From The Department of Neuroscience
Karolinska Institutet, Stockholm, Sweden

A Critic In Action?

A Functional Examination of the Striato-Pallido-Habenular Circuit

Moritz Weglage

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-637-9

A Critic In Action?

A Functional Examination of the Striato-Pallido-Habenular Circuit

Thesis presented for the
Doctoral Degree (PhD) by

Moritz Weglage

Principal Supervisor:
Professor Konstantinos Meletis
Dept. of Neuroscience
Karolinska Institutet

Co-supervisors:
Associate Professor Arvind Kumar
Dept. of Computer Science
KTH Royal Institute of Technology

Professor Per Svenningsson
Dept. of Clinical Neuroscience
Karolinska Institutet

Opponent:
Professor Bence Ölveczky
Dept. of Organismic and Evolutionary Biology
Harvard University

Examination Board:
Professor Åsa Mackenzie
Dept. of Organismal Biology
KTH Royal Institute of Technology

Associate Professor Per Petersson
Dept. of Integrative Biology
Umeå Universitet

Directeur de Recherche David Robbe
Institut de Neurobiologie de la Méditerranée

The thesis will be defended in public at the
Eva & Georg Klein lecture hall, Biomedicum, Karolinska Institutet, Stockholm
on June 10, 2022 at 13:00

Abstract

The basal ganglia (BG) and the midbrain dopamine (DA) system are considered key loci of reinforcement learning (RL), or learning by "trial and error", in the brain. The BG are implicated in action selection, and thus the "trial" part of the learning process, and the dopamine (DA) system is known to encode "error" signals. This DA error signal—the reward prediction error—is thought to adjust the BG's propensity to select a "tried" action again in the future. In RL terms, the action-selecting BG is called the "actor", and the action-critiquing DA system the "critic". Here, a candidate striato-pallido-habenular "critic pathway" upstream of the DA system is examined.

The proposed critic pathway originates in the striosome compartment of the striatum, and projects via a non-canonical internal globus pallidus (GPi) population to the lateral habenula (LHb). LHb activity has been shown to encode the inverse of the DA reward prediction error signal, and to cause inhibition within the DA system. This posits the described striato-pallido-habenular pathway as key part of the critic circuit.

To investigate the role of the striato-pallido-habenular pathway in action, we recorded and manipulated the neuronal activity of neurons of the striatal striosome (Article I) and the GPi (Article II & III) in mice performing tasks that engendered trial and error behavioral strategies. We found that the activity of striatal striosome neurons had much in common with that of neurons within the striatal "actor pathways". Moreover, all striatal neurons jointly represented the evolving behavioral context in a spatiotemporally continuous population code, undermining notions of discrete and well-defined action selection and evaluation processes. The results of our experiments on the GPi challenged its proposed role in driving the LHb's inverse reward prediction error signals, and implicated the GPi-adjacent lateral hypothalamus (LHA) in that role instead. In sum, the studies included here call into question whether the striato-pallido-habenular pathway serves as a critic in BG-mediated action.

List of Scientific Articles

I. Weglage, M.*, Wärnberg, E.*, Lazaridis, I.*, Calvigioni, D., Tzortzi, O. & Meletis, K.
Complete representation of action space and value in all dorsal striatal pathways
Cell Reports, 2021

II. Weglage, M.*, Ährlund-Richter, S.*, Fuzik, J., Skara, V., Lazaridis, I. & Meletis, K.
Sst+ GPi output neurons provide direct feedback to key nodes of the basal ganglia and drive behavioral flexibility
bioRxiv (preprint), 2022

III. Lazaridis, I., Tzortzi, O., Weglage, M., Märtin, A., Xuan, Y., Parent, M., Johansson, Y., Fuzik, J., Fürth, D., Fenno, L. E., Ramakrishnan, C., Silberberg, G., Deisseroth, K., Carlén, M. & Meletis, K.
A hypothalamus-habenula circuit controls aversion
Molecular Psychiatry, 2019

Contents

Chapter 1

Introduction

1.1 Actor-Critic Reinforcement Learning

It's a sunny Monday morning and a new ice cream vendor has opened up shop on your way to work. Spontaneously, you decide to try a scoop of the chocolate flavor. The ice cream tastes great—you'll be back tomorrow. Over the next few days, it becomes your policy to stop by the shop every day. The sun, the shop's colorful sign… you are barely aware of your actions before you find yourself back out on the street again with a cone of chocolate ice cream in hand. One popular theoretical account of how you might have acquired this (hopefully optimal) behavioral policy of buying chocolate ice cream on sunny days, and of how you are compelled to follow it almost automatically, is offered by the reinforcement learning (RL) framework.

RL emerged at the intersection of psychology and computer science, when the abstract "associative learning rules" of the behaviorists[4,5], originally devised to explain the behavioral adaptations of animals in conditioning experiments, inspired the design of artificial intelligence (AI) systems and robots capable of learning by trial-and-error[6,7]. In the simplest terms, RL is learning from experience, without explicit instruction: whenever an action leads to an unexpected improvement, that action is "reinforced", compelling the behaving agent to repeat it the next time it encounters the same situation. Conversely, if things turned out worse than expected, the agent will be less inclined to take the same course of action again in future.

A canonical AI implementation of RL is the actor-critic architecture[7,8]. An "actor" and a "critic" module both receive information about the state of the environment, which they process to different ends. The actor translates

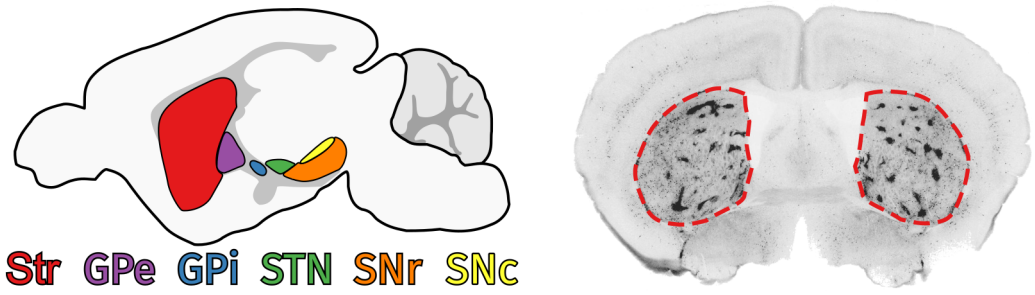

Figure 1.1 | Basal ganglia anatomy in the mouse.
Left: The basal ganglia consist of the striatum (Str), the external and internal segments of the globus pallidus (GPe & GPi), the subthalamic nucleus (STN) and the substantia nigra pars reticulata (SNr). The major nuclei of the dopaminergic midbrain are the adjacent ventral tegmental area (VTA, not shown) and the substantia nigra pars compacta (SNc).
Right: Coronal brain section depicting the striosome and matrix compartments of the striatum (red dashed line). μ-opioid receptor expression in the striosome visualized via the cre-dependent expression of the fluorescent protein tdTomato (Oprm1-cre x Ai14 mouse crossing).

state into action, determining how to act from memorized policies. The critic translates state into memorized value (or "goodness") expectations, which are used to critique the consequences of the actor's choice. Whenever an action led to a state the value of which surpassed or fell short of expectation, the critic emits a signed "prediction error" signal, which reflects the difference between the expected and the observed value return. The prediction error is used to adjust and optimize both the stored state-policy and state-value mappings for future use, thus driving learning in both the actor and the critic.

With its roots in the behaviorist study of animal learning, RL is of great interest to neuroscientists. After all, if similarities between the architectures of successful AI systems and the brain can be identified, knowledge of the computational mechanics of the former may well shed light on the workings of the latter[9–11]. In this spirit, the deeply interconnected subcortical circuitry of the basal ganglia (BG) and the dopaminergic midbrain nuclei (**Fig. 1.1 left**) have long been viewed and studied as a potential biological implementation of the actor-critic RL model[9,10,12–15]. How the BG nuclei may map onto the actor-critic architecture and drive action selection (e.g. your ice cream habit) is the subject of the next section.

1.2 The Actor-Critic Model of the Basal Ganglia

In the 1990s, Wolfram Schultz and colleagues observed that the firing of dopamine (DA) neurons in the midbrain resembled the RL prediction error[16–18], suggesting that DA neurons served as the output-end of a brain-based critic circuit[6]. The BG nuclei are profusely and in part reciprocally connected with the DA system[19–21]. The BG were consequently widely identified as the foremost candidates to instantiate both the remainder of the critic circuit, as well as the actor circuit, in a biological actor-critic system updated by the newly-discovered DA prediction error signal[12,13,22].

Before discussing how the BG nuclei may implement the RL architecture in more detail, I will describe why the BG as a whole are exquisitely-situated anatomically to recognize states and to influence action, and thus to serve as an RL agent of the brain.

1.2.1 The Basal Ganglia's Input and Output

A useful RL agent must be able to *sense* and *act* on the environment; after all, a capacity to learn how to act by trial-and-error (RL) is wasted if there is nothing to learn to respond to or to act upon. That naturally raises the question: are the BG's input and output stages—the striatum and the GPi/SNr, respectively—wired to "sense and to act" as required?

The BG's primary input structure, the striatum, receives massive excitatory (i.e. glutamatergic) input from virtually all regions of the neocortex, as well as a plethora of projections from the thalamus and the brain stem[21,23,24]. These disparate inputs supply the striatum with ample and diverse state-relevant information, ranging from the sensory, over motor activity, to the emotional, giving the BG plenty to sense and learn about. Importantly, striatal inputs are—to a degree—topographically and functionally organized, and this topography is—to a degree—preserved along the BG pathways (**Fig. 1.2 A-B**)[25–27]. At the same time, BG processing must necessarily be highly integrative, as the BG has to "funnel" its enormous input into an output (i.e. the GPi and the SNr) consisting of magnitudes fewer neurons[13,28,29].

Without a significant convergence of inputs of diverse informational content on individual BG neurons, the BG could not possibly learn to recognize complex states, as complex states are bound to be characterized by rich combinations of sense impressions and cognitive variables[12,30]. Suitably,

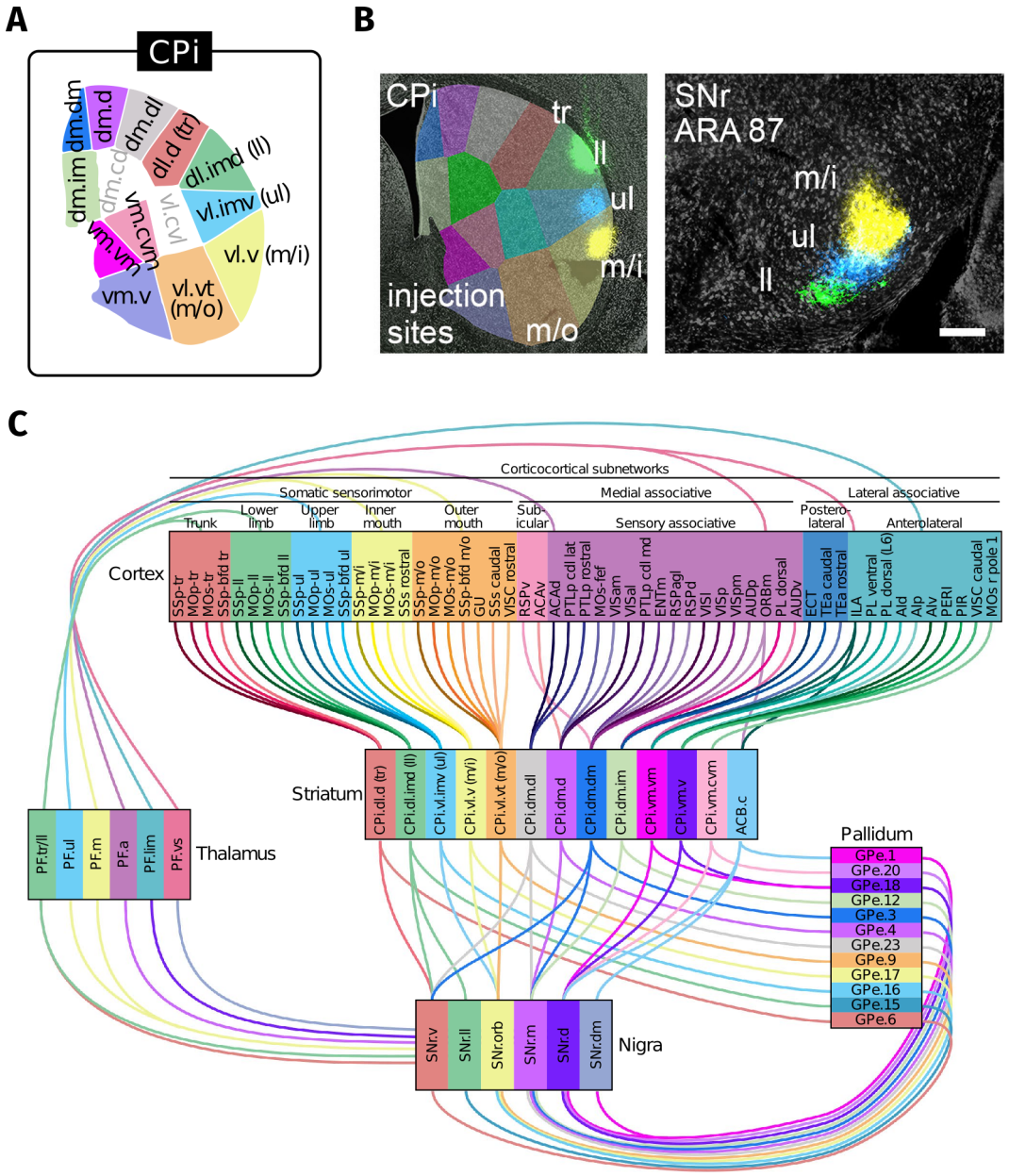

Figure 1.2 | Cortico-BG-thalamic loops.
A The striatum's dorsal part, or caudoputamen (CP), subdivided into functional domains based on the cortical sources of their input. Distinct domains in the lateral part of the dorsal striatum process information relating to the trunk (tr), lower limb (ll), upper limb (ul), and inner mouth (m/i). CPi: caudoputamen, intermediate level.
B Left: The lower limb, upper limb, and inner mouth domains of the lateral striatum were injected with an anterograde tracer. Right: The axon terminals of the striatal projection neurons located in the three injected domains are visualized in the SNr. The resulting image shows that the somatotopic organization of the striatal domains is maintained at the level of the SNr.
C Foster et al.'s[27] model of cortico-BG-thalamic loops shows the partially parallel, partially convergent topographic organization of the network. In many instances, a loop's cortical origin and final target is roughly the same region; such loops are referred to as "closed loops" (e.g. the mouth loops).

Adapted from Foster et al.[27], used under CC BY 4.0.

each cortical projection to striatum provides not only dense, focal, and topographic, but also sparse and diffuse innervation; i.e. although the dense part of a corticostriatal projection may target a clearly-delineated striatal zone, the projection's diffuse part can extend well beyond those borders. The dense and diffuse terminal fields of diverse cortical projections overlap and converge substantially, and supply the means for the BG to form the complex combinatorial associations needed to recognize state[24,31].

The BG's output structures, the GPi and the SNr, project—via the thalamus –to the cortex, as well as to a multitude of brain stem nuclei[27,32]. By targeting many of their sources of input, and by preserving input topography to some extent, the BG are known to form subcircuits of *partially-closed*, parallel loops, which subserve distinct functions related to their extra-BG origins and targets (**Fig. 1.2 C**)[26,27,32,33].

BG motor loops, involving motor effectors in the motor cortex and the brainstem, are historically best researched, owing to the fact that BG dysfunction has long been linked to movement disorders such as Parkinson's disease, and that movement is highly tractable scientifically[34–37]. Such work has established beyond a doubt that the BG are heavily involved in voluntary motor control and thus of obvious import to our ability to act on the environment. However, the existence of loops incorporating cortical regions implicated in higher cognitive and emotional functions, such as orbitofrontal or anterior cingulate cortex, strongly suggests that the BG's influence over our actions is not limited to the purely motoric[26,29,34].

It is important to note that the canonical BG output pathways are inhibitory (i.e. GABAergic) and tonically (i.e. continuously) engaged[37]. Therefore, the GPi and the SNr operate by reducing or enhancing the constant suppression they exert over their downstream targets. Presumably then, the function of the BG is *not* to intrinsically generate the neuronal activity driving our actions (motor or otherwise), but rather to constrain or outright gate such activity elsewhere in the brain[37–39]. Metaphorically-speaking, the BG must "act" by curating or directing rather than by creating. In modern BG models, including actor-critic models, actions are consequently usually initiated and executed by the (motor) effectors outside the BG—for example by motor cortex—and not by the BG itself[12,29,36,40,41].

To summarize: The BG are placed to integrate diverse, multimodal contextual information from all over the cortex and thus to accurately "sense" the

current state. They are moreover positioned to modulate or gate motor effectors across cortex and the brainstem through their partially-closed motor loops, and to do likewise for non-motor effectors through other loops. The BG may therefore exert powerful control over a variety of brain circuits shaping our actions. Judging only from their inputs and outputs, the BG appear to be an ideal "black box" to house a RL system. It is time to have a look inside that black box, at how an actor-critic architecture may be implemented within.

1.2.2 The Basal Ganglia's Matrix-Actor and Striosome-Critic

The first actor-critic account of the BG was published in 1994 by Houk, Adams and Barto[12], in the wake of the discovery of the DA prediction error[16]. It inspired a number of similar, if mechanistically more detailed, models in the late 1990s[22,40,42–44] (extensively reviewed by Joel, Niv and Ruppin in 2002[13]), and beyond[45–47]. The basic anatomical and functional prescriptions of these models are outlined next.

The early actor-critic models of the 1990s widely featured the striosome as the part of the critic circuit which contributes the state-value estimates to the prediction error computation[12,13]. The striosome is a histochemically distinctive compartment of the BG's principal input structure, the striatum. The defining feature of the striosome compartment is its dense expression of the μ-opioid receptor[48–51]. In cross-sections of the brain in which the μ-opioid receptor is fluorescently-labeled, the striosome network appears as "patches" of fluorescent tissue embedded in the many times larger, complementary "matrix" compartment (**Fig. 1.1 right**). In classic actor-critic models, the matrix implements the actor[12,13].

The striosome and the matrix were mapped onto the actor and the critic primarily due to their differential connectivity with the DA system[12,13,52]. A great number of studies of the compartments—old and new—suggest that the dorsal striatal striosome sends direct projections to the DA system, targeting in particular the DA neurons of the substantia nigra pars compacta (SNc), whereas the dorsal striatal matrix does so to a much lesser degree[53–60]. The striosome hence appears to be the origin of an alternative BG output pathway via the SNc; a pathway separate from the major matrix pathways targeting the GPi and the SNr[53]. Significantly, the SNc releases DA diffusely across the entire dorsal part of the striatum and targets both the striosome

and the matrix[20,61,62]. The hypothetical striosome-critic is hence positioned to shape *and* receive DA prediction error signals, whereas the matrix-actor is placed to modulate the primary BG output, and to receive the DA error signals shaped by the striosome—an arrangement that matches the RL actor-critic architecture perfectly.

A secondary reason why the striosome is thought especially suitable for the critic role are reports that the cortical and subcortical input to the striosome compartment favors the limbic: compared to the surrounding matrix tissue, the striosome appears to receive somewhat denser input from the emotion-processing regions of the limbic system, including the prelimbic cortex and the amygdala[53,58,63–65]. A characterization of the matrix and the striosome pathways as preferentially "sensorimotor" and "limbic" maps intuitively well onto the distinction of the "behaving" actor and the "evaluating" critic[52,53,66].

It should be noted that in many recent theoretical discussions and models of the BG (including "OpAL", discussed below) the roles of the actor and the critic are assigned by striatal region, rather than chemical compartment. In these region-based accounts, the critic is located to the ventral, and the actor to the dorsal striatum[13,15,45,67–69]. However, here we will largely focus on the original striosome-matrix actor-critic division, as the striosome-as-critic conception remains highly influential in the literature most relevant to this thesis—as will be seen later.

The anatomical arrangement of the striatal striosome and matrix compartments in relation to the dopaminergic SNc may fit the actor-critic model architecturally (**Fig. 1.3**)—but how would a striosome-critic shape the DA prediction error signal, and how a matrix-actor implement policy-based action selection, in practical, biological terms? These two questions will be explored next.

How the Striosome-Critic Shapes the DA Prediction Error

All of the striatum's output neurons, commonly referred to as the spiny projection neurons (SPNs), are inhibitory, whether they form part of the striosome or the matrix[37]. The inhibitory nature of the striosomal output is most convenient for the brain-based computation of the prediction error (δ) signal, considering that computing the δ signal requires the *subtraction* of the "value expected" ($\hat{V}$) signal from a "reward received" (r) signal, i.e. $\delta = r - \hat{V}$. As it is,

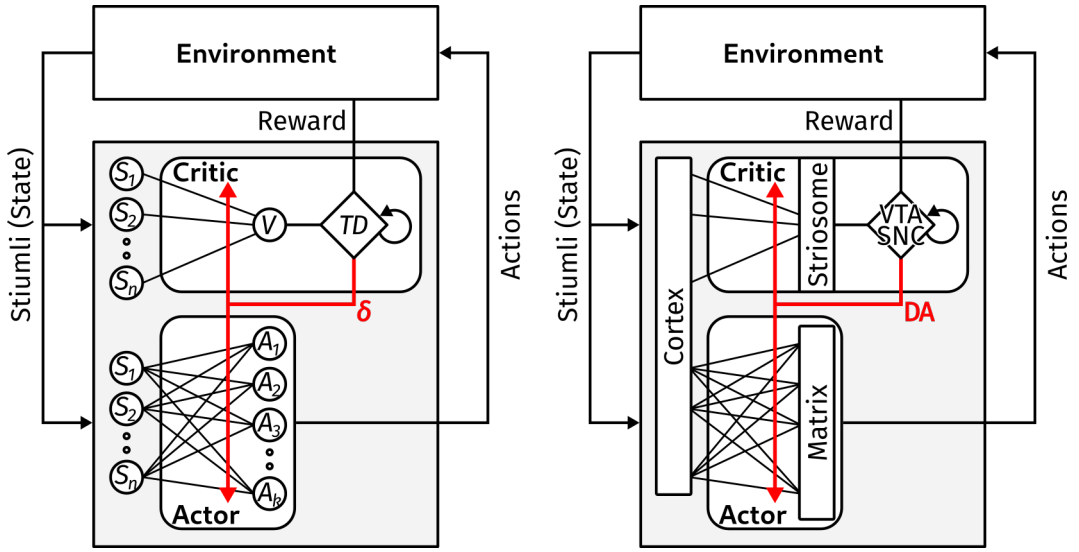

Figure 1.3 | The actor-critic architecture, implemented in the basal ganglia.
Left The actor and the critic modules receive state-related information from the environment (*S*). The actor recalls the "policy" associated with the current state—effectively a table listing the available actions and their selection probabilities (*A*). Based on these probabilities, or "action predispositions", an action is selected. Its execution affects the environment, ideally to positive effect. Whether that is the case or not is evaluated by the critic. The critic recalls the state value estimate associated with the current state (*V*). This current state value estimate, the previous state value estimate, and the current reward are used to compute the temporal difference (*TD*) prediction error signal (δ). The "temporal difference" error computation, in which the state values of adjacent time points (i.e. current and previous) are compared, can generate reinforcement signals in response to state transitions that do not involve primary reinforcers *directly*, but are predictive of them: e.g. if the actor's action triggered the onset of a tone that in the past preceded a reward, the critic can reinforce the action, despite the reward not yet being available. The TD prediction error is used to refine both the actor's policy and the critic's state value estimate.
Right In the classic actor-critic models of the BG[13], the striatal matrix compartment serves as the actor ($\rightarrow$action selection), and the striosome compartment and the SNc serve as the critic ($\rightarrow$evaluation). The striosome supplies the state value estimates, and the SNc integrates them with reward-related inputs (e.g. from the lateral hypothalamus) to compute the prediction error. The prediction error is signaled through the dopaminergic feedback to the striatum, which refines the matrix's state-to-action and the critic's state-to-value mappings via synaptic plasticity at the corticostriatal synapse.

Adapted from Takahashi et al.[68].

the SNc may compute this subtraction simply by integrating an *excitatory* $+\mathrm{r}$ input with the striosome's *inhibitory* $-\hat{\mathrm{V}}$ input, presuming that both r, $\hat{\mathrm{V}}$ and δ are encoded in the magnitude of the neuronal activation[6,12]. The source of the r signal has generally received much less attention than that of $\hat{\mathrm{V}}$. Notably, Houk and colleagues[12] speculated that it may originate from the lateral hypothalamus (LHA). Activating certain LHA inputs to the DA system indeed enhances DA release and is sufficient to reinforce behavior[70].

How the Matrix-Actor Learns Policies and Selects Actions

In RL, a policy for the selection of *discrete* actions is simply a table in which the selection probability of every action available in the current state is listed. On every trial, a stochastic process picks a single action based on this state-appropriate probability table[7]. In (models of) the brain, it is typically assumed that each discrete action available is represented by a discrete ensemble of matrix neurons (i.e. a "grandmother ensemble"[29]), and that the activity of each ensemble relative to that of the others reflects the selection probability of the action associated with that ensemble[6,13,30,39,45,71]. In short, the higher the activity of an action-specific matrix ensemble, the more likely its action is to be selected. That is because co-active action-ensembles are usually thought to resolve a "winner-take-all" competition of some sort, in order for a single action to be selected[29,30,39].

