Parallel loops of behavioural control by opponent striatal circuits

Gonçalo Dias Pereira Guiomar

A Dissertation
Presented to the Faculty
of Universidade Nova de Lisboa
in Candidacy for the Degree of
Doctor of Philosophy

Department of Neuroscience
Champalimaud Centre for the Uknonwn
ITQB

June 2023, Lisbon, Portugal

O que separa o mundo interior do exterior não é uma barreira, sim um diafragma - a superfície transparente onde afinal se anula a distinção entre interior e exterior. O que está por dentro anuncia, denuncia, prenuncia, pronuncia o que está por fora. E ao contrário. Este diafragma imaginário não põe em comunicação o mundo interior com o exterior, o que pressuporia isolamento ou ruptura de ritmo - mas comunica.
Herberto Helder - Photomaton & Vox (1979)

Para o meu avô.

Acknowledgments

Firstly, I would like to thank Joe for his supervision and support throughout this project. His strong vision and motivation were crucial, and his intuition has been a wonderful guide in the development of the core elements of my modeling work. He has taught me not only how to be a good scientist but also how to think through things when they're less clear.

I would also like to thank Christian for helping me be more precise in my thinking and for his constructive feedback during this project, which was essential for me to be able to communicate clearly the ideas in this thesis. Daniel has also been instrumental in showing me a whole ocean of neuroscientific theory that I was unaware of and on which I have taken sail.

I would also like to thank my lab, which were fundamental in allowing me to also work through this thesis. In particular, to Bruno, whose incredible dedication was always an inspiration; Daniela for being such a good presence in my initial moments in the lab; Filipe for his beautiful plots and pun mastery; Margarida for having a good ear for all my complaining; for Cristina, whose ability to pierce through things and early departure has left a huge impression on me; Ben for his wonderful artwork; and Bass for showing me things that lay beneath my blind spots.

My thesis committee was also very important throughout this process; Alfonso, through his incredible kindness, support, and occasional musical wanderings; and Eugenia, through her attentive direction and reassurance.

To my INDP class; to Juan, Tiago, Rita, Ana, Rhadika, and Oihane. The first year of the PhD was one of the most incredibly stimulating years of my life and will forever stay in my memory as one of the periods where I felt the most space to explore ideas. Doing it with all of you was great fun. Also, I'd like to thank the CCU community in general, where I've spent one of the most important periods of my life. It has profoundly shaped not only my understanding of science but also my sense of community and what is possible to build with a group of people.

To Zach, one of the most unexpected and profound friendships that I developed over these years. Our interactions have deeply affected my view of what two minds can explore together. Besides being a friend, he's also been a great mentor and an inspiration, having not only shaped my path in science, but also in life in general. I'm also thankful to Susana, who, alongside Zach, was a second family to me. I'll take in my heart all the beautiful moments we've shared, with the certainty that there will be many more.

I'd like to thank Tor for so many interesting and deep scientific conversations, as well as having shared so many great moments making music; to Gautam for being such an incredibly deep and kind and inspirational soul and lastly to Quendera and Michael, the two first friends I made in the Champalimaud, and with whom I've grown tremendously.

I'd like to thank João for his unconditional support, for being such a kind and understanding friend, and for providing reassurance that I'm not (completely) insane (sometimes). To Miguel, to whom I also share deep gratitude for his support and motivational words. When the going got tough, he was a constant inspiration for how to be a wonderful human being.

To Joana, for her deep understanding and ability to reflect back to me so many different perspectives that I missed, and that were crucial in shaping who I am today and to Jonathan and André, I'd like to thank you for your openness and willingness to take part in some weird ideas I had and thought would be fun to pursue. You've both shaped my view on life and art in ways that will stay with me forever, both as mentors and as friends.

To Betsy, I'd like to thank for being such a kind, attentive, and open person. I've been able to think more clearly by her questioning and way of seeing and experiencing things. To Inês, with whom I found an unexpected family, whose relentless support I am and will be forever grateful for. You've shaped my growth into being more kind, understanding, and loving, and your ability to see beyond unconscious walls has been a skill that I'm still trying to master.

Para os meus pais, não é possível honrar ou agradecer de uma maneira satisfatória o apoio incondicional que me deram ao longo da vida. Ficarei para sempre grato pelos sacrifícios que fizeram por mim e pela compreensão e liberdade que me deram para seguir os meus sonhos. Para a minha avó, pela sua presença e cuidado e carinho. Para a minha prima Bela que, sabendo ou não, foi quem começou tudo isto ao me ter aberto a mente e o espírito à busca do conhecimento e à beleza do mundo. Finalmente, para o meu avô, a quem esta tese é dedicada e talvez até a razão principal pela qual ela aconteceu.

As palavras nunca conseguem conter as coisas a que se associam. Para todos os nomes acima esse foi o caso. Mesmo sem palavras que cheguem, ficam sempre as marcas profundas que os outros nos deixam, que são também os lugares de onde brotamos com mais intensidade.

Título

Espirais paralelas de controlo comportamental por circuitos estriatais opostos

Resumo

A execução de ações motoras é uma forma fundamental de interação dos animais com o seu ambiente. Esta interação envolve a seleção de ações apropriadas para alcançar um objetivo relevante para o animal. As vias diretas e indiretas nos gânglios basais desempenham um papel crucial na aprendizagem, execução e inibição de sequências de ações. No entanto, os seus papéis funcionais precisos nestes processos ainda não estão claros.

Neurónios que expressam receptores D1 e formam a via direta parecem lidar com Erros de Predição de Recompensa (EPRs) positivos, com a sua excitação optogenética levando a um aumento do movimento. Em contraste, os MSNs que expressam D2, que formam a via indireta, tornam-se mais excitáveis com níveis tónicos mais baixos de dopamina (EPRs negativos) e inibem o movimento. Uma hipótese atual sobre a função estriatal propõe que estas duas populações representem sinais de Ir e Não Ir para padrões motores específicos. No entanto, esta hipótese contradiz outros resultados que mostram que ambas as populações estão ativas no início da ação, levantando a questão: como podem estas duas populações estar simultaneamente ativas com funções aparentemente opostas?

Nesta tese, fornecemos uma descrição mecanicista de como a dinâmica nestas duas vias leva à seleção e supressão de ações, conciliando estes resultados aparentemente díspares. Desenvolvemos um agente de Aprendizagem por Reforço Hierárquico Multiagente que imita o comportamento e a atividade neuronal de ratos a realizar uma tarefa de discriminação temporal baseada na supressão de movimento. Ao incorporar dados neuronais do estriado dorsolateral (DLS) e dados comportamentais de perturbações optogenéticas do estriado dorsomedial (DMS), demonstramos como diferentes laços corticoestriatais, representados por agentes de aprendizagem por reforço concorrentes, podem interagir para moldar o comportamento e a atividade neuronal observados. Por fim, teorizamos que as representações de estado subjacentes no DMS podem resultar de um processo de abstração de estado que utiliza informações das representações no DLS, sugerindo uma nova enquadramento teórico do papel funcional destes circuitos.

Abstract

The execution of motor actions is a fundamental way in which animals interact with their environment. This interaction involves selecting the appropriate actions to achieve a relevant goal for the animal. The direct and indirect pathways in the basal ganglia play a crucial role in the learning, execution, and inhibition of action sequences. However, their precise functional roles in these processes are still unclear.

Neurons that express D1 receptors and form the direct pathway appear to handle positive Reward Prediction Errors (RPEs), with their optogenetic excitation leading to increased movement. In contrast, D2-expressing MSNs, which form the indirect pathway, become more excitable with lower tonic levels of dopamine (negative RPEs) and inhibit movement. A current hypothesis about striatal function proposes that these two populations represent Go and No Go signals for specific motor patterns. However, this hypothesis contradicts other results showing both populations are active at action onset, raising the question: how can these two populations be simultaneously active with seemingly opposite functions?

In this thesis, we provide a mechanistic description of how the dynamics in these two pathways lead to action selection and suppression, reconciling these apparently disparate results. We developed a Multi-agent Hierarchical Reinforcement Learning agent that mimics the behavior and neuronal activity of mice performing a movement-suppression-based temporal discrimination task. By incorporating neuronal data from the dorsolateral striatum (DLS) and behavioral data from optogenetic perturbations of the dorsomedial striatum (DMS), we demonstrate how different corticostriatal loops, represented by competing reinforcement learning agents, might interact to shape the observed behavior and neuronal activity. Finally, we theorize that the underlying state representations in the DMS could result from a state-abstraction process utilizing information from the representations in the DLS, hinting at a potential new theoretical framing of the functional role of these circuits.

Author Contributions

Chapters 2 and 3

J. Paton. and B. Cruz conceived of the experiments. S. Soares assisted in analysis of the fibre photometry data. B. Cruz performed all experiments and analyzed the data. G. Guiomar., B. Cruz. and J. Paton conceived of the model with assistance from A. Motiwala. G.Guiomar built the multi-agent model and performed all simulations. C. Machens. provided constructive feedback about the model and helped revise the methods. J. Paton supervised all aspects of these sections.

Chapter 4

D. McNamee and G.Guiomar contributed to the conception of the theoretical framing for the state-abstraction mechanism. G. Guiomar built the model, did the derivations and performed all simulations. J. Paton supervised the conceptualization of the model.

Financial Support

This work was funded by Fundação para a Ciência e a Tecnologia (FCT) and Fundação Champalimaud. Gonçalo Guiomar is a student of the International Neuroscience Doctoral Program (INDP) 2017 and was supported by the Ph.D. fellowship PD/BD/128301/2017 from FCT.

Overview

This thesis is composed of five main chapters. In the first chapter, I present a general introduction, where a review of the literature relevant for Chapters 2 and 3 is made. In this introduction, I start by framing the general problem of action selection that biological agents are faced with and how action suppression plays a pivotal role in its solution. The remaining of this first chapter is dedicated to reviewing how a set of brain areas, so-called basal ganglia, function to support the process of action production and suppression. I will start with a short account of the basic functional anatomy of the circuit, its general organizing principles, and relevant models of its function, as well as the current state of the art of computational models for these circuits.

In Chapter 2 I present the experimental work performmed by Bruno Cruz, which will serve as the basis for all the modelling work done in the following chapters. In it, I will describe the behavioral paradigm and the cell-type-specific recordings of activity and subsequent manipulations of a set of neuronal populations chosen to probe the functional role of the direct and indirect pathways of the basal ganglia.

In Chapter 3 I'll present the modelling work I did for the dataset presented in Chapter 2. I'll build the model step by step and outline the reasoning behind the choices done. I'll show how simpler models are not sufficient to model the data and how we the new elements contribute to the successful mechanistic explanation of the observed behavior, activity of the direct and indirect pathway neurons and the optogenetic inhibition experiments performed on two distinct regions of the striatum.

In Chatper 4 I'll define a novel theoretical framing for the learning of the state representations based on an Information Theoretic perspective. I'll start with a toy model that makes explicit the underlying modelling structure that we're after and then present an integrated state-abstraction algorithm that is able to express the results obtained in Chapter 3.

In the final chapter, I'll discuss how the modelling work done in Chapters 3 and 4 fits in the overall scheme of basal ganglia theory, as well as open up a discussion to a novel approach regarding our understanding of the circuits under study and how this might be relevant not only for our understanding of the brain but also as a way of developing novel machine learning algorithms.

Table of Contents

List of Figures

Chapter 1

Introduction

(...) in the same way that each plant is a terrestrial star gazing at the sky, so also every star is a celestial plant in spiritual form, which only differs from the terrestrial ones by its matter... the celestial plants are turned towards the side of the earth and look directly at the plants they have procreated, infusing them with their particular virtues. - Crollius *Traité des signatures (1624)*

The human desire to understand oneself has historically necessitated the drawing of dichotomous lines between things. Just as the line between night and day was crucial for ancient civilizations to understand time and the coming and going of the seasons, the separation between an organism and its environment was instrumental in dissecting the mechanisms of life. Analogous to this, other lines have also been drawn between the self and the world, which have allowed us a myriad of intuitions as to the mechanisms of the world. The creation of myths and the symbols that support these myths present to us a fine thread of the constant struggle between dichotomizing and uniting concepts and entities. The Roman conception of Janus, a double-headed God who represents change and passages, is

one such example. This history of this deity traces back to previous conceptions of a dual entity such as Isimud in ancient Mesopotamia and Culsans in the later Etruscan civilization, which still preserved the same symbolic quality of being the bearers of some passageway through time and consequently, changed.

Resemblance, as Foucault puts it Foucault (1970), was one of the fundamental mechanisms by which Western knowledge was drawn up until the 16th century. It guided the interpretation of texts and the organization of symbols, by allowing the world to unfold as a set of mirrors unto themselves. Thus, representations served as pointers to what was already present in some way or form, serving as a record of a particular similarity. In this sense, change kept within itself the record of its past, and as the dual face of Janus indicates, as things change we simultaneously look forward and backward in equal measure, maintaining with it multiple identities within one single entity.

However useful this concept may have been, what we saw happening was a sharp shift from a Renaissance episteme that emphasized resemblance and similitude to a Classical episteme, which emphasizes representation, order, and identity, consisting in the insertion of specific dichotomies that separated entities which were previously united. One of the most crucial dichotomic anchors was laid with the emergence of the Cartesian framework in the 17th century, established by René Descartes. His philosophical stance initiated what would be one of the most influential frameworks used to understand experience and along with it, cognitive processes. His separation of the mind and body seemingly allowed one to ask more precise questions as to the nature of perception, given that this process was now a direct relationship with something external and objective, and as such separated two systems which were until then thought as one, allowing for specific mechanisms to be uncovered for each specific system.

This perspective, deeply rooted in Western reductionist thought, became the backbone of numerous posterior philosophical systems, which are still determinant in the ontologies used in contemporary science, with the most contemporary version of it being expressed in the stimulus-response paradigm that permeates most of neuroscience. However, even though this system has been quite productive in generating hypotheses regarding the nature of perception and the supporting neural activity, the apparent clarity also introduced a certain rigidity which arguably also limits our holistic understanding of our perception of the world and the behavior of organisms in general.

One parallel line of thought which opened up a slightly different approach was taken by Arthur Schopenhauer, later in the 18th century. He challenged

Figure 1.1: Nicolas Poussin's (1594-1665) *Dance to the music of time* with the depiction of a statue of the roman God Janus on the left.

this narrative with a profound assertion: the world, he argued, is not divisible into disparate entities such as the ones presented by Descartes, bringing back the Janusian duality mentioned in the beginning. For him, within our will to act in the world is also our perception of it, or as he put it, how we represent it internally. For Schopenhauer, these two concepts were inexorably linked and fed each other in a constructive cycle. Acting one's will influences our representation of the world as much as this in turn shapes how we'll act in the future. This brought to light an essential query: if will and representation are so interwoven, how does one's intention shape and constrain their perception of the world? If that is true, and given that each organism seems to have different goals within a certain environment, could it be that each organism also has a different perception of the world? What can we now make of objective reality and our understanding of animal behavior within it?

Jakob von Uexkull, a contemporary of Schopenhauer, proposed an interesting concept to aid in understanding this: the "Umwelt". During this period, it was thought that most non-human organisms didn't have an internal world with concrete percepts, being at their best simple reflex machines which if understood through their reflex mechanisms would not have any more complexity underneath. Uexkull's proposal recognized that organisms, due to their inherent perceptual limitations, interact with the world through a unique restricted lens, and in fact contained within themselves a *distinct perceptual universe*, tailored to their ecological niches and necessities. As an example, the colors of flowers in spring might communicate a sense of complex beauty to us, but for other organisms that share look at the same meadow, each color and shape span a highly specific and private symbolic space of action and avoidance on which all living beings must operate in order to survive.

If our interaction with the world is then so inherently constrained, it begs the question: How does our subjective reality relate to the objective world? William James, a late 20th-century psychologist and philosopher, presented an answer by suggesting that this limited empirical experience isn't just our reality — it's reality itself. Attempting to conceptualize beyond these confines, James posited is an exercise in metaphysical futility. He saw the "Umwelt" not as a limiting lens, but as a direct perceptual link to reality. James' theory naturally led to the emergence of pragmatism and catalyzed the birth of ecological psychology. Within this perspective, all perceptions become "direct accesses" to reality, defined and shaped by their ecological relevancies.

This conceptualization of experience would lead to the emergence of a field

now known as *Ecological Psychology*. This ecological interpretation of perception, i.e. one that takes into account the living function of an organism within its ecological realm, later on led to the conceptualization of "Affordances" by James Gibson. Here, the functional significance of an object to an organism is seen in relation to the ecological backdrop of its existence. Echoing Schopenhauer's intertwined notion of "*Will and Representation*," Gibson's Affordances posits that entities carry within them a dual significance: their intrinsic characteristics and their utility to the observer, both in a continuous cycle of transformation. This dynamic is shaped by the ecological web of relationships between the observer and the world at large.

The work presented in this thesis aims at continuing this line of thought whilst zooming in to the brain itself, and to the underlying circuits that might implement these action-perception cycles. In a similar way to Poussin's dance (see Figure 1), we'll focus on the music of time and how its passage determines our encoding of change within an organism's internal representation of the world. To an extent, we aim to provide a possible extension and interlinkage of the Uexkülian and Gibsonian frameworks of Umwelt and Affordances, by applying them to dynamic and task-specific representations. Instead of looking at representations themselves as mere compressive schemas of the world, we will take into account their niche ecological utility for the animal, while providing, like the two faces of Janus, a non-dichotomic expression of how thought and action might operate within the brain.

What will unfold in the following chapters is also the development of a hypothesis; the idea that within the brain itself, the same mechanisms occur i.e. not only do individual organisms of specific species have their unique worlds somewhat shielded from other species, but within these perceptual realities lies an even deeper *interior ecology of perception* shaped by each organism's nervous system. In a sense, what will be presented here will be a sort of Umweltian approach to specific neural sub-circuits, with the hopes that through building this resemblance between the system as a whole and the systems within that system, one will be able to build sharper hypothesis regarding brain function.

What we shall see in the following sections is that the brain, while a hub of interconnectedness, houses somewhat separable circuits, each responsible for distinct action-perception dynamics, each offering specific perceptual keyholes that seem to form a potential *hierarchy of abstraction* to represent the world.

This intricate design seems to suggest the existence of myriad perceptual realms even within an organism, all converging to maintain the organism's autopoietic

(self-sustaining) stability (Maturana and Varela, 1980). We will concern ourselves with one of the most over-arching circuits that allow for these cycles; the corticostriatal-thalamocortical loops.

In the following sections of Chapter 1, we will open up its anatomical intricacy to find a highly regular and modular set of circuits, on which Schopenhauer's conception of the inseparability of action and perception seems to be somewhat undeniable. We will also dive into previous attempts at modeling these circuits and what computational paradigms have emerged to explain what seems to be a vastly complex landscape of experimental results regarding animal behavior. In Chapter 2, the core experimental results on which the modeling work done in this thesis has been based on will be presented. The latter has been spread throughout Chapter 3, where we present a multi-agent model of these corticostriatal circuits and where this initial proposition of an interior ecology of perception first appeared. This idea and its relationship with Gibson's *Affordances* will be fully explored in Chapter 4, where we will further expand on the idea of an interior ecology of perception by developing a potential mechanism that could generate more abstract representations of the world by taking into account how observations are directly linked with actions.

1.1 Untangling the fabric of volition

> "*...the knowledge of the sixteenth century was a knowledge by comparison, by differentiation; it was a knowledge of identities, of designations, of assimilations. It ranged things in accordance with their resemblances. The great network of similitudes was a vast, open book that lay before the gaze; resemblances were then more than identities, marks, or even signatures—they were illuminations.*" - (Foucault, 1970)

It is likely that the existence of basal forebrain structures at the base of the cerebral hemisphere has been known since ancient times. In fact, Claudius Galenus (129-201 BC), a prominent Greek physician, anatomist, and philosopher who made significant contributions to the field of medicine and wrote extensively on various medical topics, observed the presence of these structures. However, it wasn't until Andreas Vesalius (1543), a renowned Flemish anatomist, published his treatise *De Humani Corporis Fabrica* ([1]) that a preliminary depiction of the basal ganglia

[1] translates to "On the Fabric of the Human Body"

emerged in the literature. Vesalius presented a remarkably detailed representation of the various components of the soon-to-be-denominated basal ganglia, but still without naming these structures or discussing their potential function.

Thomas Willis was the first to anatomically identify distinct subcortical structures in 1664 (Willis, 1664), using the term *corpus striatum* to describe a large group of subcortical elements. The term *striatum* comes from the Latin word "striatus," which means "striped" or "furrowed", a name that was chosen likely due to the presence of nerve fibers that run in parallel along its length, giving it a distinct banded appearance.

The corpus striatum was started to be thought of as a nodal point in the control of both motor and sensory information (matching Galen's *sensus communis*) and around this time the view of the subcortical or striatal origin of motor behavior began taking shape as Willis thought that the marked striatal atrophy he noted while dissecting the brain of patients who suffered from paralysis supported the corpus striatum's involvement in the control of voluntary movement. Willis was also the first to clearly distinguish anatomically the corpus striatum from the thalamus, which he called *Thalami nervorum opticorum.*

What is known now as the *putamen* and the *caudate nucleus* were not initially associated with each other, but instead, the *putamen* was associated with the *pallidum* in what was called the nucleus *lenticularis* or *nucleus lentiformis.* Later in the 19th century, German neuroanatomist Karl Friedrich Burdach made significant contributions to the identification and naming of various basal ganglia components in the *Anmerkungen* section of his book *Vom Baue und Leben des Gehirns* (Burdach, 1819–1826). The anatomical link of the striatum with its primary targets, the *pallidum* and the *substantia nigra* was also discovered around this time.

It is also interesting to note the evolution in the conceptual framing that this region has elicited in those who've dived in its complexity. For Burdach in particular, the basal ganglia were the site of sensory perception and consciousness. Volition, he thought, emerges from the corpus striatum, whereas sensation, particularly visual, and consciousness originate from the thalamus (Burdach, 1819–1826). Although rather simplistic for today's understanding, the idea of the striatum as being related to volition is quite remarkable given how little was known in terms of direct functional mapping of this region to any specific behavior.

Around the same period, James Parkinson started describing a particular set of conditions that expressed themselves as motor impairments, he wrote:

Involuntary tremulous motion, with lessened muscular power, in parts

not in action and even when supported; with a propensity to bend the trunk forward, and to pass from a walking to a running pace: the senses and intellects being uninjured. (Parkinson, 1817)

What would end up being described as Parkinson's disease was then separated from multiple sclerosis, (Charcot, 1877) and more atypical forms of the disease also were described for the first time (Charcot, 1888). This descriptive specificity and the evolution of staining techniques at the turn of the 19th century allowed for an initial anchoring of these symptoms with the *globus pallidus* (van Bogaert, 1930; Hunt, 1917).

The involvement of the striatum in these conditions was also about to be uncovered in the form of striatal-nigral degeneration, which was initially categorized as a distinct disease but has since been merged into the larger diagnosis of multiple system atrophy (Adams et al., 1964). Subsequently, Parkinsonian states related to striatal pathology were identified in other conditions such as Huntington's disease (Westphal, 1883) and in cases of striatal calcification (Bruyn et al., 1964)(Muenter and Whisnant, 1968).

The 20th century led to another set of remarkable discoveries that would lead to a massive paradigm shift in our understanding of this brain region and the role that neuromodulators played in shaping its function. In the 1950s, researchers identified dopamine as an intermediary in the synthesis of noradrenaline and adrenaline from tyrosine which allowed for a deeper understanding of the 1957 observations by Arvid Carlsson and colleagues that led to the unraveling of the role of dopamine as a transmitter in the nervous system (Carlsson et al., 1957).

Carlsson and his colleagues also made a critical observation that the akinetic effects of reserpine, a drug used to treat hypertension, could be reversed by injecting the dopamine precursor 3,4-dihydroxyphenylalanine (DOPA). This injection was correlated to a recovery of dopamine content in the brain, suggesting that depletion of dopamine, rather than other neuromodulators (such as noradrenaline or serotonin), caused the akinetic state in reserpine-treated animals. This finding supported the hypothesis that dopamine might have a central role in controlling motor function (Björklund and Dunnett, 2007).

In the following year, Ake Bertler and Evald Rosengren, and Sano and collaborators in Japan (Bertler and Rosengren, 1959; Sano et al., 1959), reported that dopamine in the brain was predominantly located in the striatum, which contained little noradrenaline. This discovery further supported the hypothesis that

dopamine might be involved in motor control. In 1959, Bertler and Rosengren's paper inspired Oleh Hornykiewicz and Herbert Ehringer to study the distribution of dopamine in human post-mortem brains. They found that dopamine concentration was consistently reduced in the caudate and putamen of PD patients. These observations provided the first evidence linking dopamine depletion to PD and opened up new avenues for developing treatments for this debilitating disease (Ehringer and Hornykiewicz, 1960).

Advancements in electrophysiological techniques in the 70s allowed for the first steps to be taken in recording neurons from the basal ganglia. Evarts (1968) demonstrated the feasibility of recording subcortical neurons in freely moving animals, a technique that was then used by DeLong (1972) in experiments where single-unit activity was recorded from neurons in the basal ganglia of monkeys performing voluntary limb movements. The study found that the activity of many basal ganglia neurons was related to the direction, amplitude, and velocity of limb movements. Additionally, these experiments suggested that the basal ganglia may be involved in the regulation of muscle tone and the initiation and cessation of movements.

The convergence of the detailed anatomy that was known and these electrophysiological findings sprouted the beginning of what would then turn out to be one of the first waves of functional conceptualization of the basal ganglia. Two concepts in particular, that of parallel thalamocortical loops put forward by (Alexander et al., 1986) and that of competing thalamic inhibition masks by (Mink, 1996) were crucial in defining what could potentially be the underlying motifs by which activity was segregated and functionally related amongst the different regions and nuclei of the basal ganglia, as well as providing a framework to interpret how dopamine would also affect deferentially in different regions of the striatum.

Recordings from individual dopamine neurons in the *ventral tegmental area* (VTA) of monkeys performing a task in which they learned to associate visual stimuli with a juice reward showed that dopamine neurons increased their firing rate when an unexpected reward was presented, but not when the reward was predicted or omitted. The authors also found that the magnitude of the response was correlated with the size of the reward and the degree of uncertainty in its delivery. Finally, the authors showed that dopamine neurons could learn to respond to a new, previously neutral stimulus when it was paired with a reward (Schultz et al., 1997).

