases from baseline are indicated by darker shades of black. Dots are used to mark "center out" (yellow), "outcome" (green), "first side out" (purple), "center in" (gray) events, respectively. (F) Result of neural regression using model RPE values to explain dopamine variability. Fit is measured as cross validated log-likelihood (llk_CV) relative to the RL4p model, with higher values indicating a better fit. Gray dashed line indicates the baseline of RL4p RPE fitted to dopamine measurements. (G) Dopamine response on rewarded trials binned by past history, sorted in increasing order of number and recency of rewards (note in all cases mice stayed with the same port 'a/A' for all three trials). (H) RPE predictions from different models plotted against dopamine peak values (in black). (I) Left: Relative change in dopamine as R_chosen (past rewards observed at the selected port) and R_unchosen (past rewards observed at the opposing port) change, calculated via LMER regression weights for dopamine observed in trials where the animals switched their port choices (animal switch trials). Right: Relative change in model RPE as R_chosen and R_unchosen change, calculated via regressions using model RPE predictions. (J) Similar to I, but for trials where the animal maintained their previous port selections (animal stay trials). Error bars show 95% bootstrapped confidence intervals.

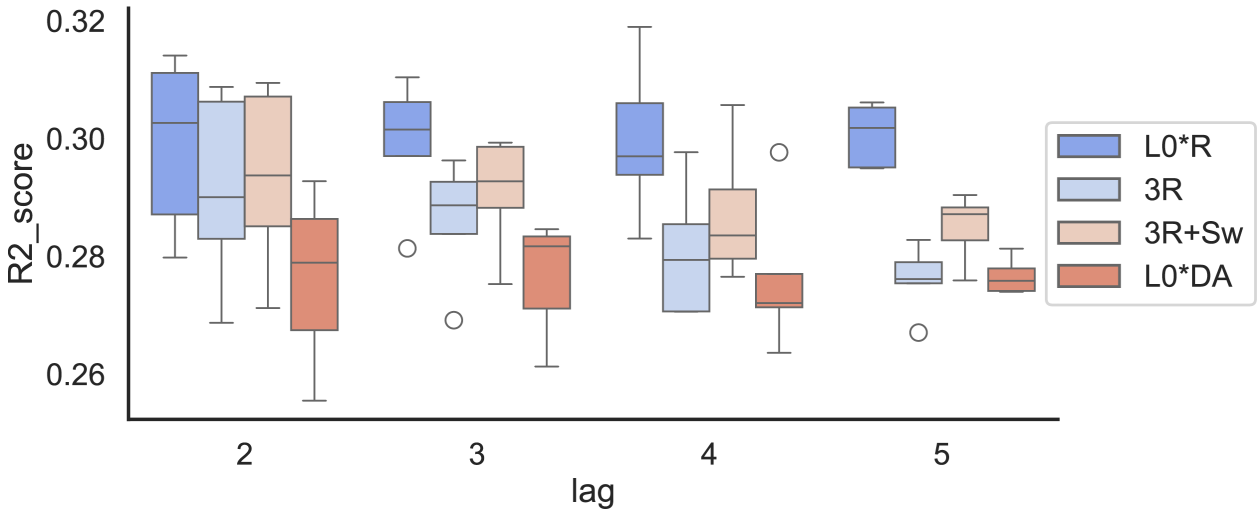

Figure 2.5: **Selection of LMER models and lags.** We compared the cross validated R2 score using different feature sets, while keeping the use of dopamine as output variable. L0:R: Standard formulation using N trial back choice and reward interactions. 3R: just R_chosen, R_unchosen, Reward features. 3R+Sw: in addition to 3R features, we included interactions of whether a trial is an animal switch trial, or animal stay trials. L0:DA: we used the interactions between past trial dopamine values and choice selections. Together we found that 3R+Sw was enough to capture similar levels of variance compared to L0:R. We used 4 lags because it allows us to capture a relatively high amount of past outcome levels, with only the expense of 1% variance explained. Error bars show 95% bootstrapped confidence intervals.

Chapter 3

Distinct Patterns of Bihemispheric Opponency in Dorsal Striatum Drives Active Rejections

3.1 Background

The circuit diagram of the dorsal striatum and basal ganglia contains an elegant opponent design both within and between hemispheres [140]. Within a hemisphere, the direct and indirect pathway formed by dopamine D1R-expressing direct pathway spiny projection neurons (dSPNs) and D2R/A2A expressing spiny projection neurons (iSPNs) respectively, are thought to compete to release or inhibit motor outputs via polysynaptic disinhibitory/inhibitory pathways [39,141]. For left vs. right orienting behaviors, it is known that push/pull competition also occurs between the basal ganglia of the two hemispheres [39]. This hemispheric competition is thought to start at input to the two sides of the dorsal striatum and then manifest downstream at the level of the colliculus [142]. Here, the hemisphere with dominant direct pathway activity wins the competition to drive orientation in the contralateral direction, or an ipsilateral hemisphere may be weakened by inhibition (or damage) allowing the opposing hemisphere to win in the balance [39, 142, 143]. This process can be charaterized as a "winner take all" competition in which inputs to the striatal dSPNs in the two hemispheres are competing to initiate outputs to the deep superior Colliculus (SC) via the substantia nigra reticualta (SNr) [39, 57, 142, 144–146].

So far, much of the basal ganglia data have been well described by opponent process model. Rewarding actions are thought to drive gains in dorsal striatal dSPN activity and 'Go' actions, while costs drive iSPN activity and 'No-Go' action [47, 51]. iSPNs are thought to preferentially encode costs not just because they inhibit the direct pathway and produce freezing stopping [42, 47], but also because, on a cellular level, they express dopamine D2 receptors that should facilitate intracellular PKA signaling when dopamine levels dip [49]. In a lateralized context, GO/No-Go dSPN/iSPN dichotomy could be neatly mapped onto re-

wards driving contralateral action through dSPNs and costs driving ipsilateral action through iSPNs. This is supported by unilateral stimulation experiments that produce dSPN contraversive actions and iSPN driven ipsi-versive actions [17,42] and insightful studies of action suppression that reveal ramps in iSPN neurons that track salient lateralized choice actions mice have been trained to suppress for over specific time intervals [46]. Despite the scientific significance of this mapping, it may not readily apply to another important class of decisions which involve active rejection, where subjects actively carry out motor behaviors to reject choices. As observed in human studies, actions or decisions involving cost encoding and avoiding costs have different behavioral characteristics, such as slower execution time and worse performance, compared to reward maximizing approach behavior [147]. For this kind of decision, it is not yet clear how costs or negative information should be integrated in the striatum, because in this special case costs should promote active rejection (moving away from a choice) rather than freezing or stopping (a no-go or suppressive action). There are also data which reveal in vivo that iSPNs and dSPNs often coactivated within a hemisphere when animals move contralaterally to the recorded hemisphere [52]. We first noted the potential importance of active rejection contexts when comparing manipulations of dSPN and iSPN function in a non-lateralized serial decision making, where we found that bilateral chemogenetic inhibition of dorsomedial striatal iSPNs did not impair a mouse's learned ability to avoid a non-rewarded choice, whereas bilateral inhibition of DMS dSPNs did disrupt this ability [81]. This led us to postulate that GABAergic inhibition of dSPNs may be a mechanism by which costs are incorporated into active decision making [81]. To further explore how costs and negative outcomes are encoded in dSPNs and iSPNs when choice rejection is active, we here used a serial decision making task, "Restaurant Row" [148], to allow mice to navigate an arena of 4 connected T-mazes and make lateralized accept or reject choices (Fig 3.1A) . At the start of each hallway, auditory offer tones indicated the probability and wait time for a reward at the upcoming T-junction right-hand arm. The serial nature of this task allowed us to isolate rejection of a choice without confounding it with accepting another choice and it allowed us to study choice rejection without asking the animal to hold still, wait, or inhibit an action, which may separately engage the indirect pathway. It also enabled us to benefit from the study of opponency and lateralized differences while matching motor action, a left turn to reject an offer vs. right turn to accept. This gave us the opportunity to potentially separate neural activity traces presumably driven by heavier weighing of costs (defined as trials that led to rejection choices) and compare them to trials where there was presumable heavier weight of benefits (defined as accept choices). Combined with our serial choice design, activity seen leading up to left turns but not right turns in Restaurant Row should therefore help us discriminate activity differences that lead up to active rejection decisions as well as motor actions.

To study neural activity, we opted to use bilateral fiber photometry in D1-Cre mice or A2A-Cre mice in this task. This method, facilitated by a commutator, allowed us to measure the two hemispheres simultaneously to capture the opponency at this bilateral level, something that is not currently feasible with bulkier mouse miniendoscope recordings. In each mouse line, we were able to use video analysis to isolate and compare activity in two

phases of each trial: a decision Phase I that started just after they heard a tone but were yet to reach the T-junction, and choice Phase II when they were making a turn left to reject or right to accept an tone offer. Using these methods our goal was to ask if hemispheric opponency was used in a similar or different fashion leading up to active accept or active reject decisions. We wanted to know if a dSPN contralateral activation bias, which we would expect for all accept choices, would also be present in active rejection choices, or if these rejection choices–based more on cost than benefit–were encoded differently. A sub-goal of this lateralization and opponency question was to ask if this contralateral bias, or alternate pattern, would be observed equally in Phase I (tone to T-junction hall way) or only in Phase II (then mice make the turn post T junction). By recording DMS activity in both D1-Cre and A2A-cre mice, we also planned to compare hemispheric opponency in two circuits with their own intra-hemispheric opponency. This allowed us to ask if iSPNs show mirror patterns of dSPNs, or show more unique opponency related to putative preferential weighing of costs by iSPNs. Finally, based on our past chemogenetic results [81], we wanted to test if there was a role for unilateral inhibition of dSPNs in active choice rejection (now lateralized in this task). To examine this, we tested if activity in one hemisphere was modulated below baseline after tone onset or entry in the T-junction, and we used unilateral optogenetic inhibition to test if reduction of dSPN activity in either the right or left hemisphere could promote active rejection behavior.

3.2 Results

To start our experiments, adult mice (n=5 D1-Cre) and n=5 A2a-Cre were injected with CaMKII GCaMP8f in the DMS and implanted with optic fibers and ferrules bilaterally. After recovery, they were trained to accept or reject offers for food pellets while running counter clockwise through four interconnected T mazes (Fig 3.1A) in a serial decision making task inspired by the previously published Restaurant Row [148]. At each individual T-maze, upon entering the straight hallway, the mice would hear an auditory tone indicating reward probabilities and wait time requirements for the upcoming right choice arm. The value of the offer tones could be clearly ranked worst to best : rank 1 (0% reward probability, 7 second wait), rank 2 (20% reward probability, 5 second wait), rank 3 (80% reward probability, 3 second wait), and rank 4 (100% reward probability, 1 second wait). Mice heard an offer tone while they traversed each of the 4 hallways before reaching the T-junction. At the T-junction, they could either turn right to accept the offer or turn left to reject the offer and advance to the next offer zone (Fig 3.1B).

When deciding to accept or reject offers, mice used offer tones, but also were informed by past trial history

Comparison of mouse behavior suggested that mice used the tones to inform their choice (OLS coef 0.095, CI: [0.088, 0.101], $p < 0.001$). On average, mice rejected 60% of the worst

offers and 15% of the best offers (Fig 3.1C).

While mice were running the Restaurant Row (RR) task, we collected video and used SLEAP offline [149] to extract the keypoint coordinates for mice's head, neck, torso, and tail head (see methods). Using the tone onset timing and mouse head location we were able to define Phase I for each trial as the time in the hallway after tone onset and before T-junction entry, and Phase II for each trial after the mice crossed the T-junction and entered the next hallway (left-reject) or reward zone (right-accept). Running speeds in the straight hallways (Phase I) varied trial by trial, but mice did not travel at a statistically different speed after hearing tones informing different offer ranks (Kruskal-Wallis test, p=0.062)(Fig 3.1D). Hallway travel time for accept and reject choices did show statistically significant yet minor (low-effect size) differences (Fig 3.1E, $p < 0.001$, cohen's D: 0.06, CI: [0.02, 0.1]) with mean travel times for accept choices mildly slower than reject. The physical movement trajectories for accept and reject choices also largely overlapped during Phase I and widely diverged during Phase II (Fig 3.1F). Given that self-paced movement may be related to vigor state, we opted to study neural data for accept and reject choices in fast, medium (mid), and slow trials separately (see below, Fig 3.1JK).

While we had good evidence that offer tone mattered to mice based on raw behavioral data (Fig 3.1C), we also considered that other latent factors from past task history may have contributed to decisions. To quantify these contributions, we fed both current trial offer rank and movement speed (indicated by `log_hall_time`) and past history variables that described previous encounters at the same restaurant, such as acceptance rate at past laps, as well as past interactions with other restaurants, like `offer_rank__b1` into a lasso logistic regression model. After careful feature selection and model comparison Fig 4.2, we found evidence that mice relied on offer rank on the current trial but also integrated previous acceptance at the same restaurant, and offer ranks and rewards at other restaurants in the past (for more details, see methods) (Fig 3.1I, Fig 4.2).

Different patterns of hemispheric opponency are observed in the dorsomedial striatum during active rejection and accepting decisions and movements

After considering behavior in the task, we next turned to analysis of bulk calcium activity in DMS SPNs of the two hemispheres. We limited analyses to imaging data from trials when mice made a clear accept/right turn (n=1488 trials, each animal has 474 ± 66 trials) or clear reject/left turn choice (n=1162 trials, each animal has 359 ± 91) without turning around or taking more than 1.5 seconds to traverse the hallway. Quit trials were excluded from analyses for this study.

For our first plot of neural data, we split data into fast, medium (mid), and slow trials and aligned baselined neural activity to tone onset, including neural activity during both Phase I hallway time and Phase II turning actions. In the plot (Fig 3.1K), the time of T-junction entry is marked with a green tick mark for each trial on the x-axis (rug plot). With

this organization we compared responses to high rank offer tones (black) and low rank offer tones (gray) in the hemisphere contralateral or ipsilateral to the upcoming turn. Here, what initially stands out is the prominent modulation after the T-junction entry during Phase II in the contralateral hemisphere in both dSPNs and iSPNs. This modulation occurs later in longer hallway movement time speeds, presumably due to the turn being made at that time. The observed increase in DMS activity contralateral to a turn is expected from the literature. The turn-related modulation did not appear to be significantly affected by high vs low offer size in contralateral side (contrastive sup-t test, see methods; p = 0.138, 0.249, 0.115 respectively for dSPN fast, mid, slow bins respectively). This offer rank comparison may have also been confounded by accept and reject choices occurring at both high and low ranks and the contributions of past trial history (Fig 3.1I).

In the ipsilateral DMS, we observed neural activity in dSPNs was signifcantly lower for low offer rank trials compared to high offer rank trials in fast and mid movement time bins (p=0.0125, $p < 0.001$ for fast and mid respectively, p=0.0788 for slow bin). In iSPNs, a significant contrast was observed in ipsilateral activity between low and high offer rank trials for the fast time bin ($p < 0.001$). The next goal of our study was to compare accept and reject choices, which were more common for low rank offer tones, but also occurred for higher offer rank tones. Before making this comparison we also worked to isolate activity in the hallway prior to the turn (Phase I) from action within the turn (Phase II) and match to the best of our ability hallway movement/path trajectories.

Isolation of Phase I and Phase II signals reveals new insights into hemispheric opponency during accept and reject decisions and actions

We next organized all trials by aligning to tone onset and then rank ordering trials by time to T-junction (Fig 3.2D-E). Then, to cleanly isolate Phase I hallway and Phase II turn activity, we used video tracking to identify time of head entry into the T-junction. We also filtered data so that movements in the hallway were deemed comparable, by using a Quadratic Discriminant Analysis (QDA) classifier to remove all trials where movement behavior from video analyses could be used to predict accept or reject choices above chance (see methods). This allowed comparison of data from similar movement trajectories.

With these isolation and filtering steps in place, we were in a position to analyze contra and ipsi hemispheric opponency the DMS during Phase I and Phase II separately, while matching movement trajectory. We found that in dSPNs in Phase I and Phase II, the opponent patterns observed for accept and reject choices were qualitatively different. In Phase I, after tone onset, the contra (right) hemi showed gain of dSPN activity when the upcoming choices was a rejection (sup-t permutation test, $p < 0.001$), whereas the ipsi hemi (left) showed gain of dSPN activity when the upcoming choice was to accept (sup-t permutation test, $p < 0.001$).

As we mentioned in the introduction, we were particularly interested in exploring a role for

unilateral downregulation of dSPN activity during active rejection, as we hypothesized that this might be an overlooked mechanism by which endogenous inhibitory activity encoding costs could drive lateralized hemispheric opponency. In Phase I during rejection trials, we found the ipsi hemisphere (left) showed modulation below baseline (sup-t permutation test, p = 0.0072). In Phase II of rejection trials, there was more dramatic modulation of ipsilateral dSPN activity below baseline ($p < 0.001$), while the contralateral hemisphere was only modestly up-regulated (p = 0.0313). Showing an interesting contrast, during accept trials in Phase II, the contralateral hemisphere showed significant positive modulation ($p < 0.001$) while the ipsilateral hemisphere was modulated downwards ($p < 0.001$, all tests in this section was corrected via Holm-Bonferroni). These data support a working model in which unilateral downregulation of dSPN activity in the ipsi hemisphere may contribute a special role to active rejection, challenging the notion that activity in contralateral hemisphere dSPNs would be present in both active accept and active reject decisions hemisphere as long as they mirror each other in the motor domain.

We next looked at iSPN activity in Phase I and Phase II during accept and reject trials. In iSPNs during Phase I in accept trials, activity in the ipsilateral hemisphere was mildly up-regulated above baseline while the contralateral hemisphere was down-regulated on average (p=0.0015, 0.0078 respectively, supT permutation test). In rejection trials, the pattern was reversed with contralateral activity upregulated and ipsilateral hemisphere downregulated (contra greater than 0, $p < 0.001$, ipsi less than 0, $p < 0.001$). In Phase II, contralateral activity was clearly upregulated relative to baseline in accept trials ($p < 0.001$) and ipsi downregulated relative to baseline in reject trials (p=0.05, after Holm correction), similar to activity patterns observed in dSPNs in Phase II. These data do not support a model in which iSPNs are more sensitive to costs than dSPNs in this context. The data are consistent with observations that iSPNs are often co-active with dSPNs [52].