The simplest mechanism proposed to help resolve the winner-take-all competition is "lateral inhibition", which entails that neurons of competing ensembles inhibit one another at the level of the striatum[22,30,71]. Additionally, the selection mechanism may incorporate BG nuclei and pathways downstream of the striatum, rendering it a competition of BG "action channels" rather than one limited to striatal "action ensembles". Interactions between BG-intrinsic pathways are held to play a very significant role in action selection, and thus feature in most BG selection models[30,39,45,72].

The pathway interactions incorporated in BG action selection models are typically based off the "classic" dual pathway model of the BG motor control circuit[30,45,72]. The classic model[34–36] will therefore be briefly sketched below, before I turn to the OpAL model[45], which combines and formalizes many popular ideas about actor policy learning and "action channel"-based, discrete action selection in the BG.

The Classic Basal Ganglia Model of Motor Control The classic BG motor control model[34–36] emerged in the late 1980s to early 1990s. The prevailing theories about the BG at that time were focused on the BG motor loop, and how it regulated the kinematics of ongoing, cortically-initiated movements, although the existence of non-motor loops was recognized.

Mechanistically, the classic model emphasized (1), the existence of two antagonistic BG pathways, and (2), DA's contrary effects on the ongoing activity within these pathways:

(1), the "direct" and "indirect" pathways are wired to inhibit and disinhibit BG output neurons, respectively, by projecting mono- versus polysynaptically from the striatal SPNs to the GPi and the SNr.

(2), DA enhances activity in the direct pathway SPNs (dSPNs), and reduces activity in the indirect pathway SPNs (iSPNs), by binding pathway-specific DA receptors (i.e. D1 versus D2 receptors).

Functionally, the direct and indirect BG pathways were proposed to facilitate and suppress movements, respectively, through their antagonistic regulation of the BG output onto the motor effectors in cortex and elsewhere. Accordingly, DA was thought to exert pro-kinetic effects by enhancing dSPN and suppressing iSPN activity.

This DA-balanced, antagonistic dual pathway architecture remains the linchpin of most contemporary BG models, including the OpAL "dual actor" model of action selection, which is the subject of the following paragraphs.

OpAL: Opponent Actor Policy Learning And Action Selection The "Opponent Actor Learning" (OpAL) model, published by Collins and Frank in 2014[45], is a recent and representative BG action selection model. OpAL is a high-level abstraction of the classic model in that unitary, discrete actions, rather than continuous movements, are facilitated and suppressed by the direct and indirect pathways. Consequently, OpAL emphasizes that the BG *gate* the activity of their targets, for the purpose of action selection (categorical, e.g. Go vs No-Go), whereas the classic model stresses that BG targets are *modulated*, in order to adjust movement kinematics (continuous, e.g. velocity). Finally, OpAL incorporates DA-mediated learning in addition to DA's acute effects on ongoing SPN activity.

Key to the OpAL model is the notion that DA prediction errors induce opposite plasticity changes (i.e. policy learning) at corticostriatal synapses

onto dSPNs and iSPNs. Specifically, a positive DA prediction error, signaling an outcome that has exceeded expectation, is thought to strengthen recently active synapses onto dSPNs and to weaken recently active synapses onto iSPNs; a negative prediction error would cause the inverse adjustments. Consequently, dSPNs learn facilitatory "Go" policies, whereas iSPNs learn suppressive "No-Go" policies.

During action selection, facilitatory dSPNs and suppressive iSPNs within the same action channel compete through their antagonistic projection pathways, and the action channel whose output neurons within the GPi and the SNr are *most net-inhibited* is selected. This has the effect of disinhibiting the extra-BG effectors of the action associated with the winning BG channel. OpAL does not specify how the inhibitory tone onto the effectors of the "losing" action channels is retained or promptly restored; however, it has been suggested elsewhere that relatively diffuse excitatory (i.e. glutamatergic) input from the STN to the GPi and SNr may negate the inhibition of all but the winning channel[29,37]. The STN is activated via collaterals of the cortical input to the striatum (i.e. the "hyperdirect pathway"), as well as disinhibited via the indirect pathway, which suppresses the STN-projecting, inhibitory GPe[73].

Importantly, OpAL's dSPN/iSPN "dual actor" representation is not redundant, for two reasons: (1), OpAL's error-driven learning process is expected to proceed asymmetrically and nonlinearly. Asymmetrically, because the learning rates at dSPN and iSPN synapses are allowed to differ. Nonlinearly, because the magnitude of the prediction error-induced plasticity changes depend on the magnitude of the SPN's recent activity, and therefore on the synaptic weights pre-update; in other words, more heavily-weighted synapses are more significantly modified by an update. (2), in OpAL, DA levels also impact the action selection process directly, by enhancing and suppressing the ongoing activity in dSPNs and iSPNs, respectively, effectively shifting the striatal population activity to favor either dSPN "Go" or iSPN "No-Go" policies. Because of these independent effects of DA on learning and action selection, OpAL does *not* imply that the activity of dSPNs and iSPNs within a given action channel is complementary and redundant.

In sum, a policy-based discrete action selection process mediated by the OpAL dual actor would proceed as follows (**Fig. 1.4**): the cortex "initiates" actions (A) and (B). In the BG, the difference between the facilitatory dSPN and suppressive iSPN activity is strongly positive within the action channel gating

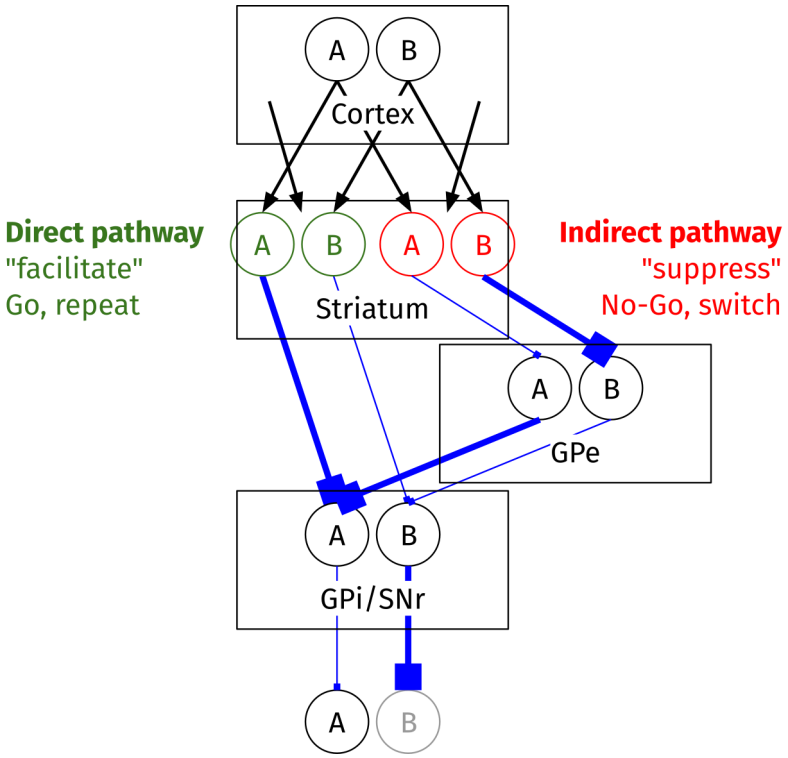

Figure 1.4 | Opponent actor action selection as envisioned by OpAL.
The OpAL model[45] reinterprets the antagonistic direct and indirect pathways of the classical BG motor model as antagonistic Go and No-Go policy networks. Each action channel—e.g. (A), (B)—consists of dSPN-"Go" (green) and iSPN-"No-Go" (red) policy neurons, which facilitate and suppress the selection of their affiliated actions, respectively. DA release balances their relative influence over ongoing behavior: if the Go pathway is favored, the OpAL-regulated agent's strategy is mostly determined by positive experiences (i.e. benefits), if the No-Go pathway is favored, negative experiences (i.e. costs) weigh more heavily on the agent's actions. This mechanism, and mechanisms enabling somewhat independent learning from the same error signals in each network, are argued to provide added flexibility over a single actor network. In the illustration, action (A) is selected through dSPN-driven disinhibition of the effectors of (A) downstream of the GPi/SNr, which is only weakly opposed by the (A)-associated iSPN neurons.

the execution of (A) but strongly negative within the action channel gating (B). The GPi and SNr output neurons belonging to channel (A) are consequently most inhibited and action (A) is selected in winner-take-all fashion. The (motor) effectors of (A) in cortex are then selectively disinhibited, leading to the successful execution of action (A), with no interference from the effectors of (B), which remain inhibited throughout. Phasic DA prediction errors in response to the outcome of action (A) induce opposite plasticity changes at activated dSPNs and iSPNs across channels. Tonic DA levels regulate the relative excitability of dSPNs versus iSPNs throughout the process, thus adjusting the relative weights of the competing Go and No-Go actor policies.

1.2.3 The Basal Ganglia Actor-Critic Model Summarized

I have now outlined the key components of actor-critic models of the BG. Distilled into a broad strokes, generic model, the BG actor-critic circuit might operate roughly as follows:

- The striatal "actor" and "critic" modules—classically the striatal matrix and striosome compartments—receive rich information about the current state, primarily via their corticostriatal inputs.
- The matrix-actor maps state to action, by modulating the inhibitory drive of the GPi and the SNr onto (motor) effectors in accordance with policies learned by trial-and-error and DA feedback. If discrete actions are to be selected, the policy is neurally-instantiated by action-specific BG channels—arising from dSPNs and iSPNs—competing to selectively disinhibit their extra-BG effector targets. Within each action channel, dSPNs act to disinhibit (facilitate) and iSPNs to inhibit (suppress) the associated action, their balance thus determining how likely the channel is to be selected.
- The striosome-critic maps state to value expectations—learned via the same DA error signals that trained the actor—and transmits them directly to the SNc. The SNc, in turn, computes the DA prediction error by integrating an excitatory "reward received" signal (of nonspecific origin) with the striosome's inhibitory "value expected" signal.
- The DA prediction error is finally used to adjust the synaptic weights on the corticostriatal inputs to the striosome-critic and the matrix-actor by

inducing DA-dependent, long-term synaptic plasticity at recently active synapses. The weights on dSPNs and iSPNs are differentially adjusted, rebalancing their activity within the action channels. By this process, the striosome's state-value and the matrix's state-action mappings are optimized.

1.3 The Striato-Pallido-Habenular Critic Pathway

In the last decade, another non-canonical BG output pathway has been linked to a critic-like, evaluative function: the GPi projection to the lateral habenula (LHb)[74–77]. A major GPi-LHb projection was first described in the 1970s[78,79]. Its discoverers noted the pathway for its confluence with a significant "limbic" input, arising from the GPi-adjacent lateral hypothalamus (LHA), because they considered such "striatal" and "limbic" system convergence a rarity, and a unique feature of the LHb. Recent interest in the GPi-LHb pathway was galvanized in the late 2000s, when the group of Okihide Hikosaka demonstrated that the LHb encoded a peculiar kind of prediction error signal[80,81], and that this signal was probably driven by excitatory input from the GPi[74]. These discoveries rekindled efforts to anatomically and functionally characterize previously-delineated "limbic" striosome-GPi-LHb and "motor" matrix-GPi-thalamus BG output pathways[66]; efforts which implicated the pathways further in critic-like "evaluation" and actor-like "motor execution" functions, respectively[76,77,82].

In the next sections, I will first highlight a few seminal studies which mapped-out the striato-pallido-habenular critic circuit, and traced its origins to the striosome. Next, I will describe how this "additional" BG-critic pathway fits into the actor-critic model described in the last section.

1.3.1 Matrix-Actor and Striosome-Critic Outputs in the GPi

In 2007 and 2009, Matsumoto and Hikosaka[80,81] showed that the LHb's responses to cued and uncued rewards and punishments were akin to "inverted" or "reward-negative" versions of the responses observed in the DA midbrain: LHb neurons were most activated by unpredicted punishments (rather than by rewards), and most inhibited by unanticipated rewards (rather than by punishments); well-predicted outcomes of either valence produced little to

no responses. This was recognized as most significant, as the LHb is known to project to[83], and to functionally inhibit[81,84,85] the DA system, suggesting the LHb error signal might contribute to the DA one.

In a follow-up paper, Hong and Hikosaka[74] identified LHb-projecting GPi neurons as the likely source of the LHb prediction error signal. They reported that a majority of task-responsive LHb-projecting GPi neurons fired like the LHb neurons do, but with an earlier onset of activity (by about 20 ms). They hence concluded that inputs from the GPi gave rise to the LHb activity, and that these inputs were presumably excitatory. This result led the authors to suggest that the GPi serves as the BG output of two functionally-distinct pathways: (1) a striato-pallido-thalamic "motor execution" pathway; and (2) a striato-pallido-habenular "reward evaluation" pathway. Hong and Hikosaka speculated that the "evaluation" pathway shapes the DA prediction error signal, and thereby reinforcement in the striatum. This hypothesis, although eschewing actor-critic terminology, arguably postulated the striato-pallido-habenular pathway to form part of a BG critic circuit.

Hong and Hikosaka's "motor" and "evaluation" pathway distinction harkens back to earlier studies of the GPi output pathways. Notably, Van Der Kooy and Carter[86] split the rat GPi into a caudal "motor" and a rostral "limbic" part, based on the regions' selective innervation of the thalamus and the LHb, respectively. In the monkey, Parent and De Bellefeuille[87] differentiated the same projection-based zones, although they found them differently arranged anatomically: in the monkey brain, the "motor" zone formed the center, and the "limbic" zone the periphery of the GPi. Both of these studies were published in the early 1980s—right around the time the striatal matrix and striosome compartments became similarly associated with "sensorimotor" and "limbic" functions, respectively, due to their biased inputs, and divergent outputs[53,65].

In retrospect, it seems a matter of course to connect the "limbic" striosome compartment with the "limbic" GPi output. In 1993, Rajakumar, Elisevich and Flumerfelt[66] became the first to do so, presenting evidence that the parallel GPi-LHb and GPi-thalamus output pathways receive preferential input from the striatal striosome and the matrix compartment, respectively. A number of studies conducted in the last decade likewise support a relatively more significant input from the striosome to the LHb-projecting GPi[75–77,88]. Hong and Hikosaka's striato-pallido-habenular "evaluation" pathway is hence

anatomically part of the "original" striosome-critic circuit of the BG[75–77].

To summarize, the GPi-LHb pathway arises from a cellularly-distinct population of ("limbic") GPi neurons which do not project to the canonical GPi targets in the thalamus and brain stem[86–91]. Moreover, GPi-LHb neurons are preferentially innervated by neurons of the striatal striosome[66,76,77,88]. Hence, the GPi provides segregated outputs for the matrix-actor and the striosome-critic circuits described in the previous section. Fittingly, functional studies of the GPi-LHb "critic" pathway, including Hong and Hikosaka's, have associated it with action and stimulus evaluation, the computation of prediction errors, and DA-driven reinforcement[74,77,92].

1.3.2 The Purpose of the Dual Striosome-Critic Pathways

The existence of a striosome-GPi-LHb-SNc pathway implies that the supposed striosome-critic may exert bidirectional control over DA release in the striatum[75]: Via the *direct* striosome-SNc projection, the striosome *inhibits* DA neurons[58,60,93], whereas via the *indirect* striosome-GPi-LHb-SNc pathway, the striosome may *disinhibit* DA neurons. This disinhibition could ostensibly be accomplished by the striosome shutting off a tonic excitatory drive from the GPi onto LHb neurons[74,76,92,94], thereby reducing LHb-mediated inhibition of the DA system[80,84,85]. Remarkably, in their classic actor-critic model, Houk, Adams and Barto[12] hypothesized the existence of a disinhibitory critic pathway—although theirs projected via the GPe and the STN to the SNc, i.e. akin to the SNr-targeting indirect pathway of the matrix. In their recent work on prediction error-coding in GPi-LHb neurons, Stephenson-Jones and colleagues[77] likewise conjectured that an indirect striosome-GPe-STN-(GPi-LHb) pathway complements the direct striosome-(GPi-LHb) one (**Fig. 1.5 A**). This begs the question: why might the striosome-critic need bidirectional control over its prediction error-computing targets, the SNc and the GPi-LHb? The answer to that question is: to drive secondary (or conditioned) reinforcement.

In order to be an effective teaching signal, the prediction error must not only reinforce events (e.g. pointing at the chocolate flavor) which *directly* led to primary rewards (e.g. ice cream), but also those events "a few steps removed", i.e. events preceding the event preceding the reward (e.g. the "open" sign in the shop window). After all, obtaining a reward may involve

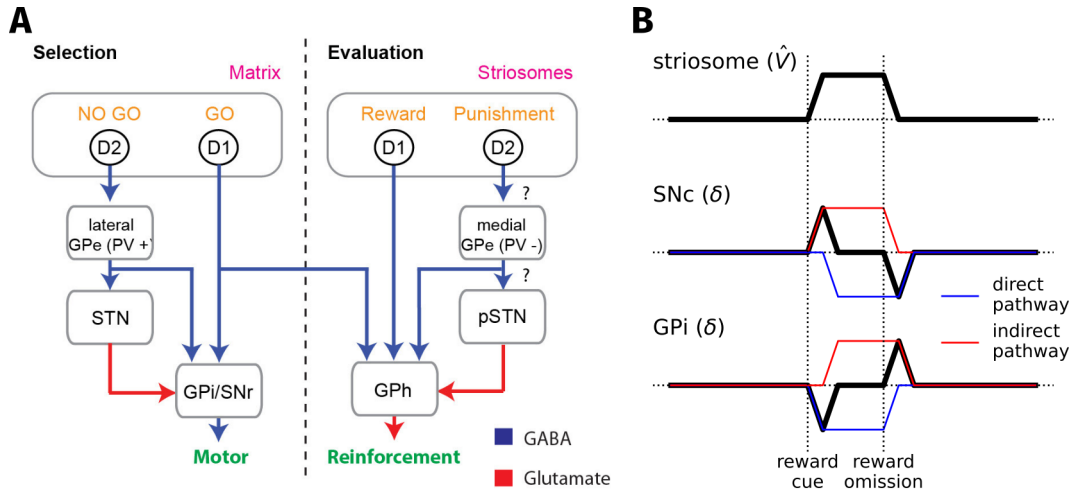

Figure 1.5 | Dual critic pathways compute TD errors and drive secondary reinforcement. **A** Stephenson-Jones et al.'s[77] recent proposal of "action selection" and "evaluation" circuits within the BG essentially recapitulates Houk et al.'s[12] classic actor-critic model, with the LHb-projecting GPi (i.e. GPh) taking the spot of the SNc to drive reinforcement. Like Houk et al., Stephenson-Jones et al. suggest a dedicated, indirect "critic" pathway via the GPe and the STN. **B** Equipped with dual pathways, the striosome may uniformly signal a reward-positive state value estimate $\hat{V}$, yet still drive opposing, reward-reward positive and reward-negative TD prediction errors δ in the SNc and GPi, respectively. All that is required is that one pathway lags behind the other by one time step t, such that one pathway represents the current value estimate $\hat{V}(t)$ and the other the previous state value estimate $\hat{V}(t-1)$. The TD computation $\delta(t) = r(t) + \hat{V}(t) - \hat{V}(t-1)$ also accounts for prediction error responses to secondary reinforcers, such as reward-predicting cues, which are needed to learn about chains of events that lead to the delivery of a primary reinforcer/reward[6,12].

A reprinted from Stephenson-Jones et al.[77], ©2016, with permission from Springer Nature.

progressing through a whole sequence of actions and states. Hence, reinforcing prediction error signals are needed not only in response to unforeseen primary rewards (i.e. primary reinforcers), but also whenever initially neutral, but reward-predictive events (i.e. secondary reinforcers) occur unexpectedly. Indeed, after some training, both SNc and GPi-LHb neurons are reported to start signaling the appearance of a reward-predicting cue—through increases and decreases in activity, respectively—whereas the delivery of the "predictable" reward ceases to incur prediction error signals[17,74]. Thus, both the SNc and the GPi-LHb do respond to secondary reinforcers, as is required to drive effective reinforcement learning, and these responses need to be accounted for in models—for example by supposing the critic exerts bidirectional control.

Houk et al.[12], and others[13], assumed that the positive DA prediction errors evoked by reward-predictive cues/secondary reinforcers were induced

via disinhibition of the SNc mediated by the striosome-critic's indirect pathway. Here is an intuition of how this could work in theory (**Fig. 1.5 B**)[6,12,17]: Prediction error signals δ caused by the unexpected appearance of a secondary reinforcer reflect the difference in the *predicted* value $\hat{V}$ of the current (i.e. time t) and the previous (time $t-1$) state with no primary reinforcers (r) directly involved. The striosome is the part of the critic circuit responsible for state-value predictions $\hat{V}$. Consequently, it can impose the correct error signal $\delta(t)$ on the SNc simply by driving "current state-value estimate"-scaled disinhibition, $+\hat{V}(t)$, and "previous state-value estimate"-scaled inhibition, $-\hat{V}(t-1)$, via its opposing pathways. Importantly, both pathways may carry the *same* state-value estimate $\hat{V}(t)$, provided that the physiological impact of the direct pathway on SNc activity lags behind that of the indirect pathway. If that is the case, the former effectively signals $-\hat{V}(t-1)$ and the latter $+\hat{V}(t)$, and by integrating these signals, the SNc computes their difference $\delta(t)$. The direct pathway's lagging state-value signal $-\hat{V}(t-1)$ moreover negates the excitation evoked by the eventual onset of the primary reinforcer r. Since state-value is defined as the (discounted) sum of all *upcoming* primary reinforcers, as $\hat{V}(t) = \sum_{i=1}^{\infty} r_{t+i}$, it is appropriate to use the *previous* state-value estimate $\hat{V}(t-1)$ in the calculation of the current prediction error $\delta(t)$, be it evoked by a primary or secondary reinforcer. Prediction errors computed in this way, as the difference of the critic's state-value predictions $\hat{V}$ at adjacent time points t, are referred to as temporal difference (TD) errors. The complete TD formula is $\delta(t) = r(t) + \gamma\hat{V}(t) - \hat{V}(t-1)$, with γ being the discount factor[6,17].

The dual pathway TD mechanism outlined above provides a hypothetical explanation of how the striosome-critic could "excite" SNc DA neurons in response to unexpected secondary reinforcers, as well as suppress their activity in response to well-predicted primary rewards (or, indeed, reward omissions), while representing nothing but the state-value estimate $\hat{V}(t)$ throughout. A practically identical dual pathway account also serves to explain how the striosome-critic—encoding the very same $\hat{V}(t)$ signal—may engender the exact opposite responses in the LHb-projecting GPi. Only one minor adjustment is necessary: to suppose that the *direct* striosome-(GPi-LHb) pathway carries $-\hat{V}(t)$ and the *indirect* pathway $+\hat{V}(t-1)$; that is, to suppose that the impact of the disinhibitory indirect pathway on GPi-LHb activity is late relative to that of the inhibitory direct pathway (**Fig. 1.5 B**).

In conclusion, in the classic actor-critic model à la Houk et al.[12], the

striato-pallido-habenular pathway could function as the striosome-critic's indirect pathway to the error-coding DA system. Alternatively, the pathway may be viewed as the striosome's direct pathway to the likewise error-coding GPi-LHb. In either case, the pathway and its opposing complement (e.g. striosome-SNc or striosome-GPe-STN-(GPi-LHb)) would be expected to jointly contribute to the computation of the TD prediction error, which is the error signal that best captures the SNc's and the GPi-LHb's responses to primary and secondary reinforcers. Finally, the dual pathway architecture presents a relatively simple explanation of how the striosome-critic could cause opposing patterns of activity in the SNc and the GPi-LHb.

1.4 Aim: Observing the Critic in Action

The principal aim of the work presented here was to examine the proposed striato-pallido-habenular critic pathway *in action*. In action is to be understood in two ways: firstly, the data included in this thesis was collected in behavioral experiments—they capture the experimental mice, the activity of the mice's critic neurons, and the effects of manipulating the activity of those neurons, in action. Secondly, we were interested in state evaluation and prediction error signals that relate to, and serve to optimize, action selection—action selection in the "common sense" sense of deciding between available paths of actions and options, such as whether to take a left or a right turn in a maze, or which ice cream flavor to choose.

An RL agent implementing the actor-critic architecture could optimize vastly different kinds of policies, as the framework is agnostic to the "level of abstraction" and nature of the state-policy associations learned[7,95,96]. Indeed, the various BG loops may well instantiate several different kinds of interacting actor-critic agents, collaborating and competing to shape our behavior[6,15]. The dorsomedial and the dorsolateral striatum are—for example—linked to the acquisition of goal-directed versus habitual behavioral strategies, respectively, and these strategies are thought to express different kinds of associations (i.e. action-outcome versus stimulus-response associations)[97]. The two parts of dorsal striatum may thus implement two different kinds of actors[15]. An actor may also learn how to best modulate and constrain movement kinematics in a given context[95,98,99], which mental representations to keep or update in working memory[46], how to correctly pitch and

sequence notes into a song[47], or—generally—how to task-appropriately distill the key information from its cortical input into a useful lower-dimensional representation[100].

Policy-based decision-making at the "chocolate ice-cream level" is but one attractive (and experimentally tractable) use of the actor-critic architecture. It is this specific use we expected to find evidence of in our examinations of the striato-pallido-habenular critic circuit, in action.