A clear link was then presented that showed how a *temporal-difference* (TD) learning rule was able to capture not only the magnitude but the temporal dynam-

ics of the activity of these neurons as the conditioning was performed. The link of dopamine to the function of the striatum opened up a highly foundational research path across various animal models, culminating with the emergence of one of the most influential theoretical frameworks that are currently used in computational neuroscience, that of the *Reward Prediction Error* (RPE) theory of dopamine.

1.2 Architecture of the Basal Ganglia

In order to be able to precisely model the basal ganglia we need to understand its underlying structure. In the following section we'll go through a detailed descriptions of the different ganglia and the underlying neuronal types as well as some of the known functional properties of each region.

The subcortical formations that we refer to as the Basal Ganglia consist of multiple sets of both feedforward connections that start from the input to the output layers of this structure which can be seen in Figure 1.2 but also a varied and complex lateral and feedback structure that connect both within the inner ganglia and brain regions outside the basal ganglia (Abdi et al., 2015).

These include input nuclei, such as the **striatum** (Str) and **subthalamic nucleus** (STN), output nuclei, such as the internal segment of the **globus pallidus** (GPi) and **substantia nigra pars reticulata** (SNr), and intrinsic nuclei, such as the external segment of the **globus pallidus** (GPe), **substantia nigra pars compacta** (SNc, and **ventral tegmental area** (VTA). Each of these nuclei plays a crucial role in the regulation of motor behavior and cognitive functions, and as mentioned before, their dysfunction has been associated with a range of neurological disorders (Parent and Hazrati, 1995).

In the following sections, we'll describe the main components of the basal ganglia, their afferent and efferent connections, and what neuronal types are more common. Due to the underlying similarities, we'll reference both primate and rodent studies and provide some detail on their functional properties to give some intuition on the role of each of these areas in the promotion or inhibition of movement.

Striatum

The **striatum** (Str), the largest component of the basal ganglia, is integral to receiving inputs from nearly all cerebral cortex areas. Functionally related cortical

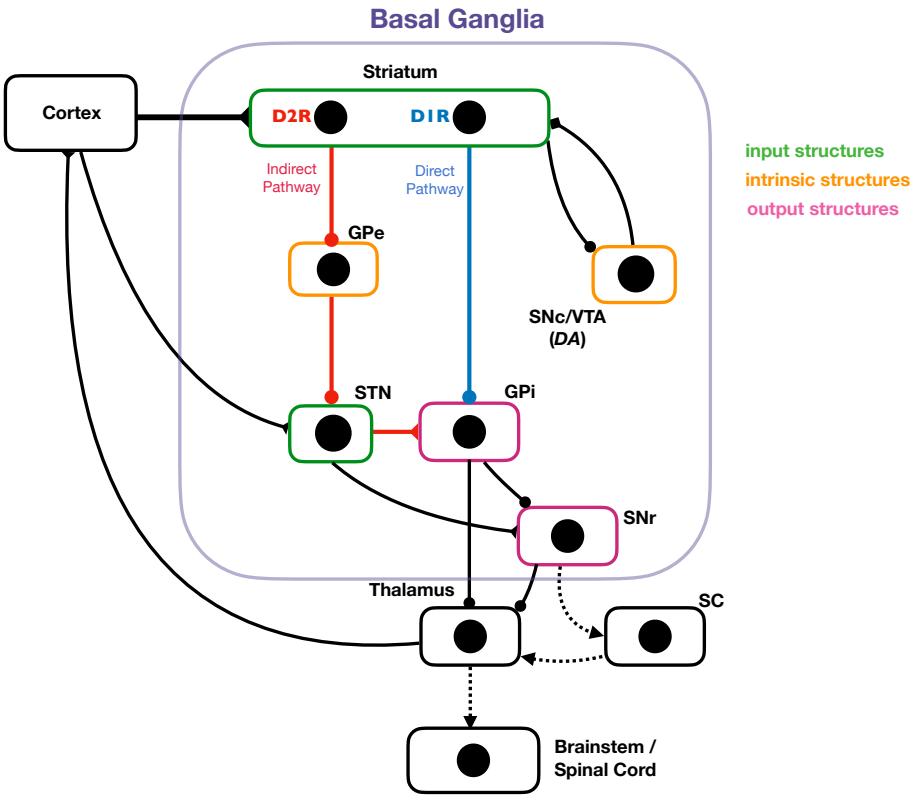


Figure 1.2: Overall structure of the basal ganglia adapted from (Mink, 1996)

areas project to overlapping striatal zones while unrelated cortical areas project to distinct zones, allowing for the integration of inputs and focused output (Mink and Thach, 1993; Smith et al., 1998). It sends inhibitory projections to the basal ganglia output nuclei GPi and SNr, as well as to the GPe (Sato et al., 2000). Two main pathways shape this circuitry and regulate the output of the circuit; the indirect pathway, which encompasses the striatum through GPe and STN, operates in opposition to the direct pathway, which connects directly to the GPi output nucleus (Kita et al., 1999).

The medium spiny neuron, which is the predominant cell type in the striatum, forms a remarkable 95 % of the neuronal population within this region (Gerfen, 1988; Kemp and Powell, 1971). Each of these neurons is capable of processing input from multiple cortical areas and of connecting to a diverse array of additional inputs; excitatory, likely glutamatergic input from thalamic regions, specifically the centromedian and parafascicular nuclei (Lapper and Bolam, 1992; Sadikot et al., 1992) and cholinergic input that originates from the large aspiny striatal neurons (Izzo and Bolam, 1988). They are also recipients of inputs from neighboring medium spiny neurons, which are characterized by the neurotransmitters gamma-aminobutyric acid (GABA), substance P, and enkephalin (Penny et al., 1986). Finally, a significant input comes from dopamine-rich neurons located in the substantia nigra pars compacta (SNc) (Carpenter, 1981; Mink, 1996).

The striatal neurons are divided into two main populations, characterized by their dopamine receptor expression and projection patterns, the D1 and D2 MSNs expressing D1R and D2R receptors, respectively (Gerfen, 1992; Gurney et al., 2015; Iino et al., 2020; Mink, 1996).

The interaction of D1 neurons with the G-protein-coupled adenylyl cyclase system via G_s/G_a G-proteins further influences the transmission of signals within these neurons. In the case of D1 neurons, their activation stimulates adenylyl cyclase, leading to an increase in cyclic AMP (cAMP) levels. This increase in cAMP activates protein kinase A (PKA), which then phosphorylates downstream targets, leading to the opening of certain ion channels which increase the neuron's excitability. Conversely, D2 neurons are coupled to Gi/o-type G-proteins, which inhibit adenylyl cyclase, leading to a decrease in cAMP levels, and thereby less activation of PKA (Steiner and Tseng, 2016).

The D1R neurons innervate the GPi and SNr, two output nuclei of the basal ganglia. Activation of these neurons initiates the release of GABA onto the GPi and SNr neurons. As these output nuclei are already inhibitory in nature, this extra layer of inhibition effectively results in a 'disinhibition' of their target neurons in

the thalamus, meaning that the inhibitory effect of the GPi and SNr on thalamic neurons is lessened (Kreitzer and Malenka, 2008). This disinhibition has a critical outcome: it allows for increased activity of the thalamocortical neurons, ultimately leading to an upregulation of cortical activity. Such effects can be particularly pronounced in motor areas of the cortex, thus leading to the initiation or facilitation of movement (Nambu, 2011).

Conversely, D2R expressing indirect pathway neurons project to the GPe, then to the subthalamic nucleus, and finally to the GPi/SNr. This longer route results in a net increase of inhibitory output from the basal ganglia to the thalamus, thus reducing cortical activity (Cazorla et al., 2014) and the increase in inhibition of motor programs (Mink, 1996).

Besides leading to opposite effects in downstream areas, the plasticity of the MSNs also changes differently depending on the receptors being expressed. Succinctly, D1R neurons increase LTP with higher amounts of dopamine, whilst D2R neurons increase LTP with lower amounts of dopamine (Gurney et al., 2015; Iino et al., 2020).

The striatum is usually divided into dorsal and ventral regions (Alexander and Crutcher, 1990). The dorsal striatum includes the caudate nucleus and the putamen, which have been associated with procedural memory and the regulation of voluntary movement respectively (Graybiel et al., 1994). It can be further subdivided into the dorsomedial striatum (DMS) and the dorsolateral striatum (DLS). The DMS is involved in goal-directed behaviors and decision-making processes (Balleine et al., 2007), while the DLS is implicated in habitual or routine behaviors, often referred to as stimulus-response learning (Yin et al., 2004a; Bornstein and Daw, 2011; Dayan and Berridge, 2014).

The ventral striatum (VS) encompasses the nucleus accumbens (NAc), a critical part of the brain's reward circuit, and the olfactory tubercle (Heimer et al., 1997). The NAc itself can be divided into a core and a shell. The core region has been suggested to play a role in the integration of motivational information, thereby facilitating decision-making and action selection (Mogenson et al., 1980). The shell, with its extensive connections with limbic structures, is involved in processing rewards and motivational states (Di Chiara, 2002). This stratified architecture of the striatum represents the complex and overlapping array of neural connections it maintains with various parts of the brain, orchestrating a broad range of cognitive, motor, and reward-related functions.

In this thesis we'll focus mainly on the DLS and DMS regions in the striatum,

not denying the essential role that the remaining structures have in producing behavior.

STN

The **subthalamic nucleus** (STN) receives inputs from various cortical regions in rodents, including the primary motor cortex, certain areas of the prefrontal cortex, the anterior and medial cingulate cortex, the primary somatosensory region, and, to a lesser extent, the granular insular cortex. These cortical inputs primarily originate from layer V neurons, and some of them have been observed to arise from collaterals of pyramidal tract fibers (Giufrida et al., 1985; Kitai and Deniau, 1981). Additionally, there are cortical neurons that project to both the subthalamic nucleus and the striatum (Féger et al., 1991).

The proportion of cortical neurons sending collaterals to these two structures varies across different cortical areas. For example, the prefrontal cortex has a higher percentage (approximately 40%) of cortical neurons projecting to the subthalamic nucleus or the striatum compared to the anterior cingulate cortex (about 15%) or the primary motor cortex (about 9%) (Féger et al., 1991).

The STN consists primarily of long-axoned projection neurons, whose sparsely spined dendrites can extend beyond 750µm (Rafols and Fox, 1976). These neurons bear resemblance to nigral and pallidal neurons, though their dendrites are considerably thinner (François et al., 1984; Percheron et al., 1984; Rafols and Fox, 1976). The morphological attributes and dimensions of the somatodendritic tree of subthalamic neurons are analogous across rats, cats, and monkeys. However, the nucleus's size varies significantly across these species, and within primates, neurons only occupy about a fifth to a ninth of the total volume (Carpenter et al., 1968b; Hammond and Yelnik, 1983; Rafols and Fox, 1976).

Regarding the projections from the subthalamic nucleus, it has been observed that subthalamic neurons projecting to the putamen and the external globus pallidus (GPe) are mainly located in the dorsolateral two-thirds of the STN. In contrast, those projecting to the caudate nucleus and the internal globus pallidus/substantia nigra pars reticulata (GPi/SNr) are confined to the ventromedial third of the nucleus (Nauta and Cole, 1978; Smith et al., 1990). The STN is known for its potent glutamate-mediated excitatory effect on the two primary output structures within the subcortical nuclei set, namely the GPi and the SNr. It serves as the primary relay nucleus in the indirect pathway, collaborating with

the GABAergic inhibitory direct pathway to modulate the activity of GPi/SNr (Hammond and Yelnik, 1983).

GPe

The **Globus Pallidus Externa** (GPe) is intrinsically linked with the STN, forming a critical loop within the basal ganglia system's overall functionality. Physiologically, this association, as previously shown in classical anatomical studies, facilitates the GPe's role as an intermediary, connecting the striatum with the output structures of the basal ganglia (Carpenter et al., 1968a; Albin et al., 1989; Alexander and Crutcher, 1990). Particularly, GPe's GABA-mediated inhibitory effect is instrumental in modulating the STN, enabling the STN to exert its potent glutamatergic drive on the basal ganglia (DeLong et al., 1985).

In terms of their cellular architecture, both the GPe and the Globus Pallidus Interna (GPi) are composed of sparsely distributed GABAergic neurons characterized by large cell bodies and high expression of the calcium-binding protein parvalbumin. The GPe is primarily innervated by two afferent systems: a GABAergic projection from D2R-containing striatal medium spiny neurons (MSNs) and a strong glutamatergic excitatory projection from STN neurons (Rosin et al., 1998; Bogenpohl et al., 2012; Shink et al., 1996; Joel and Weiner, 1997; Nambu, 2004). These afferent connections underline the integral role of the GPe in the basal ganglia's indirect pathway and its interplay with the STN. Additionally, the GPe also receives minor glutamatergic inputs from the caudal intralaminar nuclei (Kincaid et al., 1991; Marini et al., 1999; Yashukawa et al., 2004).

The GPe's efferent connections extend beyond the STN, reaching most components of the basal ganglia and significantly influencing information processing in the GPi/SNr through its substantial inhibitory projection (Hazrati et al., 1990). This feature establishes dual tiers of integration within the pallidal complex and the basal ganglia circuitry. Relatively recent studies have revealed new efferent pathways from the GPe, mainly to the output structures of the basal ganglia (GPi and SNr) (Hazrati et al., 1990; Kincaid et al., 1991).

GPi/SNr

The **Globus Pallidus internus** (GPi) and Substantia Nigra pars reticulata (SNr) are usually grouped in rodents due to their analogous cyto and chemoarchitectural

characteristics and combined role as output nuclei (Kravitz et al., 2010). In primates Ilinsky and Kultas-Ilinsky (1987), their role as high-discharge output nuclei has been shown. Regarding inputs to the GPi and SNr in rodents, two main types are recognized: D1R-containing MSNs forming the direct pathway and an excitatory glutamatergic projection from the STN, part of the indirect pathway (Mastro et al., 2014). Such diversity in inputs echoes findings in primate studies such as those by Smith and Parent (1986), where substantial inhibition from the direct pathway and excitation from the indirect pathway was observed.

These inputs impact both GPi and SNr neurons simultaneously, much like their primate counterparts (Huerta-Ocampo et al., 2014). These inhibitory GABAergic and excitatory glutamatergic afferent systems converge on basal ganglia output neurons. These output neurons then innervate thalamic and brainstem targets, specifically the superior colliculus and the pedunculopontine nucleus (PPN) (Kravitz et al., 2010; Wall et al., 2013), mirroring the findings of DeLong et al. (1983) in primate studies.

The anatomical layout of the GPi and SNr in rodents contributes to the organization of motor and limbic territories within the basal ganglia, a feature also apparent in primates (Smith and Parent, 1986). In rodents, the striatum's motor territory projects to the dorsal one-third of the SNr, thereby designating it as the SNr's motor territory. Neurons in the dorsolateral area of this region are responsive to primary motor cortex (MI) stimulation, particularly that of the orofacial region, paralleling the findings in primate studies by Kitano et al. (1998). The GPi and SNr have a continuum orofacial region, which receives inputs from the supplementary motor area (SMA) territories of the putamen and alters its activity during task performance (Kita and Kita, 2011). Mirroring primate studies such as Wichmann and Kliem (2004), the striatum's prefrontal territory projects to the rostromedial two-thirds of the SNr in rodents, encompassing the oculomotor territory (Huerta-Ocampo et al., 2014). Lastly, the limbic territory of the striatum connects to the most medial part of the SNr, echoing the primate studies by Haber et al. (1990) (Mastro et al., 2014).

Functionally, the GPi and SNr in both rodents and primates play a crucial role in facilitating and regulating movements and behaviors. At the cellular level, the firing patterns of neurons in these regions integrate various inputs, allowing for the fine-tuning of their outputs, as highlighted in studies by DeLong et al. (1983) and Wichmann and Kliem (2004). They are intimately involved in motor control, with abnormal activity in these regions associated with movement disorders such as Parkinson's disease (Kravitz et al., 2010). Furthermore, the GPi and SNr also

participate in the regulation of eye movements and saccades, with specific territories such as the oculomotor territory implicated in these processes (Hikosaka and Wurtz, 1983). Their role extends beyond the motor domain, with parts of the SNr implicated in limbic functions, including reward and motivation, which were explored in the studies by Haber et al. (1990). These functional aspects underline the critical role the GPi and SNr play within the basal ganglia network in coordinating a range of behaviors and responses.

VTA/SNc

The **ventral tegmental area** (VTA) and the substantia nigra pars compacta (SNc) are vital components of the dopaminergic system, playing crucial roles in various cognitive functions. These include associative and reward-based learning (MacInnes et al., 2016; Schultz, 2013), motivation (Berridge and Kringelbach, 2013; Ranaldi, 2014), memory (Keuken et al., 2014; Hegarty et al., 2013), and cognitive control in decision-making (Colzato et al., 2011, 2010). While some studies consider the VTA and SN/SNc as a unified midbrain DA-complex (D'Ardenne et al., 2012; Krebs et al., 2011, 2009), the majority of research delineates them as distinct structures. Despite potential overlaps, both areas are developmentally, morphologically, and functionally unique entities (Fu et al., 2016; Olszewski and Baxter, 1954). However, the VTA presents more divergence than the SNc, not only at the anatomical level (Zhang et al., 2010; McRitchie et al., 1996; Morales and Margolis, 2017; Swanson, 1982), but also in the range of functions beyond DA-related behavior.

The neuronal population of the VTA and SNc varies, implying differences in their functional complexity. SNc's neural composition is less diverse, majorly filled with dopaminergic neurons and scarcely intermingled with glutamatergic or combinatorial neurons (Root et al., 2016). Contrastingly, the VTA exhibits more heterogeneity. While over 80% of the human SNc comprises DA neurons (Root et al., 2016), only 50-80% of the cells in the human VTA are identified as DA neurons (Root et al., 2016). The remaining are GABA-ergic (30%), glutamatergic (5%) (Pignatelli and Bonci, 2015; Root et al., 2016; Zhang et al., 2017), or combinatorial neurons (Morales and Margolis, 2017; Root et al., 2016). This diversity enables GABA to mediate powerful feedback loops between NAcc and VTA, which is believed to generate motivated behavior in mice (Yang et al., 2018).

Despite sharing connections to mutual regions, the afferent and efferent connections of the SNc and VTA differ in strength, particularly to striatal and cortical

regions (Williams and Goldman-Rakic, 1998; Olszewski and Baxter, 1954; Watabe-Uchida et al., 2012). The SNc projects preferentially to the dorsomedial striatum (DMS), dorsolateral striatum (DLS) and tail of the striatum, receiving extensive GABAergic input from neurons located therein and projecting back to striatal target neurons via dense connections (Watabe-Uchida et al., 2012; Halliday et al., 2012; Parent and Hazrati, 1995). The VTA mainly targets neurons in the ventral striatum, namely the NAcc, it receives inputs from a variety of brain regions, including the dorsal raphe, pallidal regions, the central nucleus of the amygdala, hypothalamic regions, the dopaminergic retrorubral field, and the parasubthalamic nucleus (Watabe-Uchida et al., 2012). Such differentiated connectivity profiles influence the functionality of the nigrostriatal and mesocorticolimbic pathways, impacting cognition, motivation, and addiction (Aransay et al., 2015; Hauser et al., 2017; Yang et al., 2018).

Dopamine plays a crucial role in reinforcement learning and is the main feedback signal that these structures project onto the striatal cellular matrix. Reinforcement is a fundamental process by which organisms change the frequency of their behavior, depending on the valence of the reinforcement stimuli; positively valenced reinforcement signals increase the frequency, while negative valenced signals decrease it. One key aspect of this learning mechanism is that the neurons that respond to these signals seem to code not only the magnitude of this valence but the *Reward Prediction Error* (RPE). The RPE signal has been observed in the firing patterns of DA neurons and changes in DA concentration within the striatum across different species. This differential response in terms of plasticity is gated by the previously mentioned LTP mechanisms that are facilitated by the D1R and D2R expressing MSNs. Through DA-dependent plasticity, the RPE signal strengthens the synaptic connections of actions or stimuli that lead to unexpected rewards in neurons expressing D1, whilst with aversive negative stimuli, we have a sharp decrease of the response of VTA/SNc neurons, generating a decrease in overall release of DA to the striatum, which increases the plasticity of D2R MSNs. This synaptic modification increases the likelihood of repeating or inhibiting actions associated with positive or negative outcomes, respectively (Kravitz et al., 2010).

1.2.1 Disinhibition of motor programs

Looking at the individual components of this circuit has given us some evidence as to their potential effect in the modulation of movement, given the repeating

motif of cortico-motor areas connecting directly to its various nuclei. We'll briefly describe the current frameworks that have been used to understand the functioning of this circuit, mostly through the lens of disinhibition of motor programs (Mink, 1996; Grillner et al., 2005; Gurney et al., 1998).

The output from the basal ganglia (GPi/SNr) is inhibitory and projects to motor and premotor areas in the brainstem and thalamus. Given that there are no direct connections between the BG and spinal sensory or motor neurons, one influential hypothesis put forward by (Mink, 1996) is that this output creates an inhibitory mask on competing Motor Programs (MPs) that converge unto the thalamus, allowing for relevant motor programs that have lead to rewarding stimuli to be expressed, whilst those competing motor programs which have not led to positive stimuli to be suppressed (Mink and Thach, 1991; Wichmann et al., 1999). Experiments in non-human primate models of Parkinson's disease, a pathology primarily characterized by bradykinesia and caused by BG dysfunction, showed that focal application of the GABA(A) agonist muscimol in GPi improved motor symptoms (Hikosaka and Wurtz, 1985). These studies promote a view of the BG as primarily operating through disinhibition, allowing for the appropriate selection of actions (Mink, 1996; Gurney et al., 1998; Grillner et al., 2005).

The level of disinhibition exerted by the BG is believed to be modulated oppositely by its direct and indirect pathways due to the different modulatory connections of these pathways to their corresponding thalamic outputs (see Figure 1.6). For instance, the direct pathway activity has been observed to cause brief pauses in GPi/SNr, promoting action selection, while the indirect pathway activity increases discharge rates in GPi/SNr, raising the action threshold (Sippy et al., 2015; Tecuapetla et al., 2016). Such functional opposition has been documented not just concerning overall motor output but also in ongoing motor sequence production, reinforcement, and value-based decision-making (Kravitz et al., 2010; Yttri and Dudman, 2016; Tai et al., 2012).

However, the co-activation of direct and indirect pathways observed during action production (Cui et al., 2013) challenges the viewpoint of functional opposition. This led to the idea that the BG enables action production through the simultaneous and concerted activation of the direct pathway (facilitating actions) and the indirect pathway (inhibiting potentially competing actions).

The interaction between the relatively more focused inhibition from the striatum to the GPi and the excitation from the Subthalamic Nucleus (STN) has been suggested to create a "center-surround"-like filter, akin to those found in early visual systems (Kuffler, 1953). While still debated, the functional consequence of

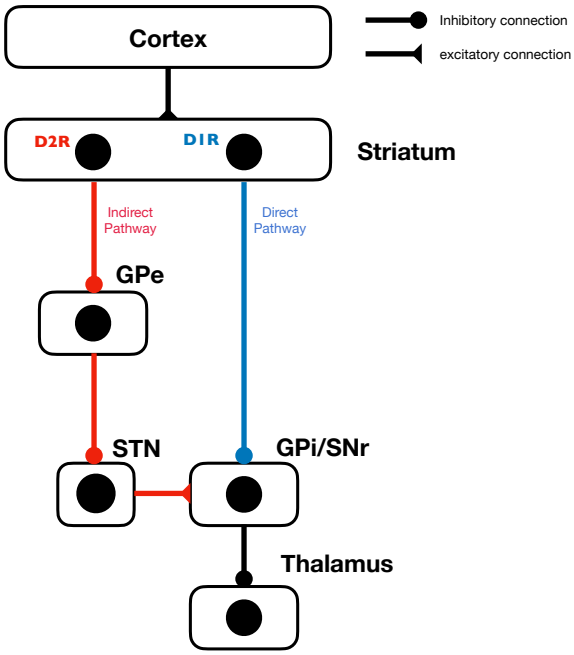


Figure 1.3: Simplified architecture of the basal ganglia.

this wiring could selectively increase the signal-to-noise ratio between actions, supporting the selection of one motor plan over others. This view has been expanded to hypothesize that a similar selection might occur in other parallel loops of the BG, potentially regulating more abstract cognitive and limbic cortical states (Doya, 1999; Hikosaka et al., 2000; Parent and Hazrati, 1995; Foster et al., 2021). Moreover, "selection through disinhibition" strengthens the argument for the BG's role in selecting, rather than generating, actions, thereby contributing to the overall vetoing of motor plans elsewhere in the brain circuitry.

It is the combination of this multiple parallel circuitry, both at a single corticostriatal loop through the direct and indirect pathway at the level of concurrent striatal circuits that connect with different associative, limbic, and motor areas that might indicate a more complex picture for the functioning of this circuit. In this work, we'll explore the possibility that elements from these two viewpoints might not only be present but might be necessary to explain the functioning of this system.

The co-activation of these circuits integrated alongside a push-pull mechanism that leads to the disinhibition of specific motor patterns and inhibition of others might be a fundamental mechanism for action selection in the basal ganglia.

1.2.2 Intracortical networks

Given the complexity of this circuitry, it is important also to understand a bit of the structure that has the highest number of connections to this region; **the cortex**. In order to do so we'll briefly outline the main cortical regions, the underlying subnetworks, how these subnetworks are internally composed, and how they connect with other subnetworks.

The intricate complexity of cognition and behavior is understood as an amalgamation of network-level phenomena, necessitating integration of various sensory inputs, synchronization of multiple motor outputs, and the harmonization of activity within large-scale networks linking these operations (Bressler and Menon, 2010; Sporns, 2010; Swanson and Bota, 2010). Given the diverse amount of sensory modalities, we can parse out the different cortical regions as pertaining to some processing of either a single sensory modality or multiple sensory modalities, which are usually denominated as associative areas. These cortical sub-regions, which form subnetworks of their own, can be classified into four main categories: somatic sensorimotor, medial associative, and lateral subnetworks (Foster et al., 2021; Zingg et al., 2014).

The somatic sensorimotor network comprises both sensory and corresponding primary motor areas. Within the medial associative, visual and auditory networks are arranged in a hierarchical structure, processing information from primary sensory regions to higher-order association areas. Medial and lateral subnetworks, organized along the longitudinal axis of the cerebrum, show distinct patterns converging onto the ventromedial half of the prefrontal cortex (Zingg et al., 2014).