Unilateral optogenetic inhibition of left hemi dSPNs selectively increased ipsiversive choice rejection for low rank offers

While opponency in the two hemispheres has long been understood to be a mechanism to generate orienting behaviors, the traditional heuristic focus has been on activation of dSPNs contralateral to an action to drive orientation and activation of iSPNs for freezing and stopping. We were interested in the possibility of a role of inhibition of dSPNs in driving choice rejection (also see [81]). Based on our calcium imaging studies we could see a role for inhibition of dSPNs in the ipsi/left hemisphere in response to low offer tones (Fig 3.1) or rejection of choices (Fig 3.2).

The next step to test this idea was to unilaterally inhibit dSPNs to test if we could drive an increase in rejection choices. As we set up these experiments we slightly modified the original task and so there were now 5 distinct offer ranks but only one single pellet flavor to bring out a wider offer rank range (Fig 3.3B). We used ArchT to unilaterally inhibit dSPNs in either the left and right DMS at the time of the offer tone onset and tapered off the laser

(250ms ramp down) after the mice either rejected and entered the next hall or accepted the offer by entering the reward zone (Fig 3.3A). When we did OLS test to regress reject rate by `Lsilence`, `Rsilence`, `offer_rank`, and interactions `Lsilence:offer_rank` and `Rsilence`, we found a significant positive effect of left hemisphere unilateral dSPN inhibition on reject rate (coef 0.4453, CI: [0.288, 0.602], $p < 0.001$), whereas right hemisphere unilateral dSPN inhibition did not have a significant beta coefficient. Interestingly, we also found a negative effect of `Lsilence:offer_rank` on rejection rate, which suggests that the impact of left dSPN inhibition reduces with higher offer rank trials. Driven by this, we zoomed in our analysis on pairwise tests, and found that left hemisphere unilateral dSPN inhibition enhanced rejection of two lowest rank offers (Fig 3.3C, after Holm correction, offer rank 1: p = 0.0015, Cohen's d: 0.177, CI: [0.08, 0.27], offer rank 2: p=0.0040, Cohen's d: 0.162, CI: [0.07, 0.26], offer rank 3: p=0.110, Cohen's d: 0.104, CI: [0.01, 0.2]). These data suggest unilateral GABAergic inhibition onto dSPNs could play a role in choice rejection but should also integrate with other ongoing activity relaying current offer and also past outcomes.

3.3 Experimental Methods

Animal protocol

Mice (all C57 Bl/6 male, bred in-house) were housed on a 12 h reversed light-dark cycle (lights on at 22:00) with nesting material. Prior to surgery they were group housed with access to food and water ad libitum. We conducted all animal procedures, which follow the principles outlined by the NIH Guide for the Care and Use of Laboratory Animals, according to the protocols approved by the University of California, Berkeley Institutional Animal Care and Use Committee (IACUC) and Office of Laboratory Animal Care (OLAC).

Surgery protocol

Mice (P83-100) were deeply anesthetized with 2.5% isoflurane in oxygen and placed on a heating pad and into a stereotactic frame (Kopf Instruments; Tujunga, CA). Anesthesia was maintained at 1%–2% isoflurane during surgery. An incision was made along the midline of the scalp and small burr holes were drilled over each injection site. Virus was delivered via microinjection using a NanojectII injector (Drummond Scientific Company; Broomall, PA). For fiber photometry experiments, 460 nL AAV5-syn-FLEX-GCaMP8f-WPRE (Addgene item #162379; titer 1.9×1013 gc/mL) was delivered bilaterally to the DMS (coordinates relative to Bregma: A/P: + 0.9 mm, M/L: ± 1.25 mm, D/V: + 3.0 mm) in A2A-Cre or D1-Cre mice. For optogenetic experiments, 460 nL AAV5-CAG-FLEX-ArchT-GFP (UNC Vector Core; titer 2x1012 vg/mL) or AAV5-CAG-ArchT-GFP (UNC Vector Core; titer 4x1012 vg/mL) was delivered bilaterally to the DMS. For all injections, pipets were held at depth for 3 min before virus was delivered at 0.1 mL/min; pipets remained at depth for 10 min before being retracted. Immediately following injections, the skull was gently scored with a scalpel and 3 mm long, 400 µm diameter optic fibers (Neurophotometrics, NA 0.37; Doric, NA 0.5; house made [Thorlabs parts], 0.48 NA) were lowered ventrally (relative to Bregma) + 2.9 mm (i.e., 0.1 mm above injection site) and cemented to the skull using Metabond. Mice were given subcutaneous injections of meloxicam (10 mg/kg) during surgery and 24 and 48 h after surgery. Mice were single-housed after surgery and allowed 3- 4 weeks for viral expression during recovery and behavioral training before experiments.

Histology and Imaging

Viral expression and placement of the optic fibers were confirmed histologically. Mice were anesthetized with 2.5% isoflurane and overdosed with an intraperitoneal injection of ketamine/xylazine before being transcardially perfused with 4% PFA. Extracted brains were kept in PFA for 16-24 h after perfusion then transferred to PB for 24 h before sectioning. Brains were coronally cut at 50 µm on a vibratome and stored in PB at 4C. Sections were stained with 640/660 deep-red nissl (1:200 dilution; Invitrogen) before being mounted and cover slipped.

Fiber Photometry (FP)

We recorded from all 5 mice with fiber implants in nucleus accumbens core (NAc) using Neurophotometrics FP3001 system at 20Hz. GCamP8f signals are measured in 470nm channel, while isosbestic control signal is measured in 410nm channel for artifact removal. During the learning switch phase, we start plugging in fiber implant without recording to get mice accustomed to moving and behaving with fibers. Starting from the first "full task" session, we alternate between a left hemisphere recording, right hemisphere recording, and no recording schedule. All FP recordings are synchronized to behavioral and video data via custom TTL systems and bonsai programs and saved for further processing. Recorded signals are preprocessed using custom python program to control for motion artifacts using 410nm reference channels [116]. Specifically, we first subtract the common DC shift by approximating a smooth trend across both channels. Then we perform a robust linear regression from reference channel to signal channel, and subtract the fitted 415nm artifact signal from 470nm channel recordings and obtain the processed dLight signals. All GCaMP8f signals are subsequently z-scored to account for across-session extraneous variations.

Optogenetic Manipulation

To investigate if we could induce rejection behaviors through optogenetically inhibiting dSPNs in striatum, we injected cre-dependent ArchT vector in D1-Cre animals. During the RR task, in 25% of trials, 30-40mW 532nm laser was turned on when the animal entered the Hall (before offer sound) and throughout Offer Zone. The laser was tapered off (250ms ramp down) when the animal enter the reward zone (Accept) or entered the next hall (Reject) or after 4 seconds.

Restaurant Row Task Training

To examine active rejection, mice were trained on an adapted version of Restaurant Row, an economic decision making task [148]. Briefly, the task requires mice to traverse a series of 4 interconnected T-mazes where all left turns are reject choices and right turns are accept choices. At the entry of each T-maze, one of four offer tones was played to signal both reward probability and wait time. We combined both reward probability and wait time costs to encourage sufficient numbers of rejection trials in a session. Before training, mice were first acclimated to the flavored pellets (banana, plain, grape, chocolate; Precision Pellets, 20 mg) for 1 week. Mice were then food restricted 24 h before the first training session and maintained above 80% initial body weight throughout the experiment. Training stages were designed to familiarize mice with, 1 retrieving pellets from the FED automated dispensers, 2) 100 % and 0 % percent offer tones, 3) the 80% offer tone, and 4) the 20 % offer tone. Additionally, and unlike [148], to encourage more rejections, mice were required to wait longer before reward availability on low probability trials (5 s for 20% and 7 s for 0 %) than on high probability trials (3 s for 80% and 1 s for 100 %). Mice were fully trained on the task in 10-14 days.

Task Behavior Analysis

All final behavioral analysis is conducted with custom implemented python packages. Behavioral data are first collected using a custom implemented Bonsai and Processing mdoule, and then preprocessed into behavioral data frames with each row representing a separate trial with various columns containing information regarding task meta data, mice choice data, and task relevant variables. For trials that mice failed to trigger camera systems, we included the data but noted the relevant variables as NaNs/null. For most behavioral analysis, outcome histories as well as covariates with null entries are dropped (which are boundary trials at the start of the sessions, or trials where animals were unmotivated and did not make a choice). All data plotting are generated via `seaborn` packages, and test statistics and various measures are generated using `statsmodels, pingouin and sklearn`.

Lasso logistic regression

To understand the contributions of past trial outcome histories to mouse decisions, we performed lasso logistic regressions and feature selections of 12 past trial and features involving previous trial offer rank, accept rate, and high offer rejection rate, etc. By carefully performing feature selection and model comparison (Fig 4.2), we identified the following model: `REJ~past_laps_ACC_rate+past_laps_high_offer_reject_rate+ past_baseline_reward_rate+offer_rank+offer_rank__b1+log_hall_time`.

Specifically, we iteratively tested the prediction accuracy of decision at trial t by using features up to k trials backwards, and remove one feature at a time to identify the changes in total accuracies. We confirmed feature importance as the differential contribution of features when we include them versus when we do not. Moreover, we evaluated lag effects on prediction accuracy and determined 12 lags provided decent information without running into low sample issues due to combinatorial problems. Then we used lasso or group lasso coefficient as a guide to remove spurious features and found a clear separation of previous visits of the same restaurants versus overall experiences with other restaurants. We then computed the features above through averaging and feature engineering, which allowed us to achieve similar accuracy than using the full set of lag features.

Neural Data Analysis

All analysis was carried out with a custom Python module for aligning, visualization, and data modeling. After baseline correction, we visually inspected both 415nm channel recording and 470nm channel recording, as well as the AUC-ROC scores between the rescaled baseline values against signal channel recordings. We picked sessions that meets the following conditions: AUC-ROC higher than 0.9, longer tail in distributions in 470nm recording compared to rescaled 415nm recordings suggesting clear fluorescence increases, and no sharp discontinuity in both channels' recordings. We use baseline-corrected and z-scored dLight signals for in-depth analysis, called Z(Ca). For neural data alignment, we interpolated Z(Ca) around event times and appended them to the trial level dataframe.

Aligned neural signals are appended to the trial level data frames as additional columns. Aligned Z(Ca^{2+}) is baselined by subtracting out its value at "outcome" time (we denote this procedure as de-base for simplicity). For trial average dopamine visualization, we first identified a list of relevant columns as grouping variables of interest, we then dropped any row with one or more NaN entries in the relevant columns. The trial averaged plots are then plotted using seaborn with the error bar being bootstrapped confidence intervals. When reporting summary statistics, we used bootstrapped confidence intervals to characterize mean estimates. For difference comparisons, we used suitable paired or unpaired tests and reported both the testing statistics as well as effect size, and the confidence intervals associated with it.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using Python with standard packages: scipy, statsmodels, pingouin. We included n=10 subjects with each around 3-5 high quality sessions. Statistical details of each analysis can be found in each figure legend, result section or correspondent method section. Error bars are 95% bootstrapped confidence intervals, and Holm-Bonferroni correction was applied when appropriate.

QDA filtering methods

To zoom in on trials in which the movement trajectories prior to motor execution Phase II were not predictive of future decisions, we used QDA filtering. This method allows us to compare neural data for trials that have similar motor sequences and rule out any lateralization motor effect in the neural data. We trained a QDA classifier to predict accept and reject trials from only the equally sampled movement trajectories. For each data point we can then obtain a prediction likelihood of the future decision. We focused on trials that "confused" the classifier and received a decision binary decision likelihood between 0.4-0.6. Then we zoomed in on these trials to perform downstream neural data visualization in Fig 3.2H-I.

Bootstrap timeseries testing

To compare the contrastive differences between two functional curve series, we developed the following testing procedure based on the sup-T statistics. To measure the maximum distinction between the two curves, we take the maximum of the time wise t statistics. To construct a null distribution, we permute across curves, and calculate a sup-T null statistics for each permutation. Then we compare our observed sup-T statistics with the null distribution to calculate the empirical p value. Detailed algoritm is given as follows.

Algorithm 1 Functional Curve Contrast 1-sided (Greater) Sup-T Permutation Test

Require:

$x_i(t),\ i \in \{1, \ldots, N_x\},\ t \in [0, \tau_i]$

$y_j(t),\ j \in \{1, \ldots, N_y\},\ t \in [0, \tau_j]$

$L \gg 0$ ▷ Permutation parameter L

$X(t) \leftarrow [x_1(t) \ldots x_{N_x}(t)]$

$Y(t) \leftarrow [y_1(t) \ldots y_{N_y}(t)]$

$u(t) \leftarrow \frac{\bar{Y}(t) - \bar{X}(t)}{\sqrt{\mathrm{Var}(Y(t))/N_y + \mathrm{Var}(X(t))/N_x}}$

$\tilde{u} \leftarrow \max_t u(t)$

$Z(t) \leftarrow [X(t) : Y(t)]$

$D \leftarrow [\,]$

for $l = 1 \ldots L$ **do**

$X_l, Y_l \leftarrow \mathrm{Permute}(Z)$

$d_l(t) \leftarrow \frac{\bar{Y_l}(t) - \bar{X_l}(t)}{\sqrt{\frac{\mathrm{Var}(Y_l(t)) + \mathrm{Var}(X_l(t))}{B}}}$

$\tilde{d_l} \leftarrow \max_t d_l(t)$

$D \leftarrow [D : \tilde{d_l}]$

end for

$p \leftarrow P_D(\tilde{u} > d)$ ▷ Empirical Quantile

3.4 Discussion

Here, we set out to study an important but poorly understood aspect of decision making: active rejection. We opted to focus on the hemispheric opponency in the dorsomedial striatum because this area is thought to be important for action selection and shows contralateral bias in lateralized orienting movements [18, 39, 57, 150]. Comparing SPN activity in the left and right hemispheres (Fig 3.2), we found that hemispheric opponency differed qualitatively between reject and accept choices as mice received offer information while approaching a T-junction decision point and then turned through the corners of the restaurant row task.

While hemispheric opponency in the dorsal striatum has long been appreciated [39,46,57], it was unclear how hemispheric opponency would manifest in a serial choice task where reward is on one side but the other side constitutes rejection of an offer that does not require freezing or stopping.

Based on past work in two alternative forced choice tasks, we would expect to observe dorsal striatum dSPNs (and possibly also iSPNs) in one hemisphere increase their activity as rodent oriented toward a contralateral choice [18, 52]. This growth in contralateral activity prior to a choice has even been tied to the accumulation of evidence for making a left right orienting decision [18]. However, it so far has been unclear how hemispheric opponency is used when decisions are motivated to avoid costs, but remain active.

Given a basic assumption of a contralateral movement bias in the striatum, one might

have predicted a priori that opponent activities in restaurant row during left and right turns would be symmetric. This symmetric opponency pattern seen in two alternative forced choice tasks should also emerge in restaurant row if opponent competition between the two hemispheres was agnostic to costs versus benefits.

An alternative plausible outcome would have been a cell-type/circuit specific outcome. Optogenetic behavior studies [151] and dopamine receptor pharmacology [49] might lead us to predict that costs may preferentially activate iSPNs while benefits will preferentially activate dSPNs. This differential activity onto iSPNs and dSPNs could be used to drive hemispheric opponency via different channels in reject vs accept trials, respectively.

Interestingly, we did not observe data consistent with what we can call "the symmetric hypothesis" or the "costs/iSPN specific encoding" hypothesis. Neural signals corresponding to accept and reject choices did show contra/ipsi asymmetry in Phase I and Phase II, but were opposite in accept and reject choices in both phases. This lack of matching in hemispheric opponency suggests the underlying neural circuit activity is very different in the two choice contexts.

Digging deeper into cell-type and circuit, when we examine iSPNs and dSPNs patterns, the observed patterns do not suggest a division of labor between iSPNs and dSPNS consistent with the costs/iSPN specific encoding hypothesis. There is opponency in dSPNs in response to costs, as well as iSPNs. In iSPNs, this opponency is also very weak in Phase I. These data highlight a putative role for dSPNs in encoding costs in an active context. Notably, these observations do not call into question the well-supported role of iSPNs in suppressing action when an organism are instructed to hold still to obtain a reward [46]. However, importantly, they do suggest we should not extrapolate this iSPN suppression function to choices where rejection involves active movement.

Going forward, our data suggest the field may pay more attention to a role for "push" and "pull" modulation of dSPN activity in hemispheric opponency, with inhibition of dSPNs in one hemisphere possibly playing an important role in successful active rejection. Here using calcium imaging, we found opponency in dSPNs was driven by downregulation of ipsi hemi dSPNs during active rejection in Phase II turns with only modest contra hemi up-regulation. In support of this possible role, unilateral optogenetic dSPN inhibition in the left hemisphere also enhanced choice rejection. This is consistent with our previous work [81], where we found that inhibiting bilateral dSPN activity disrupted learned choice suppression. One caveat is that unilateral optogenetic inhibition only enhanced rejection choices for the poorer rank offer tones. While this result was not across all tones, we consider that there may have been competing evidence for good offer tones that could not be overcome. Logistic regression models and log odds calculation from two alternative forced choice tasks [17] reveal why unilateral stimulation effects on behavior may only be detectable when mice are near 50% equivalence for the two actions but when mice are nearer ∼100% preference for one action (where high value can continue to infinity and generate an asymptotic accept rate) this evidence for an action cannot be overcome. This means equivalent shift in internal action value can occur but still not change behavior in the highest range [17]. Since our lowest offer tones were only approaching a 50% rejection rate, it may be that behavioral response rates

at these tones enables greater sensitivity to perturbation not present at higher offer tones due to the value function.