Chapter 2

Methods

2.1 Functional Circuit Interrogation Techniques

Targeting Specific Populations Using Viruses And Transgenic Mice

In the experiments described in this thesis, we always limited our neuronal activity recordings or manipulations to genetically-specified populations or cell types of interest. An example: we expressed the fluorescent neuronal activity indicator GCaMP selectively in μ-opioid receptor-expressing neurons of the dorsal striatum, rather than in all the neurons in the region[1]. In all experiments, this population-specificity was accomplished using combinations of viruses and genetically-modified (i.e. transgenic) "driver" mouse lines. The viruses expressed the engineered proteins required for the experiments, whereas the transgenic mouse lines limited the virus-mediated protein expression to the populations of interest[101].

The various viruses used (e.g. AAV, HSV) were usually injected directly into the brain region of interest (e.g. dorsal striatum, GPi). Once administered, the viruses locally transfected neurons with DNA sequences that drove the expression of the engineered proteins—proteins which were subsequently used to read out or manipulate neuronal activity during experiments (e.g. GCaMP, ChR2). In all three studies treated here, the expression of these "protein tools" was conditioned upon the presence of particular recombinase enzymes, which do not naturally occur in mice: the Cre or the Flp recombinases.

The Cre and Flp recombinases were originally derived from a bacteriophage and a yeast, respectively. Their function is to insert, remove, or invert DNA sequences which are flanked by recombinase-specific recognition

target sequences[102]. So-called DIO or FLEX viruses carry synthetic DNA sequences that include a "flanked" gene which is oriented in the antisense direction. The recognition targets in the viruses' DNA sequences are arranged in a way that cause the particular recombinase used to invert the flanked gene to its sense direction, which leads to the expression of the encoded protein. In the absence of the recombinase, the protein is not expressed, as the coding gene remains in the antisense orientation[101,102]. DIO or FLEX virus-mediated protein expression is therefore recombinase-dependent (i.e. Cre- or Flp-dependent).

The various transgenic mouse lines we employed in our studies all expressed recombinase transgenes selectively in association with defined target genes. In these "recombinase-driver" mouse lines, the use of DIO viruses consequently resulted in the restricted, recombinase-dependent expression of the engineered proteins in cells which were positive for the line's respective target gene. To return to the example provided above: to achieve the selective expression of the GCaMP calcium indicator in μ-opioid receptor-positive neurons in dorsal striatum, we injected Oprm1-Cre mice with an AAV-DIO-GCaMP virus[1]. The shorthand "Oprm1-Cre" describes transgenic mice which express the Cre recombinase wherever the protein encoded by the Oprm1 gene—that is, the μ-opioid receptor—is expressed[103], and "AAV-DIO-GCaMP" denotes a viral vector driving the Cre-dependent expression of the GCaMP protein, as indicated by the DIO tag (**Fig. 2.1 A**).

In many instances, we aimed to experimentally targeted neurons projecting to a region of interest, and not neurons with cell bodies in the area. For this purpose we employed retrograde viruses, i.e. viruses that are are transported retrogradely from the axon terminal to the cell body. Retrogradely-transported viruses include the AAVrg and HSV viruses, but not regular AAVs[101,104,105]. Thus, to express GCaMP in GPi neurons that project to the LHb and are Sst-positive, we injected a retrograde and Cre-dependent AAVrg-DIO-GCaMP virus into the LHb of Sst-Cre mice[2].

Calcium Imaging: Capturing Neuronal Activity By Proxy Using GCaMPs

Calcium serves as a key intracellular messenger that translates the electrical activation of neurons into neurotransmitter release and lasting physiological changes, such as synaptic plasticity. Calcium enters neurons through voltage-gated channels, and the concentration of intracellular free calcium is

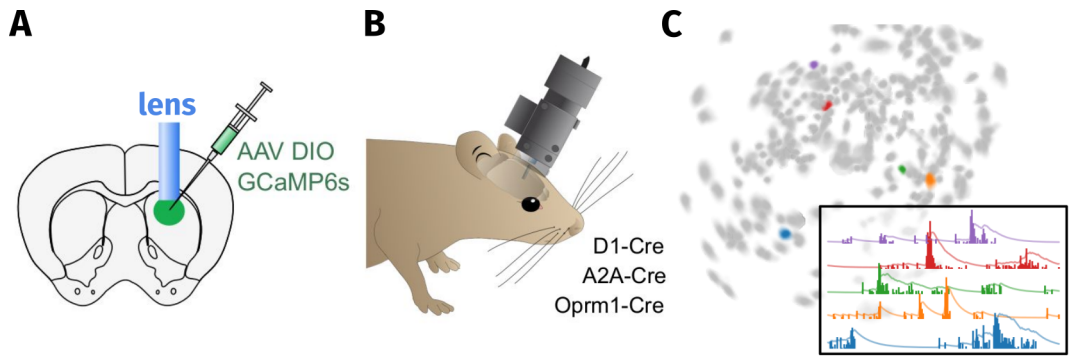

Figure 2.1 | GCaMP-based calcium imaging in transgenic mice using miniscopes.
A An AAV-DIO-GCaMP virus is injected to express the fluorescent calcium indicator GCaMP6s Cre-dependently in the dorsal striatum of a transgenic Cre driver mouse (e.g. Oprm1-Cre to target striosomal cells). For imaging purposes, a relay lens is implanted above the injection site.
B A mouse with the "miniscope" miniaturized microscope attached in order to image the neuronal activity-dependent fluorescence of GCaMP-expressing neurons.
C An illustration showing the region-of-interest (ROI) masks marking individual neurons detected in the field of view of a miniscope recording. Fluorescence transients of the color-coded ROIs are depicted in the inset (transparent line). The long fluorescence decay "tail" of the transients can be removed by means of signal deconvolution to obtain a more temporally precise estimate of the neuronal activity (opaque vertical bars).

therefore neuronal activity-dependent[106]. Consequently, calcium levels can be used as a "proxy-measurement" to approximate neuronal activity in live animals[107,108]. Technically, intracellular calcium fluctuations are commonly captured by imaging genetically-encoded, fluorescent calcium sensors, like the popular GCaMP proteins[108,109], under a microscope. The brightness of the light emitted from the green fluorescent GCaMP proteins is enhanced by calcium binding, and is thus indicative of the availability of free calcium and the activity of the imaged neurons[108,109].

In our experiments, we recorded the neuronal activity-dependent GCaMP fluorescence using two different imaging approaches. The "miniscope" approach utilizes a miniaturized microscope (i.e. "miniscopes"), which is attached to the head of the mouse during experiments and serves to film fluorescing neurons through an endoscopic relay lens chronically implanted in the brain (**Fig. 2.1**)[110,111]. The "fiber photometry" approach utilizes a chronically-implanted optical fiber instead of a lens, and the fluorescence signal is collected by a sensor at the end of a brain-external, detachable optical fiber, which is connected to brain-implanted fiber during recording sessions[112,113]. Fiber photometry hence dispenses with the head-mounted imaging sensor and replaces the implanted lens with a fiber of typically smaller diame-

ter, which makes photometry somewhat simpler, less invasive and tissue-damaging than the miniscope technique. However, these benefits come at the cost of spatial resolution, as only bulk fluorescence, aggregated over the entire target neuronal population, can be measured through the photometry fiber, whereas the miniscope's lens resolves the fluorescence emitted by individual GCaMP-expressing neurons (**Fig. 2.1 C**)[111]. Here, we used miniscopes to record three distinct populations of dorsal striatal projection neurons[1], as well as LHb-projecting LHA neurons[3] at cellular resolution, and photometry to record the population-average activity of LHb-projecting GPi neurons[2] in mice performing a number of behavioral tasks and tests.

Optogenetics: Activating Neurons With Light Using ChR2

Derived from a green alga, ChR2 (i.e. Channelrhodopsin-2) is a blue light-gated cation-selective ion channel. As it transports positively-charged ions (i.e. cations), it depolarizes cells when activated by light[114]. Critically, ChR2 evinces very fast channel opening kinetics and high conductance, allowing experimenters to use it to evoke spiking in mammalian neurons with brief pulses of blue light with millisecond precision[115]. In one study reported here[3], we employed chronically-implanted optical fibers to optogenetically-activate LHb-projecting populations of GPi and LHA neurons acutely whilst mice engaged in different behavioral tasks unrestrained.

Silencing Neurons Chronically Using TeTxLC

TeTxLC (i.e. the tetanus toxin light chain) is the fragment of the tetanus toxin protein that effects the toxin's potent inhibition of neurotransmission. TeTxLC proteolytically cleaves the vesicle-associated membrane protein (VAMP), thereby disabling neurons' synaptic vesicle release machinery and blocking all neurotransmitter release, which effectively silences the neurons without inducing cell death[116–118]. We used virally-expressed and genetically-targeted TeTxLC protein to chronically inactivate the synaptic output of GPi neurons projecting to the LHb in mice performing a maze-based reversal task[2].

2.2 Mouse Behavioral Tasks And Tests

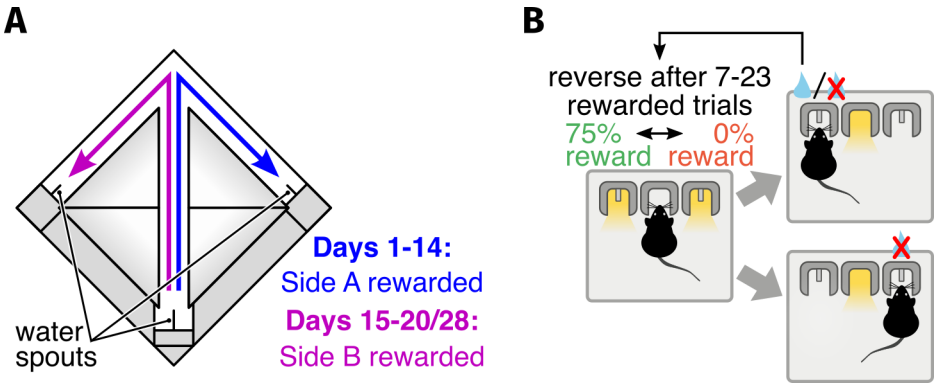

Figure 2.2 | Maze-based and nosepoke-based reversal tasks.
A The "arrow maze" reversal task. Mice were trained to shuttle back and forth between the water spout located at the start of the central corridor—the initiation spout—and the choice spouts placed at the ends of the side corridors. Initially, mice only obtained a reward if they approached the choice spout *A* first after leaving the area of the initiation spout. Approaching *B* did not yield rewards. After two weeks, the reward location was reversed; from now on, rewards were made available at spout *B* but never at *A*. After every trial, mice needed to return to the initiation spout to reactivate the choice spouts; this, too, yielded a reward.
B The nosepoke-based probabilistic reversal task[77,121]. Mice initiated trials at the center nosepoke port, then chose either the left or right side port. Nosepoking one of the two choice ports resulted in a reward with a probability of 75%, the other was not rewarded. Every 7–23 rewarded trials[2,3,77,121]—or after any reward, with a probability of 5%[1]—the location of the reward swapped sides without warning. LED light cues indicated whether the center port or the choice ports were active. To perform better than chance-level, mice had to keep track of the trial outcome history to infer whether rewards were omitted by chance or a reversal had occurred.

Reversal Learning: Assessing Behavioral Flexibility

In reversal tasks, a test subject's well-trained action-outcome associations and habitual response biases are challenged by "reversing" (i.e. swapping) the outcomes associated with the choices available in the context of the task, in many instances repeatedly. The purpose of the reversal challenge is to evaluate the subject's ability to adaptively respond to the changed contingencies, and thereby to assess reward learning, inhibitory control and other aspects of behavioral flexibility[119,120]. We recorded neuronal activity while mice performed two different kinds of reversal tasks to investigate what roles various neuronal populations in the dorsal striatum[1] and the GPi[2] might play in flexible decision-making, action execution and outcome evaluation.

Maze-Based Reversal Task In our "arrow maze" reversal task (**Fig. 2.2 A**)[2], mice were placed into a maze consisting of three corridors, each equipped with a water spout. Spout "C" was located at the starting point of the maze, at the beginning of the center corridor and opposite of a three-way "choice

junction". The spouts "A" and "B" were located at the ends of the two choice corridors branching off of the center corridor at the junction towards the left and right. Closing in on water spout "A" was consistently reinforced with water rewards, whereas approaching "B" was not rewarded. Returning to the starting point and spout "C" initiated a new trial, reactivated spouts "A" and "B", and yielded a water reward. This pre-reversal phase of the task lasted a total of 14 sessions of 20 minutes each. Subsequently, the outcome contingencies were reversed; that is, the rewards were henceforth only made available at spout "B" and no longer at "A". Mice were trained on these post-reversal contingencies for another one to two weeks. Which spout was denoted "A" and which "B" depended on the experiment and subject. The box housing the maze measured 40x40 cm and was lit by floor-level, white LED light strips. To motivate mice to perform the task for water rewards, water was not supplied outside of the task unless a mouse had consumed less than 1 ml in total over the course of the day's training session.

Nosepoke-Based Probabilistic Reversal Task In the nosepoke-based probabilistic reversal task[1–3,77,121] mice chose between nosepoke ports (i.e. circular, snout-sized openings) embedded in the front panel of an operant chamber instead of between maze corridors. The task moreover involved serial reversals and probabilistic reward contingencies, differentiating it further from the simpler maze task (**Fig. 2.2 B**).

Mice were placed into a chamber with three nosepoke ports. Reward spouts were located inside the left and right choice ports, but not the center port. The animals had a 75 % chance of obtaining a drop of sucrose solution if they poked their snout into the correct choice port, whereas incorrect choices yielded no rewards. To reactivate the choice ports and thus to initiate a new trial, the mice were required to enter the center port. Outcome contingencies reversed repeatedly and at random, but always following a reward; the exact number of rewards delivered prior to the reversal was drawn either from a uniform (7–23 trials)[2,3] or a geometric (5 % probability)[1] distribution. In one study presented here[1], the mice were required keep their snout inside the ports for at least 350 ms in order to action them, whereas in the other two[2,3] the ports were triggered instantaneously. Throughout the session, either the initiation port or both choice ports were lit by internal, white LED lights to indicate the current trial phase—"trial initiation" or "choice". Neither the

correct choice nor the occurrence of the reversal were cued. The operant chamber was circa 15x15 cm in size and illuminated by infrared LEDs. The port-mounted LEDs were the only sources of light in the visible spectrum inside the chamber. Prior to training, the mice were placed on food restriction (kept above 85 % of their free-feeding weight) to render the highly palatable sucrose solution (15 %) even more rewarding to the hungry mice.

Open Field: Activity Recordings During Spontaneous Locomotion

In the studies reported in this thesis, the open field test primarily served us to record neuronal activity in the dorsal striatum[1] and the GPi[2] from mice moving about freely and spontaneously, as opposed to guided by the rules and reward contingencies of the reversal tasks. The open field consisted in a square, dimly-lit arena measuring 49 cm a side, which mice were left to explore for 15 to 20 minutes per recording session. Infrared light, invisible to the mice, supplied additional illumination for the video camera recording the mice's locomotor behavior for later analysis. Mice were water or food restricted during the open field session, as they were during the reversal tasks.

Real-Time Place Avoidance: Is the Induced Activity Aversive?

The purpose of the real-time place avoidance test was to determine whether the optogenetic activation of either the LHb-projecting GPi or the LHb-projecting LHA was experienced as inherently "aversive" (i.e. negative)[3]. We conducted the test over the course of three consecutive days. Each day, the mice were connected to optical fibers and placed into a behavior box consisting of two identical compartments, connected by a gap in the dividing wall. The mice were free to move between and explore the two compartments for 20 minutes each session, and the time they spent in each compartment was registered. The session on day 1 served as a control, and thus no optogenetic stimulation was applied. On day 2, mice received optogenetic stimulation whenever they occupied one of the two compartments. On day 3, the "stimulated compartment" was swapped to the opposite side. If the stimulation were perceived as aversive, the mice would be expected to spend significantly less time in the stimulated compartment on days 2 and 3. On the other hand, no clear preference for either compartment should be evident in

the control session on day 1. The two box compartments were 25 cm wide and deep, and only dimly illuminated in the spectrum visible to the mice.

Auditory Fear Conditioning: Does The Population Predict Aversive Events?

We employed auditory fear conditioning to investigate whether the activity of LHb-projecting LHA neurons may contribute to the punishment prediction and prediction error signals observed in the LHb[81]. Over a period of three consecutive days, mice were exposed to a total of 15 pairings of a tone cue and a mild foot shock. The tone lasted 10 seconds, and co-terminated with the 1 second shock, delivered through the grid floor of the conditioning environment. Individual tone-shock pairings were separated by randomized inter-trial-intervals. If LHb-projecting LHA neurons acquired LHb-like prediction and prediction error signals, we expected marked increases in the magnitude of the neurons' tone responses ("prediction") as well as decreases in their shock responses ("prediction error") by the end of the fear conditioning protocol.

Chapter 3

Article I: The Striatum in Action

In **Article I**, we investigated whether the activity of striatal matrix and striosome neurons reflected their purported functions as actor and critic in decision-making. Specifically, we recorded the activity of direct pathway (dSPNs, D1+), indirect pathway (iSPNs, A2a+) and striosomal (sSPNs, Oprm1+) spiny projection neurons in the dorsomedial striatum (DMS) using miniscopes while mice performed the nosepoke-based probabilistic reversal task. Our results challenge predictions derived from action selection-focused and simplistic interpretations of the actor-critic framework. Contradicting notions of clearly segregated action channels engaged in a competitive struggle for behavioral control at the decision point, activity in all pathways was highly continuous in both space and time, and pathway differences—if present—were too subtle for us to detect with our experimental approach.

3.1 Background

We targeted the DMS, as it is the striatal region most associated with flexible decision-making and reversal task performance[122]: The DMS receives prominent input from higher-order associative regions of the frontal cortex—such as the orbitofrontal and prelimbic areas—which are likewise implicated in reward processing and decision-making[31,122]. Via the frontal cortex, the DMS has been shown to receive persistent value signals, outlasting inter-trial-intervals of up to fifteen seconds, in a probabilistic two-choice reversal task[123]. Population activity in the DMS reportedly ramps-up while animals approach the decision point of a maze, or while they anticipate a "go" cue, and peak as the choice is executed[124,125]. One study found the pre-choice activity of

more than a third of task-responsive DMS SPNs to transiently and selectively scale with the value of one of two available choices[126]. Moreover, the activity in the DMS during the reinforcement phase of a visual-motor association task reflected the rate at which the associations were acquired, and boosting the activity by means of electrical stimulation accelerated the learning process[127]. Finally, lesions of the DMS are known to impair animals' ability to adapt to reversals, or to flexibly switch between behavioral strategies[128,129]. Altogether, these pathway-agnostic studies strongly support the notion that the DMS is involved in (policy-based) action selection and action evaluation.

Matrix: Evidence for Opponent Actor Action Selection

Studies in which the two matrix pathways were selectively recorded or manipulated lend credence to the proposal that dSPNs and iSPNs function as "opponent actors"—facilitating and suppressing actions, respectively[45,130]. Nonomura and colleagues[131] described that DMS dSPNs and iSPNs encode both the selection and the reinforcement outcomes of specific actions performed in the context of a serial reversal task with probabilistic rewards. Importantly, they found that dSPNs were activated and iSPNs suppressed by the (secondary) reinforcer—a tone cue which preceded reward delivery. Conversely, a second tone, predictive of a "no-reward" trial outcome, coincided with suppression and excitation in dSPNs and iSPNs, respectively. These anti-correlated responses to trial outcomes could correspond to the DA-mediated, antagonistic plasticity updates that Collins and Frank's OpAL model[45] predicts occur in the dSPN-"Go" and iSPN-"No-Go" policy networks during reinforcement. Suggestively, enhancing the activation of dSPNs during the presentation of the outcome-signaling tone cues increased, whereas driving iSPNs decreased, the likelihood of rats repeating their choice on the next trial[131]. Relatedly, Kravitz, Tye and Kreitzer[132] triggered pathway-specific optogenetic stimulation whenever a mouse selected one of two available choices in a simple operant task. Similarly to what was observed by Nonomura et al.[131], activating dSPNs increased the likelihood of animals repeating the choice (and kept them engaged in the task), whereas activating iSPNs biased animals away from the stimulated choice (and led them to abandon the task altogether). Collins and Frank[45] successfully replicated this behavior in silico, using OpAL, by zeroing the prediction error and substituting it with a fixed "stimulation" parameter, which affected only the opponent actor

pathways, but not the critic.

Most relevant to our work presented in this chapter is evidence that pre-choice, unilateral excitation[121] (or inhibition[133]) of DMS dSPNs and iSPNs introduces opposite, lateralized biases to decision-making in mice—lateralized biases which are not inducible by the same stimulation protocol if it is applied outside the choice task, i.e. during spontaneous movement. One study providing evidence to that effect is a landmark 2012 study by Lung-Hao Tai and collaborators[121], in which they used the same nosepoke-based reversal task as later employed by us, and discussed here.

In the nosepoke-based probabilistic reversal task, mice initiate trials at a center nosepoke port, then choose either one of the two ports located to the left and right. One of these ports yields a water reward with a probability of 75 %, the other is unrewarded. Which side is rewarded, and which is not, is not signaled to the animal (other than by the actual delivery of the rewards), and its location reverses occasionally and without warning. The animals therefore have to infer if a reversal occurred by keeping track of the most recent reward outcomes (see Methods, **Fig. 2.2 B**). The behavior of the mice in the task is typically—including in Tai et al.'s study—fit with a logistic regression model, which predicts port choice on any given trial based on the outcomes of the previous trials. For mice performing the task with above-chance accuracy, the regression coefficients will reflect that rewards substantially increase the likelihood of "staying" with the rewarded choice, whereas no-reward outcomes impact choice much less markedly (as mice will persevere in their choice on some occasions, and "give up" and switch ports on others). The outcome of the most recent trial is consistently shown to affect choice the most.

In the study by Tai et al.[121], a mouse's poking into the center initiation port occasionally triggered 500 ms of unilateral optogenetic activation of dSPNs, or of iSPNs, in the DMS. This happened at random, on roughly 6 % of trials. They found that on trials in which dSPNs were activated pre-choice, mice were more likely to opt for the side port contralateral to the stimulated hemisphere than expected based on the logistic regression fit of their behavior (**Fig. 3.1 A**). Contrariwise, iSPN stimulation decreased the probability of a contraversive choice below the regression model prediction. Based on these results, Tai et al. suggest that enhanced dSPN activity translates into a transient increase, and iSPN activity into a decrease, of the "relative action

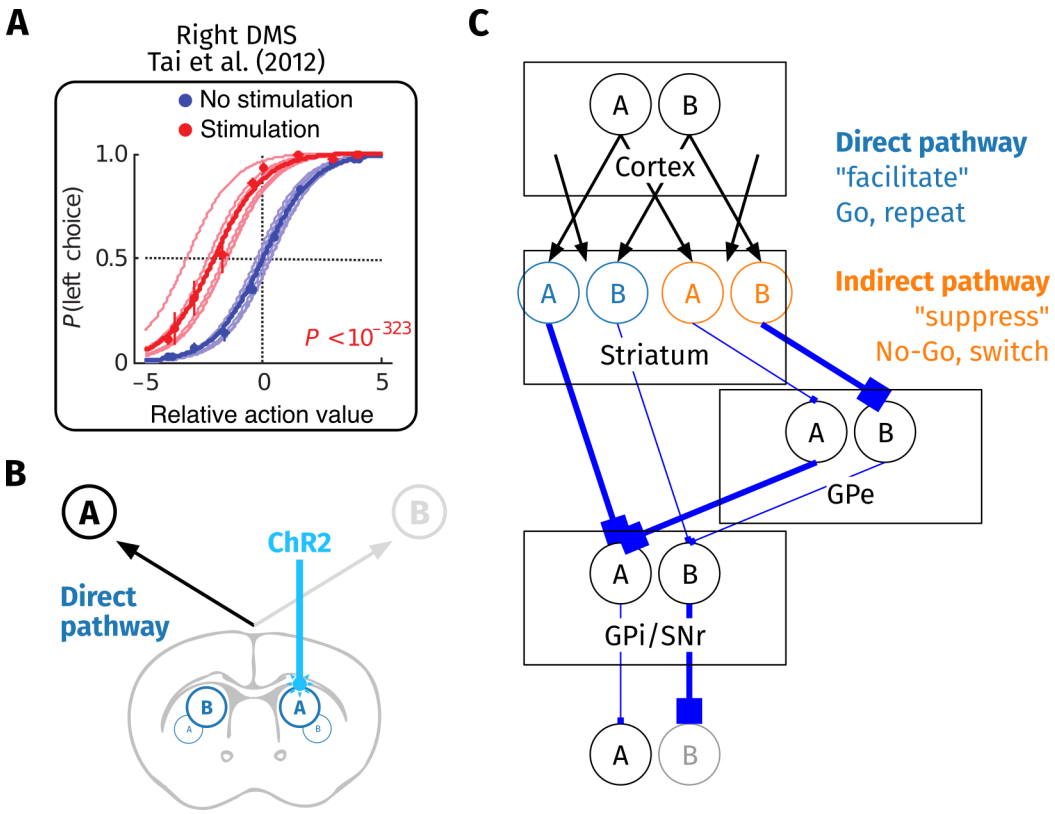

Figure 3.1 | Pathway-specific activation of matrix neurons in the DMS affects choice.
A Unilateral optogenetic activation of dSPNs in the right DMS increases the probability of contraversive, left choices in the probabilistic reversal task.
B If DMS dSPNs are policy-coding, they are presumably overrepresenting action (A)—i.e. the contraversive, left choice—in the right hemisphere.
C Collins and Frank[45] have replicated the optogenetically-induced choice bias shown in **A** using their OpAL model. They did so assuming a selective, transient increase in the synaptic weights of the "Go policy"-coding dSPNs of action channel (A) on stimulated trials.