Each subnetwork of the neocortex exhibits a unique topological organization, potentially indicating distinctive information processing strategies (Yamada et al., 2005). However, all subnetworks seem to show a high degree of interconnectivity amongst the constituting sub-regions. The somatic sensorimotor network, for instance, demonstrates strong reciprocal connections among all principal somatic nodes that directly link with orofacial, limb, and whisker sensory neurons. This structure promotes direct interactions between sensory and motor regions, bypassing the need for higher-order association areas. Such an organization could facilitate the swift integration of different sensory modalities to dynamically regulate motor actions.

The medial subnetworks chiefly mediate interactions between sensory and higher-order association areas. The first medial subnetwork facilitates sensory information transmission from visual, auditory, and somatic sensory areas and connects with higher-order association areas (Feierstein et al., 2006; Bucci, 2009; Vann et al., 2009; Weible, 2013). The second medial subnetwork differs topologically from the first, consisting of a multi-synaptic network that provides a structural foundation for transmitting information from the dorsal hippocampus which is related to spatial orientation, navigation, and episodic memory, to the medial prefrontal cortex (Vann et al., 2009; Fanselow and Dong, 2010; Weible, 2013).

The lateral subnetworks function as points of massive convergence in the cortex, focusing on two primary components: the anterolateral insula (AI) subnetwork and the temporal association (TEa), perirhinal (PERI), and ectorhinal areas in the posterolateral temporal subnetwork are the two major regions. Each of these receives input from, and projects back to, numerous cortical areas. For example, the anterolateral insular subnetwork amalgamates gustatory, visceral, and olfactory information, while the posterolateral temporal subnetwork processes predominantly visual, auditory, somatosensory, and motor information. Both then relay this information to the medial prefrontal cortex and the lateral entorhinal cortex (ENTI), supporting the suggested roles of the Al in internal state self-awareness (Craig, 2009), and the TEa/PERI/ECT in perception, object recognition, and emotion-associated contextual memory (Winters et al., 2008; Aggleton et al., 2016).

In general, there seems to be a highly complex and interconnected set of connections amongst these regions, with information flowing bidirectionally between cortical regions connected with direct sensory neurons and higher-order association areas. This recurrent connectivity might be responsible for shaping cognitive processes at multiple levels, allowing for both the higher-order abstract representations to feed and be fed by lower sensory and sensorimotor regions and vice versa (Fuster, 2001). Besides this interconnectivity, we also see a strong convergence of most of the networks to prefrontal regions, showing a convergence of multiple information modalities, which might be responsible for the creation of more abstract representations. The cells in these regions also show longer temporal constants (in contrast with shorter time constants in sensory regions) which have been associated with a functional role of allowing higher-order representations to be shaped in working memory (Fuster, 2001; Zingg et al., 2014; Murray et al., 2014).

1.2.3 The corticostriatal projectome

Ramon y Cajal (1899) was the first to describe the existence of corticostriatal projections and to suggest that these fibers represented collateral branches of one or more of the descending tracts of the internal capsule. Later work showed that the corticostriatal network is a critical pathway that modulates a vast array of neurological functions. We'll focus mainly on the cortical projections to the dorsolateral (DLS), dorsomedial (DMS), and ventral (VS) regions of the striatum.

The somatic sensorimotor cortical areas are primarily linked to the dorsolateral striatum (DLS), forming a motor circuit involved in the control of movement (Kreitzer and Malenka, 2008; Foster et al., 2021; Alexander and Crutcher, 1990). This connection mediates the translation of sensory input into motor responses and has been associated with procedural learning and habit formation. The lateral associative cortical areas, such as the prefrontal cortex, project to the DMS, constituting a cognitive circuit engaged in decision-making and goal-directed behaviors (Yin et al., 2006).

Medial associative cortical areas, including the anterior cingulate and orbitofrontal cortex, form intricate connections primarily with the DMS but also exhibit some degree of overlap with the DLS (Haber and Knutson, 2010). This overlap suggests a more nuanced circuitry, possibly reflecting the integration of different cognitive and emotional functions. This medial circuitry is particularly involved in the evaluation of reward-related information and the adjustment of behavior in response to

changing contingencies, a role highlighted by its dysregulation in conditions such as addiction and obsessive-compulsive disorder (Everitt and Robbins, 2005).

Finally, the ventral striatum, particularly the nucleus accumbens (NAc), receives projections from various cortical and subcortical regions, playing a crucial role in the integration of reward-related, motivational, and decision-making information (Russo and Nestler, 2013). Notably, projections from the prefrontal cortex (PFC) to the ventral striatum are significant, where various sub-regions of the PFC, including the medial PFC (mPFC), orbitofrontal cortex (OFC), and anterior cingulate cortex (ACC), extend their axons to the nucleus accumbens. These connections are critical in regulating cognitive functions such as decision-making, impulse control, and reward expectation (Haber and Knutson, 2010).

In addition to the PFC, the ventral subiculum of the hippocampus sends glutamatergic inputs to the ventral striatum, contributing to the regulation of affective and spatial learning processes. The amygdala, particularly its basolateral part (BLA), also projects to the ventral striatum and aids in processing emotionally salient events (Groenewegen et al., 1999). Moreover, the insular cortex, with its unique role in interoceptive awareness and emotional processing, also projects to the ventral striatum. Thus, the inputs to the ventral striatum represent a richly interconnected network, interweaving numerous aspects of cognition, emotion, and motivation (Russo and Nestler, 2013).

Beyond the striatum, the basal ganglia-thalamocortical circuit is fundamental in mediating information flow. The substantia nigra pars reticulata (SNr) serves as the primary output nucleus of the basal ganglia, relaying inhibitory GABAergic projections to the thalamus (Smith et al., 1998). Specific areas of the thalamus are then involved in the feedback loop to the cortex, particularly the motor and prefrontal areas. Notably, the somatosensory cortex's connections with the DLS and the prefrontal cortex's links to the DMS converge in the ventral anterior/ventrolateral thalamus and mediodorsal thalamus, respectively (Alexander et al., 1986). This connectivity suggests a pervasiveness of the functional segregation of the corticostriatal projections and underscores the integrative role of these circuits in coordinating a broad spectrum of behaviors.

Multiple hypothesis have shaped the way in which the field functionally interpret the connectivity between these areas. The different hypothesis focus on specific details of anatomy, particularly the inherent division of the subcircuits and the differential connectivity patterns between these areas.

The 'parallel processing' (Alexander and Crutcher, 1990; Alexander et al., 1986) hypothesis suggests that different types of cortical information are processed

through the cortico-basal ganglia-thalamo-cortical loop largely independently via multiple, parallel, segregated circuits. Up to five such parallel circuits have been identified: motor, oculomotor, orbitofrontal, dorsolateral prefrontal, lateral orbitofrontal, and anterior cingulate loops (Alexander and Crutcher, 1990). Each of these circuits appears to receive inputs from several separate cortical areas, traverse specific portions of the basal ganglia and thalamus, and project back upon one of the cortical areas providing input to the circuit, thus forming a partially 'closed' loop (Parent and Hazrati, 1995). This organization allows for the processing of distinct information types in a modular fashion while maintaining some degree of functional segregation.

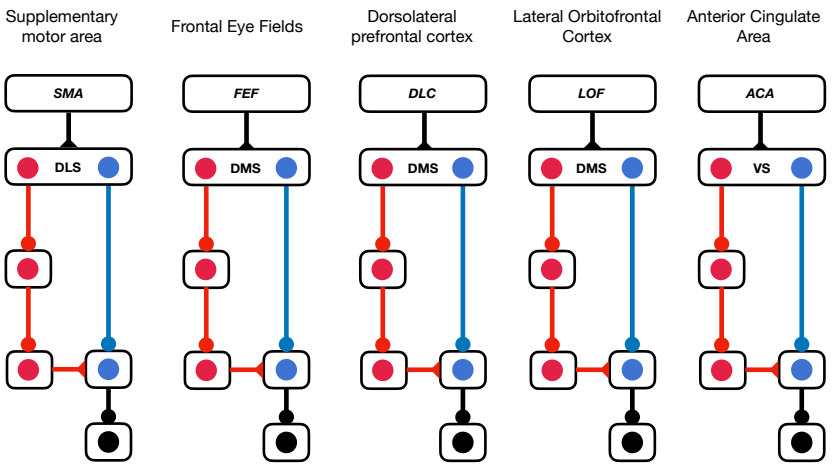


Figure 1.4: Parallel circuitry of the corticostriatal loops in primates, adapted from (Alexander and Crutcher, 1990).

In contrast, the 'information funneling'(Percheron et al., 1994) hypothesis proposes the existence of absolute convergence of cortical information along the cortico-striato-pallido/nigro-thalamo-cortical system. This funneling begins with an 'in-

trication' of axonal endings from widely separate and functionally different cortical regions within the striatum, and is further enhanced by what is called 'reception convergence' at pallidal and nigral levels. This organization suggests a greater integration of information and subsequent narrowing of outputs as they traverse the basal ganglia, ultimately facilitating the selection and control of specific motor and cognitive functions.

1.2.4 Parallel loops of behavioural control

We have so far focused on the neuronal details of each ganglia, and their overall circuit connectivity, and provided some intuitions about their potential functional role in the production of movement. In the following two sections, we'll take a bird's eye view of the structure of these circuits and try to describe a more general set of circuit motifs that will pave the way to a wider theoretical framing that will later be used in the models developed in this thesis. We'll join a series of anatomical motifs, that of the parallel circuitry, recurrent feedback circuitry, and disinhibition of motor programs in order to link the information processing performed by the relevant intracortical circuits to the potential representations learned in the striatum.

As we've seen in the section 1.2.3, cortical regions that share the same intracortical circuits project unto overlapping striatal regions, drawing a somewhat topographical map between the information classes present in the cortex and the striatum (Hunnicutt et al., 2016). This class specificity is maintained across the basal ganglia and in the thalamic projections which feedback the motor-related programs back into the cortex, drawing a set of parallel loops that are specific to each corticostriatal map (Alexander and Crutcher, 1990). However, these circuits are not isolated from each other, having multiple pathways that allow for communication across the different loops. Besides the connectivity amongst different cortical sub-circuits, striatonigrostriatal (SNS) projections across the basal ganglia seem to provide a channel that goes from the lateral to the dorsomedial regions of the striatum.

We've previously mentioned how the dorsolateral prefrontal cortex projects to the dorsomedial striatum and premotor and motor cortex projects to the dorsolateral striatum. However, midbrain projections from the NAc shell target both the VTA and ventromedial SNc. Midbrain projections from the VTA to the NAc shell form a "closed," reciprocal SNS loop. Projections from the medial SN feedforward to the core, form the first part of a spiral. The spiral continues through the SNS

projections, with pathways originating in the NAc core and projecting more dorsally. In this way, ventral striatal regions influence more dorsal striatal regions via spiraling SNS projections (Haber et al., 2000). The anatomy suggests that information seems to flow from the striatal regions usually associated with prefrontal projections such as the DMS to more motor and pre-motor related regions usually connected to the DLS, which seems to suggest, at least partially, that the higher order and potentially more abstract representations might shape the more stimulus bound representations of the direct sensory areas (Bornstein and Daw, 2011; Dayan and Berridge, 2014).

One must say, however, that the story might not be that simple. As we've shown in 1.2.2, the cortex is not only highly interconnected, but it is so in a bidirectional manner, allowing for information flow from higher to lower orders and vice versa. Even assuming that there's some separation between the different corticothalamocortical loops (Alexander and Crutcher, 1990) the high degree of connection degeneracy might allow for information to also pass from striatal areas connected to lower order sensorimotor cortical regions, such as the DLS to associative higher-order areas. The prospect of this directionality leaves a wide room for modeling information flow between different regions, something that we will focus on in the modeling done in Chapter 4.

We thus have a clearer picture of the architectural motifs present in the basal ganglia that might allow us to design more detailed predictive models of its functioning. The picture goes as follows; different cortical regions, which are highly interconnected, project unto specific striatal regions, overlapping by amount of proximity within their intracortical proximity. These striatal regions then encode deferentially the information coming from direct sensory areas and associative higher orders, feeding a set of downstream projections unto the output of the basal ganglia which, due to the direct/indirect split, encodes both excitatory and inhibitory maps for different motor programs. Simultaneously, a spiral-like set of connections project sensorimotor related information within the striatum to prefrontal associated striatal regions. Through experience, these motor programs are shaped, and the outgoing motor responses are fed back via the thalamocortical connections back into their corresponding cortical regions, leading to a new spread and mixture of the information processed in these multiple loops at different levels of abstraction (see Figure 1.5).

We will attempt at deriving a model that captures most of the basic properties of this architecture in the following chapters of this thesis.

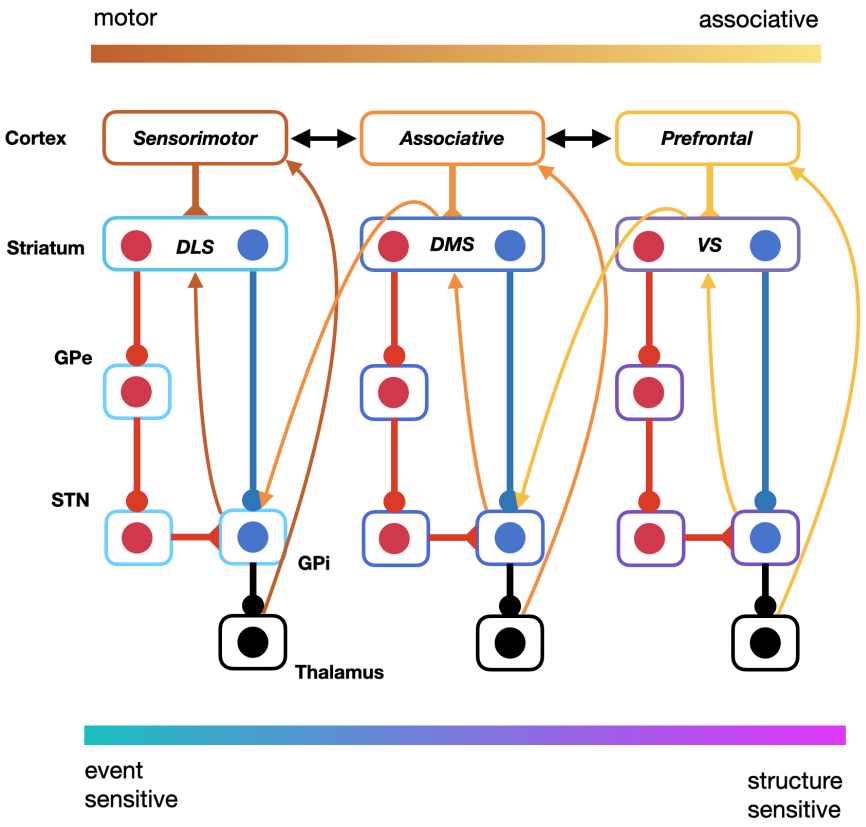


Figure 1.5: Striatonigral loops of the basal ganglia, adapted from (Haber et al., 2000) and (Alexander et al., 1986), with an added lateral to the ventral axis in yellow and an axis of representation abstraction along these circuits.

1.3 Computational models of the basal ganglia

Modelling such a complex and interconnected system requires not only attention to the biophysical details but also a broad high-level perspective as to what might be the purpose of this system. The history of our understanding of the basal ganglia is deeply interlinked with an ever-growing ecosystem of models aimed at unveiling the underlying mechanisms that might explain the functioning of this system. One potential strategy we can adopt in approaching this modeling ecosystem is by considering the highly influential work of David Marr.

Marr focused particularly on the visual system but was able to leave a deep legacy in the way in which we currently approach our understanding of brain systems as information processing systems that are cut through three fundamental *levels of analysis* (Marr, 1982). The main tenet goes as follows; the understanding of any information processing system is incomplete without insight into the problems it faces, and without a notion of the form that possible solutions to these problems can take. These fundamental levels of understanding were defined as follows Marr and Poggio (1977):

1. **computational level**: 'what' is being computed and why.
2. **algorithmic level**: 'how' is the computation being performed
3. **implementational level**: what is the 'physical medium' on which the computation is carried out.

We shall use these levels of analysis as a scaffolding for categorizing the different modeling paradigms that have emerged within the study of the basal ganglia (Gurney et al., 2004) while keeping in mind that these boundaries are not strictly applicable in cases where models present elements of various levels.

Within the **computational level** a set of paradigms have emerged; that of thinking of the basal ganglia as a reinforcement-driven dimensionality reduction (RDDR) system (Bar-Gad et al., 2000), as an action gating(Contreras-Vidal and Stelmach, 1995) or action selection (Berns and Sejnowski, 1998; Redgrave et al., 1999) system and finally as a serial processing system (Joel et al., 2002; Gillies and Arbuthnott, 2000; Bapi and Doya, 1998).

The RDDR model suggests that the basal ganglia perform an efficient reinforcement driven dimensionality reduction of the sensorimotor and higher-order

representations present in the cortex (Bar-Gad et al., 2000). This model aligns with known basal ganglia anatomy and physiology, which suggest a "funneling" structure where the number of neurons decreases from the cortex to the striatum to the GPi (Kincaid et al., 1998; Oorschot, 1996; Percheron et al., 1994) and the absence of mutual inhibition between striatal neurons despite extensive lateral connectivity (Jaeger et al., 1994; Kita et al., 1996; Yelnik et al., 1997). Bar-Gad et al. (2000)'s model suggests that the dimensionality reduction should be influenced by both the statistical properties of the input patterns and their behavioral significance. The relative importance of input is determined by novelty, incentive salience, and the ability to predict reward, signals that can be coded by DA neurons (Redgrave et al., 1999a; Berridge and Robinson, 1998; Robbins and Everitt, 1996; Suri et al., 2001).

An alternative addition to the computational role that the basal ganglia is likely to have is that it serves as a substrate for action selection/gating processes. As seen in section 1.2.1, the thalamic inhibitory tonus seems to serve as a gate that inhibits motor programs from being expressed (Mink, 1996), something that is then possible when the thalamic activity is lowered via the direct pathway inhibition. Models of this type tend to have a strong implementational axis, particularly those that are based on anatomy-inspired network models (Chakravarthy et al., 2010), linking this particular function directly to the relevant nuclei in the basal ganglia.

Simulation studies presented by (Contreras-Vidal and Stelmach, 1995) bolster the hypothesis that the pallidothalamic outputs act as gating signals, phasically modulating the initiation and execution of movement. They also theorize that Parkinson's Disease (PD) disrupts these signals, causing imbalances in the basal ganglia's direct and indirect pathways, which result in aberrations in movement speed and the premature gating of movement components. Their neural network model simulates how dopamine depletion, a characteristic of PD, mirrors the motor impairments observed in PD patients. This finding underscores the significant role the basal ganglia plays in the initiation, execution, and sequencing of motor programs.

Given the disinhibitory role that the basal ganglia seem to have in terms of motor pattern expression (Mink, 1996), action gating and action selection models are in some sense two sides of the same coin, as an action ends up being selected through the opening of the thalamic gate. This notion is grounded on the computational hypothesis that afferent signals to the striatum encode the salience of these motor requests (Redgrave et al., 1999). The basal ganglia thus encompass multiple selection mechanisms that can effectively address conflict among compet-

ing signals and ensure smooth transitions between them. Evidence exists of local inhibition within the striatum, either via local interneurons or through reciprocal inhibitory networks among the output cells (Oorschot et al., 2002)

Wickens (1997) explored the dynamics of such local striatal neuron clusters using network models, revealing a competition dynamic where more active neurons suppress less active neighbors. However, under low dopamine conditions, the network shifts from competition to co-activation, offering a potential model for the muscular rigidity seen in Parkinson's patients with dopamine deficiencies.Redgrave et al. (1999) further posit that the BG has evolved as a specialized central selection device, adept at resolving conflicts between subsystems competing for access to limited motor or cognitive resources. This selection operation favors winning competitors through selective disinhibition while maintaining or increasing inhibitory control over losing competitors. Moreover, the authors note the critical role of dopamine bursts in reallocating attention and behavioral resources towards unexpected salient events, further supporting the role of the BG as a center for resource allocation and attention switching.

In the serial processing modeling axis, (Dominey, 1995) and (Beiser and Houk, 1998) developed models suggesting these functions can exist harmoniously. Dominey (1995) proposed that an internal sequence representation is created across the prefrontal cortex, which can then be reproduced later by the basal ganglia. In the same vein, (Beiser and Houk, 1998) suggest the basal ganglia as an encoder, transforming serial events into a spatial representation of prefrontal cortex activity. In this process, the medium spiny neurons of the caudate serve as context detectors, facilitating the encoding of distinct stimuli.

Berns and Sejnowski (1998) present an alternative model for this looping process, proposing the subthalamic–pallidal loop as a site for short-term memory in the production and encoding of sequences. During the learning phase, the striatum repeatedly imposes a temporal sequence pattern on the GPe through direct inhibition. Given the significant connection delays and the fact that input to the STN arises from the GPe, the activity pattern across the STN is at least one-time step behind the GPe's. This temporal lag facilitates the learning and generation of sequences, associating the previous step in the STN's temporal sequence with the current step in the GPe. Once learned this activity pattern effectively cycles between the STN and GPe, producing and perpetuating the sequence.

At the **algorithmic level**, the models lie in the mixture of two main approaches; that of neural network models that utilize some kind of information processing defined in the level above (Gillies and Arbuthnott, 2000), or neural

reinforcement learning models which, due to the strong experimental link between *Reward Prediction Errors* and the signals from dopaminergic neurons in the VTA/SNc (Schultz et al., 1997), use value-based representations in order to learn policies which end up serving as the action selection mechanism. Given the flexibility of the latter approach, the distinction between these two approaches becomes quite fuzzy, as neural network models can be directly implemented in a reinforcement learning paradigm in order for the value representations to be learned (Chakravarthy et al., 2010; Joel et al., 2002). We will dive deeply into this framework in section 1.3.1 as it will be crucial for the model that was developed in this thesis, presented in Chapter 3.

Going back to the RDDR, one of the proposed algorithms for the dimensionality reduction step was through principal component analysis (PCA). Inspired by previous work involving neural networks performing dimensionality reduction using competitive Hebbian learning rules for interlayer connectivity and anti-Hebbian rules for lateral inhibitory intra-layer connectivity (Oja, 1982; Foldiak, 1990; Kung and Diamantars, 1990), Bar-Gad et al. (2000) studied a simulated feed-forward neural network, demonstrating that a two-fold increase in reinforcement signal value led to an almost five-fold decrease in compression reconstruction error. This hypothesis positions the basal ganglia as integral in information extraction and preprocessing from the cortex, allowing for the complexity in cortical representations to be reduced to the signals that have led to positive reinforcement.

Finally, at the **implementation level**, the intricate microcircuitry becomes relevant in order to map the mechanisms necessary for the computations defined in the levels above to be performed. In a general sense, both of the levels above also tend to refer to some circuit-level implementation details, particularly in what respects action selection and action gating models Gurney et al. (2004).

In the exploration of the basal ganglia's circuitry, models of striatal microcircuitry were originally influenced by the structure of spiny projection neurons and their collaterals (Gurney et al., 2004). These models were established with the assumption of a lateral-inhibition-type network with strong, reciprocal connections between neurons, which under high-gain conditions, would enable the most active cells to suppress activity in their weaker counterparts (Wilson and Groves, 1980; Wickens et al., 1991; Alexander and Wickens, 1993).

Most models of the striatum, which serves as the input stage of the basal ganglia, have emphasized pattern recognition or mutual competition, to form pattern classification networks (Beiser et al., 1997). Wickens et al. (1995) explored the effects of local connectivity dominated by recurrent inhibitory connections on the

response properties of striatal networks. Their model revealed unique patterns of spatio-temporal activity in different network topologies. Moreover, Fukai and Tanaka (1997) provided a more in-depth analytical treatment of selective activation in lateral inhibitory networks, demonstrating the dependence of 'winner-take-all' and 'winners-share-all' computations on the strength of collateral inhibition relative to self-inhibition. Lastly, a network model by Connolly and Burns (1995) hypothesized a functional role for characteristic patterns of subthreshold fluctuations in the membrane potential of spiny neurons and suggested that these fluctuations could be involved in coordinating obstacle-avoiding limb movements.

All in all, the diversity of basal ganglia computational models is also an expression of the underlying complexity of the system, as understanding its high-level functioning requires that we also understand all the levels below. We will now focus on a particular class of models, *Neural Reinforcement Learning* models, which have been particularly successful in explaining various characteristics of these circuits, as well as the role they play in behavior.

1.3.1 Neural Reinforcement learning

The origins of Reinforcement Learning (RL) are often associated with the early 20th century, with the seminal work by psychologists Ivan Pavlov and B.F. Skinner. Pavlov's classical conditioning (Pavlov, 1927) and Skinner's operant conditioning (Skinner, 1938) established fundamental principles regarding the role of rewards and punishments in shaping an agent's behavior. However, the idea of discounted rewards, a core concept in RL, can be linked to even earlier sources.

The notion of discounted rewards in reinforcement learning is related to the concept of time preference, the preference for immediate rewards over future ones, with early mentions found in the Talmud (Ohrenstein, 1996) and with some evidence of this process in ancient civilizational complexes such as the Sumerians, who kept records of credit based exchanges (Graeber, 2011). Later on, the evolution of mathematics that dealt with the calculation of interest led to medieval mathematicians developing the idea of compound interest to determine the accumulation of invested sums and the appropriate amounts to be paid for annuities. Although the concept was initially used as a way to calculate an expected return of a loan or investment, Martín de Azpilcueta Navarrus set forth the concept of time preference, which recognizes that a present good is worth more than future goods, thus bringing to life the idea of a discounted future reward (Azpilcueta Navarro, 1966).

From these foundations, the idea of time preference has been extensively discussed in the field of economics, specifically in the value theory of money. A notable economist, Irving Fisher, studied the time preference concept in the late 19th and early 20th centuries (Fisher, 1930). Fisher's work on interest and capital connected the idea of discounting future rewards to the notion of present value, which is a key concept in finance and economics. [2]

Coming back to a more modern era, in the 1950s Bellman introduced Dynamic Programming (DP) (Bellman, 1957), a mathematical framework for addressing complex decision-making problems in non-deterministic environments. In the 1970s, Richard Sutton presented the Temporal Difference (TD) learning algorithm (Sutton, 1988), which integrated dynamic programming and Monte Carlo methods, allowing agents to learn from incomplete sequences of experiences. Sutton's work, along with research by Andrew Barto (Barto, 1995) and others, laid the foundation for modern reinforcement learning algorithms. During the 1980s and 1990s, reinforcement learning gained increasing attention within the artificial intelligence community. Chris Watkins and Peter Dayan made notable contributions by developing the Q-learning algorithm (Watkins and Dayan, 1992), which enabled agents to learn optimal distributions of action distributions (policies) without the need for an environment model, representing a significant advancement in the field.