Interestingly, the contra-ipsi hemispheric opponency changed from the decision phase I to action phase II in accept trials. We speculate this may have reflected consideration of past or alternative outcomes in the decision phase. In contrast, active rejection trials recruited the same pattern of hemispheric opponency with the ipsi downregulation pattern begining with tone onset persisting and amplifying during motor execution. This downregualtion is measured as a bulk signal relative to baseline but could be downstream of cost encoding possibly represented in GABAergic circuits. These signals, if critical in context of active rejection, may have previously been obscured in tasks where both lateralized choices were possibly rewarded or costs were coupled to freezing and stopping. Based on these data taken from an auspicious task context, we posit a working model in which distinctive circuits encoding costs drive ipsi "pull" and encoding benefits drive contra "push" hemispheric opponency that together can contribute to online simultaneous cost/benefit evidence evaluation. We anticipate that it will be informative to add two hemispheres and this additional inhibitory cost encoding mechanism to network models of the basal ganglia.

This leads us to ask where this inhibitory signal might come from. There is growing list of sources of inhibition to the striatum. Currently this list includes local iSPN collaterals, local inhibitory interneurons, subcortical projections from arkypallidal neurons or other sources, and also GABAergic long range cortical inputs [152–154]. Our bulk imaging data are not consistent with drive from iSPNs, but activity from a few cells could play a role if they had focal access to local dSPNs. Arkypallidal neurons have already been implicated in a role for stopping ongoing action [38], but more experimental work will need to be done to test their role in lateralized action and integrative decision making. In future work, we will explore the role played by these options in active rejection. If one or more of these sources play a role in driving active rejection, they may become an attractive target for therapeutics.

3.5 Chapter Summary

This chapter demonstrated the distinct bihemispheric opponency activities in the striatal systems during accept and reject choices, contrary to conventional wisdom. This implies the potential for distinct neural circuitry for cost and benefit processing in the brain. These data point toward new circuit mechanisms for understanding decision making and behavioral control. They are most valuable if they can be translated to inform research and intervention for pathologies associated with an inability to control behavior, such as substance use disorders and obsessive compulsive disorder. In human contexts, there are many situations in which we actively reject options, driving to a coffee shop instead of a bar, for example, and we do not freeze or withhold actions to prevent them from occurring. Our observations should therefore galvanize studies of active rejection in animal models in pursuit of translational therapies.

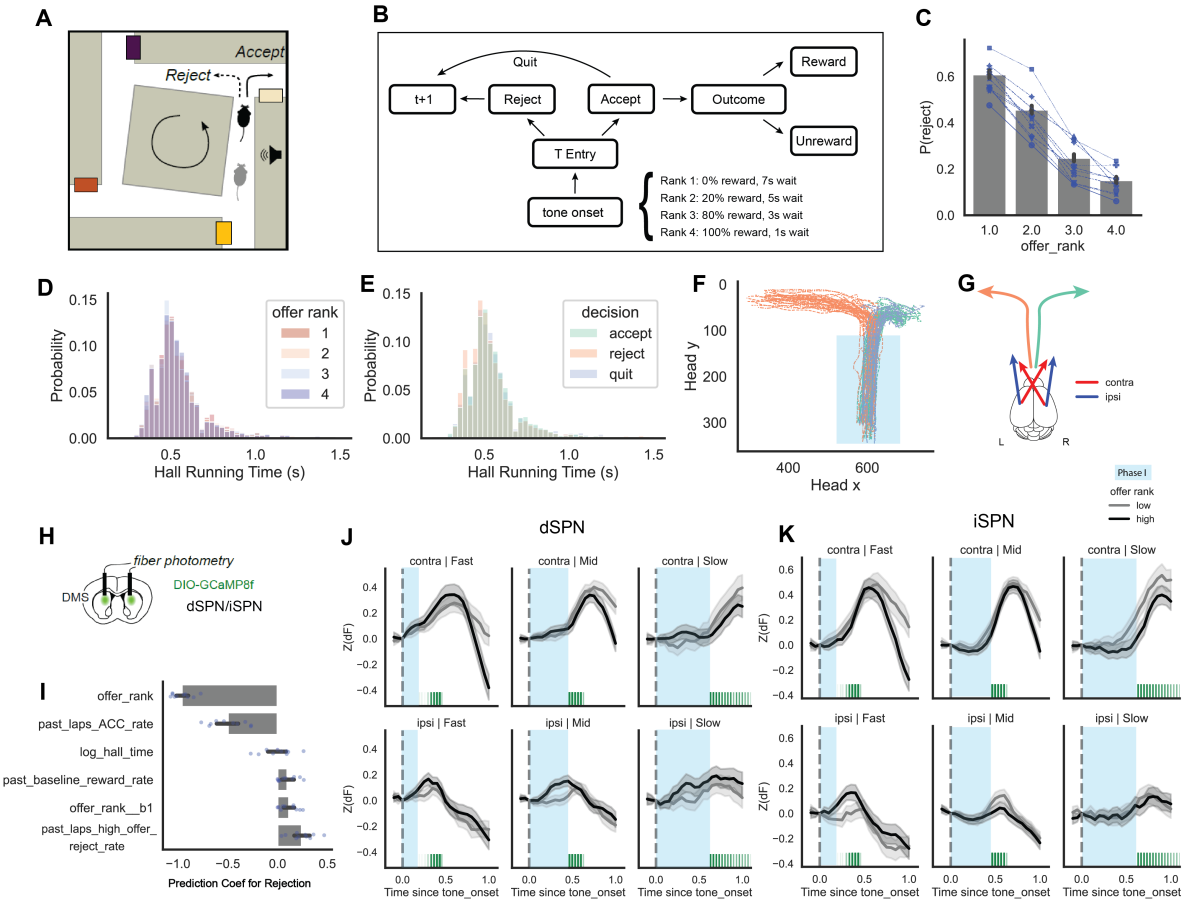
A
Accept
Reject
B
Quit
t+1
Reject
Accept
Outcome
Reward
Unreward
T Entry
tone onset
Rank 1: 0% reward, 7s wait
Rank 2: 20% reward, 5s wait
Rank 3: 80% reward, 3s wait
Rank 4: 100% reward, 1s wait
C
P(reject)
offer_rank
D
Probability
Hall Running Time (s)
offer rank
1
2
3
4
E
Probability
Hall Running Time (s)
decision
accept
reject
quit
F
Head y
Head x
G
contra
ipsi
Phase I
offer rank
low
high
H
fiber photometry
DMS
DIO-GCaMP8f
dSPN/iSPN
I
offer_rank
past_laps_ACC_rate
log_hall_time
past_baseline_reward_rate
offer_rank__b1
past_laps_high_offer_reject_rate
Prediction Coef for Rejection
J
dSPN
contra | Fast
contra | Mid
contra | Slow
ipsi | Fast
ipsi | Mid
ipsi | Slow
Z(dF)
Time since tone_onset
K
iSPN
contra | Fast
contra | Mid
contra | Slow
ipsi | Fast
ipsi | Mid
ipsi | Slow
Z(dF)
Time since tone_onset

Figure 3.1 *(preceding page)*: **Mice learn to use offer tones to accept or reject choices in Restaurant Row** (A) Diagram of the task arena for restaurant row (RR) task, where the mouse hears a tone as they enter the hallway, and then turns right to accept an offer and turns left to reject it. (B) Schematic showing task state transition structures from tone onset to reject, accept, or quit decisions. Note, we excluded complex quit decisions trials in this study. (C) Mice rejected offers with lower rank values (**higher** ranks mean **better** offers with higher reward probability and lower wait times). (D-E) The distribution for time elapsed as mice ran through the straight hall way of the T before reaching the T-junction, making a decision action, color coded by offer ranks (D) or decision categories (E). (F) Mouse head keypoint coordinates during decisions. Green tracks mark trials ending in accept choices, orange tracks mark trials ending in reject choices. Tracks ending in quit choice are not shown. Here we focus on 1 event time window: "Phase I", which notes the straight running or putative decision period. (G) Reference diagram further illustrates the color coding scheme for "contra" (red) and "ipsi" (blue) lateralized action, or "accept" (green) and "reject" (orange) decisions. (H) To record from dSPNs and iSPNs in DMS, we injected DIO-GCaMP8f bilaterally, and used fiber photometry to record bulk population activity simultaneously from both hemispheres during the task. (I) We ran a lasso logistic regression model to explore what factors beyond the current offer tone also contribute to the accept/reject decision for mice in the RR task. We found current offer tone was dominant contributor, but features like rate of accepting any offer in the previous visits at the same restaurant, "past laps ACC rate", rate of getting high offer in previous visit at the same restaurant, "past laps high offer" and "offer rank b1", offer rank one trial back, also contributed. (J) Neural activity for dSPNs during "Phase I" aligned to tone onset, split into fast, medium (mid), and slow decisions shown in three columns. Data from the hemisphere contralateral to the chosen accept/reject action are shown in the top row, while data from the hemisphere ipsilateral to the accept/reject chosen action are shown in the bottom row. Black is used to encode high offer rank ("good") offers, and gray is used to encode low offer rank ("bad") offers. Green floor markers were used to indicate the T-junction entry and end of Phase I start of Phase II (onset of lateralized decision actions). (K) Same as J but for iSPNs. From these coarse data we observe that these signals evolve with time, phase and high and low rank offers may differ in some trials depending on cell-type and speed. These data justified followup investigation in which we separate activity recorded in the T-junction hallway during Phase I, and activities during post T-junction turn in Phase II, and analyze reject and accept choices separately (averaging across effects and variation in tone offer and past history). Error bars show 95% bootstrapped confidence intervals.

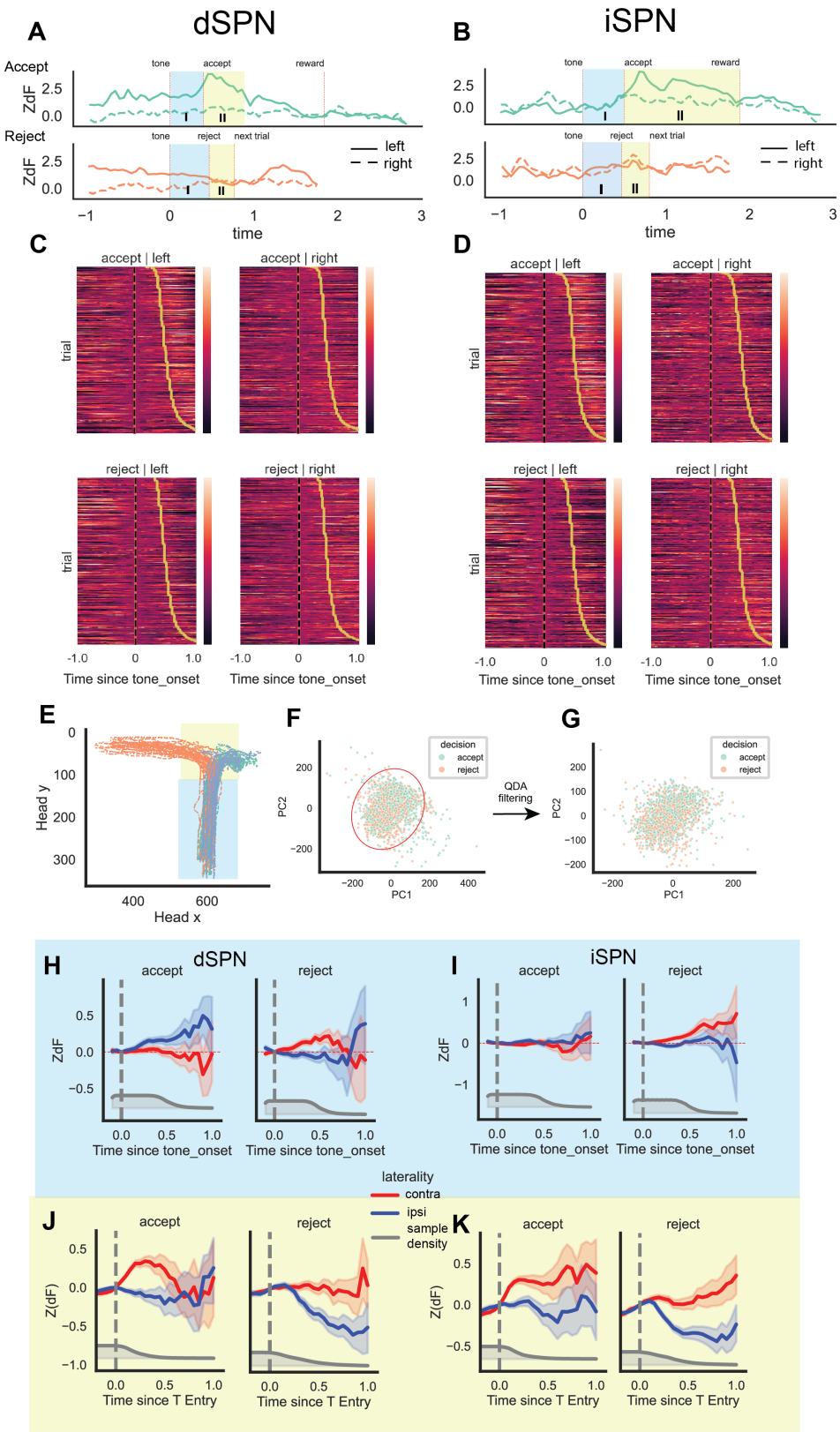
A
dSPN
B
iSPN
Accept
Reject
tone
accept
reject
reward
next trial
ZdF
left
right
time
C
D
accept | left
accept | right
reject | left
reject | right
trial
Time since tone_onset
E
F
G
Head y
Head x
decision
accept
reject
QDA filtering
PC1
PC2
H
I
dSPN
iSPN
Time since tone_onset
laterality
contra
ipsi
sample density
J
K
Z(dF)
Time since T Entry

Figure 3.2 *(preceding page)*: **dSPNs across two hemispheres displayed distinct opponency patterns during accepting and rejections** (A-B) Example dSPN activity during accept trials (top rows), and reject trials (bottom rows), and the left and right hemisphere signals are placed on the left and right column respectively. Phase I and II are indicated on the graph accordingly with light blue and light yellow. (C-D) Heatmaps for dSPN (C) and iSPN activity (D) during accept and reject trials across two hemispheres. Accept trials are in top row and reject trials in bottom rows. Trials are aligned to tone onset at time 0 and a yellow tick mark indicates T-junction entry for that trial (E-G) Schematic for movement control analysis to ensure that the trials under analysis during Phase I have overlapping movement patterns. (E) Movement was mapped with SLEAP. Phase I and II are indicated on the graph accordingly with light blue and light yellow. (F) QDA was used as a discriminator for different behavior trials and only trials where pre-T junction activity were uninformative of upcoming decisions were kept for analysis. (G) The first 2 Principal components (PCs) of the movement trajectories are plotted and labeled orange for reject and green for accept. (H) neural activity aligned to tone onset during Phase I, highlighting differences in opponency for accepting and rejection trials. dSPNs activities were plotted on the left panel and iSPNs activities were plotted on the right. The secondary plot in grey tones on the X-axis reveals how many trials were included and marks the decline in contributing data as time elapses. (I) Neural activity aligned to T junction entry during Phase II, highlighting differences in opponency during motor execution for accepting and rejection trials. dSPNs activities were plotted on the left panel and iSPNs activities were plotted on the right. Error bars show 95% bootstrapped confidence intervals.

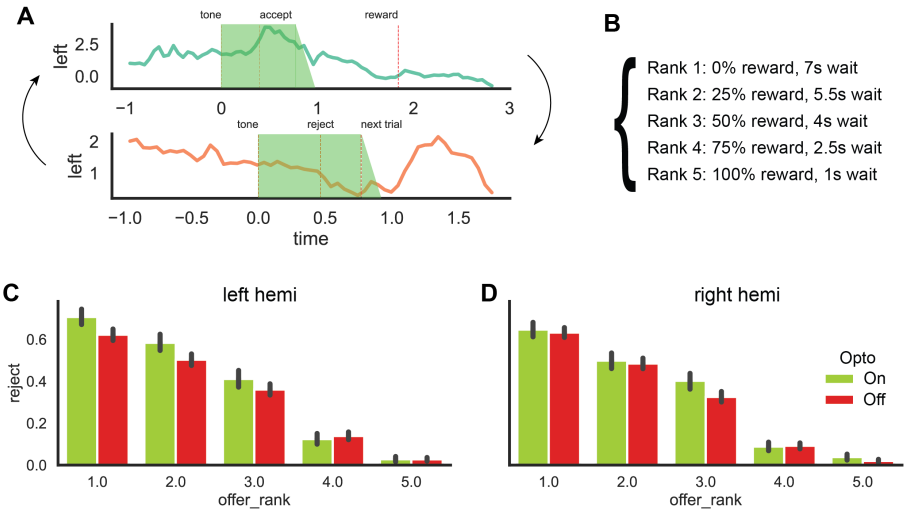

Figure 3.3: dSPN inhibition during the Restaurant Row increases choice rejection, but only for low rank offers. (A) ArchT stimulation window illustration on one example GCaMP trace. (B) Offer tone organization. (C–D) Rejection rates under optogenetic manipulation for left (C) and right (D) hemispheres.

Chapter 4

An Overview Experimental Design and Statistical Tools for Studying Decision Making Neuroscience

4.1 Variation in Task Design Have Facilitated the Study of Neural Mechanisms Underlying Decision Making

A wide variety of behavioral tasks are used in systems and circuit neuroscience. Insight into a specific function can come when the task isolates and varies specific constructs of interest. Several key task designs have proven particularly valuable for studying decision mechanisms. Two-alternative forced choice (2AFC) paradigms, where subjects must choose between two options with different reward contingencies, have revealed critical roles for various brain areas and cell types in value based decision making [5, 6, 46, 57]. This paradigm is often combined with perceptual decision making using continuous stimuli, and led to our understanding regarding evidence accumulation and decision formation [18, 155].

By introducing delays between choice and outcome, researchers can isolate neural signals related to decision confidence from those related to reward prediction errors [156]. Probabilistic reversal learning tasks, where reward contingencies periodically change without explicit cues, allowed investigation of how organisms detect and adapt to environmental volatility [17, 71, 76, 78]. By manipulating reward probability and magnitude independently, researchers can distinguish neural representations of expected value from subjective utility [157]. Effort-based decision tasks, where animals must weigh reward magnitude against effort costs, illuminate the neural basis of cost-benefit integration [158]. More recently, serial choice paradigms like "Restaurant Row" allow researchers to isolate quit decisions and regret [148] and here we use it to study such as lateralized offer evaluation and active rejection (Chapter 3).