Panel **A** adapted from Tai et al.[121], ©2012, with permission from Springer Nature.

value" of the contraversive choice.

The relative action value was defined as the log odds of a left choice on a given trial, as estimated by the logistic regression fit. Because it reflects a choice probability estimate, the "action value" is better viewed as a "policy value"[134]. Accordingly, Collins and Frank[45] replicated the Tai et al. experimental data in an OpAL-based simulation by assuming the stimulation transiently increased the synaptic weights of either the dSPN-"Go" or the iSPN-"No-Go" *actor* neurons of the "contraversive choice" action channel. That last bit is significant—Tai et al.'s interpretation of the data, and Collins and Frank's simulation, relies on the assumption that the unilaterally-targeted hemisphere's *policy neurons* preferentially represent the contraversive choice (**Fig. 3.1 B–C**).

Striosome: Evidence for a Role in (State) Evaluation

Functional data on the role of the striosome in decision-making is still very sparse, although a function in reward-processing and reinforcement learning has long been suspected. In 1998, White and Hiroi[135] demonstrated that rats will press a bar to receive direct electrical stimulation of the striosome, indicating that activity within the compartment is sufficient to drive reinforcement. The striosome has moreover been implicated in the rewarding, reinforcing and movement-enhancing properties of opioids[136], and in the development of drug-induced motor stereotypies (i.e. inflexible and repetitive behaviors) which are observed after chronic exposure to stimulants, such as cocaine[137,138]. Lesions of the striosome reduce the acquisition or expression of these drug-induced stereotypies[139,140], and of habitual operant responses, which are thought to reflect inflexible stimulus–response associations[141,142]. Such lesions further impair performance on a skill-based locomotor task—without affecting movement generally[142,143]. Taken together, these findings could suggest that the striosome, trough its connections with the DA system, regulates reward-driven reinforcement learning—at least of stimulus–response habits, of which motor stereotypies may be an extreme, aberrant form[52].

Due to its comparatively small volume and unpredictably twisting shape within the striatum, it has been difficult to reliably target the striosomal compartment in vivo. There are, however, a number of studies that accomplished the feat of selectively recording or manipulating the activity of striosomal neurons in behaving animals. The groups of Ann Graybiel[144] and Kenji Doya[145]

have imaged the activity of individual dorsal striatal striosome and matrix neurons in mice subject to classical conditioning paradigms using calcium indicators. Both groups found that a significantly greater fraction of striosome than of matrix neurons responded to reward-predicting cues. Interestingly, the Doya laboratory[145] also reported that the striosomal population was equally and similarly-selectively engaged by cues-predicting an unpleasant air puff. Ann Graybiel's group[146,147] moreover published electrophysiological recording and optogenetic manipulation data indicating that, in mice, striosomal activity affects cost-benefit tradeoffs, with greater striosomal activity correlating with an increased willingness to approach a maze location associated with a high-cost (unpleasantly bright light), high-benefit (pure chocolate milk reward) outcome. Most recently, the Graybiel team showed that striosomal population activity, recorded in the response window of an auditory discrimination task, reflected the expected value of the upcoming outcome—drops of water or bright light—to a greater degree than matrix activity[148]. Importantly, the activity within the striosome increased the larger the expected reward, and decreased the more severe the punishment. In stark contrast, a recent publication authored by the Bo Li laboratory[149] identified a subclass of striosomal neurons whose activity was strongly associated with punishment and negative reinforcement. Strikingly, inactivating these neurons selectively impaired learning to avoid punishments.

3.2 Aims and Expectations

The experimental evidence discussed above, considered through the perspective of a simple actor-critic model, led to the two major aims and the accompanying—and arguably naive—expectations listed below, and illustrated with mock data in **Fig. 3.2**.

1. Observe the opponent actor pathways of the matrix in action:

- SPN policy neurons form discrete action channels encoding contraversive or ipsiversive choice[12,29,45].
- Activity within these channels increases pre-choice—while in the center initiation port—and peaks during action execution[124–126].
- SPN policy neurons overrepresent the contraversive choice (i.e. the left choice as all recordings were performed in the right hemisphere)[45,121,125].

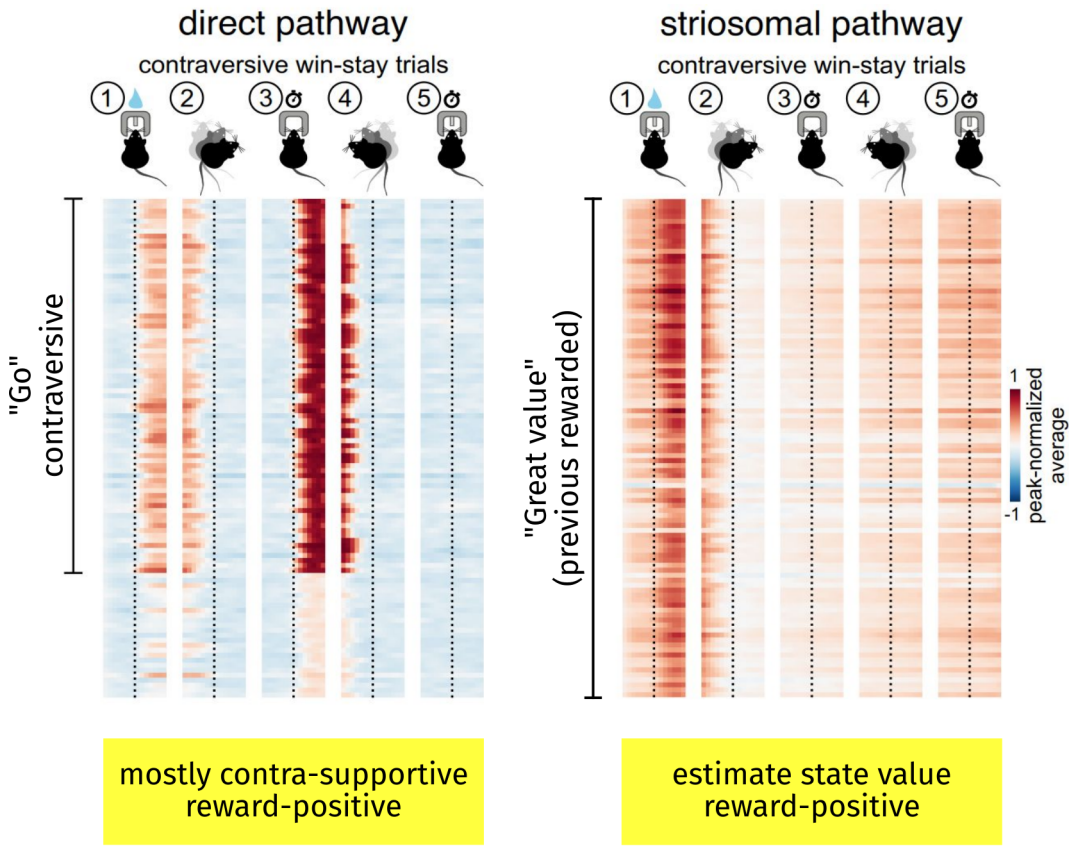

Figure 3.2 | Mock Data Average responses in contraversive win-stay trials.
Naive predictions based on the lateralized and antagonistic effects of unilateral stimulation of dSPNs and iSPNs in the probabilistic reversal task[121], the opposite responses of dSPNs and iSPNs to reinforcement outcomes[131], and the assumption that "conventional" striosome neurons signal subjective state value[148].
Actor (left): A positive (contralateral) outcome (1) excites dSPNs and inhibits iSPNs. After just receiving a reward at the contralateral port, the mostly contraversive choice-supporting dSPN-"Go" are strongly activated at trial initiation (3) to select another contraversive choice. The mostly contraversive choice-suppressing iSPN-"No-Go" neurons are much less activated pre-choice (3), "losing" the within-channel competition to control action selection. iSPN mock data not shown.
Critic (right): The striosome encodes state value, reflecting a discounted estimate of future rewards. Having just received a reward (1), another is expected in the near future. Due to the discounting of distant rewards, the activity may ramp up as the predicted moment of the next outcome approaches (2)-(5).

- dSPN neurons encode "Go" policies, supporting the action associated with their channel (i.e. mostly contralateral choices), whereas iSPNs encode "No-Go" policies, suppressing the same action[45,121,130,131].
- dSPN-"Go" and iSPN-"No-Go" neurons show opposite responses—excitation and inhibition—to reward/reinforcement outcomes, perhaps reflecting DA-mediated policy updates[45,131,132].

2. Observe the critic pathway of the striosome in action:

- sSPN critic neurons encode positive state value—the more likely it is an upcoming outcome will be a reward, the more active they are[12,148].
- sSPN critic activity is not action specific, as it reflects state—not action—value.
- State value representation is relatively persistent throughout the trial (perhaps driven by prefrontal value signals[123]), as TD errors are computed through the interaction of sSPN projections[12] (as opposed to within the compartment[29]); however, activity may ramp up as the reward gets closer (due to the discounting of distant rewards[6]).

3.3 Results

— Fig. 3.3 to 3.6 adapted from Article I —

Continuous Ensembles Encode The Entire Trial

SPNs of all three pathways activated in continuous sequences spanning the entire trial (**Fig. 3.3 A**). Could the moment-to-moment neuronal activity correspond to the ongoing behavior of the mouse, or even the abstract state of the trial? A momentary snapshot of the activity of any pathway indeed sufficed to determine at what point of, or *where* in the trial, a task-engaged mouse was. In practical terms, we were able to identify the phase of the trial, and even to predict a mouse's progress through the given phase, from phase average activity patterns (i.e. averages of consecutive imaging frames, segmented by phase) and instantaneous activity (i.e. single frames), respectively, using cross-validated SVM decoders (**Fig. 3.3 C-D**, Art. I Fig. 3 J, M-N). This worked irrespective of pathway with accuracy well above chance level. Moreover, decoders trained on neuronal activity from one

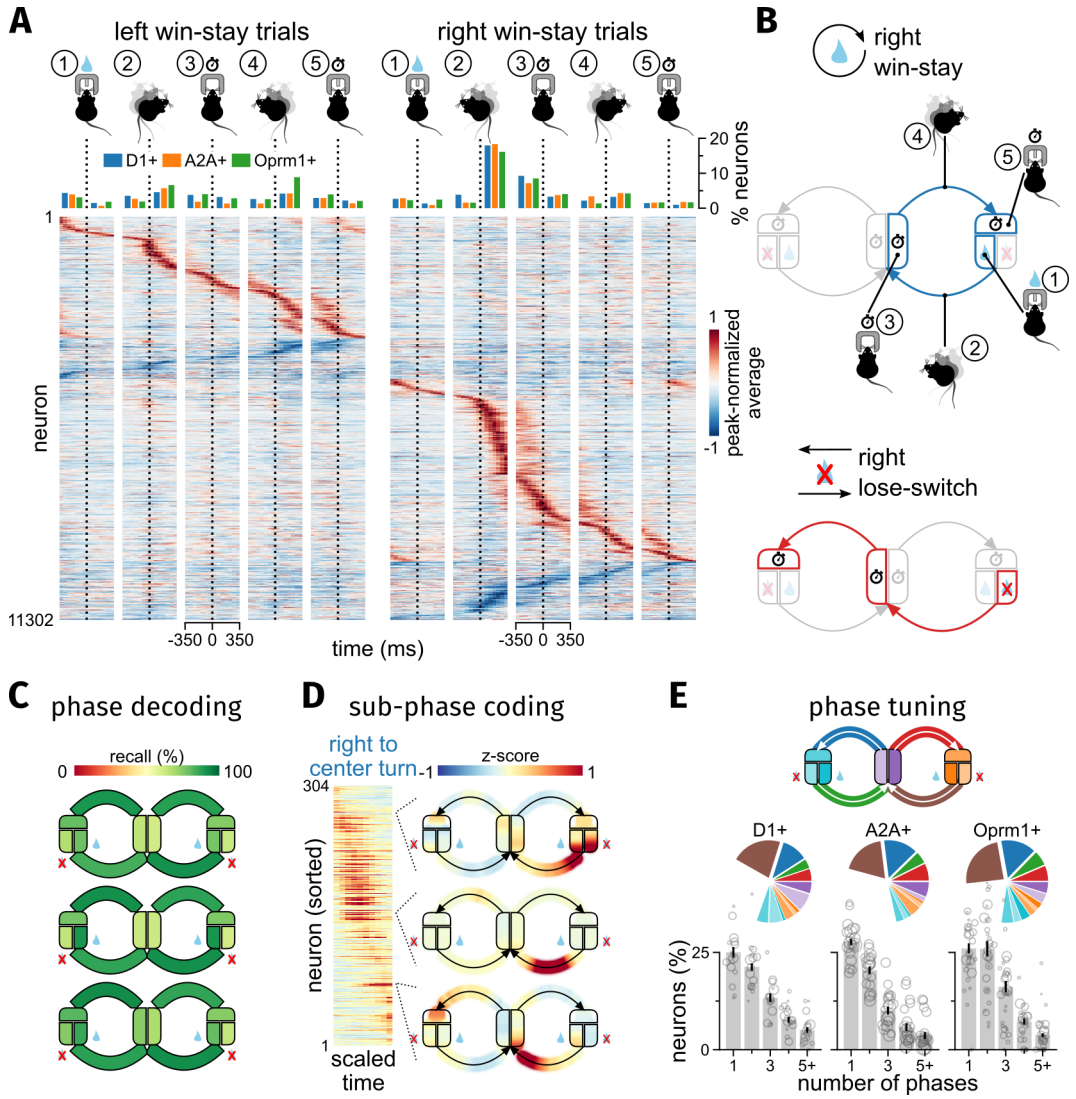

Figure 3.3 | Phase Decoding.

A Average activity of SPNs of all three pathways during win-stay trials, sorted by the activity peak (or valley). The raster columns are centered on the following trial events: (1), reward delivery; (2), return to the initiation port; (3), 350ms initiation nosepoke; (4), choice turn; (5), 350ms choice nosepoke. The bar charts above each "event column" indicate the percentage of neurons whose trial activity peak/valley fell into the pre- or post-event onset time bin of the column (i.e. peaking left or right of the time 0 dotted line), for each population.

B Illustration explaining the "Figure-8" plots. The three rounded-off boxes represent the three nosepoke ports. The choice port boxes' subdivisions represent the 350ms hold phase, and the reward and no-reward outcome phases. The center port box is split according to the direction of the upcoming movement. The movements between the ports are represented by the arrows.

C The phase decoding accuracy for each of the 12 phases we distinguished.

D Sub-phase coding in an example session (Oprm1+ mouse). The raster shows the average activity of the recorded neurons during right-to-center turns. The activity was stretched/compressed to uniform length and peak-sorted. The figure-8 plots depict the average activity of three individual neurons.

E The pie charts indicate the fraction of neurons significantly positively-tuned to each phase by pathway. Only the most significant tuning determined the grouping. Color coded as indicated by the "figure-8 key". The bar charts show the percentage of neurons significantly positively-tuned to various counts of phases, by pathway.

session proved capable of labeling phases in another session of the same animal, recorded up to 14 days later (Art. I Fig. 3 O-Q). Although the accuracy of the phase predictions dropped markedly when decoding across days, it remained above chance for data from all three pathways. Taking into account a degree of mismatching when aligning neurons across sessions, this indicates that the SPN activity patterns are at least reasonably stable over time. These findings suggest that the temporally continuous, sequential activity patterns observed in all three SPN populations reliably reflect motor behavior, or indeed cognitive processes, relating to the trial state.

The fraction of SPNs statistically significantly modulated during (i.e. "tuned to") each trial phase type, and the strengths of these modulations, did not differ markedly between pathways. That is, across pathways, significant phase tunings were similarly distributed over neurons (**Fig. 3.3 E**), and the average responses of the tuned populations were comparable in magnitude (Art. I Fig. 3 F). Moreover, the underlying distributions of tuning strengths—assessed using a continuous, standardized tuning score—proved unimodal (i.e. single-peaked) for all trial phases and pathways (e.g. Art. I Fig. 3 D). SPNs may therefore *not* be segregated into precisely-delineated, distinctly phase-tuned assemblies. This possibility meshed well with two further observations: (1) neurons of all pathways were highly heterogeneously-tuned to different counts and combinations of phases in a way that eluded functional clustering (Art. 1 Fig. S5); (2) although neurons in close spatial proximity tended to be more correlated than distant neurons, they did not form spatially-compact clusters of uniform tuning (Art. I Fig. 3 H-I & Fig. S7 E). Overall, the tuning scores for each phase appeared to be on a continuum shared across all SPNs, consistent with a stochastic process being the underlying organizational principle across pathways.

To summarize, the activity of all three SPN pathways covaried with trial phase in a qualitatively similar fashion, allowing us to decode the entire trial sequence. The underlying phase tunings were distributed in a graded, rather than discrete fashion, in all three SPN populations. Together, these findings indicate that the SPNs of all pathways are embedded in distributed, continuous, and overlapping ensembles, rather than in functionally or spatially-segregated clusters. Whether the tunings reflected processing of motor or cognitive variables, or of both is explored in the next sections.

A Value Correlate And Trial Phase Are Encoded In Conjunction

We noted that the activity of many neurons in variable subsets of phases depended on the subjective value of the trial, as roughly inferred from the previous trial's outcome and the animal's decision to "stay" or "switch" on the next trial. An example (Art. I Fig. 5 O-Q): one sSPN activated strongly during the contraversive return turn, in particular when the preceding ipsiversive choice had yielded a reward and was to be repeated on the current trial (i.e. win-stay trial); it activated much less when the choice had not resulted in a reward, yet was to remain the preferred option (lose-stay trial); and it activated barely at all on trials in which the unrewarded choice was to be abandoned (lose-switch). In **Fig. 3.4 A**, another example sSPN, recorded in the same session as the first, shows the opposite activity pattern in response to win-stay, lose-stay, and lose-switch trials.

Win-stay, lose-stay, and lose-switch trials may be loosely classified as high-value, medium-value, and low-value trials. Consequently, the described example neurons' activities corresponded, at least approximately, to trial value. We identified many similar "value-positive" and "value-negative" neurons, in all three SPN pathways, by computing a score designed to measure phase-specific, differential activity during high-value win-stay versus low-value lose-switch trials. Preference for win-stay or lose-switch trials was indicated by a positive or negative "selectivity score", respectively (**Fig. 3.4 C**). After identifying significantly selective neurons, we ascertained that these neurons also correlated with a more precise, continuous value estimate than "trial type" (i.e. win-stay, etc.), and in the manner expected: win-stay selective neurons correlated positively (**Fig. 3.4 D**), and lose-switch neurons negatively (Art. I Fig. 5 L-N) with logistic regression-based "action value" fits (as used by Tai et al.[121]).

The various phase-specific selectivities were in evidence in all pathways in comparable proportions (**Fig. 3.4 C**), as the phase tunings had been. The distributions of the selectivity scores were unimodal, like those of the tuning scores had been, and in density plots of the distributions, the curves of all three pathways overlapped closely (e.g. **Fig. 3.4 C**, Art. I Fig. S8 A). Thus, no pathway showed a sizable, distinctive bias for win-stay or lose-switch trials in any phase, as such a bias would have been reflected in pathway-unique shift of the corresponding selectivity distribution (e.g. a shift towards positive scores for a win-stay bias). In fact, the (overlapping) selectivity distributions

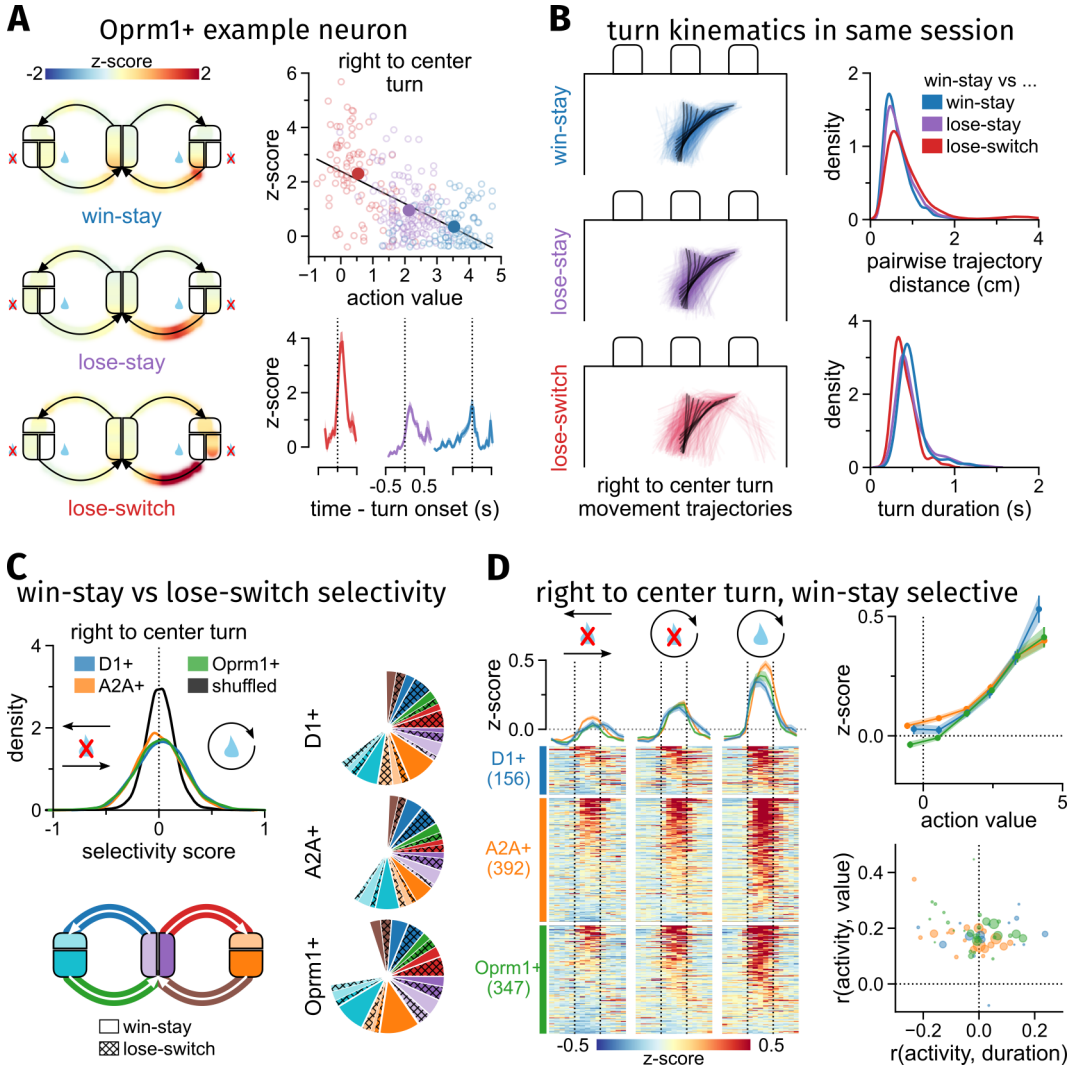

Figure 3.4 | Conjunctive phase-and-value signals in the DMS.
A Activity of an Oprm1+ example neuron in win-stay, lose-stay, and lose-switch trials. During right-to-center turns, the neuron's activity is negatively-correlated with the "action value".
B Left: Schematic top view of the operant box with lines depicting the mouse's right-to-center turn trajectories in win-stay, lose-stay, and lose-switch trials; same session as in **A**. Colored lines show the location of the spine of the mouse during 1/10th of the turn (single trial). Black lines depict the average "spine trajectory" in 10 steps. Right: Pairwise distances of single-trial trajectories (top) and turn durations (bottom) for turns shown left.
C Left, top: Selectivity scores for the right-to-center turn, by pathway. Positive and negative scores indicate preferential activation during win-stay and lose-switch trials, respectively. The curves overlap closely, indicating no distinct pathway biases. Moreover, the curves are centered on 0 and symmetrical, i.e. SPNs as a population show no clear win-stay/lose-switch bias. Right: The fraction of neurons signficiantly-selective during each of the trial phases, by pathway. The most significant selectivity determined the grouping. Color and shading: figure-8 key, left.
D Left: Raster plots of the average activity of significantly win-stay selective neurons during the right-to-center turn in lose-switch, lose-stay, and win-stay trials, by pathway. The turn activity was stretched/compressed to uniform length prior to averaging. Right, top: The average turn activity of the neurons plotted against action value, by pathway. Right, bottom: Per-session average correlation of the selective neurons' *activity X action value*, plotted against the per-session average correlation of the neurons' *activity X turn duration*. The radius of each circle represents the number of selective neurons in that session. Pathways color coded as in **C**.