The basic structure of these algorithms lies in first defining a *Markov Decision Process* (MDP), formally defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, consisting of sets of states $\mathcal{S} = \{s_1, s_2, \ldots, s_{N_s}\}$ and actions $\mathcal{A} = \{a_1, a_2, \ldots, a_{N_a}\}$ where N_s is the number of states and N_a the number of available actions at each state and $\gamma \in [0, 1]$ denotes a discount factor, for discounting future rewards.

The probability transition matrix $\mathcal{P}$ holds the underlying environment dynamics which an RL agent tries to approximate as best as possible, it is defined as a distribution

$$\mathcal{P}^a_{ss'} = \Pr\left[s_{t+1} = s' \mid s_t = s, a_t = a\right] \tag{1.1}$$

which holds the probability for a transition from s to s' when taking action a and allows for the agent to be able to execute an optimal policy $\pi^*(a|s_t) \in [0, 1]$, a probability distribution for actions given a state s_t.

[2] The connection between the present value formula and the sum of discounted rewards is based on considering rewards (R) as analogous to cash flows (CF) and the discount factor (γ) as analogous to the discount rate ($1 + r$) Both concepts involve accounting for time preference and have similar expressions: $PV = \sum_{n=0}^{N} \frac{CF_n}{(1+r)^n} \leftrightarrow G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

In an analogous manner, one can define the expected reward $\mathcal{R}$ as

$$\mathcal{R}^a_{ss'} = \mathbb{E}\left[r_{t+1}|s_t = s, a_t = a, s_{t+1} = s'\right] \tag{1.2}$$

which holds the rewards for the transition $s \to s'$ upon execution of action a. For a given state-action pair, the expected reward can be calculated to

$$\mathcal{R}(s,a) = \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'} \tag{1.3}$$

With these elements at hand, we can define state-value functions $V(s) \in \mathbf{R}^{N_s}$ or state-action value functions $Q(s,a) \in \mathbf{R}^{N_s \times N_a}$ which will represent the cumulative reward gathered at each state visit, serving as a measure of how desirable that state is. These value functions are obtained by calculating the discounted sum of returns over the whole trajectory of the MDP for a given policy π

$$V^\pi(s_t) = \mathbb{E}^\pi\big[\sum_{k=0}^{\infty} \gamma^k r_{k+t+1} | s_t = s\big] \tag{1.4}$$

for the state-value function and

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a\right] \tag{1.5}$$

for the state-action value function. The connection between the state-value function and the action-value function is given by

$$V^\pi(s) = \sum_a \pi(a|s) Q^\pi(s,a). \tag{1.6}$$

We can derive the classic Bellman equation by expanding the sum inside the expectation value (Equation 1.4)

$$\mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right] = \mathbb{E}^\pi\left[r_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^k r_{t+k+1}\right] = \mathbb{E}^\pi\left[r_{t+1} + \gamma V^\pi(s'\right) \tag{1.7}$$

which allows us to define a recursive equation for the value that depends only on the next state, more concretely

$$V^{\pi}(s) = \sum_{a} \pi(s|a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s') \right] \tag{1.8}$$

which is the so-called Bellman equation (Bellman, 1957; Sutton et al., 1998).

As solving the set of linear Bellman equations defined in 1.8 is not feasible for most MDPs, one can obtain the optimal state-value function via a different approach. We can approximate the state-value function by defining it as a linear combination of basis functions $\vec{\phi}$ with weights $\theta_i \in \mathbb{R}$, i.e.

$$V^{\pi}_{\theta}(s_t) = \sum_{i}^{N_s} \theta^i \phi_i(s_t) \tag{1.9}$$

In order to approximate this value function, we can define the least squares objective between the value function V^{π} and its linear approximation V^{π}_{θ}

$$J(\theta) = \frac{1}{2n} \sum_{t=1}^{n} (V^{\pi}_{\theta}(s_t) - V^{\pi}(s_t))^2 \tag{1.10}$$

In order to minimize this loss, one can use Stochastic Gradient Descent (SGD) on the parameters $\vec{\theta}$ whilst assuming the *Temporal Difference Learning* approximation Sutton and Barto (2018)

$$V^{\pi}(s_t) \approx \mathrm{E}[r_{t+1} + \gamma V^{\pi}_{\theta}(s_{t+1}) | s_t]. \tag{1.11}$$

The update on the parameters is then given by the gradient of the loss $J(\theta)$ with learning rate α_V

$$\theta^i_{t+1} \leftarrow \theta^i_t - \alpha_V \nabla J(\theta^i_t) \tag{1.12}$$

which expanding the gradient turns into

$$\theta^i_{t+1} \leftarrow \theta^i_t - \alpha_V (V^{\pi}_{\theta}(s_t) - V^{\pi}(s_t)) \nabla V^{\pi}_{\theta}(s_t). \tag{1.13}$$

Inserting the approximation of V^π referred to above simplifies the expression to

$$\theta^i_{t+1} \leftarrow \theta^i_t - \alpha_V(r_{t+1} + \gamma V^\pi_\theta(s_{t+1}) - V^\pi_\theta(s_t))\phi_i(s_t) \tag{1.14}$$

and the update of the value function for a given state s_t can now be given by

$$V^\pi_\theta(s_t) \leftarrow V^\pi_\theta(s_t) + \alpha_V \delta_t \phi(s_t) \tag{1.15}$$

The term in parentheses is defined as the *Reward Prediction Error* (RPE)

$$\delta_t = R_{t+1} + \gamma V_\theta(z_{t+1}) - V_\theta(s_t) \tag{1.16}$$

and represents one of the most influential hypotheses regarding the functional role of dopamine in the basal ganglia.

Put forward by Montague et al. (1996), the hypothesis suggests that the phasic activity of dopamine neurons in mammals can be explained by the signals generated through the RPEs. This theory was built on experiments led by neuroscientist Wolfram Schultz during the 90s (Schultz, 1995; Schultz et al., 1993) which, by examining the TD error signals from a model of classical conditioning, they found a striking resemblance to the patterns of phasic activity of dopaminergic neurons in the VTA.

Through simulated trials, they found that the phasic response matched the observation that RPEs also only occurred during a reward event and that the response shifted to an earlier predictive cue during the conditioning period. They also observed a decrease in response if the predicted reward is omitted. Despite some variation in dopamine neuron behaviors and mismatches between hypothesis predictions and experimental observations, the hypothesis has been widely accepted due to the compelling correspondence between the majority of neuron activities and TD errors (Schultz et al., 1997).

The link between phasic dopamine and RPEs led to an incredible amount of experimental and theoretical work that has evolved into more and more detailed models of the functional properties of the basal ganglia. One of the architectures that has had a lot of explanatory power is the actor-critic architecture, which we will dive into in the following section.

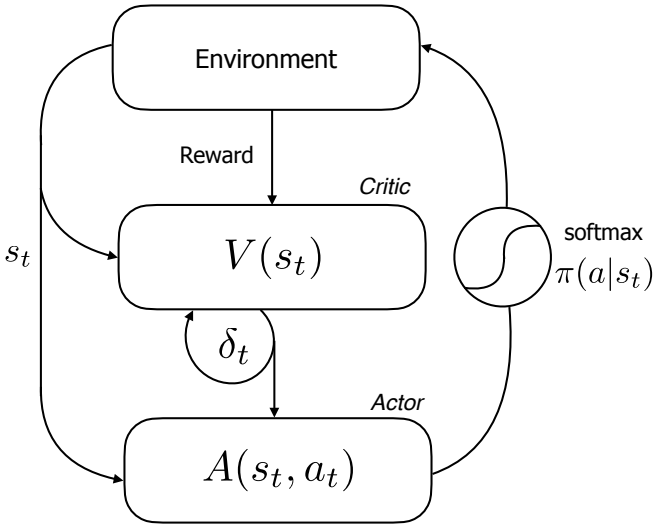


Figure 1.6: Diagram of the actor-critic architecture.

1.3.2 Learning the actor

Initially outlined in Sutton et al. (1983), the idea was of jointly learning two interacting modules: a policy unit or actor where actions would be selected and a critic, named an "Adaptive Critic Element" or ACE. The authors argued that modeling a structure as complex as a neuron at the level of a logic-gate-like unit was inadequate, hence the proposed actor-critic architecture as an initial exploration into systems that consist of a network of complex learned subunits, which laid the groundwork for the emergence of hierarchical and heterarchical RL architectures.

The basic structure of these algorithms relies on the combination of two value representations; a critic, which usually is associated with the state value function $V^{\pi}(s_t)$, and an actor, which is usually associated with an advantage or action

preference function which uses the RPEs from the critic to learn state-action values. One of the first TD-like implementations was given by Baird (1993), and we'll keep a similar notation for the actor, defining it as a state-action value function or, as we'll see later on, an action preference.

Baird (1993) defined an advantage function $A(s_t, a_t)$ that will allow the agent to choose the appropriate action at each state. This function is given by the ongoing advantage over the state value function, $V_\theta^\pi(s_t) = V(s_t)$ where we drop the dependencies on θ and π to alleviate the notation. The advantage is defined as

$$A(s_t, a_t) = \mathrm{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{k+t+1} | s_t, a_t] - V(s_t). \tag{1.17}$$

The goal of this function is to only change if doing so incurs an advantage over the previously calculated state value function. One interesting aspect of this function is that we can write it in function of the RPE δ_t

$$A(s_t, a_t) = \mathrm{E}_\pi[r_{t+1} + \gamma V(z_{t+1}) - V(s_t) | s_t, a_t] \tag{1.18}$$

now contracting the terms we see that δ_t is an unbiased estimator of the advantage function, i.e.

$$A(s_t, a_t) = \mathrm{E}_\pi[\delta_t | s_t, a_t]. \tag{1.19}$$

The choice for the update rule in this case was simply to perform an online average of the RPE, where at each iteration k of the moving average we have

$$A_k(s_t, a_t) = \frac{1}{k} \sum_{t=1}^{k} \delta_t. \tag{1.20}$$

This allows us to expand into a recursive equation

$$A_k(s_t, a_t) = \frac{1}{k}\big(\delta_k + \sum_{t=1}^{k-1} \delta_t\big) = \frac{1}{k}\big(\delta_k + \frac{k}{k-1} \sum_{t=1}^{k-1} \delta_t\big) \tag{1.21}$$

which implies the following recursive relationship, i.e. the learning rule for $A(s_t, a_t)$

$$A_k(s_t, a_t) = A_{k-1}(s_t, a_t) + \frac{1}{k}\big(\delta_k - A_{k-1}(s_t, a_t)\big) \tag{1.22}$$

which is equivalent to writing, for a learning rate α_A

$$A(s_t, a_t) \leftarrow A(s_t, a_t) + \alpha_A(\delta_t - A(s_t, a_t)). \tag{1.23}$$

Performing a tile basis function approximation to the advantage functions and factoring it as two independent functions over actions $m_i(a_t)$ and states $\phi_i(s_t)$, we get

$$A(s_t, a_t) = \sum_i^{N_s} m_i(a_t)\phi(s_t)_i\delta_t \tag{1.24}$$

which now putting into the update rule above allows us to write the update expression for the basis functions of a single action a_t

$$m_i(a) \leftarrow ((1 - \alpha_A)m_i(a) + \alpha_A\phi(s_t)_i\delta_t)\delta(a - a_t). \tag{1.25}$$

with $\delta(a - a_t)$ the kroenecker delta function, selecting only the current action a_t.

An alternative definition of an actor can be seen as a simplification of this scheme, where we do away with the balancing between the RPE δ_t and the advantage, and update directly the values through

$$A(s_t, a_t) \leftarrow A(s_t, a_t) + \alpha_A\delta_t. \tag{1.26}$$

In usual RL formalization, given the discrete observations s_t the agents have to choose an action through a policy $\pi(a|s_t)$ which represents the probability of choosing an action a given an observation s_t.

One action selection approach is that actions at each observation s_t are chosen via a softmaxxed distribution of $A(s_t, a_t)$, i.e. the probability of a given action a^j being executed can be obtained by sampling from the distribution given by

$$\pi(a|s_t) = \frac{e^{\beta A(s_t,a)}}{\sum_j e^{\beta A(s_t,a^j)}} \tag{1.27}$$

where β is the temperature parameter and the denominator contains the sum over all actions.

1.3.3 Actor–critic models of the basal ganglia

This same architecture has been used a level higher in the circuit, with actor and critic elements being associated with specific basal ganglia regions(Joel et al., 2002). This model class has shown tremendous flexibility in capturing a lot of the basal ganglia related experimental results, although not a single model has been able to model all levels: behavioral, neural and optogenetic perturbations. Other limitations, as we shall see, end up appearing due to the specificity of the regions that are trying to be modeled, leaving other highly relevant basal ganglia regions untouched. We shall do a quick pass-through of the modeling done so far and the specific elements that have been modeled by each implementation.

Suri and Schultz built upon Barto (1995) the basic actor-critic model by providing a neural representation of the actor and altering the Temporal Difference (TD) algorithm to simulate the timed depression of Dopamine (DA) activity during omitted rewards. In their model, each stimulus was depicted by a neuron set, each active for varying durations, enabling the replication of DA neuron firing patterns to reward-related stimuli (Suri and Schultz, 1998). The actor, comprising a single neuron layer representing specific actions, learned stimulus-action pairs from the critic's prediction error signal, with a winner-take-all rule ensuring only one action was chosen at any given time. While this model could solve complex behavioral tasks (Suri and Schultz, 1998, 1999), the authors did not align the critic with the known basal ganglia architecture and made arbitrary modifications to the TD algorithm to incorporate novelty responses, generalization responses, and temporal reward prediction aspects, rather than pursuing a biologically plausible implementation as attempted by Contreras-Vidal and Schultz (1999).

Houk et al. (1995b) proposed one of the earliest actor-critic models of the basal ganglia (Joel et al., 2002). Their model places striosomal modules, composed of striatal striosomes, STN, and dopaminergic neurons in the SNc, in the critic's role, while matrix modules perform as the actor. Three critical sources of input contribute to the firing patterns of DA neurons: two from the striatal striosomes and

one from the lateral hypothalamus. One striosomal input offers prolonged inhibition directly to the SNc, while the other, via the STN, provides phasic excitation. The hypothalamic input is excitatory and conveys information about primary rewards. Through dopamine-dependent strengthening of corticostriatal synapses, striatal striosomal neurons learn to fire in response to stimuli predicting future reinforcement. Once learned, a reward-predicting stimulus triggers a DA burst, with expected rewards not causing a DA response due to striosomal inhibition canceling hypothalamic excitation.

However, the model lacks a precise timing mechanism and fails to explain the timed depression of DA activity when an expected reward is omitted. Later models, using the 'complete serial compound stimulus' representation (Montague et al., 1996), addressed this issue by enabling the learning rule to select appropriate temporal representations corresponding to the stimulus-reward interval. In terms of the actor, Houk et al. (1995a) describes the implementation within the basal ganglia circuitry, assigning this role to matrix modules, which consist of the striatal matrix, subthalamic nucleus, globus pallidus, thalamus, and frontal cortex. They suggest that these modules either command actions or generate plans that organize other systems to produce actual command signals, possibly signaling the occurrence of significant contexts from a sensory perspective.

Suri and Schultz (1999) built upon (Barto, 1995) the basic actor-critic model by providing a neural representation of the actor and altering the Temporal Difference (TD) algorithm to simulate the timed depression of Dopamine (DA) activity during omitted rewards. In their model, each stimulus was depicted by a neuron set, each active for varying time intervals, enabling the replication of DA neuron firing patterns to reward-related stimuli (Suri and Schultz, 1998). The actor, comprising a single neuron layer representing specific actions, learned stimulus-action pairs from the critic's prediction error signal, with a winner-take-all rule ensuring only one action was chosen at any given time. While this model could solve complex behavioral tasks (Suri and Schultz, 1998, 1999), the authors did not align the critic with the known basal ganglia architecture and made arbitrary modifications to the TD algorithm to incorporate novelty responses, generalization responses, and temporal reward prediction aspects, rather than pursuing a biologically plausible implementation as attempted by (Contreras-Vidal and Schultz, 1999).

The implementation by Contreras-Vidal and Schultz (1999) was based on a neural network model aligned with basal ganglia anatomy that handled DA responses to stimuli, but also incorporated an additional adaptive resonance network developed by. In this model, two types of reward prediction errors, one related to timing

(TD model) and the other coding for the type and amount of reward (adaptive resonance network) were defined. Unlike the Suri and Schultz (1998) model of prolonged, variable-duration activity, their model activates striosomal neurons sequentially for a limited duration post-stimulus. In this model, learning strengthens striatonigral instead of corticostriatal synapses, something that has not been verified experimentally (Joel et al., 2002). The striatonigral synapses that are active during primary reward delivery, and post-learning, striosomes' timed inhibition would cancel the excitation of DA neurons by predicted rewards. In contrast to (Barto, 1995) critic-based models, excitation and inhibition sources to DA neurons are distinct, with phasic DA responses to reward-predicting stimuli arising from the PFC, communicated via the striatal matrix and SNr. This multi-RPE approach was also taken by Brown et al. (1999), where they hypothesized that the rapid response to conditioned stimuli and delayed, adaptive inhibition of rewarding unconditioned stimuli relies on distinct anatomical pathways.

Suri et al. (2001) expanded on the actor-critic model from Suri and Schultz (1998, 1999), integrating an extended Temporal Difference (TD) model, a basal ganglia-thalamocortical circuitry-based actor and intricate critic-actor interactions. Each striatal layer model neuron corresponds to a small population of striatal matrix neurons, capable of initiating an action. The selection of a singular action depends on direct and indirect pathway interactions connecting the striatum to the basal ganglia output nuclei, along with a winner-take-all rule at the cortical level.

The model incorporates long-term adaptation of corticostriatal transmission, and transient effects on striatal neurons' firing rates and up/down-state durations. The critic receives sensory and reward information, matching the previously mentioned corticostriatal functional hypothesis (1.2.3), as well as intended and actual action data from the thalamic and cortical levels of the actor (1.2.4), enabling the learning of stimulus-reward and action-stimulus associations.

The model demonstrated sensorimotor learning and planning capabilities, forming novel associative chains and selecting actions based on these chains' predicted outcomes. This planning ability depends on future stimuli predictions and intended action information as input to the critic. The novelty responses of DA neurons, promoting exploration behavior, are modeled through optimistic initial weights for novel stimuli. Exploratory behavior also stems from stochastic transitions between striatal neurons' up and down states.

Concluding this overview one might note that most of the underlying actor-critic architectures anchor themselves on the anatomical evidence of some critic

network that could exist in the connections of striatal striosomes with the DA system with the functional role of the remaining nuclei not being taken into consideration. Besides this, the functional role of the direct and indirect pathways expressed by the D1R and D2R MSNs is not clearly defined. One possibility for the actor component is defined in (Suri and Schultz, 1999) but it also doesn't express the distinct role of these pathways. We shall now dive into the specific class of models that have taken the direct/indirect pathways in consideration, as well as work that attempts to define a set of learning rules that process positive and negative rewards via different channels.

1.3.4 Encoding opponent reinforcement pathways

Dopamine plays a substantial role in decision-making, motivation, incentive, and effort, as numerous studies indicate that animals with heightened dopamine levels exert more effort for identical rewards (Berridge, 2012; Smith et al., 2011; Beeler et al., 2010; Salamone et al., 2005). The relationship between dopamine's effects on incentive and performance and reinforcement learning is contentious, with some suggesting reinterpretation in terms of incentive salience (Berridge, 2012) and others demonstrating learning effects, like reward-based choice preferences predicted by striatal responses to reward prediction errors during learning, modulated by dopaminergic manipulation (Jocham et al., 2011).

Most models of learning define a single value for each action, which is either incremented or decremented by dopamine-encoded reward prediction errors. This facilitates choice among different actions by comparing their current values, with higher valued actions more likely to be selected (Frank et al., 2007; Jocham et al., 2011; McClure et al., 2003; O'Doherty et al., 2004; Pessiglione et al., 2006; Samejima et al., 2005; Schonberg et al., 2007). However, these models can't fully explain the modulatory influence of dopamine on incentive choice or the asymmetric effects of dopamine on learning from positive versus negative outcomes (Berridge, 2012; Salamone et al., 2005). They also struggle to reconcile the findings that motor symptoms of Parkinson's disease can progress without further dopaminergic degeneration (Beeler et al., 2012).

Other models, which focus more on vigor, propose that tonic dopamine modulates response vigor to optimize for reward or avoid punishment per unit time (Niv et al., 2007). However, these models only address effects on vigor (e.g., response execution speed or effort), not on choices between actions with different valences/incentives. Also, they mainly consider the effects of increased dopamine

signaling, without examining cases where dopamine depletion might enhance performance.

Neurobiology and neural network models paint a more complex picture, suggesting a more intricate dual-opponent system for action and learning. As we've seen in Section 1.2, dopamine modulates activity and plasticity in two distinct populations of striatal MSNs that project to different basal ganglia output nuclei. Striatal MSNs in the direct (striatonigral) pathway predominantly express dopamine D1 receptors, facilitating actions, while those in the indirect (striatopallidal) pathway predominantly express dopamine D2 receptors, suppressing actions (Kravitz et al., 2010).

One particularly relevant model that takes this differential encoding into consideration is the Opponent Actor Learning (OpAL) model by Collins and Frank (2014). The model aims to describe dopamine's incentive, learning, and interaction effects using a simple parameterization. By utilizing variables that correlate with biologically interpretable measures like tonic or phasic dopamine level, D1 and D2-expressing striatal neuronal activity, synaptic strengths, and synaptic plasticity, the model was able to achieve a level of biophysical realism regarding these neural populations that was not present in previous models. This model hypothesizes that these pathways are not merely encoding a go or no-go message, but instead aggregate evidence in favor or against each action (Collins and Frank, 2014). The final choice is determined by the relative differences in evidence for each action, via competition throughout the corticostriatal circuit.

The model is defined through an actor-critic architecture, which associates the critic (as defined in eq. 1.4) to the ventral striatum and makes the classical assumption that phasic signals of dopamine convey the critic prediction error (Dayan and Daw, 2008; Montague et al., 1996; Roesch et al., 2007) and the actors to their corresponding D1 and D2 populations, being associated to corticostriatal synaptic weights. This split is done by splitting the corresponding action values into two oppositely signed RPE-dependent functions. The update equations for the actors are given by

$$\begin{cases} A^G(s_t, a_t) \leftarrow A^G(s_t, a_t) + \alpha_G A^G(s_t, a_t)\delta_t \\ A^N(s_t, a_t) \leftarrow A^N(s_t, a_t) - \alpha_N A^N(s_t, a_t)\delta_t \end{cases} \tag{1.28}$$

with A^G, A^N the Go and NoGo action preferences and α_G, α_N their corresponding learning rates. These two state-action values capture the notion that dopamine

has opposite effects on plasticity via stimulation of D1 and D2 receptors still allowing each population to undergo potentiation and depression. The multiplicative factor on the learning rate captures the three-factor Hebbian rule, where learning depends on pre-synaptic activation (from cortex, the stimulus and action), post-synaptic activation in striatum (proportional to the actor weight), and dopamine (Reynolds et al., 2001).

Optogenetic studies (Tai et al., 2012) have shown that the values of specific actions can be amplified or inhibited by stimulating D1 or D2 MSNs during the choice period. Stimulation of D1 MSNs in one hemisphere increased the likelihood of opting for the contralateral action, enhancing its value rather than simply spurring motor response. The level of stimulation required varied depending on the existing learned value of the action. Interestingly, stimulating D2 MSNs had the opposite effect, effectively decreasing the action's value. Thus, D1 and D2 stimulation at the time of choice respectively created additive positive and negative impacts on the action's estimated value.

Collins and Frank (2014) model the effect of optogenetic perturbations by slightly altering the learning rules, which now have an added perturbation term, specifically

$$\begin{cases} A^G(s_t, a_t) \leftarrow A^G(s_t, a_t) + \alpha_G A^G(s_t, a_t)[\delta_t + O^G] \\ A^N(s_t, a_t) \leftarrow A^N(s_t, a_t) - \alpha_N A^N(s_t, a_t)[\delta_t - O^N] \end{cases} \tag{1.29}$$

where O^G, O^N are the optogenetic perturbation weights. In either case, the actions are selected via a softmax of the sum of these two action preferences defined in eq. 1.27, i.e.

$$A^T(s_t, a_t) = \beta_G A^G(s_t, a_t) - \beta_N A^N(s_t, a_t). \tag{1.30}$$

The OpAL model was shown to account for learned choice biases and the impact of dopaminergic manipulation on these biases. It also encapsulated the interplay between performance and learning, validated by empirical data from the rotarod motor skill learning task. The model made novel predictions for reward-based decision-making, highlighting stronger incentive effects when choices were made under the same dopaminergic state as during learning. The model also suggested that learning and incentive effects could jointly influence effort-based decision-making, with their relative impacts depending on the overall effort required.

A question remains: *What could be the utility of this split reward and punishment processing?*

Relatively recent approaches have also used this idea to try to understand higher-level questions regarding cognitive function and the interlinkage between uncertainty and the valence of stimuli. (Anderson et al., 2019) propose that the uncertainty influences affective states by prompting the mental simulation of possible future outcomes, thus allowing for the construction of contextually dependent representations that might fine-tune the exploration of novel environments. (Lowe and Ziemke, 2013) explore this relationship through a reinforcement learning algorithm utilizing separate representations of reward and punishment. Their approach, based on DYNA-SARSA and policy evolution via genetic algorithms, was shown to allow for adaptive affective context-sensitive behaviors.

Okada et al. (2001) introduce a two-dimensional evaluation reinforcement learning technique that distinguishes between reward and punishment evaluation. Their approach uses the difference between these evaluations for action selection and their sum for adjusting the exploration-exploitation ratio, providing a novel means to manage the tradeoff between these crucial aspects of reinforcement learning.

This exploration of reward-punishment dynamics finds practical applications in robotics. (Navarro-Guerrero et al., 2017) explore the use of negative reinforcement signals in robot motor learning by showing that nociception can drive learning, expanding the state space and reducing task errors. On the other hand, they also suggest that punishment may be detrimental to performance if used as typically within reinforcement learning applications.