Modern task designs increasingly incorporate naturalistic elements while maintaining experimental control. Free-foraging paradigms, animals navigate complex environments with multiple reward patches, reveal how decision processes operate during continuous behavior rather than discrete trials [101,159]. Social decision tasks, where choices affect other subjects, illuminate the neural basis of prosocial and competitive behaviors [160, 161]. Virtual reality environments allow precise control of sensory input while permitting naturalistic movement, bridging the gap between controlled laboratory tasks and real-world behavior [7, 162].

Through a process of experimental manipulation of the brain and testing across a variety of tasks, the field has been able to dissociate functions and reveal that there may be multiple decision systems operating in parallel within the brain. It is thought that decision making may be initially driven by a goal-directed system, involving dorsomedial striatum and prefrontal cortex, supporting flexible behavior guided by outcome expectations [12, 163]. However, with repeat experience decisions may fall under control of the habit system, centered on dorsolateral striatum, to enable rapid responses based on stimulus-response associations [24]. A Pavlovian system, not addressed in my thesis, is also present, driving more basic approach and avoidance responses to reward-predictive and threat-predictive cues through actions in the amygdala and ventral basal ganglia [63, 164].

When combined with advanced neural recording and manipulation techniques, these task paradigms provided increasingly granular insight into the neural basis of decision making. By correlating behavioral variables with neural activity, researchers could identify the computational functions implemented by different subregions and circuits, or even specific brain cells or cellular compartments [159, 165, 166]. By causally manipulating neural function during specific task phases, researchers can establish the necessity of identified mechanisms for adaptive decision making [76, 167]. This powerful approach has revolutionized and will continue to revolutionize our understanding of how biological organisms evaluate options, select actions, and learn from experience to optimize future decisions in complex, dynamic environments.

In order to bridge between behavior changes and cellular changes at multiple time scales, we have turned to a number of statistical tools, which I used over the course of my thesis. In the following sections, I will revisit them in detail to highlight their utility and relevance for extracting meaningful insights from data in this fast developing field.

4.2 Predictive Modeling of Task Behaviors

When studying decision making, it is often of interest to first understand task behaviors of the animals, such as how behaviors change with respect to task stimuli, or what neural activity patterns are related to motor behavior generation. One common assumption underlying these analyses is that, if certain covariates like task variables or neural activity, can give reliable prediction of what the animals are going to do well ahead of time, there is a higher chance for a "Granger" causal relationship to exist [168, 169]. This approach has proven particularly valuable for identifying the factors that influence choice behaviors.

In this section, we will discuss families of distinct prediction models that are commonly used in decision making neuroscience.

Logistic Regressions

Generalized Linear Models (GLMs) can generally provide a flexible framework for capturing the relationship between covariates and behavioral outcomes while accommodating the various data distributions inherent in behavioral experiments [170, 171]. Moreover, since the mapping is inherently linear, the result is often highly interpretable, and is often the first pass modeling step by many practitioners. As many behavioral tasks require the animals to make binary decisions, like go/no go or left/right, logistic regressions are among one of the most popular modeling framework in decision making neuroscience [17, 90]. Therefore, we will first begin by introducing different variations of logistic regressions and see them in practice.

In the simplest case, logistic regressions models binary decision probability as following a monotonic relationship with other experimental variables:

$$P(y_t = 1) = \sigma(\beta_0 + \sum_{i=1}^{n} \beta_i x_{it}) \tag{4.1}$$

where y_t is the decision on trial t, x_{it} represents the various predictor variables, β_i are the regression coefficients, and $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function that maps the linear **logits** to the choice probability. This formulation can accommodate a wide range of predictors, including stimulus properties, reward history, and contextual factors [172].

A key advantage of GLMs is their ability to capture the influences of trial history on current choices. By incorporating lagged variables representing previous choices and outcomes, researchers can quantify how decision processes integrate information across trials [90, 109]. For example, win-stay/lose-shift strategies manifest as positive coefficients for terms representing the interaction between previous choices and outcomes. Perseveration appears as positive coefficients for previous choices independent of outcomes, while reward sensitivity is reflected in coefficients for previous rewards independent of choices [17, 90].

Example: Modeling 2ABT with Logistic Regression

A fundamental question in decision neuroscience is how past outcomes influence current choices. In the project from Chapter 2, we implemented logistic regression models that regressed from past choices, rewards and the interaction terms. This allowed us to capture the decaying influence of previous trials in accordance with previous literature [17, 90, 173]:

$$P(C_t = 1) = \sigma\left(\beta_0 + \sum_{i=1}^{n} \beta_i^{CR} \tilde{C}_{t-i} R_{t-i} + \sum_{i=1}^{n} \beta_i^{R} R_{t-i} + \sum_{i=1}^{n} \beta_i^{C} \tilde{C}_{t-i}\right) \tag{4.2}$$

Where:

- C_t is the choice at time t (0 or 1)
- $\tilde{C}_{t-i} = 2C_{t-i} - 1$ is the transformed previous choice
- R_{t-i} is the reward outcome
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function
- β_i^{CR}, β_i^R, and β_i^C are regression weights capturing the influence of choice-reward interactions, rewards, and choices, respectively

This model allows us to quantify the decaying influence of past experiences and identify potential blocking effects where recent rewards might modulate the impact of more distant outcomes [101]. The weights form a choice-outcome history kernel that represents how mice integrate past experiences.

Beyond simple decaying weights, we also implemented more sophisticated models that capture the "blocking" effect of recent rewards on the influence of distant outcomes (Fig 4.1):

$$P(C_t = 1) = \sigma\Bigg(\beta_0 + \sum_{i=2}^{n} \beta_i^{CR+} \tilde{C}_{t-i} R_{t-i}[R_{t-1} = 0] + \sum_{i=2}^{n} \beta_i^{CR-} \tilde{C}_{t-i} R_{t-i}[R_{t-1} = 1] + \sum_{i=1}^{n} \beta_i^R R_{t-i} + \sum_{i=1}^{n} \beta_i^C \tilde{C}_{t-i}\Bigg) \tag{4.3}$$

This formulation tests whether the effect of past choice-outcome pairs is modulated by the most recent reward, providing insight into the hierarchical organization of outcome history integration.

Considerations for reducing model complexity and increasing parameter stability

One fundamental lesson from statistical learning theory is the bias-variance tradeoff [174], which states that the prediction error of any model on new data points can be decomposed into three parts:

1. the difference between average model prediction and the average of the true data generation function, or **bias**.
2. the variation of the model prediction with different unseen data points, or **variance**. This can be seen as a measure of stability of the model prediction
3. irreducible error that is either due to the limitation of the model or the inherent noisiness of the data.

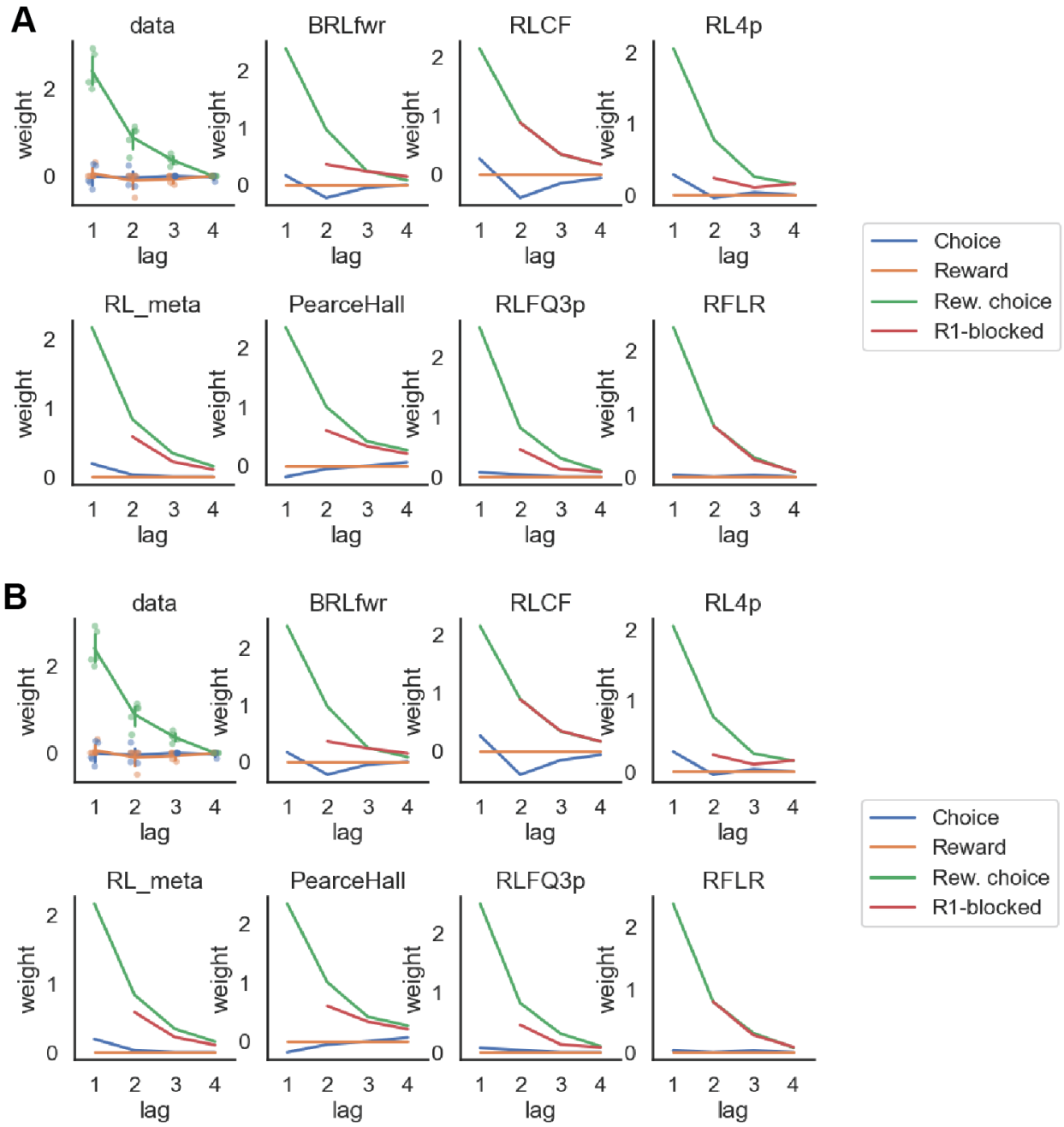

Figure 4.1: **Behavior analysis with logistic regressions and model simulations.** (A) Following the formulation from [90], we fitted logistic regression models to individual mice behavior data and identified consistent patterns of decaying choice outcome weights for more distant outcome histories. Error bars show 95% bootstrapped confidence intervals. Specifically, we used this formula ($\tilde{C}_{t-i} = 2C_{t-i} - 1$): $C_t = \sum_i \beta_i^{CR}\tilde{C}_{t-i}R_{t-i} + \beta_{t-i}^{R}R_{t-i} + \beta_{t-i}^{\tilde{C}}\tilde{C}_{t-i}$. (B) To identify the blocking effect of the reward at trial t-1 found in [101], we fitted the following regression model, where β_i^{CR+} represent the R1-blocked coefficient:
$C_t = \sum_{i=2}^{p} \beta_i^{CR+}\tilde{C}_{t-i}R_{t-i}\mathbf{1}[R_{t-1}=0] + \beta_i^{CR-}\tilde{C}_{t-i}R_{t-i}\mathbf{1}[R_{t-1}=1] + \sum_{i=1}^{p} \beta_{t-i}^{R}R_{t-i} + \beta_{t-i}^{C}\tilde{C}_{t-i}$

This is particularly important in the context of statistical modeling for decision making neuroscience, since we are building these prediction models so that we can find a reliable explanations of why and how the animals make their decisions. Ideally, we need to make sure that our model can not only explain our observed data (or **training data**) well, but also can readily generalize to future experimental observations. Due to this goal, it is often desirable to have a model that generates a stable prediction with low **variance** on unseen samples, and sometimes at the expense of increasing the **bias**.

But how do we achieve this? Some popular approaches include regularization methods like lasso, or elastic net [174,175]. The main intuition behind these methods is that, when we fit regression models to predict output like animal decisions, oftentimes a good proportion of the covariates we used actually offer little to no help in prediction. For instance, if we use the past 10 trials outcome history to predict the animals' decision, we may get some spurious and small coefficients for 7 or 10 trial back effects. More often than not, these estimations were not statistically significant but are rather due to pure data variations. Therefore, we may want to place some constraints on our regression models such that it should automatically shrink the inconsequential effect to zeros. In practice, for GLM regressions, the loss function we seek to minimize becomes:

$$\min_{\beta} \mathcal{L}(y, X\beta) + \lambda\theta||\beta||_1 + \lambda(1-\theta)||\beta||_2 \tag{4.4}$$

This describes the general formulation for elastic net. When $\theta = 1$, we only use L1 norm regularization, and problem becomes a lasso regression. When $\theta = 0$, the problem becomes a ridge regression where we try to control the L2 norm of the model parameter β. Moreover, there are situations where we would expect that different groups of lagged features should have their own sparsity constraint. For instance, we may have a good reason to believe that the amount of shrinkage penalty that we give to reward features at different trial lags should be comparable among each other, whereas such penalty may be completely different from the penalties that we would assign to past choice bias features. In this scenario, we would do group lasso, or group elastic net regression [176]. We included here an example from the project in chapter 3 to illustrate this process (Fig 4.2).

Even though regularization methods like lasso and elastic net provided a systematic statistical framework for model complexity control, sometimes domain specific insights could also help us come up with customary model "constraints". For instance, computational neuroscientists have developed GLMs that can incorporate time-varying parameters to capture learning dynamics, like adaptive forgetting by including exponentially weighted averages of past outcomes with learning rate parameters estimated from the data [177].

Cross validation and model selection

Methods like regularization allows us to alter model complexity and improve the generalizability and stability of the prediction power of the models. However, we still need to quantiatively assess the prediction ability. One particularly powerful method is cross valida-

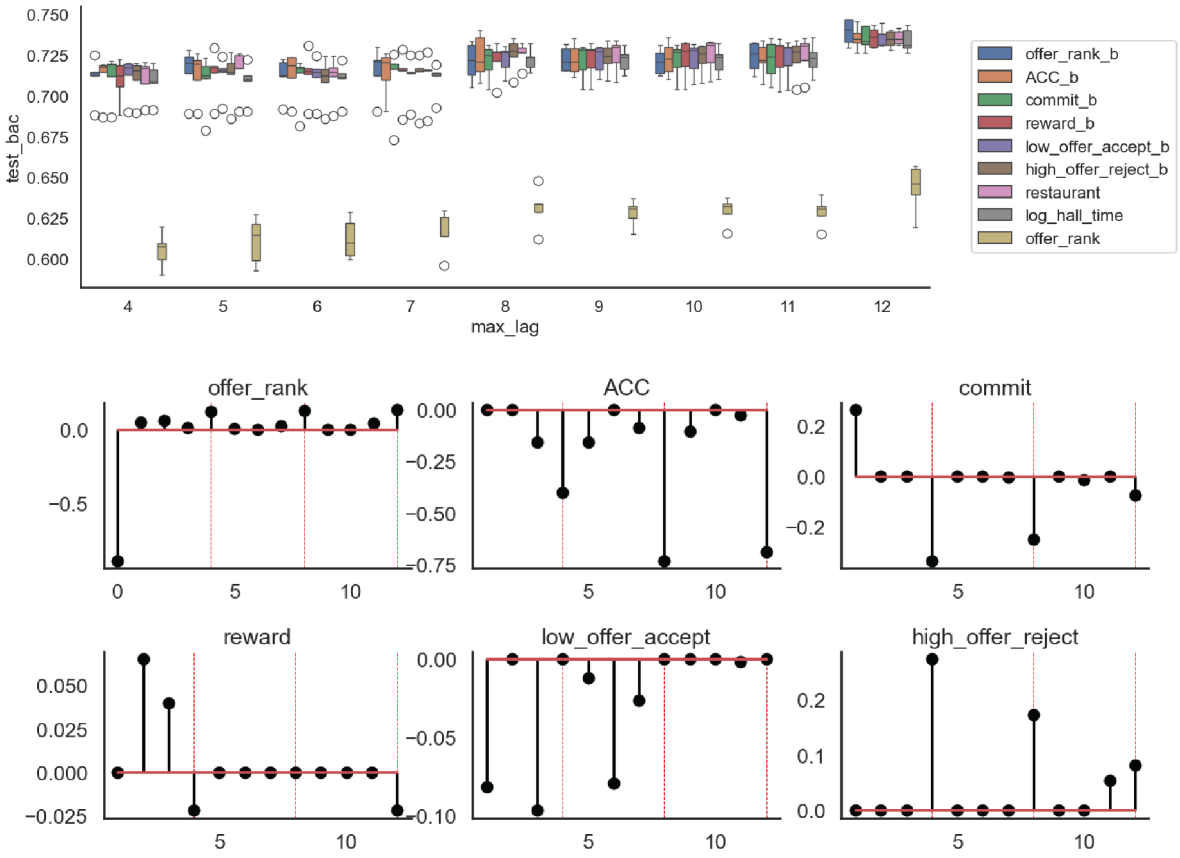

Figure 4.2: **Systematic variable selection procedure using lasso** To identify past trial outcome features that contribute to mouse decisions most, we performed step-wise feature selection. (A) Prediction accuracy of mouse decision by using feature including the past k lags, and by excluding any features. With more historical lags, we captured the behavioral variance better. (B) We then analyzed the lasso coefficients based on normalized inputs. Interestingly we found that information like offer rank, high offer reject rate, accept rate during the previous laps visiting the same restaurant, were particularly informative of future decisions. Integrating these evidence, we selected the final feature group as shown in Fig 3.1I.

tion (CV), which operates on the simple principle of splitting the dataset in two or multiple parts, training the model on one subset of the dataset, and testing the model prediction on the rest of the "unseen" datapoints [178].