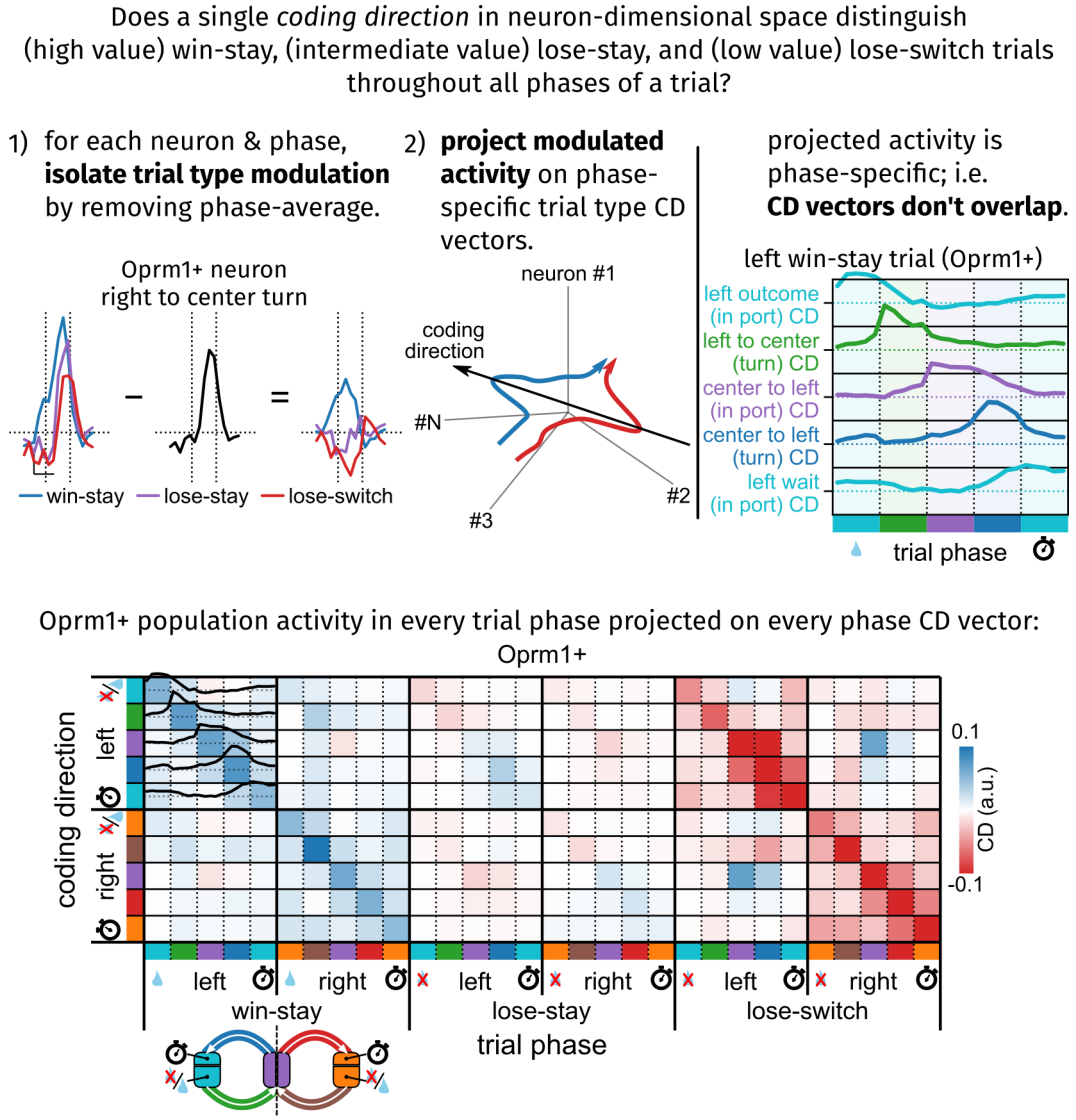

Figure 3.5 | Coding-direction analysis.
The trial type coding direction (CD) is the one-dimensional vector in the n-dimensional neuronal activity space that best separates the population activity trajectories of the different trial types. We used the win-stay vs lose-switch selectivity scores to approximate the CD vector. Thus, a projection onto a CD vector is synonymous to a selectivity score-weighted population average.

appeared approximately normal and centered on zero in all phases, indicating that win-stay vs lose-switch selectivities were very evenly distributed in the DMS in general (Art. I Fig. S8 A). In short, neither could we detect pathway-specific win-stay or lose-switch trial type preferences, nor global ones, in any of the trial phases.

The vast majority of individual neurons signaled trial type only intermittently, in varied counts and combinations of phases, and not throughout entire trials (Art. I Fig. 5 D). In support of the heterogeneity of the SPNs' selectivity profiles, we found that any selectivity-weighted population-average only reflected trial type (or value) in the specific phase for which the selectivity scores used as weights had been computed; that is, the neurons weighing into a particular "selective ensemble" only shared their selectivity for the single phase that defined the ensemble, but were inconsistent in their selectivity for any other phase (**Fig. 3.5**). Correspondingly, phase-specific SVM decoders, which successfully decoded trial type from neuronal activity in the phase they were trained on, failed to perform when applied to any other phase (Art. I Fig. S9 B).

Altogether, our analysis of the functional organization of SPNs in terms of trial type or value-coding reinforces the conclusion drawn from the preceding analysis of phase-coding: the SPNs captured in our recordings appear to exist in continuous and overlapping ensembles spanning the pathways, rather than in highly pathway-specific and functionally-discrete subpopulations. Interestingly, these cross-pathway ensembles seemed to encode value correlates dynamically and in conjunction with the current phase or ongoing action, rather than statically throughout the trial.

SPN Activity Is Not Only Sensorimotor-Related

There is an obvious potential caveat to the above analysis of trial type/value-coding: animals' movement kinematics widely depend on motivation, which is bound to covary with trial value. Therefore, our presumed phase-specifically value-correlated populations may in actuality signal sensorimotor-related information. To assuage this concern, we showed that the selective neurons were, on average, more correlated with value than action duration (and thus movement velocity), and further, that neuronal activity covaried with value even in cases in which the trajectory and the kinematics of the mouse's body appeared not to (**Fig. 3.4 A-B, D**). Consistently, phase-specific trial type

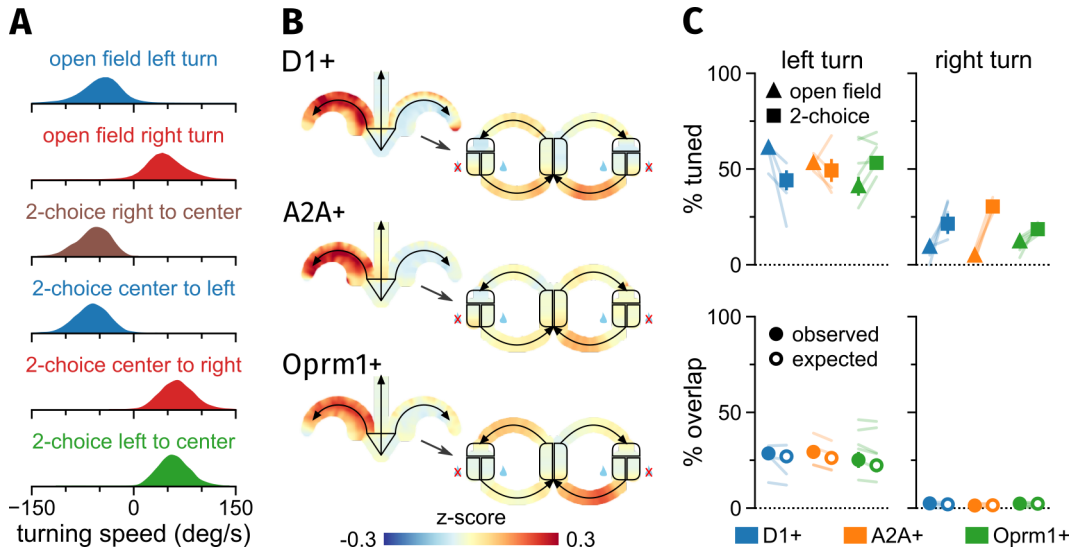

Figure 3.6 | Contra- and ipsiversive turn representation changes with task context.
A Turning speed (average per turn) in the open field and the reversal task (2-choice).
B The average activity of neurons tuned to open-field left turns (i.e. contraversive turns)—in the open field and the reversal task, by pathway. The arrows in the open field schematic represent turns, as they do in the figure-8 schematic of the reversal task.
C Top: Percentage of neurons tuned to contraversive/left and ipsiversive/right turns in the open field and the reversal task (2-choice). Bottom: The cross-task overlap in left and right turn tuning is chance level, i.e. only as many neurons are tuned to the same turn in both tasks as would be expected at random.

decoders, trained on SPN activity from any of the pathways, differentiated win-stay and lose-switch trials with high accuracy, whereas decoders trained on action duration performed no better than negative control decoders—at chance level (Art. I Fig. 6 F). On the basis of these results, we argued that we have indeed uncovered conjunctive phase-and-value signals in the activity of the SPN pathways, instead of having merely misinterpreted "plain" sensorimotor signals.

SPN Activity is Shaped by the Task Context

If the conjunctive phase-and-value signals do in fact reflect that policy or action value-related information is being processed—as we have argued—then the neuronal activity in the reversal task should be somewhat task or context-specific. Activity recorded in a context that does *not* involve reversals, outcome history-based decision-making, and rewards should be distinct from what we captured in the reversal task. To confirm such context-sensitivity in our data, we compared the activity of the same SPNs in the reversal task and the open field.

During the spontaneous, unguided and unreinforced exploration of the open field, neurons in all pathways, including the striosome, were positively modulated by movement in general, and by contraversive (i.e. left) turns in particular (**Fig. 3.6 B**, Art. I Fig. D-E, J-K). Approximately half of the neurons in each pathway were significantly positively tuned to contraversive open field turns. That is roughly the same fraction as were tuned to at least one of the contraversive turns in the reversal task (**Fig. 3.6 C, top**). This being the case, one might imagine that the neurons simply conserved a stable tuning between the two tasks, or contexts. However, the probability of a neuron being consistently-tuned across both contexts proved to be at chance level, in all three pathways (**Fig. 3.6 C, bottom**). In other words, the tunings registered in each context appeared to be independent, and thus context-specific.

In a more detailed follow-up analysis, in which we scored the kinematic similarity of individual actions (i.e. in terms of velocity, angular velocity, and body elongation), we came to the same conclusion. That is, we observed that within the same task, kinematic and neuronal similarity were associated: actions more similar in terms of their kinematics were also more similar in terms of the neuronal activity they evoked. In contrast, kinematic and neuronal activity were unrelated when comparing actions between tasks, attesting to the import of the context (Art. I Fig. 4 Q).

3.4 Conclusions

The similarity of the phase tunings and selectivity scores across pathways, and the trial-tiling sequential activity patterns stand in stark contrast with naive predictions of how the hypothesized divergent roles the three SPN types in decision-making may be reflected in neuronal activity (**Fig. 3.2**). Two key observations: firstly, the matrix actor pathways did not evidence opposing selectivities for contraversive high-value/win-stay choices (i.e. dSPNs > iSPNs). This is incongruous with Tai et al.'s[121] hypothesis that higher dSPN activity equates with a higher relative value of the contralateral choice, whereas higher iSPN activity equates with a lower relative value of the contralateral choice (**Fig. 3.1**). Secondly, the striosomal critic neurons likewise failed to display the predicted activity patterns: if they had encoded positive state value throughout the trial, we should have observed relatively persistent and

choice direction-independent striosomal selectivity for high-value/win-stay trials.

As far as we could tell, the DMS SPN activity was continuous in space and time, and even across pathways. That meant that neurons did not significantly cluster by tuning, pathway-identity, or location in the imaged tissue. Accordingly, the tuning profiles of neurons of all SPN types were suitably heterogeneous, with neurons being tuned to—or trial type selective in— diverse combinations of phases. Add to that the unimodal, and in most cases approximately normal distributions of the tuning and selectivity scores, and it seems likely that the SPNs' phase tuning and selectivity emerged from a significantly stochastic process that created rich, distributed, continuous and overlapping representations of task-relevant information. The nature of the striatal task representation we observed is hence vastly different from the temporally and spatially discrete one suggested by models emphasizing an organization of striatal neurons into competing action channels.

Chapter 4

Articles II & III: The GPi in Action

In **Article II**, we investigated the LHb-projecting GPi population, questioning the hypothesis that it serves as a prediction error-coding output of a BG critic circuit. Using fiber photometry, we recorded the population activity of the LHb-projecting "limbic" GPi (Sst+) in mice performing our maze or nosepoke-based reversal tasks, and compared it with the activity of the thalamus-projecting "motor" GPi (Pv+), i.e. the theoretical actor output. Bulk activity in the GPi-LHb pathway did not appear to encode prediction error-based choice feedback. Despite this, we found that silencing the pathway by selectively expressing TeTxLC in GPi-LHb neurons induced reversal-specific deficits in mice performing the maze task. Finding intriguing activity differences between the pathways as mice approached the decision point, we suggested that the GPi-LHb population may influence action selection through other means than prediction error signaling.

In **Article III**, we contrasted the effects of optogenetically activating the GPi inputs to the LHb with the effects of activating inputs arising from the GPi-adjacent lateral hypothalamus (LHA). These two LHb input populations are distinct, but hard to distinguish. Stimulating the GPi-LHb pathway in the real-time place avoidance paradigm or during the reinforcement phase of the nosepoke-based reversal task proved ineffective—in stark contrast to what we observed when targeting the LHA. In a final step, we showed that the LHb-projecting LHA activated in prediction error-like fashion in response to, and in anticipation of, aversive foot shocks.

4.1 Background

GABA/glutamate Co-release and Population Marker Genes in the GPi

Hong and Hikosaka[74] speculated that the GPi-LHb projection—ostensibly the source of the LHb prediction error signal—was likely excitatory, releasing either acetylcholine or glutamate. A 2012 publication by Shabel and collaborators[94] proved the latter guess correct; GPi input to the LHb is indeed glutamatergic. Perplexingly, prior work on the GPi-LHb connection, going back to the 1970s and '80s, had indicated that GPi-LHb neurons are GABAergic and therefore inhibitory[150,151]. There was also evidence that the neurons expressed the neuromodulatory peptide somatostatin (Sst)[90], which—like GABA—generally suppresses the activity of its recipient neurons[152]. Notably, Sst was found absent in the thalamus-projecting GPi population, which instead expressed the calcium-binding protein parvalbumin (Pv) selectively[89]. In a seminal 2014 publication, Shabel et al.[153] followed-up on their previous publication to resolve the "glutamatergic or GABAergic" puzzle. The team showed that GPi input to LHb is in fact glutamatergic *and* GABAergic: GPi terminals co-transmitted both neurotransmitters in the LHb, albeit to apparently net-excitatory effect.

Recent studies[77,88], including our own work[3], have confirmed the highly selective expression of Sst and Pv in the GPi-LHb and GPi-thalamus projection pathways, further emphasizing the "parallel" nature of these outputs. Taking advantage of modern RNA sequencing and molecular profiling techniques, these studies also showed that Sst-positive (Sst+) neurons co-express glutamatergic (e.g. Vglut2) and GABAergic (e.g. Vgat) markers[88], whereas the *vast majority* of Pv+ neurons only contain transcripts indicative of GABA transmission[88]. This is robust evidence that LHb-targeting Sst+ neurons co-transmit GABA and glutamate, as opposed to thalamus-targeting Pv+ neurons, which only release GABA. Moreover, Shabel et al.'s[153] functional demonstration of mixed GABA and glutamate transients in the LHb in ex vivo electrophysiological recordings replicated when exclusively Sst+, rather than all glutamatergic (i.e. Vglut2+) GPi neurons were genetically targeted[3,88,154]. Remarkably, recent evidence suggests that GABA and glutamate are not only released from the same synaptic terminals, but that they are even co-packaged into the same vesicles[154].

It should be noted that Wallace et al.[88] identified an exception to the "Sst+

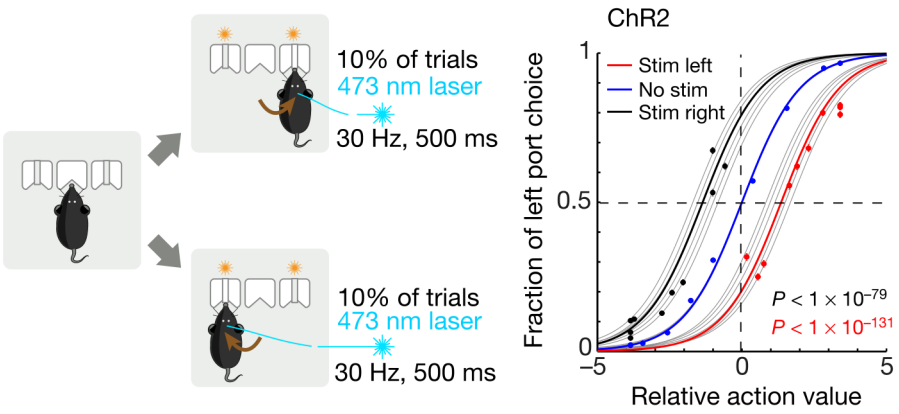

Figure 4.1 | GPi-LHb activity during outcome evaluation biases subsequent choice.
Left: Mice received bilateral optogenetic stimulation of the glutamatergic (Vglut2+) GPi-LHb population during the reinforcement phase of the probabilistic reversal task.
Right: Activating the pathway while mice nosepoked the left port lowered the probability of them returning on the next trial (red line). The same manipulation increased the likelihood of a left choice if delivered during the right-paw side nosepoke (black line).
Reprinted from Stephenson-Jones et al.[77], ©2016, with permission from Springer Nature.

equals LHb-projecting, Pv+ equals thalamus-projecting" rule. They found that a small fraction (<5 % of Pv+ neurons *did* project to the LHb, and—critically–that this population was exclusively glutamatergic, both in its molecular marker profile, as well as in terms of the responses evoked in the LHb in ex vivo slice preparations.

Behavioral Effects of GPi-LHb Activity

In their 2012 study, Shabel et al.[94] reported that activating the terminals of glutamatergic (Vglut2+) GPi neurons in the LHb was highly aversive: rats subjected to an optogenetic real-time place avoidance test shunned the stimulated half of the test environment. A 2016 study by Marcus Stephenson-Jones and colleagues[77] replicated this aversive effect in mice, then went on to show that the inhibition of the projection induced the opposite behavioral effect, i.e. place preference, and more. Most strikingly, Stephenson-Jones et al. presented evidence that optogenetic excitation, as well as inhibition, of the Vglut2+ GPi-LHb pathway affected outcome evaluation and subsequent decision-making in the nosepoke-based probabilistic reversal task.

Stephenson-Jones at al. used the same task design and logistic regression analysis as employed by Tai et al.[121] in their work on action selection in the striatum (see previous chapter). Importantly, here the investigators opted to stimulate bilaterally and time-locked to the presentation of the trial outcome

(reward or no-reward): 500 ms-long pulses of light stimulation were triggered randomly upon choice port entry on 10 % of the trials (**Fig. 4.1, left**).

Stephenson-Jones et al.[77] reported that excitation of the Vglut2+ GPi-LHb projection during the reinforcement phase of the task lowered the probability of mice repeating their port choice on the next trial considerably (**Fig. 4.1, right**). Inhibition of the pathway accomplished the opposite, enticing mice to return to the port subsequently. This data indicated—the study authors reasoned—that the induced increases and decreases in GPi-LHb activity translated into increases and decreases in the value of the stimulated action. This interpretation is in line with the notion that the GPi-LHb population drives prediction error-based reinforcement learning (be it of "action value" estimates or of "policy values"). Stephenson-Jones et al. supplied further evidence of this by showing prediction error-like responses in putative GPi-LHb neurons (all neurons that responded like opto-tagged neurons) to primary and secondary reinforcers in a classical conditioning paradigm involving tone-cued water rewards and aversive airpuffs. As expected, the recorded neurons were excited by airpuffs and airpuff predicting-cues, and inhibited by water and water-predicting cues.

4.2 Aims and Expectations

Our principal aim in the studies reported here was to reproduce and to extend the findings reported by Stephenson-Jones and colleagues[76] whilst exclusively targeting Sst+, rather than all Vglut2+ LHb-projecting neurons. We deemed this important, because the GPi-adjacent LHA is known to send a major and largely glutamatergic (Vglut2+) projection to the LHb[79,155]. Moreover, in rodents, the GPi-LHb and LHA-LHb projections are notoriously difficult to separate anatomically—e.g. prompting Parent, Gravel, and Boucher[156] to declare that "there was no distinct boundary" between the populations in 1981. Most important, however, is that many studies had strongly implicated the LHA[70,157–161], and even the LHA-LHb pathway[162] in aversion, reward-processing and reinforcement. In particular, our aims were therefore the following:

- Record and contrast the activity of the Sst+ (i.e. GPi-LHb) and the Pv+ (i.e. GPi-thalamus) GPi output populations during spontaneous loco-

motion to rule out movement-related activity as a confound when interpreting the activity of Sst+ neurons.

- In the nosepoke-based reversal task, confirm that Sst+ GPi neurons signal prediction errors as mice flexibly adapt to covert and unpredictable but—due to abundant task experience—somewhat expected serial reversals.
- In the maze-based reversal task, confirm that Sst+ GPi neurons signal prediction errors as mice *learn to* adapt to the very first action-outcome contingency reversal they experience.
- Replicate the effects of GPi-LHb pathway activation in the real-time place avoidance and nosepoke-based probabilistic reversal tasks, ensuring that only Sst+ projection neurons are targeted.
- Compare the effects of GPi-LHb (Sst+) activation with those of LHA-LHb (Vglut2+) activation.

4.3 Results

— Fig. 4.2 to 4.4 adapted from Article II; Fig. 4.5 & 4.6 adapted from Article III —

GPi-Targeting Experiments

GPi Activity in Both Pathways Is Correlated with Movement

In the open field, the population-level activity of Sst+ GPi neurons roughly reflected the kinematics of the animals' movements, as did the activity of Pv+ GPi neurons. The activity of both GPi populations grew approximately logarithmically with movement velocity (cm/s), but differed in terms of the increases observed in relation to angular velocity (°/s): while the Pv+ neurons had a marked preference for contraversive over ipsiversive turning (relative to the recording implant), the Sst+ population had no directional preference, and was only weakly modulated by turns in general (Art. II Fig. 5 a-b). Hence gross locomotion, and especially transitions from relative immobility to movement, appeared to positively impact overall GPi activity irrespective of the pathway, whilst contraversive turning selectively enhanced the activity of the Pv+ population.

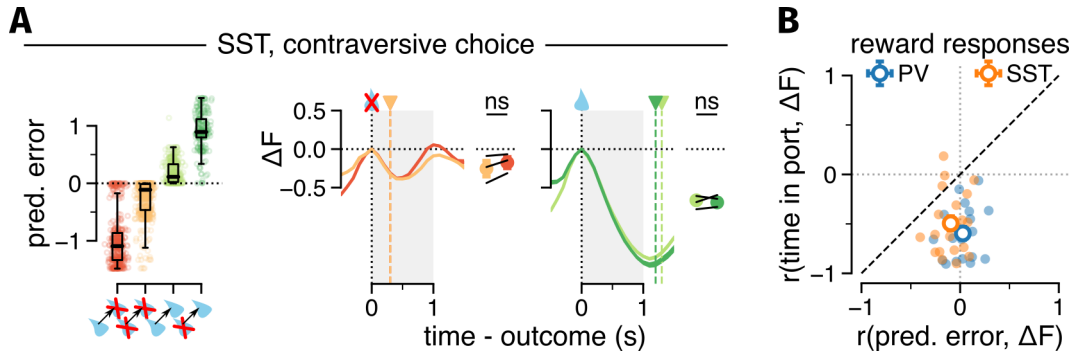

Figure 4.2 | No prediction error evident in the population activity of Sst+ GPi-LHb neurons.
A Left: Model-estimated reward prediction errors for reward and no-reward trials in the probabilistic reversal task depend on the outcome of the previous trial; e.g. a no-reward outcome is expected to elicit a larger negative prediction error if the previous trial was rewarded than if it was unrewarded. Prediction error estimates are derived from trial-to-trial changes in the action value fits. Right: Average population activity traces in response to no-reward and reward outcomes at the contralateral choice port, split based on the outcome of the previous trial. Color-coding follows the box plots on the left. Gray shading indicates the time window used to compute the means on the right of the traces. Triangles and dashed lines mark the median port exit times.
B For both Sst+ and Pv+ GPi populations, the average signal drop recorded following reward delivery and while animals occupy the reward port is more strongly associated with the time the mice spent nosepoking than with the model-estimated reward prediction error. The filled circles indicate the correlation coefficients for individual recording sessions, the open circles the group average.

No Prediction Error Signals in Response to Reward-Location Reversals

In the nosepoke-based probabilistic reversal task, there was no "reward-negative" prediction error-like modulation evident in the population activity of either the Sst+ or the Pv+ GPi neurons. Although the Sst+ neurons' activity began to fall upon entry into the rewarded nosepoke port, and kept falling for the duration of the reward consumption, it did so independently of the magnitude of the estimated prediction error. That is, the reward-phase Sst+ activity decreased the same way, whether the obtained reward was rendered more expected by a reward, or more unexpected by a non-reward outcome on the previous trial (**Fig. 4.2 A**). Moreover, the Pv+ GPi's activity dropped similarly during reward consumption (Art. II Fig. S7 f), and the average signal decreases in both pathways were more robustly associated with the amount of time spent consuming the rewards than with the prediction errors they evoked (**Fig. 4.2 B**). Taking into account that both GPi populations are positively correlated with movement, we surmised that the decreases are a consequence of the mice slowing down in order to consume the reward, rather than an evaluative signal. In line with this, the GPi activity in either path-

way decreased much less, or not at all, after unrewarded port entries, which did not feature a "reward consumption-length" interruption of movement. Critically, the two populations' "omission responses" failed to distinguish high and low prediction error trials, just as their "reward responses" had (**Fig. 4.2 A**). Therefore, there was no indication of prediction error coding by GPi neurons in the probabilistic reversal task.