Lastly, (Elfwing and Seymour, 2017) presents an alternative approach to the conventional binary system of reward and punishment in reinforcement learning systems. They introduce the MaxPain algorithm, which also predicts rewards and punishments in parallel. They demonstrate this approach can promote safer behavior and improve early exploration and learning in reinforcement learning systems across a set of grid world environments.

The parallel evaluation of positive and negative reward information seems to increase the subtlety with which agents explore environments. It's thus quite possible that the D1/D2 system evolved for that reason as well; that of generating more detailed distributions of different valence spectra, separating positive and negative regions into different encodings. Given that there's a fundamental asymmetry in terms of the frequency of positive vs negative rewards (with negative rewards being much more common than positive ones), organisms should have a

broader representation for negative stimuli and a more precise one for positive signals. This is somewhat mirrored by biology, as the connectivity patterns between striato-pallidal neurons show convergent and sparse projections and are thus hypothesized to carry action-specific information, and subthalamic projections to BG output nuclei are denser and relatively more diffuse Hazrati et al. (1990). In other words, while general spatial convergence is found, inhibition of motor programs is thought to be broader and more diffuse while excitation appears more focused in relative terms.

1.4 Opening the gates

As we've now presented the basic building blocks of the core work done in this thesis, we can outline the following chapters, which will be structured as follows:

- **Chapter 2**: We'll present the main experimental findings present in Cruz et al. (2022).
- **Chapter 3**: We'll construct a reinforcement learning model of the direct and indirect pathways within the context of the experiments in Chapter 2, carefully going through the reasoning and choices made.
- **Chapter 4**: We'll close the modeling work by proposing a potential mechanism for one of the crucial components of the model.
- **Chapter 5**: To finalize, we'll present a brief discussion of what all this might entail for the field, as well as list the multiple research avenues that might have been opened through the work presented in this thesis.

Chapter 2

Action suppression reveals opponent parallel control via striatal circuits

2.1 Introduction

The idea, presented in section 1.2.1, that the direct and indirect pathways functionally oppose each other has been a recurring hypothesis in the field. As mentioned earlier, this assumption is supported by anatomical (Smith et al., 1998) and experimental (Deniau and Chevalier, 1985; Kravitz et al., 2010; Freeze et al., 2013) findings that suggest a rapid enhancement of the BG output neurons' activity when iMSNs are activated and the rapid suppression when dMSNs are activated.

This opposite effect is also observed at the behavioral level in relation to locomotion, motor sequence production, reinforcement, and value-based decisions (Kravitz et al., 2010, 2012; Roseberry et al., 2016; Sippy et al., 2015; Tai et al., 2012; Tecuapetla et al., 2016; Yttri and Dudman, 2016). However, the observed co-activation of both dMSNs and iMSNs during action production has raised questions against the functional opposition hypothesis (Cui et al., 2013; Tecuapetla et al., 2014; Cox and Witten, 2019).

An enduring and potentially reconciling viewpoint proposes that these pathways might contribute competitively to the selection of actions (Denny-Brown and Yanagisawa, 1976; Mink, 1996; Redgrave et al., 1999). This theory suggests that action selection involves a combined effort of motor program promotion by

the direct pathway and suppression by the indirect pathway. These operations could justify the observed co-activation, even in an antagonistic setting, while also predicting that action suppression would lead to large-scale decorrelation or even anti-correlation between the two pathways. The iMSNs activity would heighten to suppress action, while dMSNs activity would be reduced until action is resumed.

We tested this hypothesis through a variant of an interval categorization task (Gouveâ et al., 2015; Soares et al., 2016) which required continuous action suppression. Using fiber photometry and electrophysiology, dMSNs and iMSNs activity in mice's DLS was recorded[1].

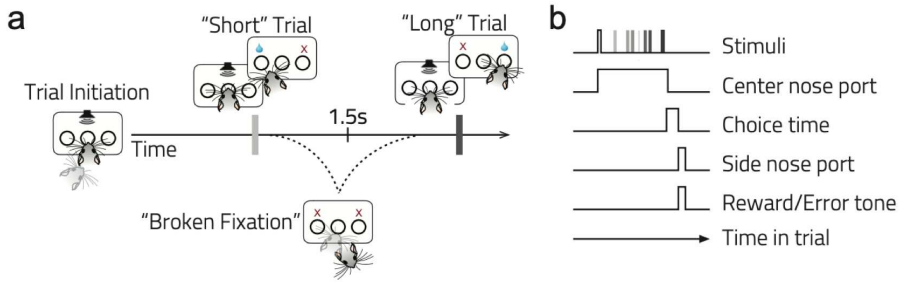


Figure 2.1: Task **(a)** and event **(b)** diagram. Subjects self-initiate each trial in a center nose port. After a variable delay, a second tone is played, and they are asked to categorize the presented interval as short or long by responding in one of two side ports. Between the two tones subjects are required to maintain position in the center port - "fixation".

The results revealed functional opponency during action suppression, distinct from the usual phasic activation observed during action production, with iMSNs showing a stronger engagement and dMSNs showing an anticorrelated activity profile. In order to understand the functional properties of these pathways, optogenetics experiments were also performed. Surprisingly, the results showed DLS circuits were not significantly involved in action promotion but primarily in action suppression.

Based on the known parallel circuit architecture of the BG (Alexander and Crutcher, 1990), we postulated that the suppressive function seen in DLS might

[1]All animal experiments and underlying data analysis and figures in this section were conducted by Bruno Cruz (Cruz et al., 2022) and adapted from (Cruz et al., 2022)

be a result of an action-promoting policy learned elsewhere in the striatum, a hypothesis that we will explore with more detail in chapter 3.

We shall, in the following chapter, go through the main experimental results that were the main anchor for the modeling work done in this thesis, outlining the main elements that were deemed as relevant for the model itself.

2.2 Mice show structured breaking fixation patterns

One of the ways in which one could unveil mechanisms of suppression is by studying in a detailed manner when that suppression fails. We did that by training mice in a modified version of a two-alternative time interval categorization task developed earlier (Gouveâ et al., 2015; Soares et al., 2016). The task involved mice self-initiating a trial by inserting their snout into a centrally located initiation nose port, which triggered a brief auditory tone. The mice were then required to maintain their position in the initiation port until a second auditory tone was presented after a random delay, ranging from 0.6s to 2.4s, as can be seen in figure 2.2

Once the second tone was delivered, the mice could then select either of the two choice ports situated at an equal distance from the initiation port. Rewards were provided based on the timing of their choice: choosing the 'short' side if the interval was less than 1.5s, and the 'long' side if the interval was more than 1.5s. Mice not only successfully learned to categorize the intervals, but did so with high accuracy for the intervals that were significantly shorter or longer than the decision boundary, drawing an expected performance psychometric curve, also exposed in figure 2.2.

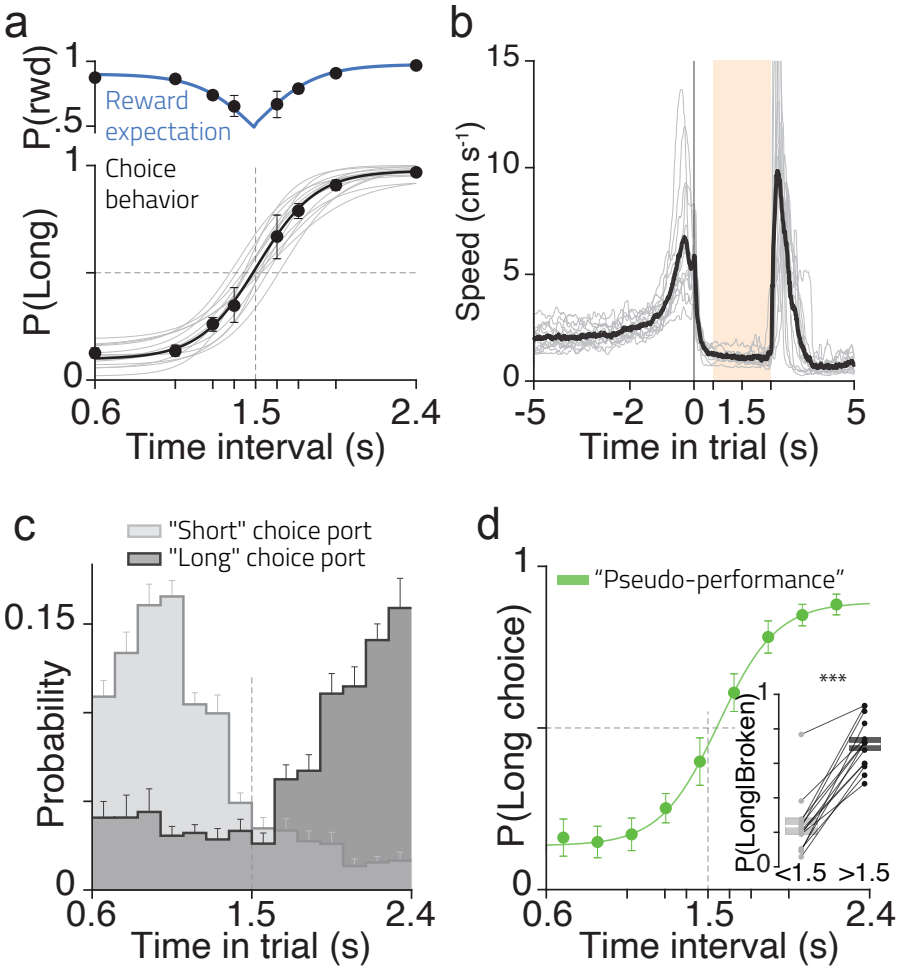


Figure 2.2: **a)** top: Psychometric fit to the performance of each mouse that underwent photometric recordings (n=14, light gray) and the logistic fit to the overall average performance across mice (black). bottom: Reward expectancy calculated from the overall performance of animals on a given stimuli (blue trace depicts the rectified psychometric fit at 1.5s) **b)** Animal's head speed as a function of time, aligned on trial initiation, for a single random session of each mouse that underwent photometric recordings, during trials wherein the longest interval (2.4 seconds) was delivered and a correct choice performed. Gray, mean of individual animals; black, average of all mice (n=14). Shaded region highlights period of immobility (0.6s to 2.4s post-trial initiation). **c)** Probability density functions of broken fixations over time during the immobility period (0.6s to 2.4s), contingent on subsequent choice of one of the side ports. **d)** Average overall pseudo-performance of all animals used in the photometric recordings calculated from broken fixation trials. To calculate the performance in broken fixation trials, we binned the times at which animals aborted the trial and calculated the proportion of reports at the long choice port over all reports. Inset depicts single animal probability of choosing long as a function of breaking fixation before ($< 1.5s$) or after ($> 1.5s$) the decision boundary

As we can see in Figure 2.2a, each trial showed a typified pattern of movement; there was an increase in movement speed leading up to trial initiation, followed by a short period of postural adjustment before a period of immobility until the second tone was played. After the second tone, movement speed increased again as the animals made their choices. If the mice failed to maintain their position in the initiation port until the second tone, thus breaking the fixation, an error tone would sound, terminating the trial.

Interestingly, even after breaking fixation and aborting the trial, the animals often chose a port (Figure 2.2c). These choices exhibited a similar timeline and trajectory to valid choices, which resulted in a very similar psychometric curve as the one obtained through valid choices. Furthermore, these were mostly 'appropriate,' leaning towards the 'short' choice port when breaking early and the 'long' port when breaking late. This pattern seems to reflect the overall reward associated with the two choices over time, rather than the likelihood of the second tone's occurrence.

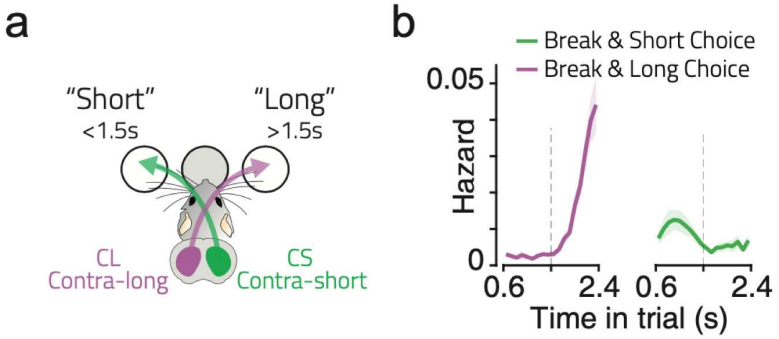


Figure 2.3: **a)** Schematic of labeling convention, the three large circles represent the three nose ports present in the behavioral task apparatus. The filled gray circle represents the trial initiation port, while the open circles represent the two choice ports. **b)** Hazard rate of broken fixations trials (see methods) wherein animals subsequently made a choice at the port corresponding to a short (green, right) or long (purple, left) choice. Error bars/boxes represent s.e.m.

This can also be seen through the hazard rate (Figure 2.3b), which gives us the probability of breaking fixation at a given time point t given that it has not been broken up until then. We see also a clear difference between the contra-side

normalized hazards, with a slight bump in the first half of the delay period and a sharp increase for the contra-right side post-second half.

We obtain these distributions via

$$H(k) = \frac{B(k)}{\sum_{j=k}^{T} B(j) + \sum_{j=k}^{T} C(j)}, t < k < t + \Delta t \tag{2.1}$$

where $B(k)$ is the number of broken fixation trials that occurred between t and $t + \Delta t$ and $\sum_{j=k}^{T} B(j)$ the sum of all broken fixations that occurred in the time greater or equal to t, up until the longest possible interval T (2.4s) and $\sum_{j=k}^{T} C(j)$ the sum of all completed trials that occurred from time t until T, and fit the psychometric functions with a 4-parameter logistic function

$$P(x) = (u - l)\frac{e^{\frac{x-b}{s}}}{1 + e^{\frac{x-b}{s}}} + l \tag{2.2}$$

where P is the performance of the animal on interval x, u and l the upper and lower asymptote of the curve, respectively, b the bias parameter and s the slope parameter.

2.3 Opposing dynamics between direct and indirect pathway activity

In order to understand in more detail the mechanics of this breaking fixation behavior, the large-scale activity of the direct and indirect pathways were probed. The direct and indirect pathway activity was recorded in the DLS while the mice were performing the task. This was achieved by merging mouse lines that expressed Cre recombinase in either dMSNs or iMSNs with cre-dependent viral expression of the calcium indicator GCaMP6f. Fiber photometry was then employed to access the pooled activity of a local population of dMSNs or iMSNs in the DLS.

The activity patterns during the trial were then examined by combining neuronal activity in both hemispheres. Sharp increases in activity in both dMSNs and iMSNs were observed during task epochs when mice were required to take action. However, distinct differences in activity were found during the delay period, when

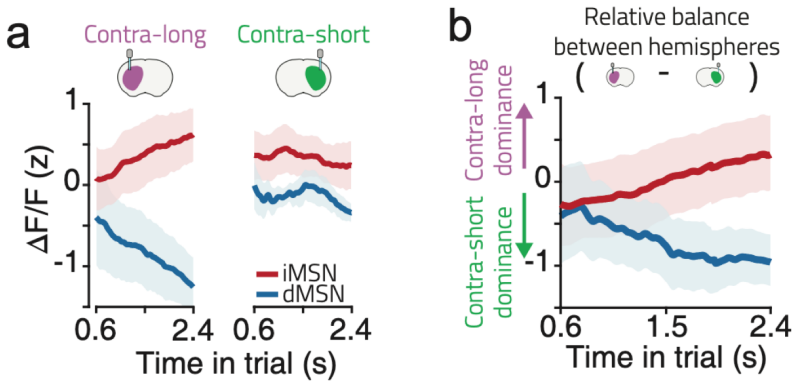


Figure 2.4: **a)** Averaged normalized activity recorded from the hemisphere contralateral to a long (left panel) or short choice (right panel) port (z-scored) across all iMSN (red) and dMSN (blue) mice. Only correct completed trials were included.**b)** Average of all pairwise differences in immobility period activity of dMSNs and iMSNs between the two hemispheres, subtracting activity recorded in hemispheres contralateral to the "long" choice port from activity recorded in hemispheres contralateral to the "short" choice port (i.e., CL activity - CS activity).

animals needed to suppress action. We observed that the activity of iMSNs was found to be significantly higher than dMSN activity (Figure 2.4a).

These patterns of activity were further confirmed by recording single units electrophysiologically from DLS and optogenetically photo-identifying iMSNs. Various patterns of responses were noted in both photo-identified and non-photo-identified putative MSNs. Consistent with photometry, the photo-identified iMSN population was significantly enriched for cells displaying increased activity during action suppression.

Lastly, pathway-specific activity was observed in the two hemispheres over time, which suggests that DLS iMSNs are engaged to dynamically suppress the temptation to act. Significant downward deviations in the rate of change of photometric signal in iMSNs preceding broken fixations as compared to time-matched control periods were detected. This may reflect the contribution of two subpopulations of iMSNs. The findings show that transient disruption of iMSN activity, as assessed both using photometry and electrophysiology, was associated with the failure to successfully suppress action.

2.4 Optogenetic inhibition experiments

Finally, in order to understand the role that each of these pathways played during the action suppression period, optogenetic experiments were undertaken. Optical fibers were implanted in the dorsolateral striatum. Cre lines used previously to label MSNs were combined with Cre-dependent viral expression of light-activated proton pump ArchT or ChR2, thus enabling inhibition or activation of iMSN or dMSN activity. The effect of photoinhibition on neuronal firing was first characterized through extracellular electrophysiological recordings from striatal neurons during a quiet awake state. Rapid and sustained inhibition of putative MSN firing was exclusively produced by illumination of striatal tissue in both iMSN-ArchT and dMSN-ArchT mice, with no evidence found for disinhibition of non-targeted MSNs.

In contrast, excitation mediated by ChR2 resulted in robust activation of MSNs but also sustained inhibition in 10-20 % of putative MSNs. In order to examine the behavioral effects of optogenetic inhibition during the task, focus was shifted from activation, as it appeared to mix both activation and inhibition, potentially across the two MSN types, complicating the interpretability of activation experiments

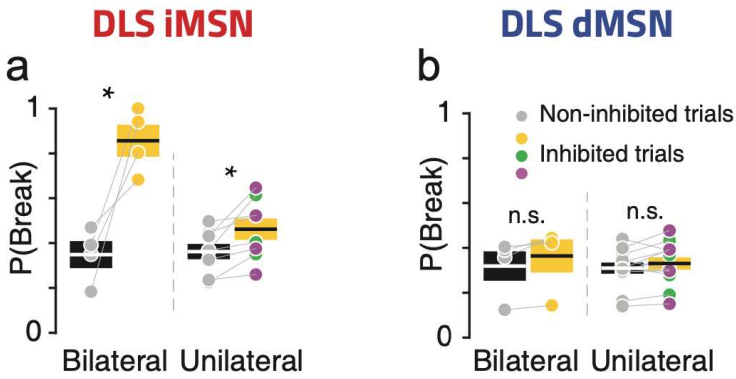


Figure 2.5: Overall probability of breaking fixation during iMSN (a) or dMSN (b) inhibition experiments. Colored and black dots represent data from laser-on and laser-off trials, respectively. "Bilateral" condition represents data from sessions wherein light was delivered bilaterally to the DLS.

concerning the role of specific cell types in the striatum. The light was subsequently delivered to the fiber(s) on a random minority of trials.

A near-complete inability of mice to suppress movement during interval presentation overall was found when iMSNs were inhibited bilaterally. In contrast, no significant effect was observed on levels of broken fixation during equivalent trial epochs by bilateral inhibition of dMSNs. An analysis of movements during the ramping off of light stimulation showed a significant increase in choice time during broken fixation trials, but only when dMSNs were inhibited.

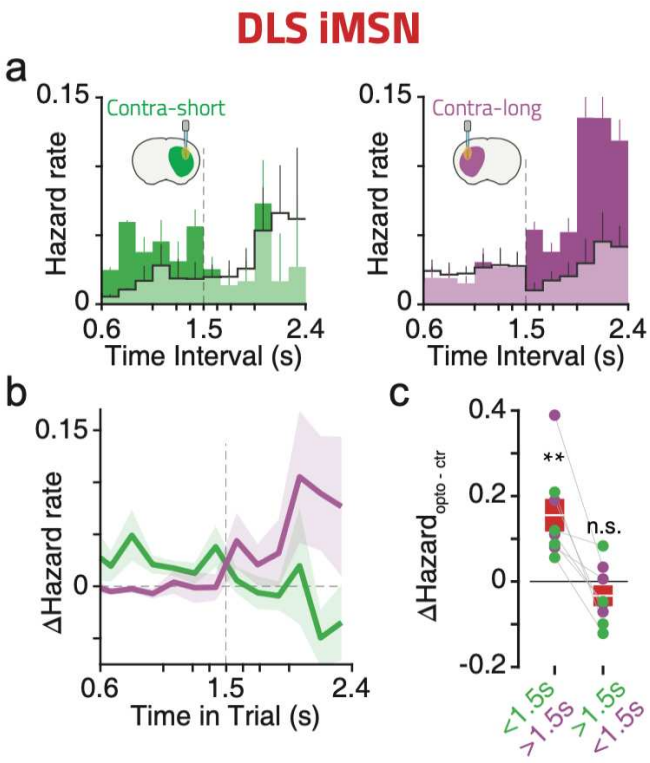


Figure 2.6: **a)** Hazard rate of breaking fixation over all included trials of a given condition, for control (black outline) and manipulated trials at the hemisphere contralateral to a short (green) or long (purple) choice port in A2a-Cre animals. **b)** Change in probability ($\Delta P = P_{Manipulation} - P_{Control}$) due to inhibition of the CS (green) or CL (purple) relative to session matched control trials. c) Quantification of the effect shown in b). We calculated the hazard of breaking fixation during the period where the choice contralateral to the site of inhibition would be correct or incorrect (before/after 1.5s and after/before 1.5s for CS and CL, respectively). Data shown are the differences between session matched controls and manipulations. Each pair of points depicts data from the same hemisphere and the color the site of manipulation.

Furthermore, a necessity for DLS dMSN activity was specifically observed for movement invigoration shortly before movement initiation. However, no effects were noticed on choice times during valid trials or in experiments where inhibition was applied during the execution of the choice movement. An increased likelihood of mice breaking fixation early or late was observed when iMSNs contralateral to the "short" or "long" choice were inhibited, respectively. This reflected a striking correlation with the pattern of hemispheric dominance observed in iMSN neural activity.

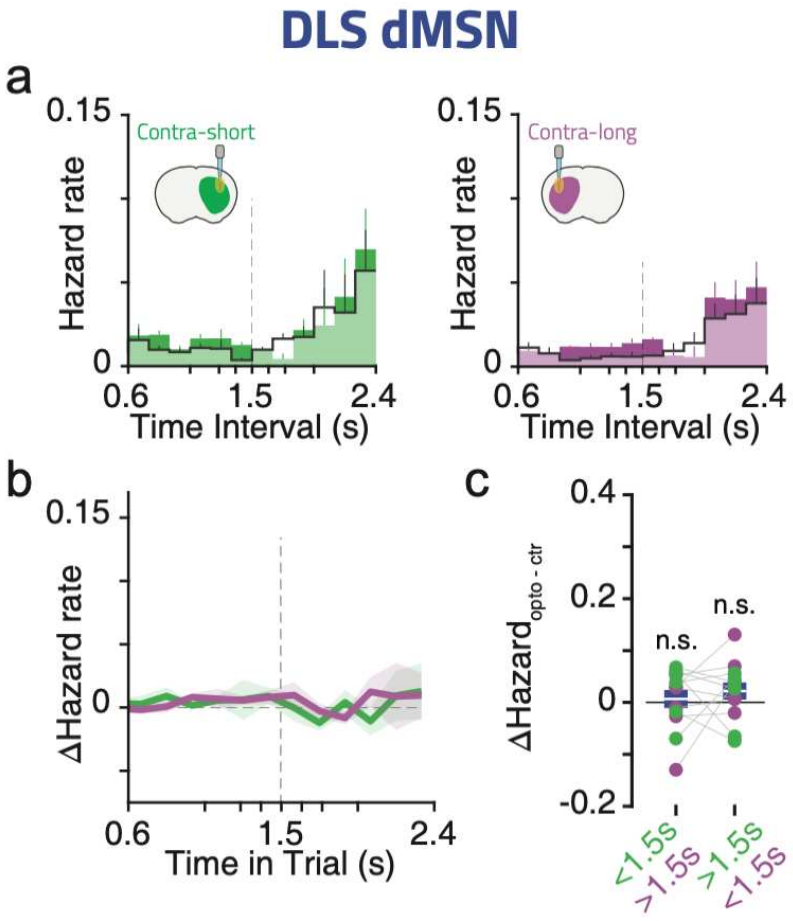


Figure 2.7: **a)** Hazard rate of breaking fixation over all included trials of a given condition, for control (black outline) and manipulated trials at the hemisphere contralateral to a short (green) or long (purple) choice port in D1-Cre animals. **b)** Change in probability ($\Delta P = P_{Manipulation} - P_{Control}$) due to inhibition of the CS (green) or CL (purple) relative to session matched control trials. c) Quantification of the effect shown in b). We calculated the hazard of breaking fixation during the period where the choice contralateral to the site of inhibition would be correct or incorrect (before/after $1.5s$ and after/before $1.5s$ for CS and CL, respectively). Data shown are the differences between session matched controls and manipulations. Each pair of points depicts data from the same hemisphere and the color the site of manipulation. (contralateral-correct: one-sample t-test, $P = 0.711, t_1 1 = 0.38$, contralateral-incorrect: $P = 0.211, t_1 1 = 1.329, N = 12$ Hemispheres). Error bars represent s.e.m.. n.s. $P \geq 0.05$.

Lastly, when iMSNs were inhibited unilaterally, a consistent increase was noticed in the probability that an animal would execute a choice to the port contralateral to the site of iMSN inhibition. Notably, no consistent difference was identified in the trajectory compared to choice movements performed without iMSN inhibition. No significant effect was observed by the unilateral inhibition of DLS dMSN on lateralized choice behavior after broken fixations, nor when dMSN inhibition was applied during the execution of the choice movement. Collectively, these data demonstrate the dynamic engagement of the indirect pathway in the DLS of a given hemisphere in suppressing contralateral movements, contributing to determining which action is selected for execution.

We shall use the aforementioned constellation of results as the basis for the model that will be developed in the following chapter, in order to understand, at least mechanistically, the underlying functional relationships between these circuits.