One popular variant of cross validation is "K-Fold" cross validation [179]. During the procedure, both the covariates, or features X, and the target variable y are divided into K approximately equal-sized portions (with only rounding sample size differences by at most 1). For every fraction of the dataset S_i, we can fit the model to the other K-1 fractions S_{-i}, and then test the model on S_i using prediction error metrics like R squared for regression problems (continuous output like movement speed), or classification accuracy for classification problems (categorical output like left or right). Then we can obtain K such

out-of-sample evaluation metrics, and compute an average prediction error as an evaluation. Oftentimes, model hyperparameters like regularization strength can be picked to minimize prediction errors.

Since neuroscience problems often deal with temporal problems, where decisions at one time point depends on past history outcomes, pure shuffle based CV methods will likely lead to biases, since random shuffling disrupts the natural temporal order in the dataset, and then mistakenly training the model with datapoints from later trials of tasks and then testing the model on earlier trials of the tasks. This creates a "look-ahead" bias. To address this issue, we can use time series split, where we simply sort the dataset by trial number within the decision task, and evenly divide the trials again by K+1 trial chunks, then we can simply train the model using trial chunk up to K and test the model on trial chunk K+1. Finally, we can again summarize the measures by averaging across the folds, as in "K-Fold" CV.

Across these different CV procedures, we can vary the hyperparameters of the model, either via methods like grid search, or optimization framework like [180], and select the set that minimizes prediction error.

Extensions to Generalized Linear Models beyond decision prediction

When predicting behavioral variables that are more complex than binary decisions, like movement speed, or number of rewards earned in the next 10 trials, we can readily adapt the above framework to using Generalized linear model, where now we have the assumption that

$$\mu_t = f(\beta_0 + \sum_{i=1}^{n} \beta_i x_{it}) \tag{4.5}$$

$$E[y_t|X_t] = \mu_t \tag{4.6}$$

To solve this, one common procedure is maximum likelihood estimation, as described above: $\min_\beta \log l(y, X\beta)$. This topic will be further expanded to adapt to time series regression problems in section 4.5.

Machine Learning Prediction Models

Despite the flexible and interpretable framework afforded by GLM models, sometimes relationships between task variables or neural data and behavioral output can be nonlinear. This often requires us to utilize machine learning models, especially for the context of neural decoding, where we wish to use neural data to predict future decisions of the animals [181]. In this section, we will discuss a few machine learning methods commonly used in computational neuroscience.

Quadratic Discriminant analysis (QDA)

In areas like V1 or M1 in cortex, we can commonly model the neural activity as operating at some low dimensional manifold, which allows us to often treat their population activities as multivariate gaussian distribution [182–184]. As a result, the neural population activities underlying two distinct motor behavior or visual representations could often be modeled as two multivariate gaussians with distinct means. Quadratic discriminative Analysis (QDA) allows us to learn two multivariate gaussian distributions from empirical data, with which we can then predict which group (for instance, decision group) does a new data point belong to by comparing its similarity to the two groups (Fig. 4.3). This provides us with a reliable way to predict behavioral variable from neural data with a fairly parsimonious model.

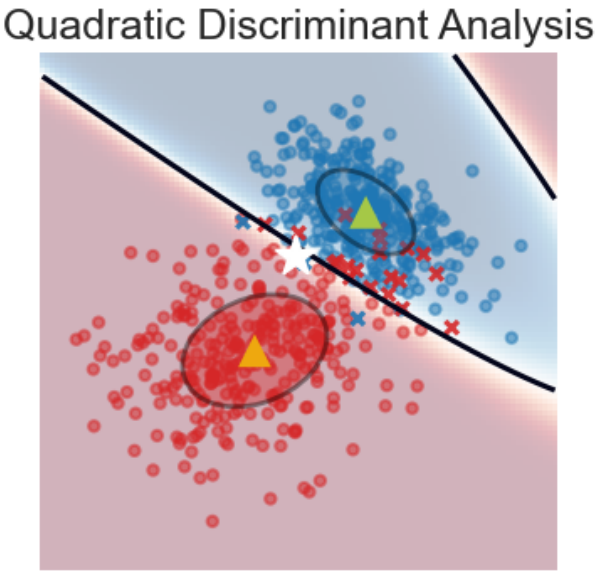

Figure 4.3: This schematic showcases an example of QDA modeling, red and blue each represent simulated gaussian neural data under a different stimuli. Using QDA, we can obtain a prediction boundary, which we can use to predict the white star datapoint to belong to the red class, therefore achieving the goal of "decision decoding"

Support Vector Machines (SVM)

Support Vector Machines provide a powerful framework for classification tasks in neuroscience, particularly when dealing with high-dimensional neural data where the relationship between neural activity and behavioral outcomes may be complex and nonlinear [185]. SVMs work by finding an optimal hyperplane that maximally separates different classes in feature space, making them well-suited for neural decoding problems where we aim to predict behavioral choices from neural activity [186, 187].

The fundamental principle of SVMs lies in maximizing the margin between decision boundaries. For a binary classification problem, the SVM seeks to find a hyperplane defined by:

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{4.7}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ represents the input features (e.g., neural firing rates), and b is the bias term. The optimization objective aims to maximize the margin $\frac{2}{||\mathbf{w}||}$ while correctly classifying training examples, leading to the constrained optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

where $y_i \in \{-1, +1\}$ are the class labels [188].

A key advantage of SVMs in neuroscience applications is their ability to handle nonlinear relationships through the kernel trick. By mapping data into higher-dimensional feature spaces using kernel functions such as the radial basis function (RBF) kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2)$, SVMs can capture complex decision boundaries that may better reflect the nonlinear dynamics of neural population activity [189]. This allows greater flexibility of the SVM to model nonlinear relationships that may exist between neural population activities, task variables and behavioral outputs.

Decision Trees and Random Forests

Decision trees offer an interpretable machine learning approach that recursively partitions the feature space based on threshold-based decisions, making them particularly valuable for understanding which neural features or task variables are most predictive of behavioral outcomes [174,190]. In neuroscience applications, decision trees can reveal hierarchical decision-making processes and identify critical neural signatures that drive behavioral choices [181].

A decision tree constructs a series of binary splits based on feature values, with each internal node representing a decision rule of the form $x_j < \theta$ for some feature j and threshold θ. The tree-building process uses impurity measures such as the Gini impurity:

$$G = \sum_{i=1}^{C} p_i(1 - p_i)$$

where p_i is the proportion of samples belonging to class i at a given node, and C is the number of classes. The algorithm selects splits that maximize information gain, defined as the reduction in impurity after the split [191].

While individual decision trees provide interpretability, they often suffer from high variance and overfitting, particularly problematic when dealing with noisy neural data. Random Forests address these limitations by combining multiple decision trees trained on bootstrap samples of the data, with each split considering only a random subset of features [192]. The final prediction is obtained through majority voting:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_B(\mathbf{x})\}$$

where $T_b(\mathbf{x})$ represents the prediction of the b-th tree and B is the total number of trees.

Random Forests have proven particularly effective for neural decoding tasks due to their ability to handle high-dimensional data, capture nonlinear relationships, and provide measures of feature importance [181]. The feature importance scores, calculated as the average decrease in impurity when a feature is used for splitting across all trees, can reveal which neural signals or task variables contribute most to behavioral predictions. This has been successfully applied to, for instance, decode movement syllables from population calcium activities [193]. The ensemble nature of Random Forests also provides natural estimates of prediction uncertainty [194], which can be valuable for assessing the reliability of neural decoding in different experimental conditions or behavioral states.

4.3 Investigating Behavioral Phenotypes with Dimensionality Reduction and Manifold Learning

Animal behavior consists of complex, high-dimensional patterns that are challenging to quantify manually. Recent advances in machine learning have revolutionized behavioral analysis by providing tools to automatically extract low-dimensional representations that capture essential behavioral structure [195, 196]. These approaches leverage dimensionality reduction and manifold learning techniques to transform raw behavioral data into interpretable features that can be linked to neural activity and task variables.

A new foundation for modern behavioral analysis is emerging from precise quantification of movement. Markerless pose estimation algorithms such as DeepLabCut, SLEAP, and LEAP can be used to track key anatomical landmarks from standard video recordings without requiring physical markers [149,197,198]. These methods generate high-dimensional time series data (e.g., x-y coordinates of multiple body parts across frames) that capture the animal's posture and movement with unprecedented detail.

To extract meaningful structure from this high-dimensional data, researchers have applied dimensionality reduction techniques that identify low-dimensional manifolds where the behavior naturally resides. Principal Component Analysis (PCA) provides a linear projection that maximizes variance in the data, capturing major axes of behavioral variability [199]. Non-linear techniques like UMAP preserve local relationships between data points, revealing clusters of similar behavioral patterns that might be obscured in linear projections [200].

Building on these representations, unsupervised methods can segment continuous behavior into discrete, stereotyped modules or "syllables". For example, Motion Sequencing (MoSeq) applies autoregressive hidden Markov models to depth camera recordings to identify behavioral syllables in freely moving rodents [195]. Keypoint-MoSeq extends this approach to work with pose tracking data, creating a modular representation of behavior that bridges the gap between raw movements and ethologically relevant actions [201]. These behavioral

syllables provide a natural vocabulary for describing complex action sequences and studying their neural correlates.

The real power of dimensionality reduction approaches emerges when behavioral representations are linked to neural data and task variables. By projecting neural activity onto behavioral manifolds, researchers can identify how specific neural circuits encode movement parameters and how this encoding is modulated by task context [202,203]. Conversely, by examining how task variables influence trajectories through behavioral state space, researchers can quantify how decisions shape action execution or learning [184, 204, 205].

Recognizing that behavior unfolds as trajectories through state space rather than isolated points, recent work has extended these approaches to incorporate dynamics. Gaussian Process Factor Analysis (GPFA), Variational Latent Gaussian Process (vLGP) models, and recurrent switching linear dynamical systems (rSLDS) capture smooth temporal dynamics while identifying low-dimensional structure [206–208]. These dynamic models reveal how behavioral states transition between one another and how these transitions are influenced by neural activity.

The integration of dimensionality reduction with reinforcement learning frameworks further illuminates how animals acquire and execute decision policies. By mapping policy representations onto behavioral manifolds, researchers can visualize how learning reshapes the landscape of possible actions and identify how specific behavioral features are selected to maximize reward [209]. This approach reveals not just what animals do, but how their behavioral repertoire is structured and optimized through experience.

These computational approaches to behavioral analysis are transforming neuroscience by providing quantitative frameworks that connect neural activity, task variables, and behavior across multiple timescales and levels of organization. By revealing the low-dimensional structure underlying complex behavioral patterns, they enable more precise investigations of the neural mechanisms that generate adaptive behavior, ultimately advancing our understanding of the biological basis of decision making.

4.4 Mechanistic Interpretation of Decision Process with Cognitive Modeling

Once we have a firm understanding of how certain aspects of tasks affect decisions, we can develop more formal views of decisions. Is the brain simply optimizing for reward? Are biases and, if so, how do they develop from experience? How do we mathematically formulate these claims?

Here we use the 2ABT task described in Chapter 2 as an example and formalize the decision problem facing animals as follows. At trial t, the animal makes a choice $c_t \in \{-1, 1\}$ and then observes a reward $r_t \in \{0, 1\}$. By convention, we take $c_t = -1$ to denote the left port, and $c_t = 1$ to denote the right port. The reward probability depends on the chosen

action and the hidden state $z_t \in \{-1, 1\}$:

$$P(r_t = 1|c_t, z_t) = \begin{cases} \rho_1 & \text{if } c_t = z_t \\ \rho_2 & \text{if } c_t \neq z_t \end{cases} \tag{4.8}$$

where ρ_1 is the probability of reward for a correct choice, which equals 0.75 during the full task, and ρ_2 is the probability of reward for an incorrect choice, which is technically 0 but which we set to 0.0001 to allow for model misspecification. The hidden state changes on each trial with probability $q = P(z_t \neq z_{t-1})$.

We assume a common functional form for the choice policy across models:

$$P(c_t = 1) = \sigma(\beta D_t + \phi c_{t-1}) \tag{4.9}$$

where $\sigma(x) = 1/(1 + e^{-x})$, $D_t = Q_t(1) - Q_t(-1)$ is the value difference, $\beta \geq 0$ is an inverse temperature parameter, $\phi \geq 0$ is a stickiness (choice perseveration) parameter, and $Q_t(c)$ is the value assigned to choice c at time t. The stickiness component captures the tendency to repeat choices independent of the reward history.

We can then compare a series of models which make different assumptions about how the values are updated based on experience. In my thesis work, I compared variations of eight model classes with different cognitive mechanisms to be more comprehensive. First, RL4p provided a baseline model with simple reinforcement learning updating, which updates the value expectations of choices iteratively. I also included models that expanded on the simple RL framework and equipped it with additional mechanisms like counterfactuals (RLCF) [8], dynamic learning rate updates (PearceHall) [102], and forgetting (RLFQ, RFLR) [90]. Finally, I included the formulation for the Bayesian inference model, as well as a hybrid model combining RL and Bayesian inference (BRL).

- **RL4p** is a standard RL model that updates Q-values according to a delta rule, with separate learning rates for positive (α^+) and negative (α^-) reward prediction error (RPE):

$$\Delta Q_t(c_t) = \alpha_t \delta_t \tag{4.10}$$

$$\alpha_t = \begin{cases} \alpha^+ & \text{if } \delta_t \geq 0 \\ \alpha^- & \text{if } \delta_t < 0 \end{cases} \tag{4.11}$$

  where $\delta_t = r_t - Q_t(c_t)$ is the RPE.

- **RLCF** is an elaboration of the RL4p model that uses "counterfactual" updates [8]. Whereas RL4p only updates the chosen action value, RLCF additionally updates the unchosen action value in the opposite direction:

$$\Delta Q_t(-c_t) = \alpha_t \tilde{\delta}_t \tag{4.12}$$

  where $-c_t$ is the unchosen action and $\tilde{\delta}_t = 1 - r_t - Q_t(-c_t)$ is the counterfactual RPE.

- **BRL** (belief state reinforcement learning) is an RL algorithm that computes the values as a linear function of the belief state $b_t(z) = P(z_t = z|c_{1:t-1}, r_{1:t-1})$, the posterior probability over the hidden state given the choice and reward history (the belief state update will be described further below):

$$Q_t(c_t) = w_1 b_t(c_t) + w_2 b_t(-c_t), \tag{4.13}$$

where w_1 and w_2 are learnable weights. Intuitively, the action value Q_t for a given action c_t is composed of two terms: the first term captures the reward weighted by probability when the animal is correct ($z_t = c_t$), and the second term captures the reward weighted by probability when the animal is incorrect ($z_t = -c_t$). If the animal has perfectly learned the task, the weights should be dictated by the ground truth reward probabilities, $w_1 = \rho_1$ and $w_2 = \rho_2$. We formalize a model of learning based on gradient descent, which allows the animal to approximate task rewards without knowing ground truth:

$$\Delta w_1 = \alpha \delta_t b_t(c_t) \tag{4.14}$$
$$\Delta w_2 = \alpha \delta_t b_t(-c_t), \tag{4.15}$$

where α is a learning rate. We considered two versions of the model. The weights were initialized to $w_1 = \rho_1$ and $w_2 = \rho_2$ (note that choosing other initial conditions had relatively little effect on model fits). Note that the weights can be interpreted as representing the Q-values of correct actions conditioned on the belief, with their update rule a confidence-weighted RL update (see e.g., [210]), hence the name BRL. We also considered a restricted version of the model (**BRLfwr**) where we fix $w_2 = \rho_2$, which we found to perform just as well as the unrestricted version, so we focused on the restricted version in the main results. Belief state updating followed Bayes' rule, with $P(z|\tilde{z})$ denoting the transition probability of the hidden state from $\tilde{z}$ to z:

$$b_{t+1}(z) \propto \sum_{\tilde{z} \in \{-1,1\}} P(z|\tilde{z}) b_t(\tilde{z}) P(r_t|c_t, \tilde{z}) \tag{4.16}$$

We assume that the belief state is initialized to $b_1(z) = 0.5$. Also we parameterize $q = P(z \neq \tilde{z}|\tilde{z})$ as the transition probability parameter of the model.

- **BIfp** (Bayesian inference with fixed parameters) makes use of the same belief state as the BRL models, but where there is no updating of the weights. Essentially, $D_t = 2b(1) - 1$. This is equivalent to $w_1 = \rho_1$ and $w_2 = \rho_2$ for BRL. Interestingly, the best fitted parameter for α in the unrestricted BRL version (BRLfw) above was zero for most subjects, establishing an equivalence, under this 2ABT task, between BIfp and and BRLfw.

- **RFLR** (recursively formulated logistic regression; [90]) is a form of RL that is mathematically equivalent to a form of logistic regression. RFLR updates the value difference

variable according to:

$$D_{t+1} = e^{-\frac{1}{\tau}} D_t + \alpha r_t c_t, \tag{4.17}$$

where τ is a timescale parameter governing the exponential decay rate of the values. Notably, we removed the non-negativity constraint on ϕ and fixed $\beta = 1$ in equation 4.9 for RFLR specifically, as it is consistent with the definition in [90] and yields slightly better performance.

- **RLFQ3p** (reinforcement learning model with forgetting, totaling 3 parameters) is a popular RL-based framework to model the forgetting, or Q value decay, of the non-chosen options. The formulation is similar to standard reinforcement learning model. However, Q value of the unchosen option decays as $Q_t(-c_t) = \zeta Q_{t-1}(-c_t)$. For model simplicity, we take $\zeta = \frac{\alpha^+ + \alpha^-}{2}$. Interestingly, the forgetting parameter ζ captures stickiness effect of past chosen options, as one could show their mathematical equivalence. Consequently, the RLFQ3p model we used here has low model complexity with only 3 parameters β, α^+, α^-.