Examination of the outcome-phase population responses after the swap of the reward location in the maze-based reversal task similarly failed to reveal the expected prediction error-like evaluation signals. Indicative of their significant reward expectations, the mice approached the unrewarded spout dozens of times, and yet we did not capture strong increases in the activity of the Sst+ GPi population when the animals found their expectations unmet. In fact, on average, the Sst+ GPi activity decreased in response to the post-reversal reward omissions (Art. II Fig. 4 b-c), and appeared to do so more markedly than it did in response to the initial rewards obtained at the new location (Art. II Fig. 4 e-f).

Analysis of the population activity in both behavioral tasks—the nosepoke-based and the maze-based reversal tasks—therefore contradicts the hypothesis that the Sst+ LHb-projecting GPi population's net output consists in a reward-negative prediction error signal.

GPi-LHb Excitation is Not Choice-Devaluating or Aversive

In another set of experiments, we manipulated the activity of the LHb-projecting GPi population rather than recording it. In the nosepoke-based reversal task, bilateral optogenetic activation of the GPi-LHb pathway during the outcome evaluation phase did not affect the mice's subsequent choices, and thus failed to interfere with the evaluation process (**Fig. 4.5 D**). Optogenetic excitation of the pathway similarly failed to induce avoidance of the stimulated compartment in the real-time place avoidance test; i.e. mice occupied the half of the test environment in which they received optogenetic stimulation about as much as the half in which they did not (**Fig. 4.5 E**). Thus, the activation of the GPi-LHb pathway did not depreciate the value of the stimulated compartment, nor was it perceived as aversive in and of itself. These manipulation experiments bolster the conclusion that the Sst+ LHb-projecting GPi neurons do not transmit intrinsically aversive, reward-negative evaluation signals—at least not uniformly on the population level.

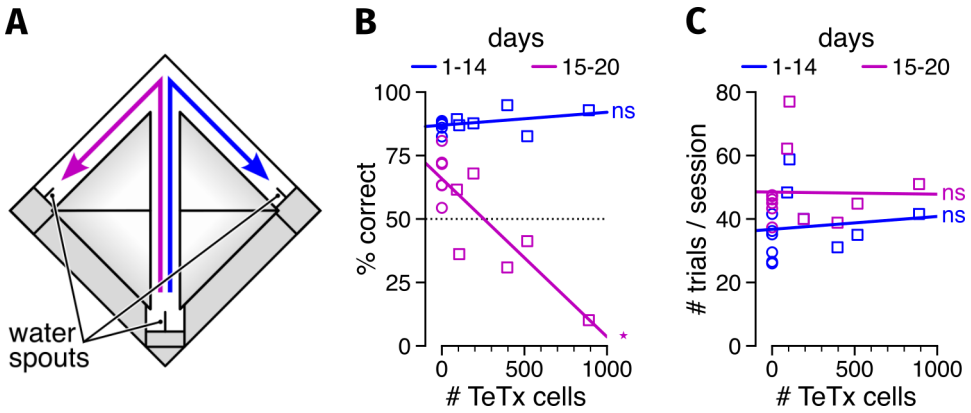

Figure 4.3 | Chronic inactivation of Sst+ GPi-LHb neurons impairs reversal performance.
A The maze-based reversal task. For 14 days, mice were trained to run down the center corridor, turn right, and approach the spout at the corridor's end to obtain a reward. On the 15th day, the reward location "reversed" from the right to the left-paw side maze corridor.
B The more Sst+ GPi neurons expressed TeTxLC, the worse the average performance on the five days following the reversal. Pre-reversal performance was unaffected by the treatment.
C TeTxLC expression did not predict the average number of trials a mouse performed per session pre- or post-reversal.

Chronic Inactivation of GPi-LHb Neurons Impairs Reversal Performance

Interestingly, TeTxLC-mediated chronic inactivation of the Sst+ LHb-projecting GPi population selectively impaired the ability of mice to adapt to the reward location reversal in the maze-based reversal task. The number of Sst+ GPi neurons found to express TeTxLC, and hence the extent of the inactivation, strongly predicted the overall error rate in the six days following the reversal: whereas mice with only a few or no TeTxLC-transfected cells erroneously returned to the formerly rewarded arm in around 30 % of the trials they performed, the mouse with the most widespread TeTxLC expression went wrong in about 90 % of its trials (**Fig. 4.3 B**). Importantly, the number of TeTxLC-positive cells did not predict the choice accuracy prior to the reversal, nor the average number of trials completed per session pre- or post-reversal (**Fig. 4.3 C**). Therefore, it is unlikely that TeTxLC expression affected learning generally, nor is it probable that it blunted motivation or locomotion, or reduced task-engagement or learning opportunities.

GPi-LHb Neurons Are Distinctly Modulated During the Choice Phase

If not via prediction error-driven reinforcement learning, motivation or locomotion, how might the Sst+ GPi population facilitate accurate behavior post-reversal in the maze task? Is it possible that the chronic inactivation

of Sst+ GPi neurons impaired *action selection* rather than *action evaluation*? In other words, are the TeTxLC-expressing mice able to learn from their mistakes, but unable to flexibly deviate from the previously (overtrained) behavioral pattern?

The evolution of the choice turn-related population activity of intact Sst+ GPi neurons over the course of the maze reversal experiment is suggestive of this hypothesis. In the first couple of training days, the activity at the onset of the choice turn was usually above the session mean activity—as expected, considering that the animal is running into the turn, and movement is positively correlated with GPi activation. However, by training day seven, the Sst+ activity recorded as an animal entered the turn was typically *below* the session mean early-on in the session (i.e. suppressed), only to recover to its former above-mean level (i.e. excited) as the session progressed (**Fig 4.4 A**). This pattern—"early-session suppression to late-session excitation"—is unexpected, especially considering that the reward-thirsty animals moved quickest early on, and slowed towards the later parts of the session, when their thirst was quenched. The within-session development of the Sst+ GPi activity at turn onset is thus roughly *anti-correlated* with movement velocity—unlike what we observed in the open field (**Fig. 4.4 C**).

Remarkably the "early-session suppression to late-session excitation" activity pattern disappeared after the reversal, just as the mice adopted the alternative turn response, which now yielded rewards. That is, after the successful reversal, the activity at the choice turn onset was yet again above the mean for the entire session—as it was in the first few days of training—and so it remained for all subsequent sessions recorded (**Fig 4.4 A-B**).

Importantly, the Sst+ activity modulations were (1) specific to the choice turn phase and (2) the Sst+ population: (1), that the modulations did not reflect general changes in "baseline activity" (i.e. affecting the entire trial) was evident from turn onset-aligned, average activity traces of the maze runs: depending on the trial type and the session's progress, the averages fell or ramped up toward the turn, with their highest peak or lowest valley shortly after the turn onset (**Fig. 4.4 C**). (2), the modulations were specific to Sst+ GPi neurons, as Pv+ neurons did not show any modulation in the choice phase, neither across nor within-sessions (**Fig. 4.4 B**). Curiously, the Pv+ neurons' activity even appeared insensitive to turn direction: the choice turn onset-aligned, average activity traces of contra- and ipsiversive choice maze runs, though

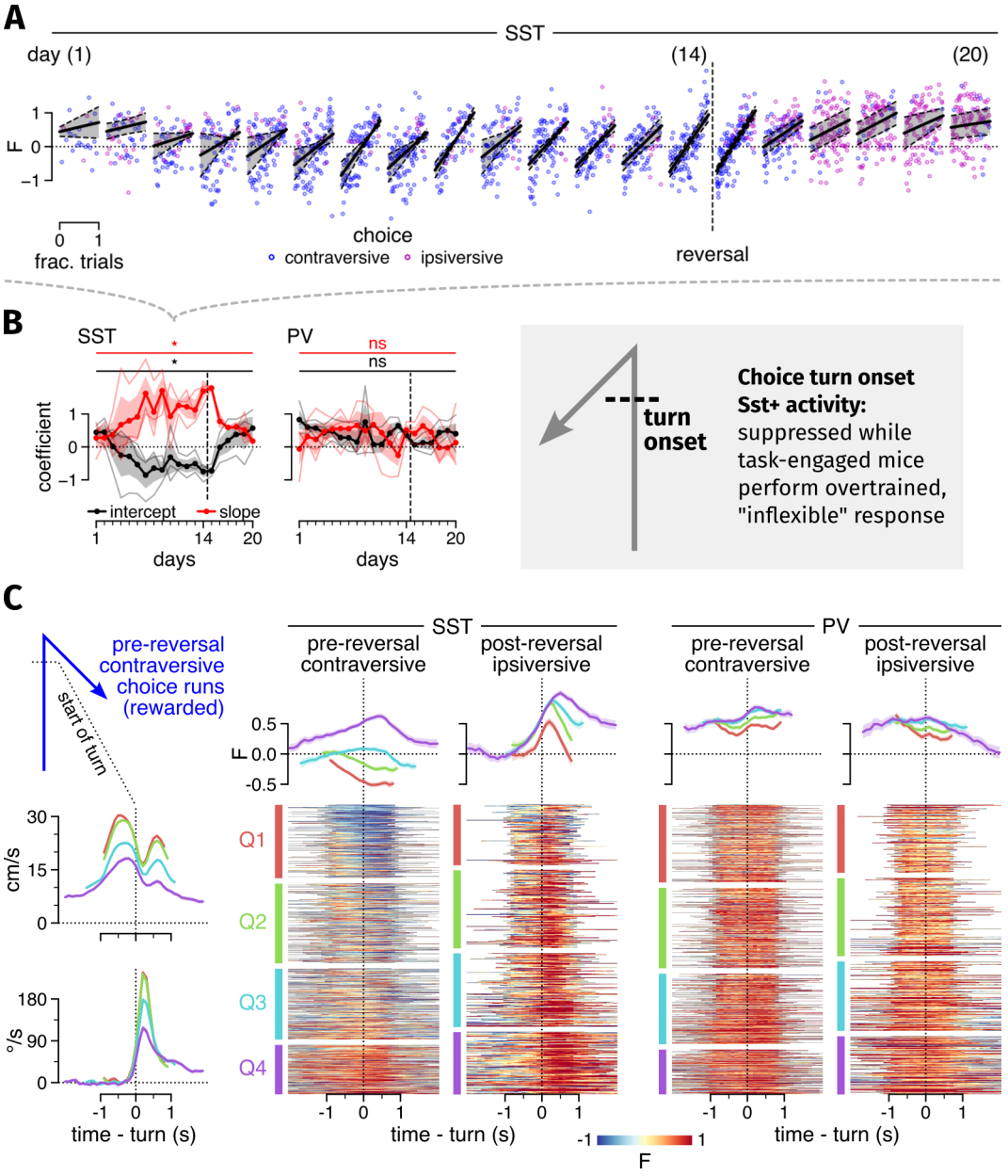

Figure 4.4 | The Sst+ GPi-LHb population is distinctly modulated during the choice phase.
A The evolution of the Sst+ GPi's population activity recorded at the onset of the choice turn (see gray box). Each "session subplot" (days 1–20) presents the turn-onset activity plotted against the fraction of trials completed of the session total. The dots represent single contraversive (blue) and ipsiversive (magenta) trials, pooled over animals. The black lines are the mean of per-animal regression fits.
B Mean coefficients for regressions fitted—per animal and day—to the turn-onset activity data (as shown for Sst+ mice in **A**), presented by pathway. The mean intercept (black line) approximates the turn-onset activity at the start of a session; the mean slope (red) approximates the change in turn-onset activity between the start and end of a session.
C Left: Average velocity (cm/s) and angular velocity (°/s) of animals entering the contraversive choice turn (time 0), pre-reversal. Trials were averaged by session quartile; i.e. Q1 to Q4 are the 1st to 4th 1/4 of trials of the session total (see color coding of the raster plot). Right: Raster plots of the population activity recorded during individual maze runs, aligned to the onset of the choice turn (time 0). In each raster, trials were grouped by session quartile, and sorted from earliest to latest. Per-quartile average traces are shown above the rasters. Only correct maze runs were included in each raster (i.e. pre-reversal = contraversive, post-reversal = ipsiversive).

elevated above the session mean, proved similarly flat throughout the run, in stark contrast to the contra-preferring responses captured in the open field (**Fig. 4.4 C**).

One may speculate widely as to the meaning of these modulations; we observe, that the GPi activity is suppressed when the mice behave at their most habitual (after a week of training), and high when the behavior is more flexible and exploratory (early in training and during/after reversal), or even when the mice are disengaged from the task (towards the end of a session). Perhaps the GPi-LHb pathway's function in reversal tasks is indeed to enable the selection of alternative actions or behavioral strategies over the most prepotent responses.

LHA-Targeting Experiments

LHb-projecting and mainly glutamatergic LHA neurons border the GABA and glutamate co-transmitting Sst+ GPi[155,162]. Without visualizing the Sst expression, or the overlap of GABAergic and glutamatergic markers, the populations are difficult to segregate anatomically (**Fig. 4.5 A-C**)[156]. Critically, excitatory transmission at LHA-LHb axon terminals has been implicated in acute[162] and learned behavioral avoidance[163]. To investigate whether the evaluative and aversive functions previously attributed to the GPi-LHb projection may instead arise from the neighboring LHA-LHb projection, we conducted optogenetic and calcium imaging experiments targeting the latter pathway. These experiments finally yielded the results we had initially hypothesized to obtain in our GPi-LHb studies.

LHA-LHb Excitation is Choice-Devaluating and Aversive

In the nosepoke-based reversal task, optogenetic activation of the LHA-LHb projection significantly decreased the likelihood of the mice repeating their choice on the next trial, indicating that the choice had been devalued by the manipulation (**Fig. 4.5 D**). Moreover, we found LHA-LHb stimulation to be extremely potent and aversive in the real-time place avoidance assay: the experimental mice spent only a small fraction of time in the stimulated chamber, usually exiting the chamber soon after the onset of the stimulation (**Fig. 4.5 E**). This matched prior data from the laboratory of Garret Stuber[162], who had reported the same in 2016. Both of these strong and negative-

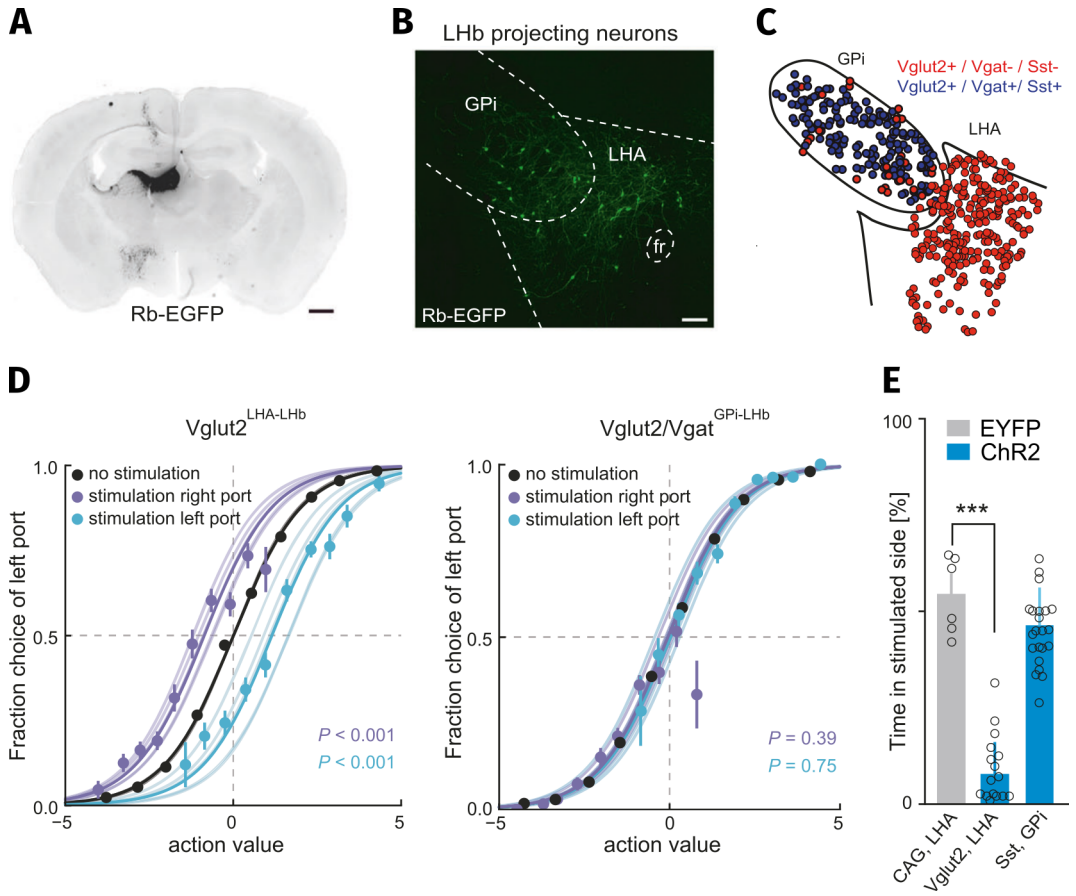

Figure 4.5 | LHA-LHb stimulation is devaluating and aversive, GPi-LHb stimulation is not.
A Coronal brain section of a mouse that had a retrograde rabies virus (Rb-EGFP) injected into the LHb for the purpose of identifying glutamatergic (Vglut2+) inputs to the area.
B Vglut2+ neurons that project to the LHb in the GPi and the LHA.
C Vglut2+ neurons in the GPi also express the GABAergic marker Vgat and Sst (blue). Vglut2+ neurons in the LHA are Vgat- and Sst- (red).
D Left: Activating the LHA-LHb pathway (Vglut2+) while mice nosepoked the left or right choice port lowered the probability of the mice returning on the next trial. Right: The same manipulation of the GPi-LHb pathway (Vglut2+/Vgat+) had no such effect.
E In the real-time place avoidance test, mice avoided the chamber paired with Vglut2+ LHA-LHb axon terminal stimulation. The stimulation of Sst+ GPi-LHb terminals had no effect.

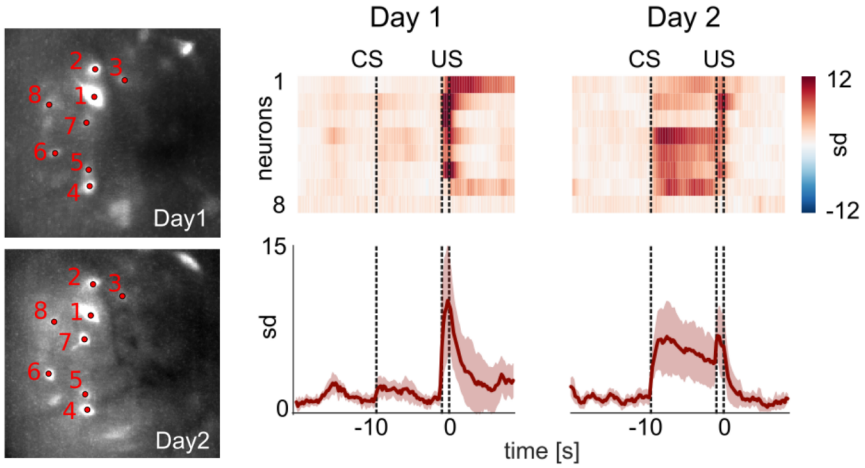

Figure 4.6 | LHA-LHb neurons show prediction error-like responses.
Left: Z-projected video stills (miniscope recording) show 8 Vglut2+ LHA-LHb projection neurons which we followed across two days of fear conditioning. Right: The raster shows the average responses of the 8 neurons to the shock-predicting cue (conditioned stimulus, CS) and the foot shock (unconditioned stimulus, US), on conditioning days 1 and 2. Per-day population averages of all 8 are traced-out below.

valued effects contrast with the null results observed in the same experiments when targeting the GPi-LHb pathway.

LHA-LHb Neurons Activate in Prediction Error-Like Fashion

Using miniscopes, we also showed that a subset of LHb-projecting excitatory LHA neurons signal negative prediction error-like signals in an auditory fear conditioning paradigm. Over the course of repeated pairings of an auditory tone cue with a mild electrical foot shock, in which the onset of the tone consistently precedes that of the shock, these neurons acquired a strong response to the predictive tone whilst concomitantly losing much of their sensitivity to the shock (**Fig. 4.6**). Unlike the Sst+ GPi-LHb population, LHA-LHb neurons therefore appear to be modulated by outcome expectations in the manner expected of a negative prediction error-signaling population.

4.4 Conclusions

The data presented here suggests that the LHb-projecting GPi does *not* drive the prediction error signals of the LHb—at least not the major Sst+ GABA and glutamate co-transmitting population. The unexpected delivery or omission of rewards did not evoke prediction error signals in the population activity

of Sst+ GPi neurons, neither in the context of the overtrained serial reversal nosepoke task—in which the expert subjects "expected" the randomized omissions and reversals to some degree—nor in response to the singular reversal in the maze task—for which subjects were wholly unprepared. Furthermore, we found outcome-concomitant optogenetic manipulations of the Sst+ population activity in the probabilistic nosepoke reversal task to be ineffective, hence failing to replicate Stephenson-Jones et al.'s[77] results (**Fig. 4.1**) in our more accurately-targeted experiment (Sst+ only, rather than all Vglut2+). Optogenetic activation of the Sst+ GPi-LHb pathway also failed to induce real-time place avoidance, suggesting it was not experienced as aversive by the mice. However, the same experiments yielded positive results when we targeted the Vglut2+ LHA-LHb pathway. Neurons of the latter pathway also encoded a prediction error-like signal in response to the shock-predicting tone cues and shocks of a fear conditioning paradigm. We concluded that the Vglut2+ LHA-LHb projection—not the adjacent Vglut2+ and Sst+, GABA and glutamate co-transmitting GPi-LHb projection—gives rise to the negative reinforcement-related effects and activity observed in the region. We cannot rule out, however, that the minor Pv+, and exclusively glutamatergic GPi-LHb population Wallace and colleagues[88] observed (see Background section) serves a similar role to that of the Vglut2+ LHA-LHb projection.

Despite their apparent non-involvement in outcome evaluation, we did find that chronically silencing the Sst+ GPi population impaired the ability of overtrained mice to consistently shift their efforts to the alternative choice, post-reversal in the arrow maze, whereas it did not affect the pre-reversal acquisition or performance of the task. When comparing pre- and post-reversal sessions, we observed differential population activity as mice approached the decision point in the first half of a session: in the days prior to the reversal, the activity of Sst+ GPi neurons in mice persistently and rapidly executing well-trained contraversive choice runs appeared suppressed. In contrast, post-reversal, the activity was elevated while the mice prepared and executed ipsiversive choice turns. Turn direction alone could explain the disparity of activity observed pre- versus post-reversal, as there was no such difference in Sst+ GPi neuronal activity between ipsiversive and contraversive turns in the open field. These data led us to suspect that the Sst+ GPi-LHb pathway may be involved in the action selection process, especially if a prepotent re-

sponse needs to be suppressed (perhaps by mediating between competing goal-directed and habitual actor circuits).

A significant caveat to the latter speculations must be noted: the use of unilaterally implanted photometry fibers and a single maze training schedule (first contraversive, then ipsiversive) precluded us from fully assessing whether the maze choice-related Sst+ GPi activity was—unlike in the open field—lateralized, or became so with training. Follow-up experiments are thus needed to support a role for the Sst+ GPi-LHb population in reversal *execution*, rather than *evaluation*, i.e. a role in the actor, rather than the critic.

Chapter 5

Conclusion and Outlook

In this thesis, I have explored data that captured major, genetically-identified BG pathways at work at both ends of the BG network—i.e. at the input end, in the striatum, and the output end, in the GPi. The major aim, alluded to in the thesis title, was to show a BG critic circuit *in action*, first evaluating (in DMS, Art. I), and then critiquing (in GPi, Art. II & III) the consequences of actions selected by a complementary and largely parallel actor circuit. The hypothetical critic pathway of interest here originates in the dorsal striatal striosome (Oprm1+), and projects via a distinct GPi population (Sst+) to the LHb, a region well-known for its error signals and involvement in (negative) reinforcement. With regards to the ambitious aim of showing this particular critic "critiquing action", the data presented here was negative, both in the DMS and the GPi.

BG-mediated policy-driven action selection—rightfully *en vogue* since the discovery of the DA prediction error—is typically conceptualized as the BG gating between relatively discrete *units* of behavior, to be found relatively high-up in the behavioral control hierarchy—the chocolate ice cream-level—and thus proceeding a relatively glacial time scale, if compared to "lower-level" processes such as the regulation of movement kinematics. This is not only the simplest and most natural way to think about "actions", it is also the level of analysis behavioral neuroscience inherited from behaviorist theory and conditioning methodology, in which actions are associated with outcomes, and stimuli with other stimuli, and stimuli with responses, and so forth. Indeed, the DA prediction error itself is commonly understood to reinforce associations at this particular level. This conception of actions is therefore most natural to the behavioral neuroscientist's mind. Not coinci-

dentally, it is also most experimentally tractable: discrete actions are easily observable in well-established behavioral paradigms; discrete decision processes are straightforward to analyze and reason about; and—perhaps most importantly—neuroscience has the tools to record, identify, and manipulate discrete ensembles, e.g. by means of optogenetics or similar techniques. That is to say, the full might of the modern circuit neuroscience toolkit can be brought to bear to "dissect" a system—if what is to be dissected is, conveniently, discrete.

In Article I, we reported a just so pitched study of the DMS in decision-making. Imaging the two matrix pathways, we expected to see discrete SPN ensembles ("something that clusters") representing the two available port choices in the run up to the decision—i.e. the actor in action; imaging the striosome, we expected state value representations—i.e. the critic in action. We found the SPN activity to carry on at much faster clip than anticipated, and lacking clear breaks between the events deemed important at the "action-level". The signal was much more complex within each pathway, but also more uniform across them than anticipated. The principal thing we could tell was: SPNs represent the "where" and "why" of the task in cross-pathway ensembles that were spatially and temporally continuous more or less at the spatiotemporal resolution we could technically resolve.