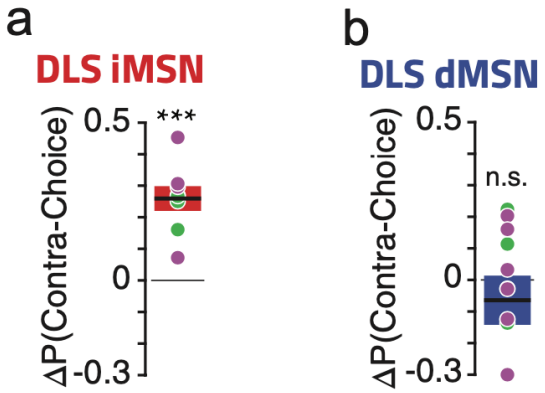


Figure 2.8: Change in probability of registering a choice at the port contralateral to the hemisphere manipulated, after breaking fixation ($\Delta P = P_{Manipulation} - P_{Control}$, in A2a-Cre **(a)**, one-sample t-test, $P \leq 0.001, t_7 = 6.605, N = 8$ hemispheres) or D1-Cre animals **(b)**,$P = 0.428, t_1 1 = -0.824, N = 12$ hemispheres). Error bars represent s.e.m.. n.s. $P \geq 0.05$. $***$ $P \leq 0.001$

2.5 Discussion

Animals were trained in a time interval categorization task wherein the relative importance of left versus right movements varied over time. This unique setup allowed us to observe and study the specific neuronal populations that become active in both suppression and promotion of actions, showing a clear functional opponency between the two major pathways present in the BG (Schwarting and Huston, 1996; Kravitz et al., 2010).

We observed functional opponency between direct and indirect pathway neurons during a state of voluntary movement suppression. This suppressive state was characterized by the proactive suppression of actions, meaning that it was carried out in anticipation of an external stimulus, rather than in response to it. Besides this, when animals failed to perform such suppression, their breaking fixation distributions showed the actual reward structure of the task, indicating a potential latent model or plan that could be suppressed.

During this state, we noticed an overall higher level of activity in the iMSNs situated contralaterally to the direction of the movement that should be performed, compared to their ipsilateral counterparts, a result that is consistent with functional magnetic resonance imaging data in humans and electrophysiological data in non-human primates (Majid et al., 2013; Watanabe and Munoz, 2010; Ford and Everling, 2009; Amita and Hikosaka, 2019).

Importantly, we observed that even though both populations of iMSNs and dMSNs were co-active during the suppression, as shown in (Cui et al., 2013), that iMSNs were necessary for the suppression to take place whilst dMSNs were not necessary. This highlights a distinct role of iMSNs and dMSNs in motor functions that extend beyond mere proactive action suppression (Kravitz et al., 2010; Tecuapetla et al., 2014; Markowitz et al., 2018).

Finally, a set of optogenetic experiments that inhibited iMSNs did not result in less vigorous movements. Rather, it disrupted the suppression of actions. Specifically, this inhibition seemed to disrupt the overall suppression of actions enacted by body parts on the side contralateral to the hemisphere. This suggests a degree of action specificity in iMSN function, with pathways more active on the contralateral side during movement (Albin et al., 1989; Parker et al., 2018). Conversely, inhibiting the activity of DLS dMSNs slowed down movements. This finding is consistent with previous studies, underscoring the vital role dMSNs play in invigorating movement (Panigrahi et al., 2015).

In order to make sense of the constellation of results shown, we'll develop a computational model in the next chapter that will aim at reproducing the behavioral, neuronal recordings and optogenetic experimental data, to unveil a potential global set of mechanisms that might be responsible for the production and suppression of behavior in the BG.

Chapter 3

Multi-agent actor-critic model of basal ganglia suppression

3.1 Introduction

Given the experimental results shown in the previous section, our goal will be to build a computational model that attempts to explain the neural mechanisms underlying interval the behavioral and neurophysiological dynamics ongoing in this temporal discrimination task. In order to provide maximal clarity as to how the model works we will build each one of its components and show their underlying mechanics so that the full model becomes intuitive.

As a historical note, the development of this model was anchored in a set of fundamental questions that now lay bare what we didn't fully understand about this system. In the process of designing the model the questions that guide it were the following:

1. *How can one explain and reproduce the co-activation of both pathways during action onset?*
2. *How can one explain the dynamics of both pathways during the delay period?*
3. *How are the dynamics of these pathways related with the underlying breaking fixation behavior?*
4. *How can one explain the effect of the optogenetic inhibition experiments performed in both pathways?*

In order to answer these questions we will start by building an equivalent MDP to the task that is solved by mice, aiming at modelling the most relevant components of their behavior. Then, we will delve into the details of how direct and indirect pathway plasticity can be modeled and their respective contributions to learning.

With these components we built an actor-critic model that is able to contain representations analogous to those of DLS iMSNs and dMSNs, using that information to shape a policy that allows it to fully solve the IDT, based on the intuitions of a potential functional role of the indirect pathway in suppressing movement. We will show that a single two-pathway actor-critic agent is insufficient to explain the data, which leads us to introduce a multi-agent architecture. In this architecture, we will introduce a new agent containing a different and more abstract state representation.

This will allow us to model the opponent dynamics between the direct and indirect pathways and the predictions that the model gives as to what would be the response to optogenetic perturbations in the two different agents. To finalize we'll compare how these perturbations relate to actual optogenetic experiments done in mice in two regions of the striatum.

3.2 Modeling components

3.2.1 Task representation

In the Interval Discrimination Task (IDT) (Gouveâ et al., 2015; Motiwala et al., 2022), animals begin a trial by choosing a central nose port and receiving an onset tone that demarcates the beginning of the trial. After this tone they need to voluntarily suppress movement until a second offset tone is delivered and a reward is available at the corresponding short or long nose-port. The length of this delay period between tones is sampled randomly from a set of different intervals and the trial identity is given by the relative position of the tone to the decision boundary; short intervals for second tones before the decision boundary and long intervals for second tones after that boundary.

The animal's behavior in the task and the underlying neural activities of neurons of the striatal populations was modeled as a Reinforcement Learning (RL) problem. RL models require the definition of a Markov Decision Process (MDP)

(Sutton et al., 1998) which consists of a set of states $e \subseteq E \in \mathbb{N}$ of the environment, a set of actions $a \in A$ and a transition matrix $P(e_{t+1}, r_t | e_t, a_t)$ that gives us the probability of moving to a new state e_{t+1} and receiving a reward r_t after performing action a_t, with $t \in \mathbb{Z}$.

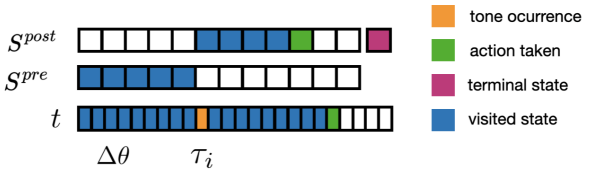


Figure 3.1: Basic scheme of the Interval Discrimination Task with an example trajectory.

We model the discrimination task by considering a linearly evolving variable $\theta \in \mathbb{R}$ representing the passage of time in milliseconds (ms) since the onset of a trial. This variable is constrained to the interval $\theta \in [\theta_0, \theta_f] \in \mathbb{R}^+$ where θ_0 is the interval onset, where animals are already at the center port and θ_f the time point at which, if not action is taken a timeout is produced and the trial is over. We discretize this interval into a set of N_T bins of duration $\Delta\theta = \frac{\theta_f - \theta_0}{N_T}$ (in milliseconds). The environment transitions $P(e_{t+1}|e_t, a_t)$ are defined as an off-diagonal transition matrix were for each true time state $e_t = t$ we move to $e_{t+1} = t + 1$ and where the corresponding true time is given by $\theta = t\Delta\theta$.

In each trial, the second tone is defined in a set of temporal markers $\tau_i \in \mathbb{R}$ all defined in the range $[\theta_0, \theta_f]$ and given by the values [0.6, 1.05, 1.26, 1.38, 1.62, 1.74, 1.95, 2.4] s which will signal to the agents the occurrence of the second tone (interval offset) on a given trial i at real time t identified by the true time state e_t.

At each e_t the environment will accept one of 3 actions; choice of the short port ($a = S$), choice the long port ($a = L$) or perform no action, which we define as a HOLD action ($a = H$). If no action other than H is performed during the delay period and the correct choice is made after the tone, the agents will receive a reward. If the incorrect action is performed in the choice period or if an action other than H is performed during the delay period then a punishment is delivered, with both reward and punishment being defined as $\{r_t^+, r_t^-\} \in \mathbb{R}$ respectively.

3.2.2 Internal state representation

In order to model the passage of time from the perspective of an organism, we need to create a state representation for an artificial agent that has similar enough mechanics to those of the model organism under study. One crucial aspect of temporal perception that we take into account is that animals also seem to show scalar variability in temporal perception of elapsed time. This scalar variability is seen in linear relationship between the variance in the estimate of time elapsed and the time since a given event (Gibbon, 1977; Gouveâ et al., 2015).

We model this by considering that the agents do not have direct access to the "true time" state e_t but instead rely on an internal estimate of elapsed time given by another set of states. This mapping is performed by a function $K : E \rightarrow S$ that goes from the experimenter's clock $e_t \in E$ to the time as felt from an observer, defined as $s_t \subseteq S = [0, ..., N_K] \in \mathbb{Z}$, where N_K is the total number of states. The function K also maps the occurence of a *tone event* within the internal temporal estimate by creating two possible sets of states; post 1st tone states (or pre 2nd-tone states) and post 2nd tone states i.e. the states are split into two subsets

$$S = \begin{cases} S^{pre} \subseteq \left[0, ..., \frac{N_k}{2}\right] \in \mathbb{Z}, \text{if } \ e_t < \tau_i \\ S^{post} \subseteq \left[\frac{N_k}{2}, ..., N_K\right] \in \mathbb{Z}, \text{if } \ e_t \geq \tau_i \end{cases} \tag{3.1}$$

with the transition between these two sets triggered by the occurrence of a tone at time τ_i, where i is the index of the *second-tone event*.

This distinction between the sets $\{S^{pre}, S^{post}\}$ is what allows the agent to correctly discern in which states it should perform an action or remain "still" through the HOLD action. This can be made clear by the fact that if we only had states representing time after first-tone and did not perform this split into post second-tone states, one state would potentially signify both the correct and incorrect temporal region for an action to be performed, which would not allow for informative reward signals regarding the correct actions to reach the agent.

This internal estimate was constructed so as to produce scalar variability in temporal estimates around the true time. To achieve this, we define a central dwell time μ that corresponds to a sequence of states that accurately follows the elapsed time from the experimentalist's viewpoint; that is, true time. We then draw the dwell times at each trial from a gaussian distribution given by $\Delta\theta_i^d \sim \mathcal{N}(\mu, \sigma)$,

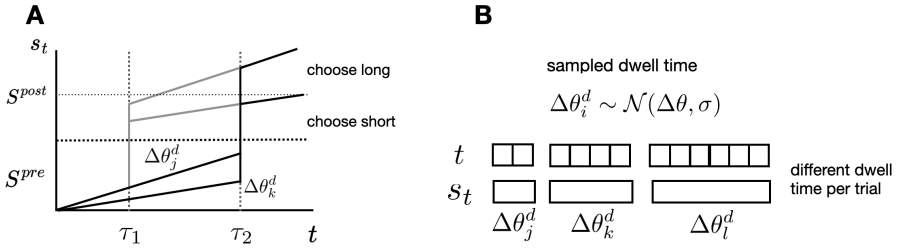


Figure 3.2: Representation of how temporal variability is generated in the model. On panel A we show how different trajectories stemming from different dwell times $\Delta\theta_i^d$ emerge and how after each tone $\{\tau_1, \tau_2\}$ there's a transition from S^{pre} to S^{post}. In Panel B we see how the dwell time relates with the experimenter's time $e_t = t$.

with the variance of the distribution σ defining the range of potential dwell times for a given agent.

The variable $\Delta\theta_i^d$ represents the dwell time for the current trial i, becoming discretized in a number of 'true time' states, e_t that elapse in a single internal state s_t, and is responsible for the scalar variability of internal time estimates expressed by the model across trials. Because the effect of trial to trial dwell time variability accumulates as the model progresses through states, internal variability in time estimates grows in aggregate as true time elapses.

The mapping K contains the transition dynamics between pre and post-second tone, as it is defined in the the full set of states $K : \{\theta, \Delta\theta_i^d, \tau_i\} \rightarrow \{S^{pre}, S^{post}\}$ and is given by

$$s_t = K(\theta, \Delta\theta_i^d, \tau_i) = \begin{cases} \left\lfloor \frac{\theta}{\Delta\theta_i^d} \right\rfloor, \text{if } \theta < \tau_i \\ \left\lfloor \frac{\theta}{\Delta\theta_i^d} \right\rfloor + J, \text{if } \theta \geq \tau_i \end{cases} \tag{3.2}$$

An important aspect of this state representation is that the agents only have access to their internal estimate of time s_t and are not aware of the tone identity at any given trial. Their experience of the environment is completely determined by the dynamics of s_t and how reward and punishment appear once an action is performed. Previous models of this task have defined the temporal variability as each trial depending on a stochastic variable that would determine the dwell time

at each clock tick (Motiwala et al., 2022). The state representation defined here is inspired by such previous models and portrays a more simplified version, whilst also allowing for the expression of the fundamental behavioral properties that we want to produce in our agents.

3.3 A two pathway actor-critic model

In order to model the recorded neural activities, we will modify an actor-critic algorithm(Motiwala et al., 2022)(Baird and C., 1993) into a multi-agent paradigm where different striatal pathways are represented by different actors with different observation dynamics and different Reward Prediction Error (RPE) sensitivities. We take the Action Preferences framework as a starting point so as to define the learning rules for the different pathways and then combine the learned action-preferences into a single probabilistic policy $\pi(a|s)$ from which a decision will be sampled.

The choice of action preference algorithms for this problem is twofold; actor-critic models have been mapped onto striatal anatomy as a way of explaining the potential relationship between these areas (Joel et al., 2002) Secondly, due to the fact that the action value functions are computed through the reward prediction error δ_t, we'll be able to map the the plasticity dynamics of dopaminergic neurons to our learning rules as has been done in previous work (Collins and Frank, 2014; Grondman et al., 2012).

3.3.1 Learning the critic

In the usual actor-critic paradigm the critic is given by a state value function, defined as the expected value over a policy $\pi(a|s_t)$ of the sum of discounted rewards and is written as

$$V(s_t) = \mathrm{E}_\pi\Big[\sum_{k=0}^{T-k-1} \gamma^k r_{t+k+1}\Big| s_t\Big] \tag{3.3}$$

with $\gamma < 1$ being the discount factor. We use the Temporal Difference (TD) method to learn the state value function, with an update given by (Sutton and Barto, 2018)

$$V(s_t) \leftarrow V(s_t) + \alpha_V(s_t)\delta_t. \tag{3.4}$$

with $\alpha_V(s_t)$ the learning rate for the critic and δ_t the *Reward Prediction Error* (RPE) which is defined as

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t). \tag{3.5}$$

This RPE will serve as a global learning signal for all the action-value functions that will correspond to the direct and indirect pathway values. In order to facilitate convergence we define the learning rate as a function of the number of visits to a given state s_T, $N(s_t)$

$$\alpha_V(s_t) = \frac{1}{N(s_t)} \tag{3.6}$$

The learning rates for both the critic and actor components of the model will be defined equally across all models developed.

3.3.2 Modeling direct and indirect pathway plasticity

Given that we want to reproduce the joint activity of D1-MSNs and A2A-MSNs observed in the data, we'll build an actor-critic architecture that supports these two pathways in the form of specific value functions for each pathway. However, in order to make that separation, we need to distinguish which processes might differentiate each one of these pathways.

We take inspiration from biology and the known results regarding the underlying plasticity dynamics in striatal neurons to model the activities of the direct and indirect pathways. In-vitro work performed in (Gurney et al., 2015) has shown that positive changes in overall plasticity of D1-MSNs are directly correlated with an increase in dopamine (DA) in the cellular matrix, whilst decreases in DA concentration show a decrease in plasticity.

Conversely, for D2-MSNs higher DA values have the opposite effect as it reduces the change in plasticity while lower concentrations of DA potentiate plasticity. This result seems to also be present in the volume change of dendritic spines for D1 and D2 MSNs. High DA amounts potentiate an increase of volume for D1 dendritic spines (D1-SPN) whilst decreasing volume for the spines of D2-SPN. Conversely,

a decrease in DA concentration increases the volume for D2-SPN and reduces that volume for D1-SPN (Iino et al., 2020). One important aspect is that these changes in plasticity show a non-linear response to the amount of dopamine present in the cellular matrix, corresponding to a non-linear saturation function (Gurney et al., 2015).

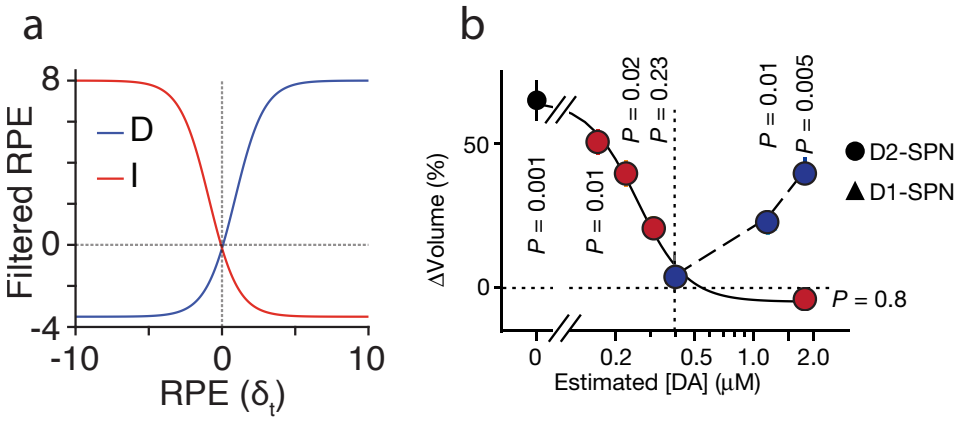


Figure 3.3: **a)** Transfer functions for the direct and indirect pathways, mapping the RPE to the change in action preference in the corresponding actor. **b)** Spine volume changes plotted against estimated DA concentrations, adapted from Iino et al. (2020)

We'll attempt to capture these plasticity rules by defining a set of non-linear functions f^p that we'll associate with individual pathway $p = \{D, I\}$, direct or indirect, respectively. As the *Reward Prediction Error* (RPE) has been linked with activity in VTA, the projections from this region to the striatum will sense a corresponding release of dopamine, thus linking RPE with a change in the plasticity of the striatal MSNs. In order to model this, we'll make this non-linear function dependent on the RPE of the agent experiencing the states s_t ,i.e. $f^p(\delta_t)$. In this manner we're able to associate a given pathway to a specific combination of a non-linear function f^p that is modulated by the same RPE $\delta_t^{D,I}$ but going through different "transfer" functions, i,e. for the direct pathway D1-MSNs we have $f^D(\delta_t)$ and indirect pathway D2-MSNs will be affected by $f^I(\delta_t)$. Concretely, these functions are defined for each pathway p as

$$f^p(\delta_t) = b_0^p + \frac{b_1^p}{1 + b_2^p e^{(1-b_3^p \delta_t)}} \tag{3.7}$$

where the k's are adjusted for each individual pathway. The definition of these non-linear functions as adjustable sigmoid type functions allows for the expression of the observed co-activation of the direct and indirect pathways neurons, both observed in the data and also in (Cui et al., 2013). We achieve this by parameterizing these f^p in such a way that there's an overlap of the RPE sensitivity. Although other formulations that rely on rectified reward units (Collins and Frank, 2014) or reward sensitivity filters (Tano et al., 2020) have been proposed and do in fact model some aspects of these populations, they do not allow for pathway co-activation, which we try to address by allowing more flexibility in the above defined RPE sensitivity functions.

3.3.3 Learning the actors

In actor-critic formalisms one needs to learn the policy independently from the state value function. In our case we chose to learn the action preferences $A(s_t, a_t)$ at each state-action pair that will allow the agent to choose the appropriate action at each state. In the usual definition of these actor-critic models, for an action a_t and state s_t the action preferences are learned through the RPE that is calculated at the current time δ_t.

The update rule for the action-preference for a single agent receiving RPEs from a critic is given by (Sutton et al., 1998)

$$A(s_t, a_t) \leftarrow A(s_t, a_t) + \alpha_A \delta_t \tag{3.8}$$

In simple terms, at each update if the action executed at state s_t results in a positive RPE then that action will become more likely to be performed in the future, if the RPE is negative then the likelihood of the action decreases for that state s_t.

The equivalence in the change of plasticity observed in (Gurney et al., 2015)(Iino et al., 2020) with the action-preference change in the model is done by slightly altering the action-preference updating learning rules defined above. We do this by applying the pathway transfer function $f^p(\delta_t)$ to each pathway learning rule, i.e.

$$A^p(s_t, a_t) \leftarrow A^p(s_t, a_t) + \alpha_A f^p(\delta_t) \tag{3.9}$$

One can see that given an RPE δ_t, different pathways will change their corresponding action-preference in different ways, depending on the particular filtering that is being applied to the global RPE. In the most general case, for a given number N_p of pathways the total action preference A^T would be given by the weighted sum of all the pathways and can be expressed as

$$A^T(s_t, a_t) = \sum_p \omega_p A^p(s_t, a_t) \tag{3.10}$$

where ω_p represents the weighting of each actor. In order for the agent to correctly navigate the environment, this function needs to be such that the correct actions are expressed or suppressed at any given observed state.

3.3.4 Policy definition

In our framework, the direct and indirect pathways are efined by two oppositely tuned RPE sensitivity functions within each corresponding action-preference function. This opposition allows us to express the inhibitory nature of the indirect pathway and the excitatory nature of the direct pathway at the level of action execution.

A policy defined only through direct and indirect pathway agents would be given by the total action preference

$$A^T(s_t, a_t) = \omega_D A^D(s_t, a_t) - \omega_I A^I(s_t, a_t) \tag{3.11}$$

where the negative sign in the indirect pathway allows for it to serve as an inhibitory gate for action execution. This idea was first proposed in (Mink, 1996), where it is shown that the expression of motor patterns seem to be direcly correlated with the amount of inibition present in the outputs of the Basal Ganglia.

Given that the indirect pathway is a fully inhibitory signal in the total action-preference A^T, it serves as the only action "break" signal that is available to the agent. Therefore, all of the inhibitory information which codes the errors learned through experience is coded in the A^I action-preference. One way of looking at

this function is that it represents, instead of the policy which tells agents what to do the **anti-policy**, contaning information about what the agent should **not do**. It serves as a dynamic map that evolves through the environment, releasing and increasing inhibitory pressure on the available positively tuned signals that push for the actions to be expressed.

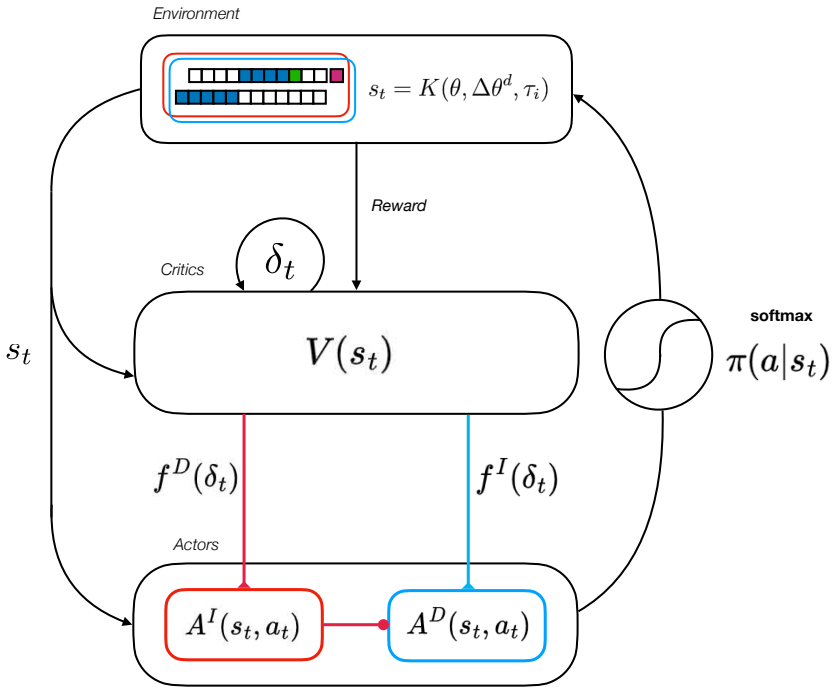


Figure 3.4: Architecture of the two pathway actor-critic model of the basal ganglia.

We consider that both SHORT and LONG actions are taken explicitly by the agent as in the vanilla RL case through a probabilistic policy $\pi(a_t|s_t)$ but the HOLD action will be triggered by a dependency on the joint activity of the direct-indirect pathway action-preferences so as to mimic the inhibitory role of the indirect

pathway seen experimentally(Kravitz et al., 2010).

$$\begin{aligned} \text{if} \quad A^T(s_t, a_t) < 0 \quad \forall \quad a_t = \{SHORT, LONG\} \\ \Rightarrow A^T(s_t, HOLD) = 0, \\ \text{else} \\ \Rightarrow A^T(s_t, HOLD) = -\infty \end{aligned} \tag{3.12}$$

Given that sampling from the policy implies a softmax operation, this set of conditions will increase the probability of an HOLD action being taken proportionally to the total amount of inhibition that is present in the indirect pathway action preference A^I. For the remaining case and in order for the agent to be able to choose an action, this probability is sharply reduced when the inhibition of the indirect pathway fails to be higher than the value held by the direct pathway action preference A^D.

The policy for this case is then defined as,

$$\pi(a|s_t) = \frac{e^{\beta A^T(s_t,a)}}{\sum_j e^{\beta A^T(s_t,a^j)}}. \tag{3.13}$$

3.3.5 A single two-pathway agent is insufficient to explain the data

We now have all the ingredients to build a simulation of an agent in the task environment defined in section 3.2.1. The experimental results give us a clear cut list of elements that we should see expressed in our model, we can define them as follows:

1. Activity of the indirect pathway should be higher than direct pathway during the delay period
2. The activity of the indirect pathway should increase as time passes in the delay period
3. Agents should show a bimodal breaking fixation distribution, with choice matching the correct choice in the correspondent delay period region

We first train the agent until convergence of both the critics and the action preferences by calculating the relative difference of all functions at each iteration step.