- **RL_meta** (reinforcement learning model with "meta learning" [91] adopts a similar general framework as RL model with forgetting (ζ is the forgetting parameter), but adapts the learning rate parameter by unexpected uncertainty ν (the rate of this adaption is controlled via ψ). In other words, α^- varies, subjected to non-negativity, as a function of how surprising recent outcomes were:

$$\alpha_t^- = \begin{cases} \alpha_{t-1}^- & \text{if } \delta_t \geq 0 \\ \psi(\nu(t) + \alpha_0^-) + (1-\psi)(\alpha_{t-1}^-) & \text{otherwise} \end{cases} \tag{4.18}$$

$$\nu_t = |\delta_t| - \omega_{t-1} \tag{4.19}$$

$$\omega_t = \omega_{t-1} + \alpha_\nu \nu_t \tag{4.20}$$

Moreover, expected uncertainty variable ω, initialized from 0, further mediates the update of Q values as following, on top of learning rate adaptations:

$$Q_{t+1}(c_t) = \begin{cases} Q_t(c_t) + \alpha^+ \delta_t & \text{if } \delta_t \geq 0 \\ Q_t(c_t) + \alpha_t^- \delta_t & \text{otherwise} \end{cases} \tag{4.21}$$

- **PearceHall** (Pearce Hall model) is one of the most established reinforcement learning model with dynamic learning mechanism [102]. It adopts a similar general framework as RL model with forgetting (ζ is the forgetting parameter), but adapts a dynamic learning component α_ν with a adjustment rate ψ. Intuitively, α_ν functions as a running estimate of magnitude of RPEs. Formally, we have:

$$\delta_t = R_t - Q_t(c_t) \tag{4.22}$$

$$\alpha_\nu = \alpha_\nu + \psi(|\delta_t| - \alpha_\nu) \tag{4.23}$$

And Q values are estimated as follows, combining both dynamic learning, as well as forgetting:

$$Q_{t+1}(c_t) = \begin{cases} Q_t(c_t) + \alpha_\nu \alpha^+ \delta_t & \text{if } \delta_t \geq 0 \\ Q_t(c_t) + \alpha_\nu \alpha_t^- \delta_t & \text{otherwise} \end{cases} \tag{4.24}$$

$$Q_{t+1}(-c_t) = Q_0 + \zeta(Q_t(-c_t) - Q_0) \tag{4.25}$$

Model Comparison and identification

From the Q values, one can readily compute the decision likelihood and fit them to empirical data. Then we can aggregate the likelihood and estimate the best fitted model parameters by using maximum likelihood estimation (MLE) procedure. After obtaining the estimated parameters, there are three key considerations for arbitrating between model parameter recovery, model identifiability test, and model comparison. [211] provided a detailed discussion on this topic.

Model parameter recovery is useful to assess that whether the model fitting procedures can effectively estimate the ground truth parameters. We can calculate them with the following procedures:

1. With the estimate parameters across different subjects, we can construct an empirical range of parameters D
2. For each model, we can simulate behavioral data from randomly sampled model parameters within the empirical ranges D obtained above
3. Then we can refit the model to the simulated data to obtain the model-fitted parameters.
4. For each simulated experiment we can calculate the difference between ground truth parameter and fitted parameters, and then aggregate the errors. Parameter recovery can then be evaluated based on the aggregated errors and be used to inform the interpretability of the model parameters.

The model identification tests go through three stages: model behavior simulation, model cross-fitting, and confusion matrix construction. We denote our total model set as $\mathcal{M}$.

1. **Model behavior simulation**: Similar to the previous section, sampling randomly from the empirical range of parameters constructed from the mice data fitting, we simulated behavioral data $D_k(m_i)$ for 1000 times for each model $m_i \in \mathcal{M}$, where k indexes simulation runs.
2. **Model cross-fitting**: For each simulated data set $D_k(m_i)$, we fit each $m_j \in \mathcal{M}$, and get a best fitting model $m^*_{ki}|D_k(m_i)$ based on lowest AIC measure.

3. **Confusion matrix**: For all 1000 sessions, we can compute an empirical probability that the model m_j best fits to $D(m_i)$, or $P(m_j|m_i)$. Then we construct a confusion matrix from $P(m_j|m_i)$, $\forall m_i, m_j \in \mathcal{M}$, where each element at row i, column j represent $P(m_j|m_i)$.

To arbitrate and compare model fitness to the experimental data, we can perform model comparison by computing information criterion measures. These measures model fitness by considering both the model log likelihood as well as the model complexity. Here we discuss the equations for some common example measures: the Akaike Information Criterion (AIC), Bayeian Information Criterion (BIC), or Watanabe-Akaike Information Criterion for each model $\mathcal{M}_i$ [212, 213].

$$\text{AIC} = 2k - 2\sum_t \big(\ln P(c_t|\hat{\theta}, \mathcal{M}_i)\big), \tag{4.26}$$

$$\text{BIC} = k\ln n - 2\sum_t \ln P(c_t|\hat{\theta}, \mathcal{M}_i), \tag{4.27}$$

$$\text{lppd}(y, \Theta) = \sum_i \log \frac{1}{S}\sum_s p(y_i \mid \Theta_s) \tag{4.28}$$

$$\text{WAIC}(y, \Theta) = -2\left(\text{lppd} - \underbrace{\sum_i \text{Var}_\theta \log p(y_i \mid \theta)}_{\text{penalty term}}\right) \tag{4.29}$$

where $P(c_t|\hat{\theta}, \mathcal{M}_i)$ is the likelihood of the data c_t conditional on the maximum likelihood estimate ($\hat{\theta}$). These IC scores can be used to compare model fitness, where more negative scores indicate that a model explains the data better.

Cognitive modeling allows us to form mechanistic hypotheses and interpretation of the behavioral and neural data, and could share insights into the complex computational algorithm that the brain may employ during decision making.

4.5 Explaining Neural Representations of Behavioral and Decision Variables with Functional Regressions

Once we have a firm grasp of factors that affect decision making, we can start to address how these factors are represented in brain circuitry.

General Linear Models (GLMs) for Neural Data Analysis

To relate neural activity to behavioral variables, we could first employed conventional General Linear Models (GLMs) with temporal lags. There temporal lags allow us to capture lasting representations of specific behavioral variables in neural activity.

$$y(t) = \sum_{\tau} \beta_j(\tau) x_j(t-\tau) + \epsilon_t$$

where x_js representated individual behavioral covariates, like keypoints, offer values, past trial outcomes, etc. These models allowed us to identify the main effects of task variables and their temporal effects on neural activity and to test for interactions between these variables.

Functional GLMs for Temporal Dynamics

A key limitation of conventional GLMs with temporal lags is the increasing model complexity as we increase the length of the temporal window. Since neural signals often exhibit complex temporal patterns in response to behavioral events, with different predictors potentially influencing the signal over different timescales, it is useful to adopt a parsimonious representation of these regressive relationship by taking advantage of the temporal dependency structure in the data. Seminal work in computational neuroscience like [214,215] have introduced the usages of cosine spline basis or other basis functions to address this problem. However, without proper regularization, the regression problem may still encounter identifiability issue or generate unstable temporal predictions, especially under low-sample settings.

To address this, we could adopt methodologies from Functional Data Analysis (FDA). FDA provides a series of analysis methods that deals with functions or curves, with the assumptions that the empirically observed time series data have underlying "smooth" functions [216–219]. Here smoothness is commonly defined as the integral (or sum for discrete systems) of the squared second order derivative. Now, if the function is more wiggly with rapid changes in magnitudes, it would have higher rate of changes in the first order derivatives and therefore higher Ω.

$$\Omega(\beta) = (\int [L\beta(t)]^2 dt) \tag{4.30}$$

One of the simpliest formulation of the functional regression is functional scalar regression, which we could use to measure how a target variable, like dopamine level m seconds after the animal hears obtains a reward, changes as a function of past history or time. Functional scalar regression models the time series data as follows, while solving a PENSSE problem described below:

$$y_i = \alpha + \int \beta(t) x_i(t) dt \tag{4.31}$$

$$\text{PENSSE}_\lambda(\beta) = \sum_i \left(y_i - \alpha - \int x_i(t)\beta(t)dt \right)^2 + \lambda(\int [L\beta(t)]^2 dt) \tag{4.32}$$

This framework can also be readily extended with mixed effect methods to model trial-by-trial variability of neural data while accounting for subject-level variations [220].

However, this paradigm can be particularly limiting if we wish to model more complex temporal dependencies between neural signals and other continuous covariates, like functional connectivity effect from other neurons to target neuron. [221] proposed Functional Convolutional Models (FCMs) that model the neural response as a convolution of predictor time series with response kernels [221, 222]:

$$y(t) = \alpha + \sum_j \sum_\tau h_j(\tau)\ x_j(t-\tau) + z(t) + \epsilon(t) \tag{4.33}$$

where:

- $y(t)$ is the neural signal at time t
- α is a constant offset
- $h_j(\tau)$ is the response kernel for predictor j at lag τ
- $x_j(t-\tau)$ is the value of predictor j at time $t-\tau$
- $z(t)$ represents slow drift components
- $\epsilon(t)$ is the residual error

The response kernels $h_j(t-\tau)$ can often be approximated with basis functions like B-splines [216] to ensure smoothness and computational tractability as $h_j(\tau) = \sum_l \beta_{jl} \cdot B_l(\tau)$, where $B_l(\tau)$ are B-spline basis functions and β_{jl} are coefficients to be estimated.

With these flexible functional models, we would be able to identify effects of neural responses to different behavioral and neural events at different time scales, while maintaining a parsimonious model and generate stable predictions.

4.6 Chapter Summary

This chapter describes a wide range of computational methods commonly used to study decision making in the brain, from classial statistical models like GLMs, mechanistic models like Bayesian inference RLs, to manifold methods and functional regressions. As the neuroscience community collects neural recordings with higher density and temporal resolution,

these methods will allow us to unravel more mysteries regarding the computational processes underlying the brain's decision making processes.

Chapter 5

Conclusion

The classic actor-critic framework has provided a foundational understanding of how the brain implements reinforcement learning for decision making [3]. However, as detailed throughout this thesis, growing experimental evidence revealed that biological implementations of decision making involve further circuit and computational complexity that extend well beyond this traditional framework. Rather than employing a temporal difference learning based critic system, the brain utilizes Bayesian inference for state estimation and credit assignment. Moreover, instead of adopting a cost-benefit symmetric policy updating mechanism in the actor system, the bihemispheric SPNs in the striatum exhibit distinct opponency patterns that differ between cost-driven and benefit-driven decisions during action selection. This work demonstrates that while the core insight of separating value estimation from action selection remains valid, the specific computational and neural mechanisms underlying each component require substantial refinement to capture the full richness of biological decision making.

Experimental evidence demonstrating that dopamine reflects reward prediction error [4] and parallel processing across dorsal and ventral striatum performing distinct functions [5] has firmly established reinforcement learning as a viable biological learning paradigm. To date, a growing number of studies of the basal ganglia have found evidence of inference-based signals in both dorsal and ventral pathways [9, 71, 75–77]. In Chapter 2, we integrated reinforcement learning and Bayesian inference to propose a framework for biological learning and decision making that matches better with the experimental data, where the brain uses Bayesian inference to perform state inference and uses dopamine to guide value updating and reward predictions. Our experimental and computational work demonstrated that dopamine release in the nucleus accumbens provides evidence consistent with cognitive models that combine Bayesian inference and reinforcement learning [71, 78].

The dorsal striatum has long been considered the site of the actor component for action selection [39]. Classic evidence suggests that lateralized movement requires an increase in the activity of the striatal hemisphere contralateral to an orienting action [39]. More recently, it has been appreciated that an increase in unilateral dSPN activity is paralleled by an increase in iSPN activity in the same contralateral hemisphere [52]. There is also evidence

that this concurrent iSPN activity may serve to suppress undesired actions [46, 52]. This model has been very successful for 2AFC and Go/No-Go tasks but it was unclear how it should apply to serial choice tasks where responses to costs drive active rejection, not freezing and stopping. In Chapter 3, we demonstrated in a serial decision making task that the two hemispheres in dorsomedial striatum exhibit distinct opponency patterns when costs drive mice to turn to either reject or accept a choice option. The difference in opponency we observed was striking because it was asymmetric. In the classic model, we would predict that a contralateral activity bias would be present for both accept and reject choices, making both left-reject and right-accept choices symmetric in a restaurant row. However, given that the bihemispheric patterns for these two kinds of choices were not symmetric, it implies that cost and benefit processing in striatum may incorporate distinct circuit mechanisms, where cost encoding requires ipsilateral inhibition, or a "pull", while benefit encoding requires a contralateral increase in striatal activities, or a "push". This finding suggests that the neural actor model should be further explored for the possibility of distinct circuit mechanisms involved in cost and benefit processing.

Collectively, this work demonstrates that the brain's implementation of decision-making extends far beyond the classical actor-critic framework, incorporating sophisticated Bayesian inference mechanisms in the "critic" and asymmetric lateralized cost/benefit specialization in the "actor". By revealing these computational and circuit-level refinements, we move closer to a complete understanding of how biological systems achieve adaptive behavior in complex, uncertain environments. The integration of advanced experimental techniques with principled computational modeling offers a roadmap for future investigations into the neural basis of decision-making, allowing not only deeper insights into normal brain function but also new approaches to understanding and treating various disorders like addiction. As we continue to unravel the intricate neural computations underlying adaptive behavior, the integration of computational tools into complex brain systems will undoubtedly yield critical insights into the neural circuitry, and provide useful insights for advancement in artificial intelligence.

Bibliography

1. Joel D, Niv Y, Ruppin E. (2002). Actor–critic models of the basal ganglia: new anatomical and computational perspectives. Neural Networks;15(4-6):535–547. doi:10.1016/S0893-6080(02)00047-3.

2. Rescorla RA. (1971). Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. Learn Motiv;2(2):113–123. doi:10.1016/0023-9690(71)90002-6.

3. Sutton RS, Barto AG. (2018). Reinforcement learning: an introduction. Second edition ed. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press. Available from: `https://dl.acm.org/doi/10.5555/3312046`.

4. Schultz W, Dayan P, Montague PR. (1997). A Neural Substrate of Prediction and Reward. Science;275(5306):1593–1599. doi:10.1126/science.275.5306.1593.

5. Lee D, Seo H, Jung MW. (2012). Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci;35(1):287–308. doi:10.1146/annurev-neuro-062111-150512.

6. Lee AM, Tai LH, Zador A, Wilbrecht L. (2015). Between the primate and "reptilian" brain: Rodent models demonstrate the role of corticostriatal circuits in decision making. Neuroscience;296:66–74. doi:10.1016/j.neuroscience.2014.12.042.

7. Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, et al.. (2020). A Unified Framework for Dopamine Signals across Timescales. Cell;183(6):1600–1616.e25. doi:10.1016/j.cell.2020.11.013.

8. Eckstein MK, Master SL, Dahl RE, Wilbrecht L, Collins AGE. (2022). Reinforcement learning and Bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. Developmental Cognitive Neuroscience;55:101106. doi:10.1016/j.dcn.2022.101106.

9. Babayan BM, Uchida N, Gershman SJ. (2018). Belief state representation in the dopamine system. Nat Commun;9(1):1891. doi:10.1038/s41467-018-04397-0.

10. Barto AG, Sutton RS, Anderson CW. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern;SMC-13(5):834–846. doi:10.1109/TSMC.1983.6313077.

11. Grondman I, Busoniu L, Lopes GAD, Babuska R. (2012). A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews);42(6):1291–1307. doi:10.1109/TSMCC.2012.2218595.

12. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science;304(5669):452–454. doi:10.1126/science.1094285.

13. Glimcher PW. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences;108(3):15647–15654. doi:10.1073/pnas.1014269108.

14. Ito M, Doya K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr Opin Neurobiol;21(3):368–373. doi:10.1016/j.conb.2011.04.001.

15. Day JJ, Roitman MF, Wightman RM, Carelli RM. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci;10(8):1020–1028. doi:10.1038/nn1923.

16. Samejima K, Ueda Y, Doya K, Kimura M. (2005). Representation of action-specific reward values in the striatum. Science;310(5752):1337–1340. doi:10.1126/science.1115270.

17. Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci;15(9):1281–1289. doi:10.1038/nn.3188.

18. Yartsev MM, Hanks TD, Yoon AM, Brody CD. (2018). Causal contribution and dynamical encoding in the striatum during evidence accumulation. Elife;7. doi:10.7554/eLife.34929.

19. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, et al.. (2016). Asynchronous Methods for Deep Reinforcement Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR. Available from: `https://proceedings.mlr.press/v48/mniha16.html`.

20. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. (2017). Proximal Policy Optimization Algorithms. arXiv;1. doi:10.48550/arXiv.1707.06347.

21. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.. (2022). Training language models to follow instructions with human feedback. arXiv;1. doi:10.48550/arXiv.2203.02155.

22. Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, et al.. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv;1. doi:10.48550/arXiv.2402.03300.

23. Graybiel AM. (2008). Habits, rituals, and the evaluative brain. Annu Rev Neurosci;31(1):359–387. doi:10.1146/annurev.neuro.29.051605.112851.

24. Yin HH, Knowlton BJ. (2006). The role of the basal ganglia in habit formation. Nat Rev Neurosci;7(6):464–476. doi:10.1038/nrn1919.

25. Alexander GE, DeLong MR, Strick PL. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci;9(1):357–381. doi:10.1146/annurev.ne.09.030186.002041.

26. Haber SN. (2003). The primate basal ganglia: parallel and integrative networks. J Chem Neuroanat;26(4):317–330. doi:10.1016/j.jchemneu.2003.10.003.

27. Roseberry T, Lee AM, Lalive AL, Wilbrecht L, Bonci A, Kreitzer A. (2016). Cell-Type-Specific Control of Brainstem Locomotor Circuits by Basal Ganglia. Cell;164(3):526–537. doi:10.1016/j.cell.2015.12.037.

28. Voorn P, Vanderschuren LJMJ, Groenewegen HJ, Robbins TW, Pennartz CMA. (2004). Putting a spin on the dorsal-ventral divide of the striatum. Trends Neurosci;27(8):468–474. doi:10.1016/j.tins.2004.06.006.

29. Balleine BW, Delgado MR, Hikosaka O. (2007). The role of the dorsal striatum in reward and decision-making. J Neurosci;27(31):8161–8165. doi:10.1523/JNEUROSCI.1554-07.2007.