Our article is but one in a list of articles similarly showing just how good the striatum is at representing the evolving behavioral context mice find themselves in; continuous activity in the striatum keeps track of animals' movements[164,95,98,165], of their location[166–171], and of time[172–174], moment-by-moment, if the task requires it. Ironically, what tripped us up when confronted with the experimental results is precisely what made the BG attractive for RL-flavored models in the first place—their capacity to distill their enormous convergent input into state representations which could then be mapped to action-policies and value estimates. It tripped us up chiefly, I would argue, because our expectations of how states, actions, and values would be represented in the DMS were off.

The caveats to a single study like ours are—of course—countless: perhaps, we did not "hit" any DMS action channels (what are, after all, the chances?). Or we did not target the right BG loop; the fairly repetitive behavior could be habitual, and driven by stimulus-response associations localized in the lateral part of the striatum[133]. Perhaps the use of a task and analysis strategy

little suited to reveal differences between policy and value representations obscured differences between the striosome and matrix[134]. Perhaps we were led astray by possibly coincidental movement-related activity that can be picked-up all over the brain[175,176]…

However, in the light of our data, as well as data presented in the studies cited above, I wish to suggest that—perhaps we are too focused on finding *discrete* or *explicit* action policy and state value representations, and to neatly parcel them out to "parallel" BG pathways. Let me explore some of the issues with the discrete action selection-focused actor-critic model I have outlined in this thesis.

It seems clear that there is substantial anatomical and likely functional overlap between the BG pathways that complicate the picture I have drawn over the past chapters to support the actor-critic BG model. Stephenson-Jones et al.[77] found barely more than half of the input to the LHb-projecting GPi to arise from the striosome, around 55 %, whereas in the study by Wallace and colleagues[88], only about 25 % of the striatal input neurons to the GPi-LHb pathway were located in the striosome. The latter was nevertheless described as a "patch-biased" input, by comparison to a tracing that showed the striosomal input to a GPi-thalamus pathway being close to zero. Similarly, Jared Smith and collaborators[177] found the percentage of striosomal neurons among the SNc inputs in the striatum to number below even a quarter. Moreover, while many or most striosomal neurons may project to the SNc, the majority of them appear to collateralize and to simultaneously project to targets primarily linked with the matrix pathways, such as the SNr and the GPe[56,58,149]. The SNr and GPe themselves project to and inhibit the SNc[32,60,178], which has led to the suggestion that the *matrix* pathways may drive the computation of the TD prediction errors within the DA midbrain[179]. Some dSPNs—i.e. neurons of the supposed facilitatory "Go" pathway—project like neighboring iSPN "No-Go" neurons, and, like the latter, suppress movement[180,181]. It has also been reported that selective optogenetic manipulation of either dSPNs or iSPNs evoked excitation and inhibition in about equal fractions of the responsive SNr BG output units[182]. Inhibition of the two striatal pathways similarly failed to induce the expected, opposing effects in the SNr[183]. Clearly, the much-touted parallel pathways are not *quite* that parallel.

Another important issue for our DMS striosome-matrix actor-critic model, with the DMS-innervating SNc providing action selection-relevant update

signals, arises from the fact that the DA input from the SNc frequently appears more related to movement than reward processing and prediction error signaling[184,185]. Critically, optogenetic stimulation of SNc neurons failed to confer incentive value onto a neutral, stimulation-paired cue, as would be expected if SNc signals shaped the value-expectations of a striosome-critic; that is, rather than serving as a secondary reinforcer in an operant task, the SNc excitation-paired cue evoked vigorous rotational movements[186].

The point is, both the anatomical and functional assumptions of the model, at least with regards to *discrete action selection*, are arguably very strained. Perhaps greater attention should be paid to the continuous and pathway-crossing patterns we—and others before us—have observed. Yoo, Hayden, and Pearson[96] formulated it perfectly in a 2021 opinion piece titled "Continuous decisions":

> Traditional 'box-and-line' approaches to cognitive neuroscience presume the existence of discrete cognitive functions with intuitive easy-to-name roles. These functions are assumed to be reified in neuroanatomy. For example, if choice consists of evaluation, comparison and selection, then these three conceptually discrete functions ought to correspond to discrete anatomical substrates. An alternative viewpoint is distributed; it imagines that choice reflects an emergent process arising from multiple brain regions whose functions may not correspond to nameable processes, and/or that may largely overlap.

In a similar vein, Eberhard Fetz[187] wrote in 1992 about the question "Are movement parameters recognizably coded in the activity of single neurons?"…

> [T]he search for explicit coding may actually be misleading, and may divert our understanding of distributed neural mechanisms that operate without literal representations.

…implying that individual motor cortex neurons do not *explicitly* code for movement parameters, including force and limb displacement, and that evidence suggesting it does may in fact arise from conceptually-biased and selective analysis of the data[187]. If a major corticostriatal input to the BG

might not explicitly encode key behavioral variables[188,189], then perhaps the BG itself does not either?

A major advantage of a RL system in which policy and value are learned by separate modules—i.e. of the actor-critic system—is that it can learn continuous policies[7], which allow the agent to deal with continuous action spaces. That is, for example, useful if you are faced with the (admittedly obscure) question "*How bitter* would you like your chocolate ice cream?" rather than "Chocolate or Vanilla?". For the BG, an ability to acquire continuous policies may also serve to distill complex, dynamical cortical states into equally dynamical, but presumably somewhat less complex and more contextually-useful, "BG states"—which the BG's cortical and subcortical targets may find highly actionable[30,100]. What is the nature of cortical population-level computations?, and what is that of their BG "distillate", as returned through the BG's various loops? Investigating these questions, and comparing their answers, is what I find most intriguing today. If we do, there is hope that someday we might know how a BG actor-critic—if one more distributed than outlined here—has shaped your ice cream habit.

Acknowledgements

First of all, thank you **Dinos Meletis** for taking me on as a Master student back in 2015, upgrading me to a PhD by 2017, and for all the years of ergocentric supervision, Whiskycentric after hours, and—of course—Greececentric retreats, which combined the former two to brilliant effect. It's been a blast (and I think I've learned a thing or two, too)!

A big thanks to my co-supervisors **Arvind Kumar** and **Per Svenningsson**! I've consistently enjoyed being grumpy in your journal club, Arvind, and I am happy it kept you entertained, too ;).

Thank you, **Marie Carlén**. Not a supervisor on paper, but definitely in spirit. I'm grateful for all the advice and support.

Next up, I have to thank **Iakovos**. You inspired the lab-MacGyver in me, duct tape, Arduinos, soldering fumes, and all. You've been a great mentor, even greater collaborator, and tons of fun to be around and argue with. This PhD would not have been possible without you.

Emil, I can't thank you enough either. I have learned so much from our collaboration, and all the conversations over lunch or coffee/tea. You're endlessly insightful. Thank you also for all the patient explanations, and the feedback on anything I ever got nervous about!

And thanks, to all the past and present DMC lab members and friends at KI, who I've had the pleasure to share this journey with!

Bibliography

1. Weglage, M., Wärnberg, E., Lazaridis, I., Calvigioni, D., Tzortzi, O. & Meletis, K. Complete representation of action space and value in all dorsal striatal pathways. *Cell Reports* **36,** 109437. doi:10.1016/j.celrep.2021.109437 (2021).
2. Weglage, M., Ährlund-Richter, S., Fuzik, J., Skara, V., Lazaridis, I. & Meletis, K. Sst+ GPi output neurons provide direct feedback to key nodes of the basal ganglia and drive behavioral flexibility. *bioRxiv.* doi:10.1101/2022.03.16.484460 (2022).
3. Lazaridis, I., Tzortzi, O., Weglage, M., *et al.* A hypothalamus-habenula circuit controls aversion. *Molecular Psychiatry* **24,** 1351–1368. doi:10.1038/s41380-019-0369-5 (2019).
4. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Theory and Research* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, New York, New York, 1972).
5. Pearce, J. M. & Bouton, M. E. Theories of Associative Learning in Animals. *Annual Review of Psychology* **52,** 111–139. doi:10.1146/annurev.psych.52.1.111 (2001).
6. Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 215–232 (The MIT Press, Cambridge, Massachusetts, 1994). doi:10.7551/mitpress/4708.003.0018.
7. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* Second Edi (The MIT Press, Cambridge, Massachusetts, 2018).
8. Barto, A. G., Sutton, R. S. & Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-13,** 834–846. doi:10.1109/TSMC.1983.6313077 (1983).
9. Maia, T. V. & Frank, M. J. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience* **14,** 154–162. doi:10.1038/nn.2723 (2011).
10. Averbeck, B. & O'Doherty, J. P. Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology* **47,** 147–162. doi:10.1038/s41386-021-01108-0 (2022).
11. Richards, B. A., Lillicrap, T. P., Beaudoin, P., *et al.* A deep learning framework for neuroscience. *Nature Neuroscience* **22,** 1761–1770. doi:10.1038/s41593-019-0520-2 (2019).
12. Houk, J. C., Adams, J. L. & Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 249–270 (The MIT Press, Cambridge, Massachusetts, 1994). doi:10.7551/mitpress/4708.003.0020.
13. Joel, D., Niv, Y. & Ruppin, E. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks* **15,** 535–547. doi:10.1016/S0893-6080(02)00047-3 (2002).

14. Bergman, H. in *The Hidden Life of the Basal Ganglia* 117–132 (The MIT Press, Cambridge, Massachusetts, 2021). doi:10.7551/mitpress/14075.003.0015.

15. Bornstein, A. M. & Daw, N. D. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology* **21,** 374–380. doi:10.1016/j.conb.2011.02.009 (2011).

16. Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R. & Dickinson, A. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 233–248 (The MIT Press, Cambridge, Massachusetts, 1994). doi:10.7551/mitpress/4708.003.0019.

17. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275,** 1593–1599. doi:10.1126/science.275.5306.1593 (1997).

18. Montague, P., Dayan, P. & Sejnowski, T. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience* **16,** 1936–1947. doi:10.1523/JNEUROSCI.16-05-01936.1996 (1996).

19. Beckstead, R. M., Domesick, V. B. & Nauta, W. J. Efferent connections of the substantia nigra and ventral tegmental area in the rat. *Brain Research* **175,** 191–217. doi:10.1016/0006-8993(79)91001-1 (1979).

20. Gerfen, C., Herkenham, M. & Thibault, J. The neostriatal mosaic: II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. *The Journal of Neuroscience* **7,** 3915–3934. doi:10.1523/JNEUROSCI.07-12-03915.1987 (1987).

21. Haber, S. The place of dopamine in the cortico-basal ganglia circuit. *Neuroscience* **282,** 248–257. doi:10.1016/j.neuroscience.2014.10.008 (2014).

22. Suri, R. E. & Schultz, W. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research* **121,** 350–354. doi:10.1007/s002210050467 (1998).

23. Hintiryan, H., Foster, N. N., Bowman, I., *et al.* The mouse cortico-striatal projectome. *Nature Neuroscience* **19,** 1100–1114. doi:10.1038/nn.4332 (2016).

24. Hunnicutt, B. J., Jongbloets, B. C., Birdsong, W. T., Gertz, K. J., Zhong, H. & Mao, T. A comprehensive excitatory input map of the striatum reveals novel functional organization. *eLife* **5,** 1–32. doi:10.7554/eLife.19103 (2016).

25. Flaherty, A. W. & Graybiel, A. M. Corticostriatal transformations in the primate somatosensory system. Projections from physiologically mapped body-part representations. *Journal of Neurophysiology* **66,** 1249–1263. doi:10.1152/jn.1991.66.4.1249 (1991).

26. Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex. *Annual Review of Neuroscience* **9,** 357–381. doi:10.1146/annurev.neuro.9.1.357 (1986).

27. Foster, N. N., Barry, J., Korobkova, L., *et al.* The mouse cortico–basal ganglia–thalamic network. *Nature* **598,** 188–194. doi:10.1038/s41586-021-03993-3 (2021).

28. Dudman, J. T. & Krakauer, J. W. The basal ganglia: from motor commands to the control of vigor. *Current Opinion in Neurobiology* **37,** 158–166. doi:10.1016/j.conb.2016.02.005 (2016).

29. Berns, G. S. & Sejnowski, T. J. in *Neurobiology of Decision-Making* (eds Damasio, A. R., Damasio, H. & Christen, Y.) 101–113 (Springer, Berlin, 1996). doi:10.1007/978-3-642-79928-0_6.

30. Houk, J. C. & Wise, S. P. Distributed Modular Architectures Linking Basal Ganglia, Cerebellum, and Cerebral Cortex: Their Role in Planning and Controlling Action. *Cerebral Cortex* **5,** 95–110. doi:10.1093/cercor/5.2.95 (1995).

31. Mailly, P., Aliane, V., Groenewegen, H. J., Haber, S. N. & Deniau, J.-M. The Rat Prefrontostriatal System Analyzed in 3D: Evidence for Multiple Interacting Functional Units. *Journal of Neuroscience* **33,** 5718–5727. doi:10.1523/JNEUROSCI.5248-12.2013 (2013).

32. McElvain, L. E., Chen, Y., Moore, J. D., Brigidi, G. S., Bloodgood, B. L., Lim, B. K., Costa, R. M. & Kleinfeld, D. Specific populations of basal ganglia output neurons target distinct brain stem areas while collateralizing throughout the diencephalon. *Neuron* **109,** 1721–1738.e4. doi:10.1016/j.neuron.2021.03.017 (2021).

33. McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V. & Redgrave, P. Subcortical loops through the basal ganglia. *Trends in Neurosciences* **28,** 401–407. doi:10.1016/j.tins.2005.06.006 (2005).

34. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends in Neurosciences* **12,** 366–375. doi:10.1016/0166-2236(89)90074-X (1989).

35. DeLong, M. R. Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences* **13,** 281–285. doi:10.1016/0166-2236(90)90110-V (1990).

36. Alexander, G. E. & Crutcher, M. D. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences* **13,** 266–271. doi:10.1016/0166-2236(90)90107-L (1990).

37. Mink, J. W. The Basal Ganglia: Focused Selection and Inhibition of Competing Motor Programs. *Progress in Neurobiology* **50,** 381–425. doi:10.1016/S0301-0082(96)00042-1 (1996).

38. Park, J., Coddington, L. T. & Dudman, J. T. Basal Ganglia Circuits for Action Specification. *Annual Review of Neuroscience* **43,** 485–507. doi:10.1146/annurev-neuro-070918-050452 (2020).

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

40. Suri, R., Bargas, J. & Arbib, M. Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience* **103,** 65–85. doi:10.1016/S0306-4522(00)00554-6 (2001).

41. Yttri, E. A. & Dudman, J. T. A Proposed Circuit Computation in Basal Ganglia: History-Dependent Gain. *Movement Disorders* **33,** 704–716. doi:10.1002/mds.27321 (2018).

42. Suri, R. & Schultz, W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* **91,** 871–890. doi:10.1016/S0306-4522(98)00697-6 (1999).

43. Contreras-Vidal, J. L. & Schultz, W. Predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience* **6,** 191–214. doi:https://doi.org/10.1023/A:1008862904946 (1999).

44. Brown, J., Bullock, D. & Grossberg, S. How the Basal Ganglia Use Parallel Excitatory and Inhibitory Learning Pathways to Selectively Respond to Unexpected Rewarding Cues. *The Journal of Neuroscience* **19,** 10502–10511. doi:10.1523/JNEUROSCI.19-23-10502.1999 (1999).

45. Collins, A. G. E. & Frank, M. J. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review* **121,** 337–366. doi:10.1037/a0037015 (2014).

46. O'Reilly, R. C. & Frank, M. J. Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation* **18,** 283–328. doi:10.1162/089976606775093909 (2006).

47. Chen, R. & Goldberg, J. H. Actor-critic reinforcement learning in the songbird. *Current Opinion in Neurobiology* **65,** 1–9. doi:10.1016/j.conb.2020.08.005 (2020).

48. Pert, C. B., Kuhar, M. J. & Snyder, S. H. Autoradiographic localization of the opiate receptor in rat brain. *Life Sciences* **16,** 1849–1853. doi:10.1016/0024-3205(75)90289-1 (1975).

49. Pert, C. B., Kuhar, M. J. & Snyder, S. H. Opiate receptor: autoradiographic localization in rat brain. *Proceedings of the National Academy of Sciences* **73,** 3729–3733. doi:10.1073/pnas.73.10.3729 (1976).

50. Graybiel, A., Ragsdale, C., Yoneoka, E. & Elde, R. An immunohistochemical study of enkephalins and other neuropeptides in the striatum of the cat with evidence that the opiate peptides are arranged to form mosaic patterns in register with the striosomal compartments visible by acetylcholinesterase staining. *Neuroscience* **6,** 377–397. doi:10.1016/0306-4522(81)90131-7 (1981).

51. Herkenham, M. & Pert, C. B. Mosaic distribution of opiate receptors, parafascicular projections and acetylcholinesterase in rat striatum. *Nature* **291,** 415–418. doi:10.1038/291415a0 (1981).

52. Graybiel, A. M. Habits, Rituals, and the Evaluative Brain. *Annual Review of Neuroscience* **31,** 359–387. doi:10.1146/annurev.neuro.29.051605.112851 (2008).

53. Gerfen, C. R. The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature* **311,** 461–464. doi:10.1038/311461a0 (1984).

54. Gerfen, C. R. The neostriatal mosaic. I. compartmental organization of projections from the striatum to the substantia nigra in the rat. *The Journal of Comparative Neurology* **236,** 454–476. doi:10.1002/cne.902360404 (1985).

55. Jiménez-Castellanos, J. & Graybiel, A. Compartmental origins of striatal efferent projections in the cat. *Neuroscience* **32,** 297–321. doi:10.1016/0306-4522(89)90080-8 (1989).

56. Fujiyama, F., Sohn, J., Nakano, T., Furuta, T., Nakamura, K. C., Matsuda, W. & Kaneko, T. Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector. *European Journal of Neuroscience* **33,** 668–677. doi:10.1111/j.1460-9568.2010.07564.x (2011).

57. Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A. & Uchida, N. Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. *Neuron* **74,** 858–873. doi:10.1016/j.neuron.2012.03.017 (2012).

58. McGregor, M. M., McKinsey, G. L., Girasole, A. E., Bair-Marshall, C. J., Rubenstein, J. L. & Nelson, A. B. Functionally Distinct Connectivity of Developmentally Targeted Striosome Neurons. *Cell Reports* **29,** 1419–1428.e5. doi:10.1016/j.celrep.2019.09.076 (2019).

59. Crittenden, J. R., Tillberg, P. W., Riad, M. H., Shima, Y., Gerfen, C. R., Curry, J., Housman, D. E., Nelson, S. B., Boyden, E. S. & Graybiel, A. M. Striosome–dendron bouquets highlight a unique striatonigral circuit targeting dopamine-containing neurons. *Proceedings of the National Academy of Sciences* **113,** 11318–11323. doi:10.1073/pnas.1613337113 (2016).

60. Evans, R. C., Twedell, E. L., Zhu, M., Ascencio, J., Zhang, R. & Khaliq, Z. M. Functional Dissection of Basal Ganglia Inhibitory Inputs onto Substantia Nigra Dopaminergic Neurons. *Cell Reports* **32,** 108156. doi:10.1016/j.celrep.2020.108156 (2020).

61. Poulin, J.-f., Caronia, G., Hofer, C., Cui, Q., Helm, B., Ramakrishnan, C., Chan, C. S., Dombeck, D. A., Deisseroth, K. & Awatramani, R. Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. *Nature Neuroscience* **21,** 1260–1271. doi:10.1038/s41593-018-0203-4 (2018).

62. Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., Arai, R. & Kaneko, T. Single Nigrostriatal Dopaminergic Neurons Form Widely Spread and Highly Dense Axonal Arborizations in the Neostriatum. *Journal of Neuroscience* **29,** 444–453. doi:10.1523/JNEUROSCI.4029-08.2009 (2009).

63. Ragsdale, C. W. & Graybiel, A. M. Fibers from the basolateral nucleus of the amygdala selectively innervate striosomes in the caudate nucleus of the cat. *The Journal of Comparative Neurology* **269,** 506–522. doi:10.1002/cne.902690404 (1988).

64. Gerfen, C. R. The Neostriatal Mosaic: Striatal Patch-Matrix Organization Is Related to Cortical Lamination. *Science* **246,** 385–388. doi:10.1126/science.2799392 (1989).

65. Eblen, F. & Graybiel, A. Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *The Journal of Neuroscience* **15,** 5999–6013. doi:10.1523/JNEUROSCI.15-09-05999.1995 (1995).

66. Rajakumar, N., Elisevich, K. & Flumerfelt, B. A. Compartmental origin of the striatoentopeduncular projection in the rat. *The Journal of Comparative Neurology* **331,** 286–296. doi:10.1002/cne.903310210 (1993).

67. O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R. J. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science* **304,** 452–454. doi:10.1126/science.1094285 (2004).

68. Takahashi, Y., Schoenbaum, G. & Niv, Y. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience* **2,** 86–99. doi:10.3389/neuro.01.014.2008 (2008).

69. Van der Meer, M. A. & Redish, A. D. Ventral striatum: a critical look at models of learning and evaluation. *Current Opinion in Neurobiology* **21,** 387–392. doi:10.1016/j.conb.2011.02.011 (2011).

70. Nieh, E. H., Vander Weele, C. M., Matthews, G. A., Presbrey, K. N., Wichmann, R., Leppla, C. A., Izadmehr, E. M. & Tye, K. M. Inhibitory Input from the Lateral Hypothalamus to the Ventral Tegmental Area Disinhibits Dopamine Neurons and Promotes Behavioral Activation. *Neuron* **90,** 1286–1298. doi:10.1016/j.neuron.2016.04.035 (2016).

71. Burke, D. A., Rotstein, H. G. & Alvarez, V. A. Striatal Local Circuitry: A New Framework for Lateral Inhibition. *Neuron* **96,** 267–284. doi:10.1016/j.neuron.2017.09.019 (2017).

72. Dunovan, K., Vich, C., Clapp, M., Verstynen, T. & Rubin, J. Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. *PLOS Computational Biology* **15** (ed Bogacz, R.) e1006998. doi:10.1371/journal.pcbi.1006998 (2019).

73. Polyakova, Z., Chiken, S., Hatanaka, N. & Nambu, A. Cortical Control of Subthalamic Neuronal Activity through the Hyperdirect and Indirect Pathways in Monkeys. *The Journal of Neuroscience* **40,** 7451–7463. doi:10.1523/JNEUROSCI.0772-20.2020 (2020).

74. Hong, S. & Hikosaka, O. The Globus Pallidus Sends Reward-Related Signals to the Lateral Habenula. *Neuron* **60,** 720–729. doi:http://dx.doi.org/10.1016/j.neuron.2008.09.035 (2008).

75. Hong, S., Amemori, S., Chung, E., Gibson, D. J., Amemori, K.-i. & Graybiel, A. M. Predominant Striatal Input to the Lateral Habenula in Macaques Comes from Striosomes. *Current Biology* **29,** 51–61.e5. doi:10.1016/j.cub.2018.11.008 (2019).

76. Stephenson-Jones, M., Kardamakis, A. A., Robertson, B. & Grillner, S. Independent circuits in the basal ganglia for the evaluation and selection of actions. *Proceedings of the National Academy of Sciences* **110,** E3670–E3679. doi:10.1073/pnas.1314815110 (2013).

77. Stephenson-Jones, M., Yu, K., Ahrens, S., Tucciarone, J. M., van Huijstee, A. N., Mejia, L. A., Penzo, M. A., Tai, L.-H., Wilbrecht, L. & Li, B. A basal ganglia circuit for evaluating action outcomes. *Nature* **539,** 289–293. doi:10.1038/nature19845 (2016).

78. Nauta, H. J. W. Evidence of a pallidohabenular pathway in the cat. *The Journal of Comparative Neurology* **156,** 19–27. doi:10.1002/cne.901560103 (1974).

79. Herkenham, M. & Nauta, W. J. H. Afferent connections of the habenular nuclei in the rat. A horseradish peroxidase study, with a note on the fiber-of-passage problem. *The Journal of Comparative Neurology* **173,** 123–145. doi:10.1002/cne.901730107 (1977).

80. Matsumoto, M. & Hikosaka, O. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447,** 1111–1115. doi:10.1038/nature05860 (2007).

81. Matsumoto, M. & Hikosaka, O. Representation of negative motivational value in the primate lateral habenula. *Nature Neuroscience* **12,** 77–84. doi:10.1038/nn.2233 (2009).

82. Grillner, S. & Robertson, B. The Basal Ganglia Over 500 Million Years. *Current Biology* **26,** R1088–R1100. doi:10.1016/j.cub.2016.06.041 (2016).

83. Herkenham, M. & Nauta, W. J. H. Efferent connections of the habenular nuclei in the rat. *The Journal of Comparative Neurology* **187,** 19–47. doi:10.1002/cne.901870103 (1979).

84. Christoph, G., Leonzio, R. & Wilcox, K. Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *The Journal of Neuroscience* **6,** 613–619. doi:10.1523/JNEUROSCI.06-03-00613.1986 (1986).