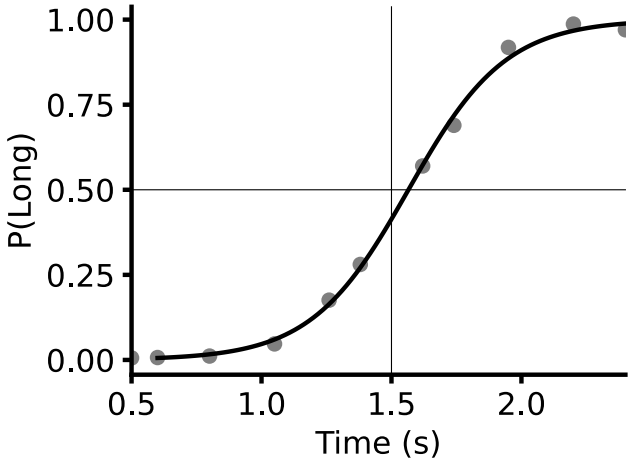


Figure 3.5: Fitted psychometric curve for the two pathways agent in the IDT

The psychometric curve for this agent shows that the model is not only capable of solving the task but shows very similar behavior in terms of the success rate for each of the tones, as can be seen in Figure3.5.

However, one of the crucial elements that we need to model, namely the breaking fixations do not show the bi-modality that we were looking for. The distributions show a strong bias to breaking and choosing with similar probability for both SHORT and LONG actions. In fact it seems that the agents show practically no urge to break and choose the LONG action in the second half of the delay period.

In order to understand this we need to look at the activities of each pathway throughout the delay period and for each action, which will correspond to the contralateral hemisphere of the mice i.e. SHORT action will correspond to the contra short side and LONG action to the contra long side.

Noting that the agent only has access to the states s_t and is trained only in these variables, we need to map the values of the action preferences back to the time variable θ so that we can plot them in a reasonable way. We do this by

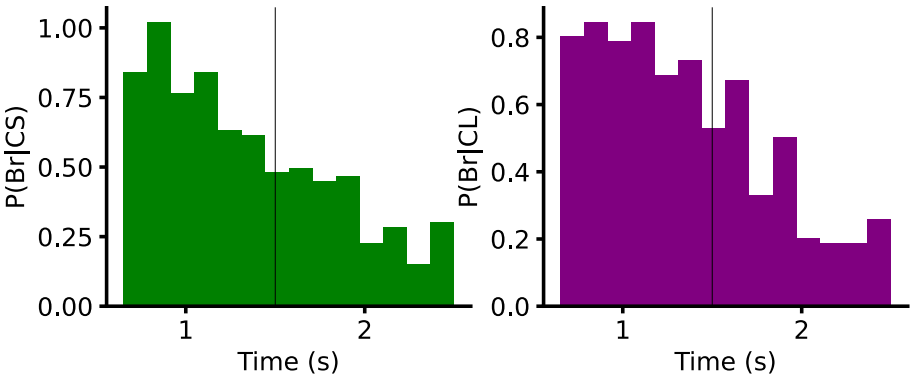


Figure 3.6: Broken fixations of the agents during training for the "contra-short" CS (left panel) and "contra-long" CL (right panel)

running over the time bins corresponding to the delay period and sampling from the states that are closest to that bin for each trial i.e. for a given trial $i \in N_I$ where N_I is the total number of trials we'll have a "sampled time" t_i which will be in the experimenter's frame of reference, this is similar to an inversion of the transformation $K(\theta, \Delta\theta_i^d, \tau_i)$ that was done previously, so we can define it as

$$t_i = K^{-1}(s_t)_i \tag{3.14}$$

and the action preference at a given time point t is the average over trials

$$A^p(t, a_t) = \langle A(t_i, a_t) \rangle_i. \tag{3.15}$$

By plotting the action preferences for both pathways in the delay period we notice the cause for the unimodal breaking fixations. Although the first goal is achieved, that of a higher activity of the indirect pathway over the direct pathway, the signal is constant throughout the whole delay period as one can see in Fig 3.8, creating no differential "urge" to break fixation and choose the action corresponding to the delay period region. What the breaking fixations seem to show is only the decrease in the probability of reaching the latter states in the environment, which is expected due to the nature of the task.

However, we need to ask. Can this be a result of a bad convergence of the model? We can check for convergence by observing the following limits

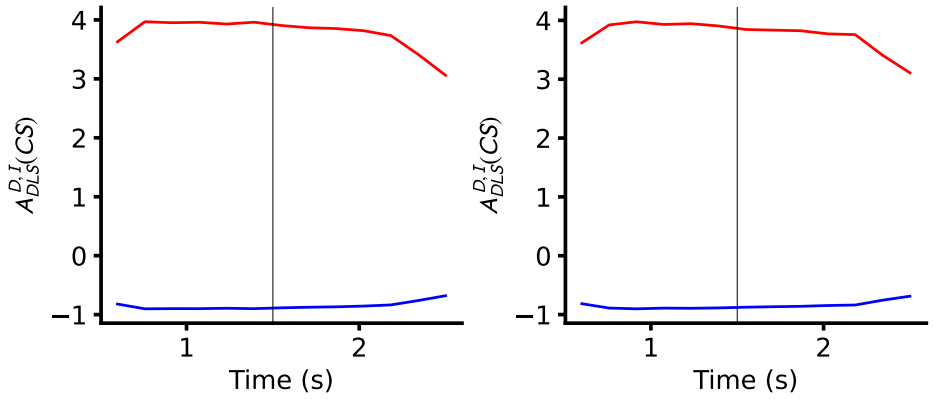
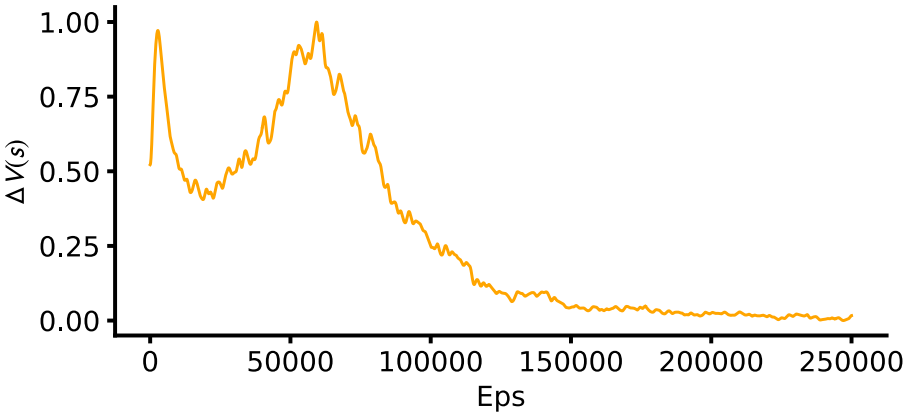


Figure 3.7: Signals of the action preference functions during the delay period of the IDT.

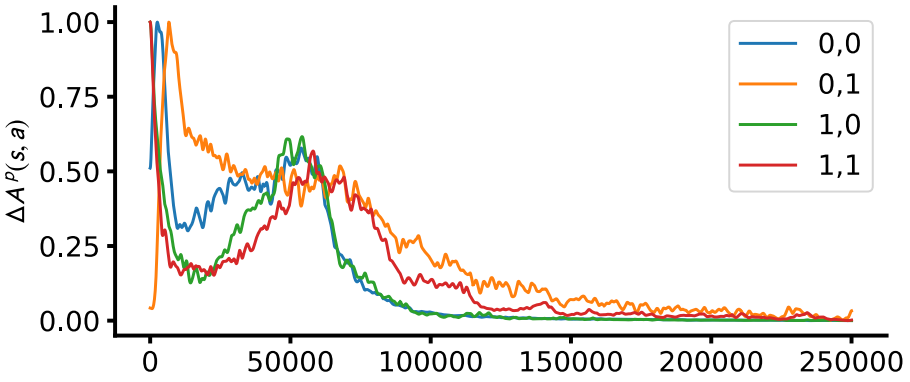
$$\begin{cases} \lim_{e\to\infty}(\Delta V)^2 = \lim_{e\to\infty}(V(\vec{s})_e - V(\vec{s})_{e-1})^2 \approx 0 \\ \lim_{e\to\infty}(\Delta A^p)^2 = \lim_{e\to\infty}(A(\vec{s},\vec{a})_e - A(\vec{s},\vec{a})_{e-1})^2 \approx 0 \end{cases} \tag{3.16}$$

where $V(\vec{s})_e$ and $A(\vec{s},\vec{a})_e$ are the state and advantage value functions vector and matrix (respectively), for the episode e of the simulation.

If we plot these curves for all the value functions present in the model, we see that the model is able to converge swiftly towards a very small change after 250k episodes. Given this, we state that although this simple model is able to solve the task it does so too optimally, creating no internal urge to break fixations as mice do in the experiments shown earlier.



(a) Convergence of the critic value function $V(s_t)$ for the two pathway agent



(b) Convergence of the action preferences $A^p(s_t, a_t)$ for the two pathway agent. Indexes are (p, a), pathway and action respectively

Figure 3.8: Convergence measure of the two pathway agent in the IDT (250k episodes)

3.4 Introducing a new agent

We've now shown that the simplest of models that could potentially contain all the known experimental results was not able to reproduce the data. In order to solve this problem we will start from a very basic question:

what could be the reason for the animals to break fixation?

Our reasoning is that, due to the fact that we're able to draw the same psychometric curve with the choices after a breaking fixation as with the correct choice behavior, that there's some level of understanding of the task structure by the mice.

One potential reason for this behavior is, that through experimenting with all the trials multiple times, they've gotten to know the task structure and thus might be predicting where the reward would be located at different points of the delay period. Because there's a sense of where the reward will be, either on the SHORT or the LONG port, they have an *urge* to choose that port at a given point in time, and sometimes this urge is too strong, causing a breaking fixation.

In order for this urge to act *specifically* at different regions of the delay period we hypothesize that there is potentially a more abstract task representation living in parallel with the more detailed representation shown in section 3.2.1.

In order to test this hypothesis, we will add a new agent to the model, one that has a more abstract state representation for the task and that would correspond to a different, but unknown region in the striatum. The basic hypothesis is that this "urge" signal must come from somewhere, so we will generate a new agent that will have different action preference signals and will compete, as in a multi-agent scenario but generating a single policy out of which actions will be selected as in the previous single agent model.

3.4.1 Introducing a new state representation

We build this more abstract state representation by constructing a scheme where the agent does not have perceptual access to all the information present in the environment. Specifically, we prevent the second-tone from occurring and keep the agent in a pre-second tone sequence of states, whilst still having access to the reward or punishment received.

The dynamics of this environment are straightforward, as they do not contain the second tone jump in states present in the original state representation, shaping

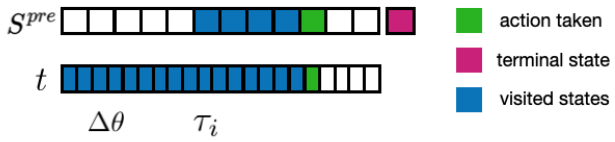


Figure 3.9: State representation of the new agent.

a similar situation to that of information asymmetry present in other RL models (Galashov et al., 2019). Importantly, these state transitions share the same dwell time $\Delta\theta_i^d$. In order to differentiate from the previously defined state variable we will call these states, s_t^X and define the mapping as K^X via

$$s_t^X = K^X(\theta, \Delta\theta_i^d, \tau_i) = \left\lfloor \frac{\theta}{\Delta\theta_i^d} \right\rfloor \tag{3.17}$$

As we will see in the next sections, although the representation seems simple, it has a lot of expressive power in terms of allowing the model to reproduce the underlying behavior and activity patterns of the two pathways.

3.4.2 Multi-agent architecture

The multi-agent representation used in this model consists in two agents with the same underlying pathway architecture defined in the previous section but each with a different state representation. In order to make the notation more consistent, we'll add a superscript L to the initial state representation defined by Eq. 3.2 i.e. $s_t \rightarrow s_t^L$, and define the complete set of sub-agents by the superscript $c \in \{L, X\}$.

We can thus write the critic update equations for both agents

$$V^c(s_t^c) \leftarrow V^c(s_t) + \alpha_V(s_t^c)\delta_t^c. \tag{3.18}$$

and the RPE as

$$\delta_t^c = R_{t+1} + \gamma V^c(s_{t+1}^c) - V^c(s_t^c). \tag{3.19}$$

noting that now the RPE is different for each agent, as well as the learning rate, which counts visits for the states corresponding to each representation.

We now have in total $p \times c$ action preferences, which depend respectively on the state representation s_t^c and the pathway transfer function $f^p(\delta_t^c)$; they can be written as

$$A^{c,p}(s_t^c, a_t) \leftarrow A^{c,p}(s_t^c, a_t) + \alpha_A f^p(\delta_t^c) \tag{3.20}$$

and are joined in a total action preference, now dependent on both state representations

$$A^T(s_t^L, s_t^X, a_t) = \sum_p \sum_c \omega_{pc} A^p(s_t^c, a_t) \tag{3.21}$$

where ω_{pc} is the weight of the pathways of each sub-agent.

By having the total action preference A^T we can now define the policy for the global agent, made out of the two sub-agents

$$\pi(a|s_t^L, s_t^X) = \frac{e^{\beta A^T(s_t^L, s_T^X, a)}}{\sum_j e^{\beta A^T(s_t^L, s_T^X, a^j)}}. \tag{3.22}$$

Because of the way in which we define the policy, each actor-critic sub-agent is now in a competition for expression in terms of providing the correct action-value output for the policy to be optimal. We choose this architecture of a single policy as it maps with the convergent and lower dimensional output of the Basal Ganglia.

Another thing of note is that the mechanics of the HOLD action are still the same as before, except now the indirect pathways of both sub-agents need to adjust so that inhibition is enough for $A^T < 0$, i.e. the condition for withholding movement now becomes

$$\sum_c A^{c,D}(s_t^c, a_t) < \sum_c A^{c,I}(s_t^c, a_t), \forall a_t \in \{SHORT, LONG\} \tag{3.23}$$

All of the sub-agents are trained in parallel, as one can see from the algorithm (ref-appendix) with the same convergence conditions as the ones used in the single sub-agent case (see Eq. 3.16).

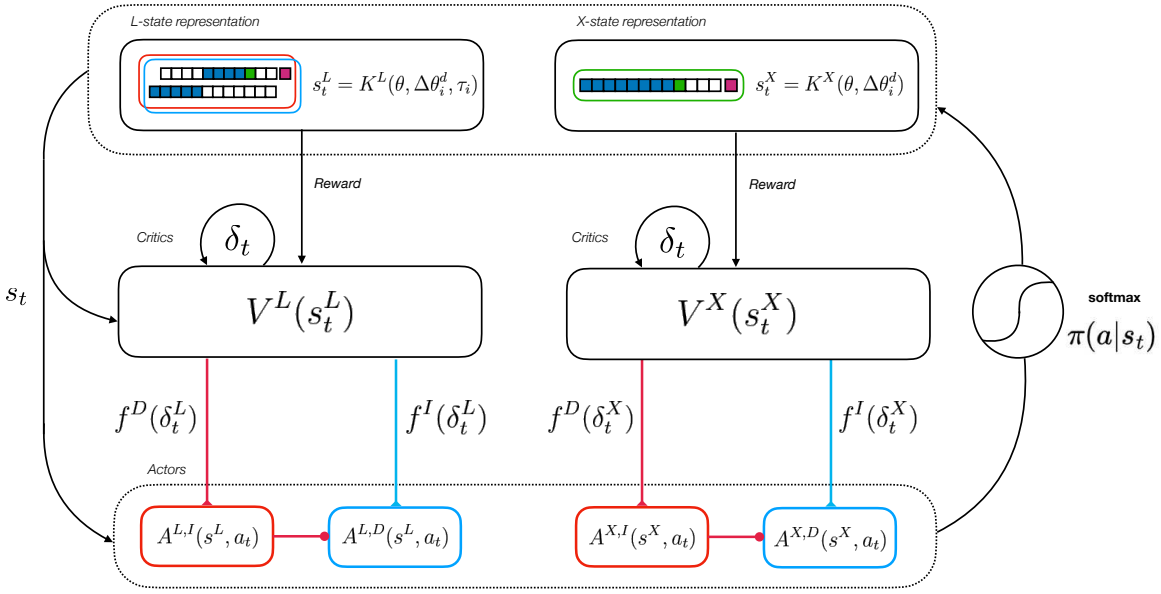


Figure 3.10: Architecture of the multi-agent model with direct and indirect pathway.

3.4.3 Reproducing breaking fixation behaviour

In order to preserve symmetry amongst agents, we endowed both agents with the same RPE transfer functions for the direct and indirect pathways as can be seen in Figure3.11. We will test this new architecture in the task environment containing the two state representations. As with the previous model, the multi-agent model is able to learn the task, producing an equivalent psychometric curve as the previous iteration (see panel b Figure3.11). However, a sharp difference now occurs in the breaking fixation distributions; the bi-modality is now present and is qualitatively similar to the one observed experimentally.

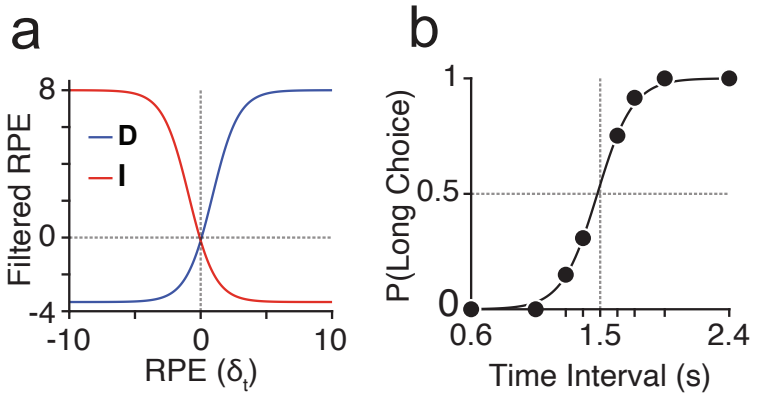


Figure 3.11: **a)**: Transfer functions for the direct and indirect pathways of both the DLS and the X agents. **b)**: Psychometric fit (black line) to the performance of the model for each interval (black dots) used in the time interval categorization task.

The agents are now not only breaking fixation with a higher frequency, but they're also choosing what would be the correct answer for that region of the delay period i.e. breaking and choosing SHORT in the first half of the period, followed by breaking and choosing LONG in the second half, as was observed in the data. The hazard rate (Figure 3.12) of breaking for each contra-lateral side also matches the observations, with a slight leftward hazard in the first half of the delay period and an almost increasing hazard in the second half.

The mechanics of this new behavior are a consequence of the new sub-agent containing a different and more compact state representation. As we'll see in the following sections, this new representation is responsible for shaping a new set of action-preferences which are able to positively enforce the correct action selection and simultaneously force the DLS inhibition to adjust itself to this new action execution impetus.

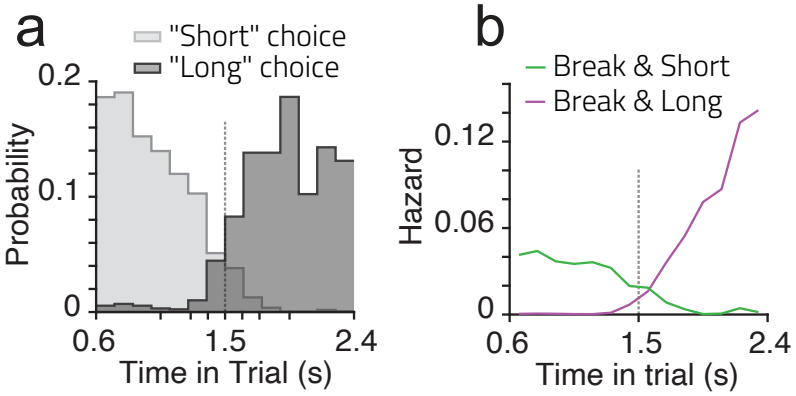


Figure 3.12: **a)**: Probability density functions of broken fixations, produced by the model agent, over time (0.6s to 2.4s), conditioned on subsequent choice at one of the side ports (light grey: SHORT choice, dark grey: LONG choice). **b)**: Hazard rate of broken fixation conditioned on subsequent choice of one of the side ports (SHORT and LONG as green and purple, respectively).

3.4.4 Modeling opponent dynamics in direct and indirect pathway

If one now looks at the profiles of the action preferences, the flat-lined profile of the corresponding DLS (L agent) agent now gives way to a non-monotonic one. In particular, the action preferences that correspond to the iMSN neurons for the left and right actions now show an increase in inhibition for the first and second halves of the delay period, respectively. The corresponding DLS direct pathway action

preferences are oppositely tuned to their indirect counterparts, showing a slight increase in the short-action action preference and a decrease for the long-action.

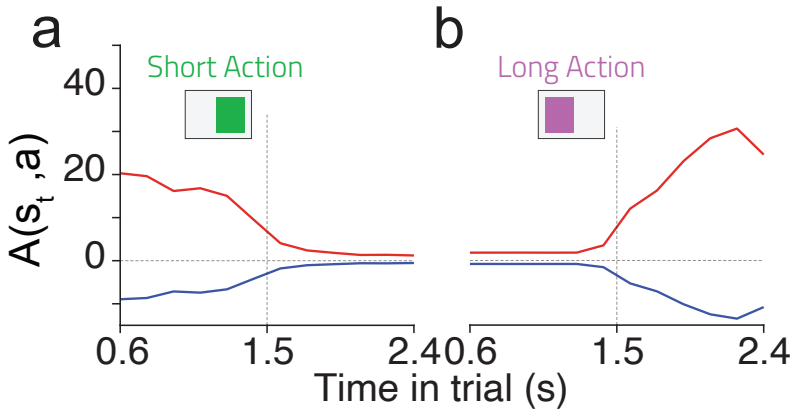


Figure 3.13: Action preferences for two actors ($A^{L,D}(s_t^L, a_t)$, in blue and $A^{X,I}(s_t^L, a_t)$, in red) that make up the L agent. Left and right showthe action preferences for the LONG, and SHORT actions, respectively.

The dynamics of the indirect pathway action-preferences can be better understood if one now looks at the X-pathway action preferences, also for the direct and indirect pathways. In this case we see a flip in terms of the relevant pathwhays that exert action-selection pressure, now with the X agent direct pathway generating selection pressure for the left-action in the first half of the DP and for the right action in the second half.

As can be seen from the policy definition in equation 3.22, each set of action preferences corresponding to a different agent shape a different sub-policy. These sub-policies interact so that the total action preference A^T can approximate the optimal policy, with the action-promoting signals emerging from the X pathway direct pathway agent are then counteracted by the inhibition produced by the L agent indirect pathway action-preferences (corresponding to the DLS).

The main difference in these two agents is the state-representation, which is the main mechanism that allows for the accumulation of different RPE signals in the states corresponding to the delay period in each one of the agents. As both

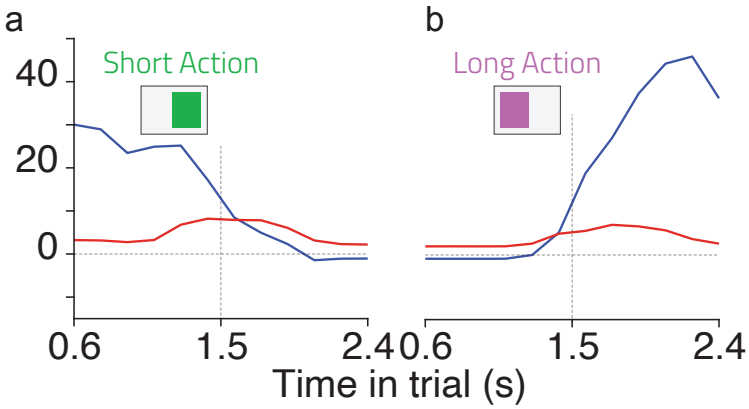


Figure 3.14: Action preferences for two actors ($A^{X,D}(s_t^X, a_t)$, in blue and $A^{X,I}(s_t^X, a_t)$, in red) that make up the X agent. Left and right show the action preferences for the LONG, and SHORT actions, respectively.

agents move through the delay period, the correct execution of an action after a second tone is emitted leads to a positive RPE, which will update the action-preference in different states for each agent. In the case of the L agent, the RPEs will accumulate in the pre and post-second tone states, leading to an increase of direct pathway action preferences only in post-second tones, when the agents should actually choose an action. For the X-pathway, having only one set of states without pre/post-second tone distinction, the accumulation of both positive and negative RPEs occurs throughout the delay period, leading to the emergence of a positive drive signal that matches the optimal post-second tone state policy.

3.4.5 Modeling DLS indirect pathway optogenetic experiments

Given that we were able to reproduce the basic profiles observed in the DLS direct and indirect pathway recordings, can we also reproduce the effect of the optogenetic experiments on the DLS indirect pathway? In the model, these perturbations can be achieved by lowering the pathway weights ω_I, which due to the fact that this

corresponds to an effective decrease of the indirect pathway relevance in the policy, is equivalent to an "inhibition" of the corresponding pathway.

By now fitting this perturbation parameter to the experimental data, we can see that the model reproduces with precision the effect of the inhibition (Figure3.15). As in mice, this led to the failure to suppress action. Furthermore, inhibiting the 'short' action preference led to failed action suppression early, whereas inhibiting the long action preference led to failed action suppression late, in the delay period. In addition, unilateral inhibition also caused an increase in contralateral action, as in the experimental data (Figure 2.6). The results are both valid for the probability of breaking fixation and the hazard, the latter can be seen in Figure3.16.

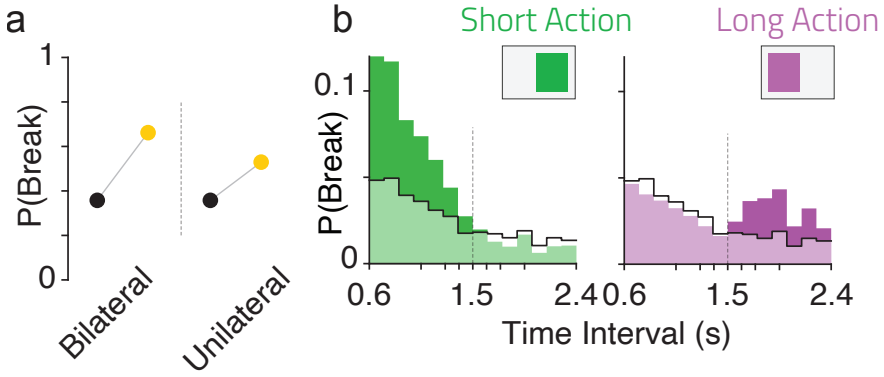


Figure 3.15: **a)** Effect of action preference inhibition of the indirect pathway for the two actions simultaneously or for only short or long (control and manipulated trials shown as black and yellow, respectively) **b)** Histogram of normalized broken fixation counts for control (black outline) and inhibition of short (green) and long (purple) DLS sub-agent indirect-pathway action preferences.