30. Gerfen CR, Surmeier DJ. (2011). Modulation of striatal projection systems by dopamine. Annu Rev Neurosci;34(1):441–466. doi:10.1146/annurev-neuro-061010-113641.

31. Hikosaka O, Kim HF, Yasuda M, Yamamoto S. (2014). Basal Ganglia Circuits for Reward Value–Guided Behavior. Annu Rev Neurosci;37(1):289–306. doi:10.1146/annurev-neuro-071013-013924.

32. DeLong MR. (1990). Primate models of movement disorders of basal ganglia origin. Trends Neurosci;13(7):281–285. doi:10.1016/0166-2236(90)90110-v.

33. Alexander GE, Crutcher MD. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends in Neurosciences;13(7):266–271. doi:10.1016/0166-2236(90)90107-L.

34. Middleton F. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Res Rev;31(2-3):236–250. doi:10.1016/s0165-0173(99)00040-5.

35. Nambu A. (2008). Seven problems on the basal ganglia. Curr Opin Neurobiol;18(6):595–604. doi:10.1016/j.conb.2008.11.001.

36. Haynes WIA, Haber SN. (2013). The organization of prefrontal-subthalamic inputs in primates provides an anatomical substrate for both functional specificity and integration: implications for Basal Ganglia models and deep brain stimulation. J Neurosci;33(11):4804–4814. doi:10.1523/JNEUROSCI.4674-12.2013.

37. Smith Y, Galvan A, Ellender TJ, Doig N, Villalba RM, Huerta-Ocampo I, et al.. (2014). The thalamostriatal system in normal and diseased states. Front Syst Neurosci;8:5. doi:10.3389/fnsys.2014.00005.

38. Mallet N, Schmidt R, Leventhal D, Chen F, Amer N, Boraud T, et al.. (2016). Arkypallidal cells send a stop signal to striatum. Neuron;89(2):308–316.

39. Redgrave P, Prescott TJ, Gurney K. (1999). The basal ganglia: a vertebrate solution to the selection problem? Neuroscience;89(4):1009–1023. doi:10.1016/S0306-4522(98)00319-4.

40. Cisek P, Kalaska JF. (2010). Neural mechanisms for interacting with a world full of action choices. Annu Rev Neurosci;33(1):269–298.

41. Jin X, Costa RM. (2015). Shaping action sequences in basal ganglia circuits. Curr Opin Neurobiol;33:188–196.

42. Kravitz AV, Freeze BS, Parker PRL, Kay K, Thwin MT, Deisseroth K, et al.. (2010). Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. Nature;466(7306):622–626.

43. Albin RL, Young AB, Penney JB. (1989). The functional anatomy of basal ganglia disorders. Trends Neurosci;12(10):366–375.

44. Gerfen CR, Engber TM, Mahan LC, Susel Z, Chase TN, Monsma FJ Jr, et al.. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. Science;250(4986):1429–1432.

45. Freeze BS, Kravitz AV, Hammack N, Berke JD, Kreitzer AC. (2013). Control of basal ganglia output by direct and indirect pathway projection neurons. J Neurosci;33(47):18531–18539.

46. Cruz BF, Guiomar G, Soares S, Motiwala A, Machens CK, Paton JJ. (2022). Action suppression reveals opponent parallel control via striatal circuits. Nature;607(7919):521–526.

47. Kravitz AV, Tye LD, Kreitzer AC. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nature neuroscience;15(6):816–818. doi:10.1038/nn.3100.

48. Yagishita S, Hayashi-Takagi A, Ellis-Davies GCR, Urakubo H, Ishii S, Kasai H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science;345(6204):1616–1620. doi:DOI: 10.1126/science.1255514.

49. Lee SJ, Lodder B, Chen Y, Patriarchi T, Tian L, Sabatini BL. (2021). Cell-type-specific asynchronous modulation of PKA by dopamine in learning. Nature;590(7846):451–456. doi:10.1038/s41586-020-03050-5.

50. Frank MJ. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci;17(1):51–72. doi:10.1162/0898929052880093.

51. Collins AGE, Frank MJ. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev;121(3):337–366. doi:10.1037/a0037015.

52. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, et al.. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. Nature;494(7436):238–242. doi:10.1038/nature11846.

53. Klaus A, Alves da Silva J, Costa RM. (2019). What, if, and when to move: Basal ganglia circuits and self-paced action initiation. Annu Rev Neurosci;42(1):459–483. doi:10.1146/annurev-neuro-072116-031033.

54. Labouesse MA, Torres-Herraez A, Chohan MO, Villarin JM, Greenwald J, Sun X, et al.. (2023). A non-canonical striatopallidal Go pathway that supports motor control. Nat Commun;14(1):6712. doi:https://doi.org/10.1038/s41467-023-42288-1.

55. Cazorla M, de Carvalho FD, Chohan MO, Shegda M, Chuhma N, Rayport S, et al.. (2014). Dopamine D2 receptors regulate the anatomical and functional balance of basal ganglia circuitry. Neuron;81(1):153–164. doi:10.1016/j.neuron.2013.10.041.

56. Kupchik YM, Brown RM, Heinsbroek JA, Lobo MK, Schwartz DJ, Kalivas PW. (2015). Coding the direct/indirect pathways by D1 and D2 receptors is not valid for accumbens projections. Nat Neurosci;18(9):1230–1232. doi:https://doi.org/10.1038/nn.4068.

57. Hikosaka O, Takikawa Y, Kawagoe R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. Physiological reviews;80(3):953–978. doi:10.1152/physrev.2000.80.3.953.

58. Foster NN, Barry J, Korobkova L, Garcia L, Gao L, Becerra M, et al.. (2021). The mouse cortico-basal ganglia-thalamic network. Nature;598(7879):188–194. doi:10.1038/s41586-021-03993-3.

59. Clapp M, Bahuguna J, Giossi C, Rubin JE, Verstynen T, Vich C. (2025). CBGTPy: An extensible cortico-basal ganglia-thalamic framework for modeling biological decision making. PLoS One;20(1):e0310367. doi:10.1371/journal.pone.0310367.

60. Yin HH, Mulcare SP, Hilário MRF, Clouse E, Holloway T, Davis MI, et al.. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. Nat Neurosci;12(3):333–341. doi:10.1038/nn.2261.

61. Gremel CM, Costa RM. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat Commun;4(1):2264. doi:10.1038/ncomms3264.

62. Cox J, Witten IB. (2019). Striatal circuits for reward learning and decision-making. Nat Rev Neurosci;20(8):482–494.

63. Soares-Cunha C, Coimbra B, Sousa N, Rodrigues AJ. (2016). Reappraising striatal D1- and D2-neurons in reward and aversion. Neurosci Biobehav Rev;68:370–386. doi:10.1016/j.neubiorev.2016.05.021.

64. Gelman A, Carlin JB, Stern HS, Rubin DB. (1995). Bayesian Data Analysis. 0th ed. Chapman and Hall/CRC. Available from: `https://www.taylorfrancis.com/books/9781135439415`.

65. Gershman SJ, Norman KA, Niv Y. (2015). Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences;5:43–50. doi:10.1016/j.cobeha.2015.07.007.

66. Ma WJ, Beck JM, Latham PE, Pouget A. (2006). Bayesian inference with probabilistic population codes. Nat Neurosci;9(11):1432–1438. doi:10.1038/nn1790.

67. Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, et al.. (2008). Probabilistic Population Codes for Bayesian Decision Making. Neuron;60(6):1142–1152. doi:10.1016/j.neuron.2008.09.021.

68. Gershman SJ, Uchida N. (2019). Believing in dopamine. Nat Rev Neurosci;20(11):703–714. doi:10.1038/s41583-019-0220-7.

69. Hennig JA, Romero Pinto SA, Yamaguchi T, Linderman SW, Uchida N, Gershman SJ. (2023). Emergence of belief-like representations through reinforcement learning. PLOS Computational Biology;19(9):e1011067.

70. Sohn H, Narain D, Meirhaeghe N, Jazayeri M. (2019). Bayesian Computation through Cortical Latent Dynamics. Neuron;103(5):934–947.e5. doi:10.1016/j.neuron.2019.06.012.

71. Mishchanchuk K, Gregoriou G, Qü A, Kastler A, Huys QJM, Wilbrecht L, et al.. (2024). Hidden state inference requires abstract contextual representations in the ventral hippocampus. Science;386(6724):926–932. doi:10.1126/science.adq5874.

72. Garvert MM, Saanum T, Schulz E, Schuck NW, Doeller CF. (2023). Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. Nat Neurosci;26(4):615–626. doi:10.1038/s41593-023-01283-x.

73. Elston TW, Wallis JD. (2025). Context-dependent decision-making in the primate hippocampal-prefrontal circuit. Nat Neurosci;28(2):374–382. doi:10.1038/s41593-024-01839-5.

74. Bartolo R, Averbeck BB. (2020). Prefrontal Cortex Predicts State Switches during Reversal Learning. Neuron;106(6):1044–1054.e4. doi:10.1016/j.neuron.2020.03.024.

75. Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O. (2010). A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. Journal of Neurophysiology;104(2):1068–1076. doi:10.1152/jn.00158.2010.

76. Costa VD, Tran VL, Turchi J, Averbeck BB. (2015). Reversal Learning and Dopamine: A Bayesian Perspective. J Neurosci;35(6):2407–2416. doi:10.1523/JNEUROSCI.1989-14.2015.

77. Starkweather CK, Babayan BM, Uchida N, Gershman SJ. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. Nat Neurosci;20(4):581–589. doi:10.1038/nn.4520.

78. Qü AJ, Tai LH, Hall CD, Tu EM, Eckstein MK, Mishchanchuk K, et al.. (2025). Nucleus accumbens dopamine release reflects Bayesian inference during instrumental learning. PLoS Comput Biol;21(7):e1013226. doi:10.1371/journal.pcbi.1013226.

79. Langlois TA, Charlton JA, Goris RLT. (2025). Bayesian inference by visuomotor neurons in the prefrontal cortex. Proceedings of the National Academy of Sciences;122(13):e2420815122. doi:10.1073/pnas.2420815122.

80. Tecuapetla F, Jin X, Lima SQ, Costa RM. (2016). Complementary contributions of striatal projection pathways to action initiation and execution. Cell;166(3):703–715. doi:10.1016/j.cell.2016.06.032.

81. Delevich K, Hoshal B, Zhou LZ, Zhang Y, Vedula S, Lin WC, et al.. (2022). Activation, but not inhibition, of the indirect pathway disrupts choice rejection in a freely moving, multiple-choice foraging task. Cell Reports;40(4):111129. doi:10.1016/j.celrep.2022.111129.

82. Gläscher J, Hampton AN, O'Doherty JP. (2009). Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. Cerebral Cortex;19(2):483–495. doi:10.1093/cercor/bhn098.

83. Hauser TU, Iannaccone R, Walitza S, Brandeis D, Brem S. (2015). Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. NeuroImage;104:347–354. doi:10.1016/j.neuroimage.2014.09.018.

84. Santacruz SR, Rich EL, Wallis JD, Carmena JM. (2017). Caudate Microstimulation Increases Value of Specific Choices. Current Biology;27(21):3375–3383.e3. doi:10.1016/j.cub.2017.09.051.

85. Montague P, Dayan P, Sejnowski T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci;16(5):1936–1947. doi:10.1523/JNEUROSCI.16-05-01936.1996.

86. Eshel N, Tian J, Bukwich M, Uchida N. (2016). Dopamine neurons share common response function for reward prediction error. Nat Neurosci;19(3):479–486. doi:10.1038/nn.4239.

87. Takahashi YK, Stalnaker TA, Mueller LE, Harootonian SK, Langdon AJ, Schoenbaum G. (2023). Dopaminergic prediction errors in the ventral tegmental area reflect a multithreaded predictive model. Nat Neurosci;26(5):830–839. doi:10.1038/s41593-023-01310-x.

88. Blanco-Pozo M, Akam T, Walton ME. (2024). Dopamine-independent effect of rewards on choices through hidden-state inference. Nat Neurosci;27. doi:10.1038/s41593-023-01542-x.

89. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Current Biology;27(6):821–832. doi:10.1016/j.cub.2017.02.026.

90. Beron CC, Neufeld SQ, Linderman SW, Sabatini BL. (2022). Mice exhibit stochastic and efficient action switching during probabilistic decision making. Proc Natl Acad Sci USA;119(15):e2113961119. doi:10.1073/pnas.2113961119.

91. Grossman CD, Bari BA, Cohen JY. (2022). Serotonin neurons modulate learning rate through uncertainty. Current Biology;32(3):586–599.e7. doi:https://doi.org/10.1016/j.cub.2021.12.006.

92. Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, et al.. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nature Neuroscience;19(6):845–854. doi:10.1038/nn.4287.

93. Foerde K, Braun EK, Shohamy D. (2012). A Trade-Off between Feedback-Based Learning and Episodic Memory for Feedback Events: Evidence from Parkinson's Disease. Neurodegenerative Diseases;11(2):93–101. doi:10.1159/000342000.

94. Collins A, Koechlin E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. PLOS Biology;10(3):1–16. doi:10.1371/journal.pbio.1001293.

95. Rmus M, McDougle SD, Collins AG. (2021). The role of executive function in shaping reinforcement learning. Current Opinion in Behavioral Sciences;38:66–73. doi:10.1016/j.cobeha.2020.10.003.

96. Beier K, Steinberg E, DeLoach K, Xie S, Miyamichi K, Schwarz L, et al.. (2015). Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. Cell;162(3):622–634. doi:10.1016/j.cell.2015.07.015.

97. Keiflin R, Janak P. (2015). Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. Neuron;88(2):247–263. doi:https://doi.org/10.1016/j.neuron.2015.08.037.

98. van der Meer MA, Redish AD. (2011). Ventral striatum: a critical look at models of learning and evaluation. Current Opinion in Neurobiology;21(3):387–392. doi:10.1016/j.conb.2011.02.011.

99. Patriarchi T, Cho JR, Merten K, Howe MW, Marley A, Xiong WH, et al.. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. Science;360(6396):eaat4422. doi:10.1126/science.aat4422.

100. Lin WC, Liu C, Kosillo P, Tai LH, Galarce E, Bateup HS, et al.. (2022). Transient food insecurity during the juvenile-adolescent period affects adult weight, cognitive flexibility, and dopamine neurobiology. Current Biology;32(17):3690–3703.e5. doi:10.1016/j.cub.2022.06.089.

101. Vertechi P, Lottem E, Sarra D, Godinho B, Treves I, Quendera T, et al.. (2020). Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas. Neuron;106(1):166–176.e6. doi:https://doi.org/10.1016/j.neuron.2020.01.017.

102. Pearce JM, Hall G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review;87(6):532–552. doi:10.1037/0033-295X.87.6.532.

103. Ito M, Doya K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. Journal of Neuroscience;29(31):9861–9874. doi:10.1523/JNEUROSCI.6157-08.2009.

104. Eckstein MK, Collins AGE. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. Proc Natl Acad Sci USA;117(47):29381–29389. doi:10.1073/pnas.1912330117.

105. Yechiam E, Busemeyer JR, Stout JC, Bechara A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. Psychological science;16(12):973–978. doi:10.1111/j.1467-9280.2005.01646.x.

106. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of Sciences;104(41):16311–16316. doi:10.1073/pnas.0706111104.

107. Niv Y, Edlund JA, Dayan P, O'Doherty JP. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. Journal of Neuroscience;32(2):551–562. doi:10.1523/JNEUROSCI.5498-10.2012.

108. Gershman SJ. (2015). Do learning rates adapt to the distribution of rewards? Psychonomic Bulletin & Review;22:1320–1327. doi:10.3758/s13423-014-0790-3.

109. Sugawara M, Katahira K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. Scientific Reports;11(1):3574.

110. Boorman ED, Behrens TE, Rushworth MF. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biology;9(6):e1001093. doi:10.1371/journal.pbio.1001093.

111. Palminteri S, Kilford EJ, Coricelli G, Blakemore SJ. (2016). The computational development of reinforcement learning during adolescence. PLoS Computational Biology;12(6):e1004953. doi:10.1371/journal.pcbi.1004953.

112. Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. PLoS computational biology;13(8):e1005684.

113. Hampton AN, Bossaerts P, O'doherty JP. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. Journal of Neuroscience;26(32):8360–8367. doi:10.1523/JNEUROSCI.1010-06.2006.

114. Bayer HM, Glimcher PW. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron;47(1):129–141. doi:10.1016/j.neuron.2005.05.020.

115. Starkweather CK, Uchida N. (2021). Dopamine signals as temporal difference errors: recent advances. Current Opinion in Neurobiology;67:95–105. doi:10.1016/j.conb.2020.08.014.

116. Martianova E, Aronson S, Proulx CD. (2019). Multi-Fiber Photometry to Record Neural Activity in Freely-Moving Animals. JoVE;1(152):e60278. doi:10.3791/60278.

117. Daw N, Gershman S, Seymour B, Dayan P, Dolan RJ. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron;69(6):1204–1215. doi:10.1016/j.neuron.2011.02.027.

118. Collins AGE, Ciullo B, Frank MJ, Badre D. (2017). Working Memory Load Strengthens Reward Prediction Errors. J Neurosci;37(16):4332–4342. doi:10.1523/JNEUROSCI.2700-16.2017.

119. Collins AGE. (2018). Learning Structures Through Reinforcement. In: Goal-Directed Decision Making. Elsevier. p. 105–123. Available from: `https://linkinghub.elsevier.com/retrieve/pii/B978012812098900005X`.

120. Yoo AH, Keglovits H, Collins AGE. (2023). Lowered inter-stimulus discriminability hurts incremental contributions to learning. Cogn Affect Behav Neurosci;doi:10.3758/s13415-023-01104-5.

121. Donoso M, Collins AGE, Koechlin E. (2014). Foundations of human reasoning in the prefrontal cortex. Science;344(6191):1481–1486. doi:10.1126/science.1252254.

122. Heald JB, Lengyel M, Wolpert DM. (2021). Contextual inference underlies the learning of sensorimotor repertoires. Nature;600(7889):489–493. doi:10.1038/s41586-021-04129-3.

123. Collins AGE, Frank MJ. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological Review;120(1):190–229. doi:10.1037/a0030852.