85. Hong, S., Jhou, T. C., Smith, M., Saleem, K. S. & Hikosaka, O. Negative Reward Signals from the Lateral Habenula to Dopamine Neurons Are Mediated by Rostromedial Tegmental Nucleus in Primates. *Journal of Neuroscience* **31,** 11457–11471. doi:10.1523/JNEUROSCI.1384-11.2011 (2011).

86. Van Der Kooy, D. & Carter, D. A. The organization of the efferent projections and striatal afferents of the entopeduncular nucleus and adjacent areas in the rat. *Brain Research* **211,** 15–36. doi:10.1016/0006-8993(81)90064-0 (1981).

87. Parent, A. & De Bellefeuille, L. Organization of efferent projections from the internal segment of globus pallidus in primate as revealed by flourescence retrograde labeling method. *Brain Research* **245,** 201–213. doi:10.1016/0006-8993(82)90802-2 (1982).

88. Wallace, M. L., Saunders, A., Huang, K. W., Philson, A. C., Goldman, M., Macosko, E. Z., McCarroll, S. A. & Sabatini, B. L. Genetically Distinct Parallel Pathways in the Entopeduncular Nucleus for Limbic and Sensorimotor Output of the Basal Ganglia. *Neuron* **94,** 138–152.e5. doi:10.1016/j.neuron.2017.03.017 (2017).

89. Rajakumar, N., Elisevich, K. & Flumerfelt, B. A. Parvalbumin-containing GABAergic neurons in the basal ganglia output system of the rat. *The Journal of Comparative Neurology* **350,** 324–336. doi:10.1002/cne.903500214 (1994).

90. Vincent, S. R. & Brown, J. C. Somatostatin immunoreactivity in the entopeduncular projection to the lateral habenula in the rat. *Neuroscience Letters* **68,** 160–164. doi:10.1016/0304-3940(86)90134-5 (1986).

91. Miyamoto, Y. & Fukuda, T. Immunohistochemical study on the neuronal diversity and three-dimensional organization of the mouse entopeduncular nucleus. *Neuroscience Research* **94,** 37–49. doi:10.1016/j.neures.2015.02.006 (2015).

92. Li, H., Pullmann, D. & Jhou, T. C. Valence-encoding in the lateral habenula arises from the entopeduncular region. *eLife* **8,** 1–17. doi:10.7554/eLife.41223 (2019).

93. Nadel, J. A., Pawelko, S. S., Scott, J. R., McLaughlin, R., Fox, M., Ghanem, M., van der Merwe, R., Hollon, N. G., Ramsson, E. S. & Howard, C. D. Optogenetic stimulation of striatal patches modifies habit formation and inhibits dopamine release. *Scientific Reports* **11,** 19847. doi:10.1038/s41598-021-99350-5 (2021).

94. Shabel, S. J., Proulx, C. D., Trias, A., Murphy, R. T. & Malinow, R. Input to the Lateral Habenula from the Basal Ganglia Is Excitatory, Aversive, and Suppressed by Serotonin. *Neuron* **74,** 475–481. doi:10.1016/j.neuron.2012.02.037 (2012).

95. Dhawale, A. K., Wolff, S. B. E., Ko, R. & Ölveczky, B. P. The basal ganglia control the detailed kinematics of learned motor skills. *Nature Neuroscience* **24,** 1256–1269. doi:10.1038/s41593-021-00889-3 (2021).

96. Yoo, S. B. M., Hayden, B. Y. & Pearson, J. M. Continuous decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **376,** 20190664. doi:10.1098/rstb.2019.0664 (2021).

97. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* **7,** 464–476. doi:10.1038/nrn1919 (2006).

98. Rueda-Orozco, P. E. & Robbe, D. The striatum multiplexes contextual and kinematic information to constrain motor habits execution. *Nature Neuroscience* **18,** 453–460. doi:10.1038/nn.3924 (2015).

99. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533,** 402–406. doi:10.1038/nature17639 (2016).

100. Bar-Gad, I., Havazelet-Heimer, G., Goldberg, J., Ruppin, E. & Bergman, H. Reinforcement-Driven Dimensionality Reduction - A Model for Information Processing in the Basal Ganglia. *Journal of Basic and Clinical Physiology and Pharmacology* **11,** 305–320. doi:10.1515/JBCPP.2000.11.4.305 (2000).

101. Nectow, A. R. & Nestler, E. J. Viral tools for neuroscience. *Nature Reviews Neuroscience* **21,** 669–681. doi:10.1038/s41583-020-00382-z (2020).

102. Branda, C. S. & Dymecki, S. M. Talking about a Revolution. *Developmental Cell* **6,** 7–28. doi:10.1016/S1534-5807(03)00399-X (2004).

103. Märtin, A., Calvigioni, D., Tzortzi, O., Fuzik, J., Wärnberg, E. & Meletis, K. A Spatiomolecular Map of the Striatum. *Cell Reports* **29,** 4320–4333.e5. doi:10.1016/j.celrep.2019.11.096 (2019).

104. Fenno, L. E., Mattis, J., Ramakrishnan, C., *et al.* Targeting cells with single vectors using multiple-feature Boolean logic. *Nature Methods* **11,** 763–772. doi:10.1038/nmeth.2996 (2014).

105. Tervo, D. G. R., Hwang, B.-Y., Viswanathan, S., *et al.* A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron* **92,** 372–382. doi:10.1016/j.neuron.2016.09.021 (2016).

106. Siegelbaum, S. A., Kandel, E. R. & Südhof, T. C. in *Principles Of Neural Science* (eds Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A. & Hudspeth, A. J.) 5th ed., 260–288 (McGraw-Hill, New York, 2013).

107. Svoboda, K., Denk, W., Kleinfeld, D. & Tank, D. W. In vivo dendritic calcium dynamics in neocortical pyramidal neurons. *Nature* **385,** 161–165. doi:10.1038/385161a0 (1997).

108. Dana, H., Sun, Y., Mohar, B., *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods* **16,** 649–657. doi:10.1038/s41592-019-0435-6 (2019).

109. Nakai, J., Ohkura, M. & Imoto, K. A high signal-to-noise Ca2+ probe composed of a single green fluorescent protein. *Nature Biotechnology* **19,** 137–141. doi:10.1038/84397 (2001).

110. Ghosh, K. K., Burns, L. D., Cocker, E. D., Nimmerjahn, A., Ziv, Y., Gamal, A. E. & Schnitzer, M. J. Miniaturized integration of a fluorescence microscope. *Nature Methods* **8,** 871–878. doi:10.1038/nmeth.1694 (2011).

111. Resendez, S. L., Jennings, J. H., Ung, R. L., Namboodiri, V. M. K., Zhou, Z. C., Otis, J. M., Nomura, H., McHenry, J. A., Kosyk, O. & Stuber, G. D. Visualization of cortical, subcortical and deep brain neural circuit dynamics during naturalistic mammalian behavior with head-mounted microscopes and chronically implanted lenses. *Nature Protocols* **11,** 566–597. doi:10.1038/nprot.2016.021 (2016).

112. Adelsberger, H., Garaschuk, O. & Konnerth, A. Cortical calcium waves in resting newborn mice. *Nature Neuroscience* **8,** 988–990. doi:10.1038/nn1502 (2005).

113. Kim, C. K., Yang, S. J., Pichamoorthy, N., *et al.* Simultaneous fast measurement of circuit dynamics at multiple sites across the mammalian brain. *Nature Methods* **13,** 325–328. doi:10.1038/nmeth.3770 (2016).

114. Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P. & Bamberg, E. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proceedings of the National Academy of Sciences* **100,** 13940–13945. doi:10.1073/pnas.1936192100 (2003).

115. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience* **8,** 1263–1268. doi:10.1038/nn1525 (2005).

116. Schiavo, G. G., Benfenati, F., Poulain, B., Rossetto, O., de Laureto, P. P., DasGupta, B. R. & Montecucco, C. Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature* **359,** 832–835. doi:10.1038/359832a0 (1992).

117. Yamamoto, M., Wada, N., Kitabatake, Y., Watanabe, D., Anzai, M., Yokoyama, M., Teranishi, Y. & Nakanishi, S. Reversible Suppression of Glutamatergic Neurotransmission of Cerebellar Granule Cells In Vivo by Genetically Manipulated Expression of Tetanus Neurotoxin Light Chain. *The Journal of Neuroscience* **23,** 6759–6767. doi:10.1523/JNEUROSCI.23-17-06759.2003 (2003).

118. Murray, A. J., Sauer, J.-F., Riedel, G., McClure, C., Ansel, L., Cheyne, L., Bartos, M., Wisden, W. & Wulff, P. Parvalbumin-positive CA1 interneurons are required for spatial working but not for reference memory. *Nature Neuroscience* **14,** 297–299. doi:10.1038/nn.2751 (2011).

119. Izquierdo, A. & Jentsch, J. D. Reversal learning as a measure of impulsive and compulsive behavior in addictions. *Psychopharmacology* **219,** 607–620. doi:10.1007/s00213-011-2579-7 (2012).

120. Izquierdo, A., Brigman, J., Radke, A., Rudebeck, P. & Holmes, A. The neural basis of reversal learning: An updated perspective. *Neuroscience* **345,** 12–26. doi:10.1016/j.neuroscience.2016.03.021 (2017).

121. Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience* **15,** 1281–1289. doi:10.1038/nn.3188 (2012).

122. Sharpe, M. J., Stalnaker, T., Schuck, N. W., Killcross, S., Schoenbaum, G. & Niv, Y. An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. *Annual Review of Psychology* **70,** 53–76. doi:10.1146/annurev-psych-010418-102824 (2019).

123. Bari, B. A., Grossman, C. D., Lubin, E. E., Rajagopalan, A. E., Cressy, J. I. & Cohen, J. Y. Stable Representations of Decision Variables for Flexible Behavior. *Neuron* **103,** 922–933.e7. doi:10.1016/j.neuron.2019.06.001 (2019).

124. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. *Neuron* **66,** 781–795. doi:10.1016/j.neuron.2010.04.036 (2010).

125. Lauwereyns, J., Watanabe, K., Coe, B. & Hikosaka, O. A neural correlate of response bias in monkey caudate nucleus. *Nature* **418,** 413–417. doi:10.1038/nature00892 (2002).

126. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of Action-Specific Reward Values in the Striatum. *Science* **310,** 1337–1340. doi:10.1126/science.1115270 (2005).

127. Williams, Z. M. & Eskandar, E. N. Selective enhancement of associative learning by microstimulation of the anterior caudate. *Nature Neuroscience* **9,** 562–568. doi:10.1038/nn1662 (2006).

128. Ragozzino, M. E. The Contribution of the Medial Prefrontal Cortex, Orbitofrontal Cortex, and Dorsomedial Striatum to Behavioral Flexibility. *Annals of the New York Academy of Sciences* **1121,** 355–375. doi:10.1196/annals.1401.013 (2007).

129. Castañé, A., Theobald, D. E. & Robbins, T. W. Selective lesions of the dorsomedial striatum impair serial spatial reversal learning in rats. *Behavioural Brain Research* **210,** 74–83. doi:10.1016/j.bbr.2010.02.017 (2010).

130. Bariselli, S., Fobbs, W., Creed, M. & Kravitz, A. A competitive model for striatal action selection. *Brain Research* **1713,** 70–79. doi:10.1016/j.brainres.2018.10.009 (2019).

131. Nonomura, S., Nishizawa, K., Sakai, Y., *et al.* Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. *Neuron* **99,** 1302–1314.e5. doi:10.1016/j.neuron.2018.08.002 (2018).

132. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience* **15,** 816–818. doi:10.1038/nn.3100 (2012).

133. Bolkan, S. S., Stone, I. R., Pinto, L., *et al.* Opponent control of behavior by dorsomedial striatal pathways depends on task demands and internal state. *Nature Neuroscience* **25,** 345–357. doi:10.1038/s41593-022-01021-9 (2022).

134. Elber-Dorozko, L. & Loewenstein, Y. Striatal action-value neurons reconsidered. *eLife* **7,** 1–34. doi:10.7554/eLife.34248 (2018).

135. White, N. M. & Hiroi, N. Preferential localization of self-stimulation sites in striosomes/patches in the rat striatum. *Proceedings of the National Academy of Sciences* **95,** 6486–6491. doi:10.1073/pnas.95.11.6486 (1998).

136. Cui, Y., Ostlund, S. B., James, A. S., *et al.* Targeted expression of μ-opioid receptors in a subset of striatal direct-pathway neurons restores opiate reward. *Nature Neuroscience* **17,** 254–261. doi:10.1038/nn.3622 (2014).

137. Canales, J. J. & Graybiel, A. M. A measure of striatal function predicts motor stereotypy. *Nature Neuroscience* **3,** 377–383. doi:10.1038/73949 (2000).

138. Saka, E., Goodrich, C., Harlan, P., Madras, B. K. & Graybiel, A. M. Repetitive Behaviors in Monkeys Are Linked to Specific Striatal Activation Patterns. *Journal of Neuroscience* **24,** 7557–7565. doi:10.1523/JNEUROSCI.1072-04.2004 (2004).

139. Murray, R. C., Gilbert, Y. E., Logan, A. S., Hebbard, J. C. & Horner, K. A. Striatal patch compartment lesions alter methamphetamine-induced behavior and immediate early gene expression in the striatum, substantia nigra and frontal cortex. *Brain Structure and Function* **219,** 1213–1229. doi:10.1007/s00429-013-0559-x (2014).

140. Murray, R. C., Logan, M. C. & Horner, K. A. Striatal patch compartment lesions reduce stereotypy following repeated cocaine administration. *Brain Research* **1618,** 286–298. doi:10.1016/j.brainres.2015.06.012 (2015).

141. Jenrette, T. A., Logue, J. B. & Horner, K. A. Lesions of the Patch Compartment of Dorsolateral Striatum Disrupt Stimulus–Response Learning. *Neuroscience* **415,** 161–172. doi:10.1016/j.neuroscience.2019.07.033 (2019).

142. Nadel, J. A., Pawelko, S. S., Copes-Finke, D., Neidhart, M. & Howard, C. D. Lesion of striatal patches disrupts habitual behaviors and increases behavioral variability. *PLOS ONE* **15** (ed Beeler, J. A.) e0224715. doi:10.1371/journal.pone.0224715 (2020).

143. Lawhorn, C., Smith, D. & Brown, L. Partial ablation of mu-opioid receptor rich striosomes produces deficits on a motor-skill learning task. *Neuroscience* **163,** 109–119. doi:10.1016/j.neuroscience.2009.05.021 (2009).

144. Bloem, B., Huda, R., Sur, M. & Graybiel, A. M. Two-photon imaging in mice shows striosomes and matrix have overlapping but differential reinforcement-related responses. *eLife* **6,** e32353. doi:10.7554/eLife.32353 (2017).

145. Yoshizawa, T., Ito, M. & Doya, K. Reward-Predictive Neural Activities in Striatal Striosome Compartments. *eneuro* **5,** ENEURO.0367–17.2018. doi:10.1523/ENEURO.0367-17.2018 (2018).

146. Friedman, A., Homma, D., Gibb, L. G., Amemori, K. I., Rubin, S. J., Hood, A. S., Riad, M. H. & Graybiel, A. M. A corticostriatal path targeting striosomes controls decision-making under conflict. *Cell* **161,** 1320–1333. doi:10.1016/j.cell.2015.04.049 (2015).

147. Friedman, A., Homma, D., Bloem, B., *et al.* Chronic Stress Alters Striosome-Circuit Dynamics, Leading to Aberrant Decision-Making. *Cell* **171,** 1191.e28–1205.e28. doi:10.1016/j.cell.2017.10.017 (2017).

148. Friedman, A., Hueske, E., Drammis, S. M., *et al.* Striosomes Mediate Value-Based Learning Vulnerable in Age and a Huntington's Disease Model. *Cell* **183,** 918–934.e49. doi:10.1016/j.cell.2020.09.060 (2020).

149. Xiao, X., Deng, H., Furlan, A., *et al.* A Genetically Defined Compartmentalized Striatal Direct Pathway for Negative Reinforcement. *Cell* **183,** 211–227.e20. doi:10.1016/j.cell.2020.08.032 (2020).

150. Nagy, J., Carter, D., Lehmann, J. & Fibiger, H. Evidence for a GABA-containing projection from the entopeduncular nucleus to the lateral habenula in the rat. *Brain Research* **145,** 360–364. doi:10.1016/0006-8993(78)90869-7 (1978).

151. Araki, M., McGeer, P. & McGeer, E. Retrograde HRP tracing combined with a pharmacohistochemical method for GABA transaminase for the identification of presumptive GABAergic projections to the habenula. *Brain Research* **304,** 271–277. doi:10.1016/0006-8993(84)90330-5 (1984).

152. Song, Y.-H., Yoon, J. & Lee, S.-H. The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. *Experimental & Molecular Medicine* **53,** 328–338. doi:10.1038/s12276-021-00580-4 (2021).

153. Shabel, S. J., Proulx, C. D., Piriz, J. & Malinow, R. GABA/glutamate co-release controls habenula output and is modified by antidepressant treatment. *Science* **345,** 1494–1498. doi:10.1126/science.1250469 (2014).

154. Kim, S., Wallace, M. L., El-Rifai, M., Knudsen, A. R. & Sabatini, B. L. Co-packaging of opposing neurotransmitters in individual synaptic vesicles in the central nervous system. *Neuron,* 1–14. doi:10.1016/j.neuron.2022.01.007 (2022).

155. A glutamatergic projection from the lateral hypothalamus targets VTA-projecting neurons in the lateral habenula of the rat. *Brain Research* **1507,** 45–60. doi:10.1016/j.brainres.2013.01.029 (2013).

156. Parent, A., Gravel, S. & Boucher, R. The origin of forebrain afferents to the habenula in rat, cat and monkey. *Brain Research Bulletin* **6,** 23–38. doi:10.1016/S0361-9230(81)80066-4 (1981).

157. Olds, J. Self-Stimulation of the Brain. *Science* **127,** 315–324. doi:10.1126/science.127.3294.315 (1958).

158. Huston, J. P. Relationship between motivating and rewarding stimulation of the lateral hypothalamus. *Physiology & Behavior* **6,** 711–716. doi:10.1016/0031-9384(71)90259-9 (1971).

159. Ono, T. & Nakamura, K. Learning and integration of rewarding and aversive stimuli in the rat lateral hypothalamus. *Brain Research* **346,** 368–373. doi:10.1016/0006-8993(85)90872-8 (1985).

160. Jennings, J. H., Rizzi, G., Stamatakis, A. M., Ung, R. L. & Stuber, G. D. The Inhibitory Circuit Architecture of the Lateral Hypothalamus Orchestrates Feeding. *Science* **341,** 1517–1521. doi:10.1126/science.1241812 (2013).

161. Wise, R. A. Dual Roles of Dopamine in Food and Drug Seeking: The Drive-Reward Paradox. *Biological Psychiatry* **73,** 819–826. doi:10.1016/j.biopsych.2012.09.001 (2013).

162. Stamatakis, A. M., Van Swieten, M., Basiri, M. L., Blair, G. A., Kantak, P. & Stuber, G. D. Lateral Hypothalamic Area Glutamatergic Neurons and Their Projections to the Lateral Habenula Regulate Feeding and Reward. *Journal of Neuroscience* **36,** 302–311. doi:10.1523/JNEUROSCI.1202-15.2016 (2016).

163. Trusel, M., Nuno-Perez, A., Lecca, S., Harada, H., Lalive, A. L., Congiu, M., Takemoto, K., Takahashi, T., Ferraguti, F. & Mameli, M. Punishment-Predictive Cues Guide Avoidance through Potentiation of Hypothalamus-to-Habenula Synapses. *Neuron* **102,** 120–127.e4. doi:10.1016/j.neuron.2019.01.025 (2019).

164. Klaus, A., Martins, G. J., Paixao, V. B., Zhou, P., Paninski, L. & Costa, R. M. The Spatiotemporal Organization of the Striatum Encodes Action Space. *Neuron* **95,** 1171–1180.e7. doi:10.1016/j.neuron.2017.08.015 (2017).

165. Sjöbom, J., Tamtè, M., Halje, P., Brys, I. & Petersson, P. Cortical and striatal circuits together encode transitions in natural behavior. *Science Advances* **6,** eabc1173. doi:10.1126/sciadv.abc1173 (2020).

166. Van der Meer, M. A. A., Johnson, A., Schmitzer-Torbert, N. C. & Redish, A. D. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67,** 25–32. doi:10.1016/j.neuron.2010.06.023 (2010).

167. Mizumori, S. J. Y., Ragozzino, K. E. & Cooper, B. G. Location and head direction representation in the dorsal striatum of rats. *Psychobiology* **28,** 441–462. doi:10.3758/BF03332003 (2000).

168. Hinman, J. R., Chapman, G. W. & Hasselmo, M. E. Neuronal representation of environmental boundaries in egocentric coordinates. *Nature Communications* **10,** 2772. doi:10.1038/s41467-019-10722-y (2019).

169. Berke, J. D., Breck, J. T. & Eichenbaum, H. Striatal Versus Hippocampal Representations During Win-Stay Maze Performance. *Journal of Neurophysiology* **101,** 1575–1587. doi:10.1152/jn.91106.2008 (2009).

170. Wiener, S. Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *The Journal of Neuroscience* **13,** 3802–3817. doi:10.1523/JNEUROSCI.13-09-03802.1993 (1993).

171. Yeshenko, O., Guazzelli, A. & Mizumori, S. J. Y. Context-Dependent Reorganization of Spatial and Movement Representations by Simultaneously Recorded Hippocampal and Striatal Neurons During Performance of Allocentric and Egocentric Tasks. *Behavioral Neuroscience* **118,** 751–769. doi:10.1037/0735-7044.118.4.751 (2004).

172. Zhou, S., Masmanidis, S. C. & Buonomano, D. V. Neural Sequences as an Optimal Dynamical Regime for the Readout of Time. *Neuron* **108,** 651–658.e5. doi:10.1016/j.neuron.2020.08.020 (2020).

173. Akhlaghpour, H., Wiskerke, J., Choi, J. Y., Taliaferro, J. P., Au, J. & Witten, I. B. Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. *eLife* **5,** 1–20. doi:10.7554/eLife.19507 (2016).

174. Bakhurin, K. I., Goudar, V., Shobe, J. L., Claar, L. D., Buonomano, D. V. & Masmanidis, S. C. Differential Encoding of Time by Prefrontal and Striatal Network Dynamics. *The Journal of Neuroscience* **37,** 854–870. doi:10.1523/JNEUROSCI.1789-16.2017 (2017).

175. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience* **22,** 1677–1686. doi:10.1038/s41593-019-0502-4 (2019).

176. Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M. & Harris, K. D. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364,** eaav7893. doi:10.1126/science.aav7893 (2019).

177. Smith, J. B., Klug, J. R., Ross, D. L., Howard, C. D., Hollon, N. G., Ko, V. I., Hoffman, H., Callaway, E. M., Gerfen, C. R. & Jin, X. Genetic-Based Dissection Unveils the Inputs and Outputs of Striatal Patch and Matrix Compartments. *Neuron* **91,** 1069–1084. doi:10.1016/j.neuron.2016.07.046 (2016).

178. Tepper, J. M. & Lee, C. R. in *Progress in Brain Research* 189–208 (Elsevier, 2007). doi:10.1016/S0079-6123(06)60011-3.

179. Morita, K., Morishima, M., Sakai, K. & Kawaguchi, Y. Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends in Neurosciences* **35,** 457–467. doi:10.1016/j.tins.2012.04.009 (2012).

180. Cui, Q., Du, X., Chang, I. Y. M., *et al.* Striatal Direct Pathway Targets Npas1 + Pallidal Neurons. *The Journal of Neuroscience* **41,** 3966–3987. doi:10.1523/JNEUROSCI.2306-20.2021 (2021).

181. Okamoto, S., Sohn, J., Tanaka, T., Takahashi, M., Ishida, Y., Yamauchi, K., Koike, M., Fujiyama, F. & Hioki, H. Overlapping Projections of Neighboring Direct and Indirect Pathway Neostriatal Neurons to Globus Pallidus External Segment. *iScience* **23,** 101409. doi:10.1016/j.isci.2020.101409 (2020).

182. Freeze, B. S., Kravitz, A. V., Hammack, N., Berke, J. D. & Kreitzer, A. C. Control of basal ganglia output by direct and indirect pathway projection neurons. *Journal of Neuroscience* **33,** 18531–18539. doi:10.1523/JNEUROSCI.1278-13.2013 (2013).

183. Tecuapetla, F., Matias, S., Dugue, G. P., Mainen, Z. F. & Costa, R. M. Balanced activity in basal ganglia projection pathways is critical for contraversive movements. *Nature Communications* **5,** 4315. doi:10.1038/ncomms5315 (2014).

184. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535,** 505–510. doi:10.1038/nature18942 (2016).

185. Lee, R. S., Mattar, M. G., Parker, N. F., Witten, I. B. & Daw, N. D. Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. *eLife* **8.** doi:10.7554/eLife.42992 (2019).

186. Saunders, B. T., Richard, J. M., Margolis, E. B. & Janak, P. H. Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. *Nature Neuroscience* **21,** 1. doi:10.1038/s41593-018-0191-4 (2018).

187. Fetz, E. E. Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and Brain Sciences* **15,** 679–690. doi:10.1017/S0140525X00072599 (1992).

188. Churchland, M. M. & Shenoy, K. V. Temporal Complexity and Heterogeneity of Single-Neuron Activity in Premotor and Motor Cortex. *Journal of Neurophysiology* **97,** 4235–4257. doi:10.1152/jn.00095.2007 (2007).

189. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503,** 78–84. doi:10.1038/nature12742 (2013).