124. Collins AGE, Frank MJ. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci USA;115(10):2502–2507. doi:10.1073/pnas.1720963115.

125. Howe MW, Tierney PL, Sandberg SG, Phillips PEM, Graybiel AM. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature;500(7464):575–579. doi:10.1038/nature12475.

126. Farrell K, Lak A, Saleem AB. (2022). Midbrain dopamine neurons signal phasic and ramping reward prediction error during goal-directed navigation. Cell Reports;41(2):111470. doi:10.1016/j.celrep.2022.111470.

127. Nicola SM. (2010). The Flexible Approach Hypothesis: Unification of Effort and Cue-Responding Hypotheses for the Role of Nucleus Accumbens Dopamine in the Activation of Reward-Seeking Behavior. Journal of Neuroscience;30(49):16585–16600. doi:10.1523/JNEUROSCI.3958-10.2010.

128. Morrison SE, McGinty VB, du Hoffmann J, Nicola SM. (2017). Limbic-motor integration by neural excitations and inhibitions in the nucleus accumbens. Journal of Neurophysiology;118(5):2549–2567. doi:10.1152/jn.00465.2017.

129. Hamilos AE, Spedicato G, Hong Y, Sun F, Li Y, Assad JA. (2021). Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements. Elife;10. doi:10.7554/eLife.62583.

130. Azcorra M, Gaertner Z, Davidson C, He Q, Kim H, Nagappan S, et al.. (2023). Unique functional responses differentially map onto genetic subtypes of dopamine neurons. Nat Neurosci;26(10):1762–1774. doi:10.1038/s41593-023-01401-9.

131. Kutlu MG, Zachry JE, Melugin PR, Cajigas SA, Chevee MF, Kelly SJ, et al.. (2021). Dopamine release in the nucleus accumbens core signals perceived saliency. Current Biology;31(21):4748–4761.e8. doi:10.1016/j.cub.2021.08.052.

132. Kutlu MG, Zachry JE, Melugin PR, Tat J, Cajigas S, Isiktas AU, et al.. (2022). Dopamine signaling in the nucleus accumbens core mediates latent inhibition. Nat Neurosci;25(8):1071–1081. doi:10.1038/s41593-022-01126-1.

133. Collins AL, Saunders BT. (2020). Heterogeneity in striatal dopamine circuits: Form and function in dynamic reward seeking. Journal of Neuroscience Research;98(6):1046–1069. doi:https://doi.org/10.1002/jnr.24587.

134. Taira M, Millard SJ, Verghese A, DiFazio LE, Hoang IB, Jia R, et al.. (2024). Dopamine Release in the Nucleus Accumbens Core Encodes the General Excitatory Components of Learning. Journal of Neuroscience;44(35). doi:10.1523/JNEUROSCI.0120-24.2024.

135. Wang Y, Lak A, Manohar SG, Bogacz R. (2024). Dopamine encoding of novelty facilitates efficient uncertainty-driven exploration. PLOS Computational Biology;20(4):1–27. doi:10.1371/journal.pcbi.1011516.

136. Coddington LT, Lindo SE, Dudman JT. (2023). Mesolimbic dopamine adapts the rate of learning from action. Nature;614(7947):294–302. doi:10.1038/s41586-022-05614-z.

137. Jeong H, Taylor A, Floeder JR, Lohmann M, Mihalas S, Wu B, et al.. (2022). Mesolimbic dopamine release conveys causal associations. Science;378(6626):eabq6740. doi:10.1126/science.abq6740.

138. Akam T, Costa R, Dayan P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLOS Computational Biology;11(12):1–25. doi:10.1371/journal.pcbi.1004648.

139. Miller KJ, Botvinick MM, Brody CD. (2017). Dorsal hippocampus contributes to model-based planning. Nat Neurosci;20(9):1269–1276. doi:10.1038/nn.4613.

140. Korponay C, Stein EA, Ross TJ. (2022). Laterality hotspots in the striatum. Cerebral Cortex;32(14):2943–2956. doi:10.1093/cercor/bhab392.

141. Hart G, Leung BK, Balleine BW. (2014). Dorsal and ventral streams: The distinct role of striatal subregions in the acquisition and performance of goal-directed actions. Neurobiology of Learning and Memory;108:104–118. doi:10.1016/j.nlm.2013.11.003.

142. Lee J, Sabatini BL. (2021). Striatal indirect pathway mediates exploration via collicular competition. Nature;599(7886):645–649. doi:10.1038/s41586-021-04055-4.

143. Essig J, Hunt JB, Felsen G. (2021). Inhibitory neurons in the superior colliculus mediate selection of spatially-directed movements. Communications Biology;4(1):719. doi:10.1038/s42003-021-02248-1.

144. Essig J, Felsen G. (2021). Functional coupling between target selection and acquisition in the superior colliculus. Journal of Neurophysiology;126(5):1524–1535. doi:10.1152/jn.00263.2021.

145. Giossi C, Rubin JE, Gittis A, Verstynen T, Vich C. (2024). Rethinking the external globus pallidus and information flow in cortico-basal ganglia-thalamic circuits. European Journal of Neuroscience;60(9):6129–6144. doi:10.1111/ejn.16348.

146. Grillner S, Robertson B, Stephenson-Jones M. (2013). The evolutionary origin of the vertebrate basal ganglia and its role in action selection. J Physiol;591(22):5425–5431. doi:10.1113/jphysiol.2012.246660.

147. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, et al.. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol;7(4):e1002028.

148. Sweis BM, Thomas MJ, Redish AD. (2018). Mice learn to avoid regret. PLoS Biol;16(6):e2005853. doi:10.1371/journal.pbio.2005853.

149. Pereira TD, Tabris N, Matsliah A, Turner DM, Li J, Ravindranath S, et al.. (2022). SLEAP: A deep learning system for multi-animal pose tracking. Nat Methods;19(4):486–495. doi:10.1038/s41592-022-01426-1.

150. Thorn CA, Atallah H, Howe M, Graybiel AM. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. Neuron;66(5):781–795. doi:10.1016/j.neuron.2010.04.036.

151. Tye KM, Mirzabekov JJ, Warden MR, Ferenczi EA, Tsai HC, Finkelstein J, et al.. (2013). Dopamine neurons modulate neural encoding and expression of depression-related behaviour. Nature;493(7433):537–541.

152. Burke DA, Rotstein HG, Alvarez VA. (2017). Striatal local circuitry: a new framework for lateral inhibition. Neuron;96(2):267–284. doi:10.1016/j.neuron.2017.09.019.

153. Giossi C, Bahuguna J, Rubin JE, Verstynen T, Vich C. (2024). Arkypallidal neurons in the external globus pallidus can mediate inhibitory control by altering competition in the striatum. Proceedings of the National Academy of Sciences;121(47):e2408505121. doi:10.1073/pnas.2420815122.

154. Melzer S, Gil M, Koser DE, Michael M, Huang KW, Monyer H. (2017). Distinct corticostriatal GABAergic neurons modulate striatal output neurons and motor activity. Cell Rep;19(5):1045–1055. doi:10.1016/j.celrep.2017.04.024.

155. Gold JI, Shadlen MN. (2007). The Neural Basis of Decision Making. Annu Rev Neurosci;30(1):535–574. doi:10.1146/annurev.neuro.29.051605.113038.

156. Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, et al.. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. Neuron;105(4):700–711.e6. doi:10.1016/j.neuron.2019.11.018.

157. Stauffer WR, Lak A, Kobayashi S, Schultz W. (2016). Components and characteristics of the dopamine reward utility signal. J Comp Neurol;524(8):1699–1711. doi:10.1002/cne.23880.

158. Walton ME, Bannerman DM, Rushworth MFS. (2002). The role of rat medial frontal cortex in effort-based decision making. J Neurosci;22(24):10996–11003. doi:10.1523/JNEUROSCI.22-24-10996.2002.

159. Lottem E, Banerjee D, Vertechi P, Sarra D, Lohuis MO, Mainen ZF. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. Nat Commun;9(1):1000. doi:10.1038/s41467-018-03438-y.

160. Apps MAJ, Rushworth MFS, Chang SWC. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. Neuron;90(4):692–707. doi:10.1016/j.neuron.2016.04.018.

161. Zhang W, Yartsev MM. (2019). Correlated neural activity across the brains of socially interacting bats. Cell;178(2):413–428.e22. doi:10.1016/j.cell.2019.05.023.

162. Dombeck DA, Reiser MB. (2012). Real neuroscience in virtual worlds. Curr Opin Neurobiol;22(1):3–10. doi:10.1016/j.conb.2011.10.015.

163. Balleine BW, O'Doherty JP. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology;35(1):48–69. doi:10.1038/npp.2009.131.

164. Dayan P, Berridge KC. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. Cogn Affect Behav Neurosci;14(2):473–492. doi:10.3758/s13415-014-0277-8.

165. Musall S, Kaufman MT, Juavinett AL, Gluf S, Churchland AK. (2019). Single-trial neural dynamics are dominated by richly varied movements. Nat Neurosci;22(10):1677–1686. doi:10.1038/s41593-019-0502-4.

166. Urai AE, Doiron B, Leifer AM, Churchland AK. (2022). Large-scale neural recordings call for new insights to link brain and behavior. Nat Neurosci;25(1):11–19. doi:10.1038/s41593-021-00980-9.

167. Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. Neuron;93(3):480–490. doi:10.1016/j.neuron.2016.12.041.

168. Granger CWJ. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica;37(3):424. doi:10.2307/1912791.

169. Shojaie A, Fox EB. (2021). Granger causality: A review and recent advances. arXiv;1. doi:10.48550/arXiv.2105.02675.

170. McCullagh P, Nelder JA. (1989). Generalized Linear Models. 2nd ed. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Philadelphia, PA: Chapman & Hall/CRC. Available from: `https://doi.org/10.1201/9780203753736`.

171. Nelder JA, Wedderburn RWM. (1972). Generalized Linear Models. J R Stat Soc Ser A;135(3):370. doi:10.2307/2344614.

172. Lau B, Glimcher PW. (2008). Value representations in the primate striatum during matching behavior. Neuron;58(3):451–463. doi:10.1016/j.neuron.2008.02.021.

173. Lau B, Glimcher PW. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav;84(3):555–579. doi:10.1901/jeab.2005.110-04.

174. Hastie T, Tibshirani R, Friedman J. (2003). The elements of statistical learning. 1st ed. Springer series in statistics. New York, NY: Springer. Available from: `https://link.springer.com/book/10.1007/978-0-387-84858-7`.

175. Zou H, Hastie T. (2005). Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol;67(2):301–320. doi:10.1111/j.1467-9868.2005.00503.x.

176. Yang J, Hastie T. (2024). A fast and scalable pathwise-solver for group lasso and elastic net penalized regression via block-coordinate descent. arXiv;1. doi:10.48550/arXiv.2405.08631.

177. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. (2007). Learning the value of information in an uncertain world. Nat Neurosci;10(9):1214–1221. doi:10.1038/nn1954.

178. Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Available from: `https://dl.acm.org/doi/10.5555/1643031.1643047`.

179. Raschka S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv;1. doi:10.48550/arXiv.1811.12808.

180. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. KDD '19. New York, NY, USA: Association for Computing Machinery. Available from: `https://doi.org/10.1145/3292500.3330701`.

181. Glaser JI, Benjamin AS, Chowdhury RH, Perich MG, Miller LE, Kording KP. (2020). Machine learning for neural decoding. eNeuro;7(4):ENEURO.0506–19.2020. doi:10.1523/ENEURO.0506-19.2020.

182. Pagan M, Simoncelli EP, Rust NC. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with V1-like computation. Neural Comput;28(11):2291–2319.

183. Li Q, Ko H, Qian ZM, Yan LYC, Chan DCW, Arbuthnott G, et al.. (2017). Refinement of learned skilled movement representation in motor cortex deep output layer. Nat Commun;8(1):15834. doi:10.1038/ncomms15834.

184. Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, et al.. (2014). Neural constraints on learning. Nature;512(7515):423–426. doi:10.1038/nature13665.

185. Cortes C, Vapnik V. (1995). Support-vector networks. Mach Learn;20(3):273–297. doi:10.1007/BF00994018.

186. Kamitani Y, Tong F. (2005). Decoding the visual and subjective contents of the human brain. Nat Neurosci;8(5):679–685. doi:10.1038/nn1444.

187. Haynes JD, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. (2007). Reading hidden intentions in the human brain. Curr Biol;17(4):323–328. doi:10.1016/j.cub.2006.11.072.

188. Decoste D, Schölkopf B. (2002). Training invariant support vector machines. Mach Learn;46(1-3):161–190. doi:10.1023/A:1012454411458.

189. Chang CC, Lin CJ. (2011). LIBSVM. ACM Trans Intell Syst Technol;2(3):1–27. doi:10.1145/1961189.1961199.

190. Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984). Classification And Regression Trees. Routledge.

191. Quinlan JR. (1986). Induction of decision trees. Mach Learn;1(1):81–106. doi:10.1007/BF00116251.

192. Breiman L. (2001). Random forests. Mach Learn;45(1):5–32. doi:10.1023/A:1010933404324.

193. Markowitz JE, Gillis WF, Beron CC, Neufeld SQ, Robertson K, Bhagat ND, et al.. (2018). The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. Cell;174(1):44–58.e17. doi:10.1016/j.cell.2018.04.019.

194. Mentch L, Hooker G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. Journal of Machine Learning Research;17(26):1–41. doi:10.48550/arXiv.1404.6473.

195. Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, et al.. (2015). Mapping sub-second structure in mouse behavior. Neuron;88(6):1121–1135. doi:10.1016/j.neuron.2015.11.031.

196. Berman GJ, Choi DM, Bialek W, Shaevitz JW. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. J R Soc Interface;11(99):20140672. doi:10.1098/rsif.2014.0672.

197. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al.. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci;21(9):1281–1289. doi:10.1038/s41593-018-0209-y.

198. Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, et al.. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. Elife;8. doi:10.7554/eLife.47994.

199. Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS. (2008). Dimensionality and dynamics in the behavior of C. elegans. PLoS Comput Biol;4(4):e1000028. doi:10.1371/journal.pcbi.1000028.

200. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al.. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol;37(1):38–44. doi:10.1038/nbt.4314.

201. Weinreb C, Pearl JE, Lin S, Osman MAM, Zhang L, Annapragada S, et al.. (2024). Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. Nat Methods;21(7):1329–1339. doi:10.1038/s41592-024-02318-2.

202. Bukwich M, Campbell MG, Zoltowski D, Kingsbury L, Tomov MS, Stern J, et al.. (2023). Competitive integration of time and reward explains value-sensitive foraging decisions and frontal cortex ramping dynamics.

203. Batty E, Whiteway M, Saxena S, Biderman D, Abe T, Musall S, et al.. (2019). BehaveNet: nonlinear embedding and Bayesian neural decoding of behavioral videos. vol. 32. Available from: `https://papers.nips.cc/paper_files/paper/2019/hash/a10463df69e52e78372b724471434ec9-Abstract.html`.

204. Shenoy KV, Sahani M, Churchland MM. (2013). Cortical control of arm movements: a dynamical systems perspective. Annu Rev Neurosci;36(1):337–359. doi:annurev-neuro-062111-150509.

205. Saxena S, Cunningham JP. (2019). Towards the neural population doctrine. Curr Opin Neurobiol;55:103–111. doi:10.1016/j.conb.2019.02.002.

206. Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. vol. 102. American Physiological Society. Available from: `https://papers.nips.cc/paper_files/paper/2008/hash/ad972f10e0800b49d76fed33a21f6698-Abstract.html`.

207. Zhao Y, Park IM. (2017). Variational latent Gaussian process for recovering single-trial dynamics from population spike trains. Neural Comput;29(5):1293–1316.

208. Linderman S, Johnson M, Miller A, Adams R, Blei D, Paninski L. (2017). Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. vol. 54. PMLR. Available from: `https://proceedings.mlr.press/v54/linderman17a.html`.

209. Niv Y. (2019). Learning task-state representations. Nat Neurosci;22(10):1544–1553. doi:10.1038/s41593-019-0470-8.

210. Doya K, Samejima K, Katagiri Ki, Kawato M. (2002). Multiple model-based reinforcement learning. Neural computation;14(6):1347–1369. doi:10.1162/089976602753712972.

211. Wilson RC, Collins AG. (2019). Ten simple rules for the computational modeling of behavioral data. eLife;8:e49547. doi:10.7554/eLife.49547.

212. Watanabe S. (2012). A widely applicable Bayesian information criterion. arXiv;1. doi:10.48550/arXiv.1208.6338.

213. Bozdogan H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika;52(3):345–370. doi:10.1007/BF02294361.

214. Paninski L, Pillow J, Lewi J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. In: Progress in Brain Research. Progress in brain research. Elsevier. p. 493–507.

215. Park IM, Meister MLR, Huk AC, Pillow JW. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. Nat Neurosci;17(10):1395–1403. doi:10.1038/nn.3800.

216. Ramsay JO, Silverman BW. (2005). Functional Data Analysis. 2nd ed. Springer series in statistics. New York, NY: Springer.

217. Ramsay J, Hooker G, Graves S. (2009). Functional Data Analysis with R and MATLAB. 2009th ed. Use R!. New York, NY: Springer.

218. McLean MW, Hooker G, Staicu AM, Scheipl F, Ruppert D. (2014). Functional generalized additive models. J Comput Graph Stat;23(1):249–269. doi:10.1080/10618600.2012.729985.

219. Meyer MJ. (2024). Advances in estimation and inference for historical functional linear models. Wiley Interdiscip Rev Comput Stat;16(1).

220. Loewinger G, Cui E, Lovinger D, Pereira F. (2024). A statistical framework for analysis of trial-level temporal dynamics in Fiber photometry experiments. eLife;doi:10.7554/eLife.95802.3.

221. Asencio M, Hooker G, Gao HO. (2014). Functional convolution models. Statistical Modelling;14(4):315–335. doi:10.1177/1471082X13508262.

222. Gertheiss J, Rügamer D, Liew BXW, Greven S. (2024). Functional data analysis: An introduction and recent developments. Biom J;66(7):e202300363. doi:10.1002/bimj.202300363.