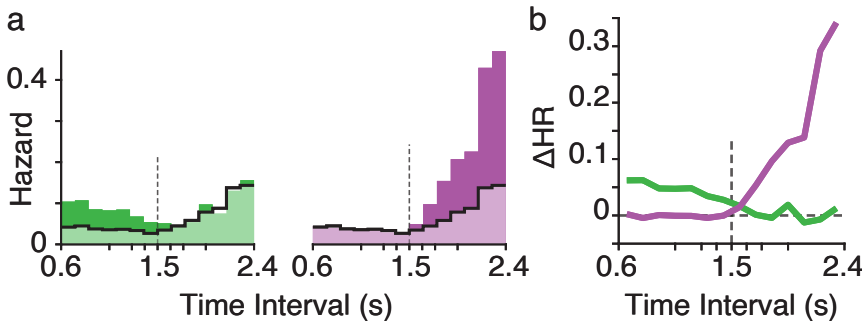


Figure 3.16: **a)** Hazard of breaking fixation for the control (black outline) and inhibition conditions resulting from the selective reduction of action preference values for the Indirect pathway $A^{L,I}(s_t^L, a_t)$ for the SHORT (green) or LONG (purple) actions. **b)**, Change in hazard rate $\Delta HR = HR_{Manipulation} - HR_{Control}$.

3.4.6 Prediction of an effect on DMS direct pathway optogenetic perturbations

As we've seen in Figure3.14, the drive for action-selection and its execution seems to not be a consequence of the L-agent direct pathway, but of the direct pathway drive being expressed in the X-agent direct pathway. Given this, the model would predict that if we now inhibit the $A^{X,D}$ action preferences the amount of breaking fixations should decrease, as the drive for action promotion would decrease as well.

Previous data have shown that optogenetic activation of DMS MSNs can affect action selection (Tai et al., 2012). We thus wondered whether more medial circuits in the DMS might contribute to action promotion similarly to what has been observed in the X-pathway of the model. We repeated the optogenetic inhibition experiments carried out in DLS, but now targeting the DMS dMSNs in separate groups of mice. In general, we observed a modest but significant decrease in the probability of contralateral choice after broken fixations when inhibiting DMS dMSNs, as can be seen in Figure 3.17.

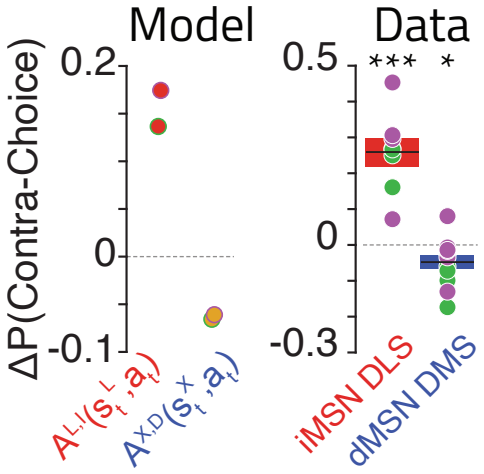


Figure 3.17: **a)** Change in probability of generating a contralateral action when decreasing the magnitude of action preference values, of the indirect pathway, $A^{L,I}(s_t^L, a_t)$ (red outline), or direct pathway $A^{X,D}(s_t^X, a_t)$ (gold outline), for one of the two actions (SHORT or LONG) before the second tone observation (i.e. broken fixations). **b)** Change in probability of registering a choice at the port contralateral to the hemisphere manipulated when inhibiting iMSNs in DLS or dMSNs in DMS, after breaking fixation.

3.5 Discussion

The basal ganglia is characterized by a parallel circuit architecture, which manages distinct sets of sensorimotor, associative, and limbic information. This architecture is key to its largely segregated operation (Alexander et al., 1986; Alexander and Crutcher, 1990; Prescott et al., 2006; Lau and Glimcher, 2007), which converges at the level of action promotion and behavioral expression.

Building on this concept, we designed a model that integrates these parallel circuits and varying plasticity dynamics in the direct and indirect pathways. Our model, based on a multi-agent Reinforcement Learning paradigm, successfully captures both the behavior and the neural profiles of populations in the dorso-lateral striatum. Moreover, it can emulate the behavioral effect of inhibiting certain populations in different striatal regions, namely, the DLS and DMS, while also offering a strong prediction for the potential profiles of the DMS activity profiles (Figure 3.14) (Cruz et al., 2022).

Different functions have been proposed for the parallel circuits, with emphasis on arbitration between types of control as per context or task demands, such as model-based versus model-free (Daw et al., 2005), or Pavlovian versus instrumental control (Dorfman and Gershman, 2019). Our model integrates these functional hypotheses by anchoring each agent in a different state representation. Each state representation leads to different RPE signals being generated as the animal experiences the task environment.

Given the corticostriatal projectome structure presented in section 1.2.2 of Chapter 1, we hypothesize that these different state representations would also lie in different (but still connected) cortical regions. Sensorimotor areas tend to be more event specific, thus containing second tone information, while the dorsomedial regions tend to project to more prefrontal areas, which might lead to the learning of more abstract and compressed state representations of the environment.

These different representations generate multiple, anticipatory, and possibly conflicting drives on behavior, (Dayan et al., 2006) which the system needs to manage via the promotion and suppression of actions. This provides a rationale for a suppressive indirect pathway, linking RL-based control to traditional BG circuit function proposals (Mink, 1996) and advancing previous work that has proposed a single-agent architecture multi-pathway approach to modeling basal ganglia functionality (Collins and Frank, 2014).

Adaptive behavior involves balancing action-promoting and suppressing mechanisms. Our data show that in the sensorimotor striatum, elements of the direct

and indirect BG pathways can express antagonistic modulation patterns, underlying opposing yet distinct aspects of motor function. Understanding the interplay in other brain circuits is crucial for grasping animal behavior at a fundamental level.

This model provides a strong link between the multi-agent reinforcement learning paradigm and basal ganglia circuits, opening the way towards the development of even more detailed functional architectures which have as a theoretical anchor the idea of multiplicity in state-representations and a converging interaction in the policy, which generates different control signals throughout the striatum and the underlying direct and indirect populations, each working towards some aspect of promotion and suppression of actions.

Chapter 4

Representational multiplicity in the basal ganglia

4.1 Introduction

Organisms in general have access to limited amounts of energy and time to perform the necessary computations for their survival. Given this constraint, there is an underlying need to be as efficient as possible in processing environment-related information. One hypothesis is that this processing is done via the compression of state and action-related information via structuring predictive models of the world (Tishby and Polani, 2010).

As we've seen in the model developed in Chapter 3, the reproduction of the iMSN activity requires that the urge to act is formed in a parallel circuit, with a different state representation. Given that the mice choose the correct port after breaking fixation, we assumed that this state representation should allow for the formation of a more abstract perception of the reward locations, inciting an urge to act and select the left or right port at the corresponding region of the delay period.

One way to represent this is to consider an alternative version of the model presented in Chapter 3. Whereas as the X-model presented in Chapter 3 had two pathways, we can alternatively build a similar model with just one positively tuned pahthway with transfer functions that filter out only positive RPEs; this results in a set of action preferences for the X-agent which only contain the driving functions for choosing the short and long nose-ports during the delay period.

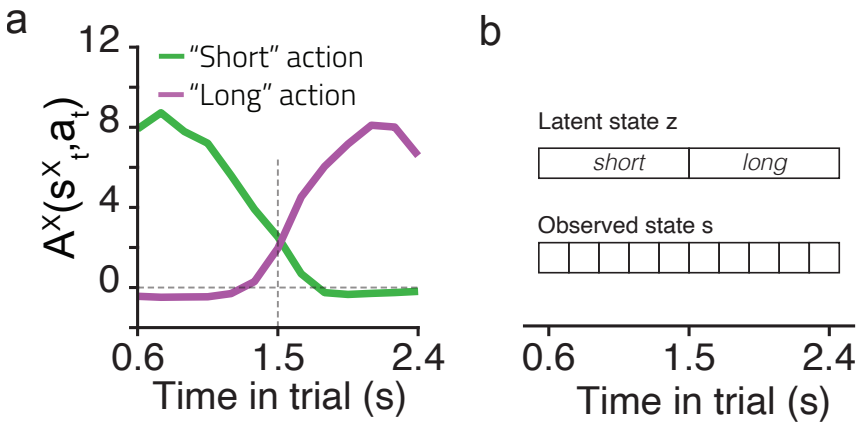


Figure 4.1: **a)**: Action preferences $A^X(s_t^X, a_t)$ for an alternative formulation of the model one single positively tuned pathway **b)**: Diagram of the abstraction present in the learned action preferences that could be represented in a set of two latent states z. Half of the delay period is compressed into a *short* action state and the remaining part in the *long* action state.

As one can see in *panel a* of Figure 4.1 the action preferences for this agent have a bimodal structure that associates each half of the delay period with the corresponding action. The action preferences reflect short vs long state abstractions which intuitively could be thought of as clustering of states according to the specific actions leading to reward in those states. This mechanism could be thought of as an abstraction over the set of observed states s, which depending on their location in the delay period could be abstracted or compressed into two latent states corresponding to the *short* action selection region and the *long* action selection region, as one can see in *panel b* of Figure 4.1.

However, the question remains of how the underlying state representation could be learned. The following chapter will explore two possibilities for such a process; one based on the *information bottleneck* (IB) method (Tishby et al., 2000) as a means for state compression in conjunction with direct policy learning. Subsequently, we provide a more general theoretical perspective integrating the IB methodology with the variational inference approach to reinforcement learning (Levine, 2018).

4.2 State abstraction for control via the Information Bottleneck method

Given the structure of the interval discrimination task, how could one construct a useful latent state space for action control? In order to address this question, we leverage the information bottleneck perspective (Tishby et al., 2000).

This ubiquitous framework has been applied in many distinct domains of machine learning and computational neuroscience Schneidman et al. (2001), as it allows a partially supervised construction of compressed representations, which is also referred to as latent representations, from representations lying in higher dimensional spaces.

The *Information Bottleneck (IB)* principle is a method in information theory for extracting relevant information that a random variable X contains about another random variable Y. It's particularly useful in scenarios where you want to compress the information in X while retaining its predictive power about Y.

The key idea is to find a compressed representation Z of X that maximizes the mutual information with Y, while minimizing the mutual information between

Z and X. This is akin to passing X through a 'bottleneck,' keeping only the information that is most relevant for predicting Y.

The formal objective function in the IB principle is a trade-off between two mutual information quantities, expressed as follows:

$$p^* \coloneqq \arg\min_{p(t|x)} I(X:Z) - \beta I(Z:Y) \tag{4.1}$$

Where $I(X:Z)$ is the mutual information between X and Z, which we want to minimize. as it represents the compression of X into Z. $I(Z:Y)$ is the mutual information between Z and Y, which we want to maximize. This term quantifies the amount of relevant information about Y that is preserved in Z and β is a Lagrange multiplier that controls the trade-off between compression and preservation of relevant information. A higher value of β places more emphasis on retaining information about Y.

The mutual information terms can be defined as:

$$I(X:Z) = \sum_{x,z} p(x,z) \log \frac{p(x,z)}{p(x)p(z)} = \left\langle \log \frac{p(x|z)}{p(x)} \right\rangle_{p(x|z)p(z)} \tag{4.2}$$

$$I(Z:Y) = \sum_{z,y} p(z,y) \log \frac{p(z,y)}{p(z)p(y)} = \left\langle \log \frac{p(z|y)}{p(z)} \right\rangle_{p(z|y)p(y)} \tag{4.3}$$

Where $p(x,z)$, $p(z,y)$, $p(x)$, $p(z)$, and $p(y)$ are the joint and marginal probability distributions of X, Z, and Y and we use the relationship between the joint and the conditional probability to define the the averages in angular brackets[1]

In practice, solving the IB problem often involves iterative algorithms such as the *Blahut-Arimoto* algorithm (which we will present in the following section), as it's not straightforward to directly solve the optimization problem due to the dependencies between $p(z|x)$ and the marginal distributions.

Adapting the base formalism defined in (Tishby et al., 2000), we can bring this framework to the MDP setting by defining a set of world or perceived states $s \in \mathcal{S}$, a set of latent states $z \in \mathcal{Z}$ (usually with a lower dimensionality) and a set of actions $a \in \mathcal{A}$. The number of states and actions are given by n_s, n_z, and n_a, respectively. Previous work has studied the IB method in the MDP context has focused on

[1] $p(x,z) = p(x|z)p(z)$

exploring policy compression Tishby and Polani (2010); Lai and Gershman (2021). Here, the IB method is applied to the dependencies between state, action, and reward $S \to A \to R$ which a view to compressing action encoding conditional on state while attempting to simultaneously maximize reward obtained.

In contrast, we consider the dependencies between random variable $S \to Z \to A$ such that we seek to compress observed states into latent states and while retaining the ability to execute the optimal actions accurately. That is, given an optimal policy $\pi^*(a|s)$, we aim to learn a latent space z such that the induced policy $\pi^*(a|z) \approx \pi^*(a|s)$.

An essential component of this framework is to construct a *perceptual policy* $\rho(z|s)$ that encodes a given set of states s in a set of states with a lower dimensional z. In other words, the perceptual policy aims at representing the internal compressed construction that an organism or agent creates of the world, due to informational and perceptional constraints inherent in its bounded interaction with the world. We can define a loss function to optimize the construction of ρ. More explicitly, we can construct the following functional (notation adapted from Tishby et al. (2000))

$$\rho^* \coloneqq \arg\max_{\rho} [\mathcal{I}_\pi(Z:A) - \lambda \mathcal{I}_\rho(S:Z)] \tag{4.4}$$

where we have the following definitions of mutual information between latent states and actions

$$\mathcal{I}_\pi(Z:A) = \sum_{z,a} p(z,a) \log \frac{p(z,a)}{p(z)p(a)} = \left\langle \log \frac{\pi(a|z)}{p(a)} \right\rangle_{\pi(a|z)p(z)} \tag{4.5}$$

and latent states with the observed states s

$$\mathcal{I}_\pi(Z:S) = \sum_{z,s} p(z,s) \log \frac{p(z,s)}{p(z)p(s)} = \left\langle \log \frac{\rho(z|s)}{p(z)} \right\rangle_{\rho(z|s)p(s)}. \tag{4.6}$$

We'll concern ourselves with how these latent representations could be learned and how their dynamic evolution might lead to observable changes in behavior.

We also note that the agent can learn a policy over the observable states first while simultaneously learning the configuration of the latent states. We can expand the loss in 4.4 to contain a cost referent to the state-action mutual information

$\mathcal{I}(\mathcal{S}:\mathcal{A})$. This mutual information can be directly expanded into observable state policy terms and action-only dependent terms via

$$\mathcal{I}(S:A) = \left\langle \log \frac{\pi(a|s)}{p_\pi(a)} \right\rangle_{\pi(a|s)p(s)} \tag{4.7}$$

so that the complete loss can be written as a multi-objective optimization procedure over both the behavioral and perceptual policies π, ρ

$$\{\pi^*, \rho^*\} \coloneqq \arg\max_{\pi,\rho} [\mathcal{I}_\pi(S:A) + \mathcal{I}_\pi(Z:A) - \lambda \mathcal{I}_\rho(S:Z)] \tag{4.8}$$

Intuitively, this problem can be seen as a multi-objective information bottleneck problem such that a given triplet of random variables (e.g. $S \to Z \to A$) needs to be solved for optimal π and ρ. However, in order to simplify our presentation in the following examples, we will assume that the optimal action policy π^* has been obtained, serving as the anchor to learning the perceptual policy ρ.

4.2.1 Blahut-Arimoto equations

Knowing the optimal policy $\pi^*(a|s)$, the behavioral policy over latent states $\pi(a|z)$ now needs to match the same solution. Given that these are two probability distributions over actions, one can define a distance metric and find the perceptual policy configuration ρ that minimizes this distance. We can solve the iterative IB problem over two policies defined over the set of observable $s \in S$ and latent $z \in Z$ states by defining such distance metric as the *Kullback-Liebler* distance between them, i.e.

$$d(s,z) = D_{KL}[\pi(a|s)||\pi(a|z)] \tag{4.9}$$

where the goal is to fit the target policy over the latent states $\pi(a|z)$ to the optimal policy over observed states $\pi(a|s)^*$.

The KL distance can be expressed as

$$D_{KL}[\pi(a|s)||\pi(a|z)] = \sum_{s,a,z} \pi(a|s) \log \frac{\pi(a|s)}{\pi(a|z)} \tag{4.10}$$

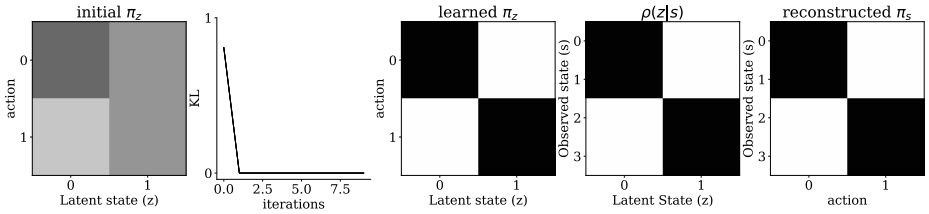


Figure 4.2: Example simulation of the IB optimisation for $n_s = 4, n_z = 2, n_a = 2$, with initial conditions for the latent space policy π_z on panel **a**, the KL divergence on panel **b** and the final latent space policy π_z (panel **c**), perceptual policy $\rho(z|s)$ (panel **d**) and the reconstructed observed state policy π_s (panel **e**).

In order to perform the iterative updates for a given s and a given z, we define $d(s, z)$ as a sum only over the the allowed actions, i.e.

$$d(s, z) = \sum_a \pi(a|s) \log \frac{\pi(a|s)}{\pi(a|z)}. \tag{4.11}$$

The full set of iterative IB equations read as follows (Tishby et al., 2000)

$$\begin{cases} \rho_{t+1}(z|s) = \frac{p_t(z)}{Z(s,\beta)} e^{-\beta d(z,s)} \\ p_{t+1}(z) = \sum_s p(s)\rho_t(z|s) \\ \pi_{t+1}(a|z) = \sum_s \pi(a|s) \frac{\rho_t(z|s)p(s)}{p_t(z)} \end{cases} \tag{4.12}$$

where the normalization factor $Z(s, \beta)$ is given by

$$Z(s, \beta) = \sum_z p(z) e^{-\beta d(s,z)} \tag{4.13}$$

and from Bayes conditional rule,

$$p(s|z) = \frac{\rho(z|s)p(s)}{p(z)}. \tag{4.14}$$

we can extract the perceptual policy $\rho(z|s)$.

4.2.2 Simulating a toy example

We can now apply this set of equations to a simplified version of the temporal discrimination task. If we look at the action preferences obtained in Section 3.4.4, particularly for the X-agent, one can see that the action preferences are clustered in each of the delay period regions, for the left and right action, respectively.

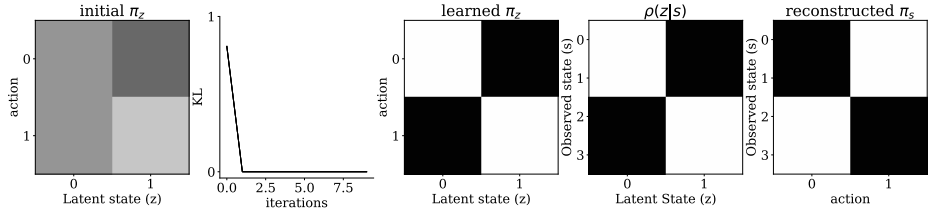


Figure 4.3: Same simulation as in Figure B.1 but with different initial conditions.

In this example, we'll attempt to learn the optimal latent state policy $\pi(a|z)$ and the perceptual policy $\rho(z|s)$ for the case where the number of states is $n_s = 4, n_z = 2, n_a = 2$.

As one can see in Figure B.1, the IB algorithm can converge to the optimal policy for the latent state policy. However, this is dependent on initial conditions. If one starts from a set of initial states that are inversely biased, the perceptual policy converges to a degenerate solution that matches this new configuration (see Figure 4.3) while still being consistent with the optimal policy. This degeneracy is observed for multiple initial conditions[2].

With this toy example, we were able to show that the Information Bottleneck objective can learn an optimal policy for a simplified version of the interval discrimination task. However, the prerequisites of this formalism imply that the number of states is already defined, which is a very strong constraint for a process that is adaptive by nature. Even though an optimal policy for this latent state space can be learned, the question remains as to how a biological system could know *a priori* the number of states to perform the compression.

Due to this, we will now explore a different but related approach, based on variational inference, which will learn a new latent state representation based on the initial state representation defined for the L-agent defined in 3.

[2]See Appendix A

4.3 Affordances as Perceptual Policies

Although the previous formulation gives us the right answer as to the behavioral policy of a lower-dimensional latent state-representation, its definition requires that we preemptively choose the number of states that are relevant for a given task, which already imposes a very strong constraint on the structure of the latent representation z.

In the following sections, we'll derive an alternative formulation of this problem, one that is less strongly constrained and also easier to implement in the original algorithm of the multi-agent model presented in Chapter 3. We will base ourselves in the *variational inference* formulation of reinforcement learning (Levine, 2018) using the underlying optimization procedures as the backdrop for the computations needed to shape the state representation that is used by the X-agent.

RL as Inference

Our main goal is to define a global loss function for this problem that takes into account both the behavioral policy optimization and the optimization of the perceptual policy $\rho(z|s_t)$ but allows flexibility in choosing which method is used to optimize either one. We will focus mostly on the optimization of ρ, assuming that an optimal behavioral policy π has been obtained via a single two-pathway agent as the one presented in Section 3.3.

The intuition behind this formulation is that behavioral policies benefit from being **maximum entropy** regularized (Levine, 2018; Ghugare et al., 2023) whilst perceptual policies benefit from being **minimum entropy** regularized. This distinction can be justified from a biological perspective. For any organism to learn the reward structure of a given environment, its state representation has to be as stable as possible, i.e. the learned latent state identities shouldn't change as every time a new stimulus is presented, as then the credit-assignment problem would be intractable i.e. the latent states sampled from $\rho(z|s)$ should not be highly variable at a given time point.

The MDP for this problem is similar to the one defined for the variational inference RL framework defined in Levine (2018), where the optimization is guided by an optimization variable $o_t \sim e^{r_t}$ which defines which state-action pairs are optimal for a given path. We define a similar optimality condition for the latent state z_t, given by an *affordance* function g over the future latent states z_t.

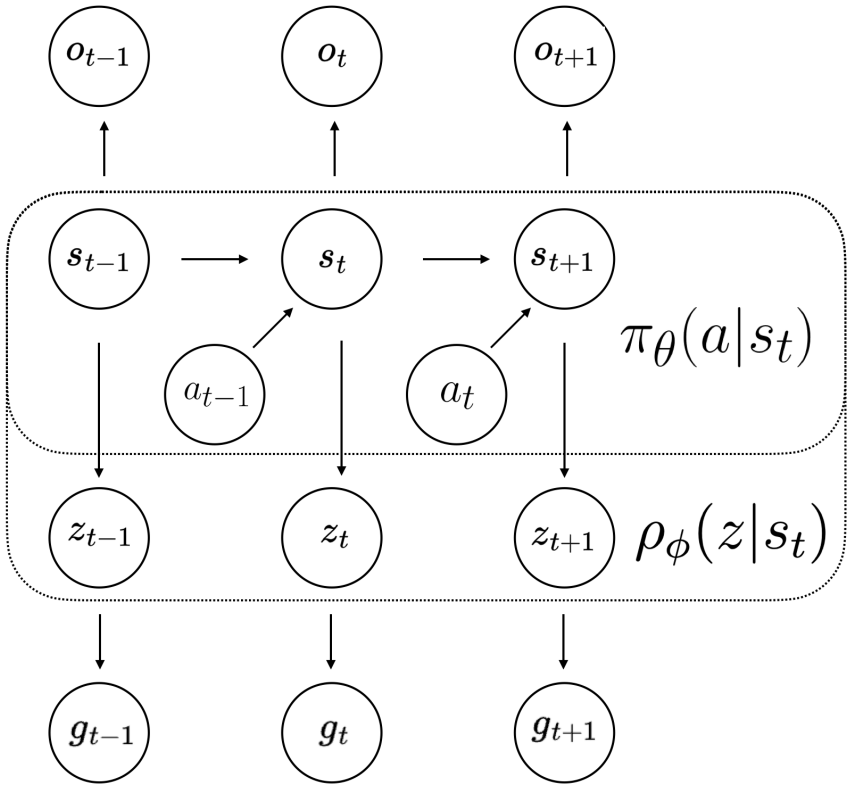


Figure 4.4: Markov Decision Process structure with a latent variable z_t with optimality variables g_t, defined similarly to the optimality variables o_t defined for the observable state representation s_t.

Affordances

The concept of affordance, first defined by James Gibson in the 70s (Gibson, 1979), was a theoretical concept that aimed at portraying how object classes in the world were not static but a collection of use-case projections that depended on the organism's relationship with the world, as he put it

> The fact that a stone is a missile does not imply that it cannot be other things as well. It can be a paperweight, a bookend, a hammer, or a pendulum bob. It can be piled on another rock to make a cairn or a stone wall. These affordances are all consistent with one another. The differences between them are not clear-cut, and the arbitrary names by which they are called do not count for perception. If you know what can be done with a graspable detached object, and what it can be used for, you can call it whatever you please.
>
> - Gibson (1979)

As one can already intuit from this description, the concept of affordance itself is a latent property; the use cases of something in the immediate future are latent in the sense of being possibilities for a given organism with a goal in a given environment. The utility of something in the immediate future thus entail some amount of prediction, so we intuit that it's dependent on a learned model of the world and the evaluation of some sort of policy in a **future state**.

This model should learn the statistics of the occurring transitions as a given path in the state space of state-action pairs $s_t \in \mathcal{S}, a_t \in \mathcal{A}$ and state-latent-state pairs $z_t \in \mathcal{S}$ is sampled from the space of all possible paths, i.e. τ_s, τ_z are sampled paths for each set of trajectories [3]

$$\tau_s = \{s_0, a_0, ..., s_k, a_k, ...s_T, a_T\} \tag{4.15}$$
$$\tau_z = \{s_0, z_0, ..., s_k, z_k, ...s_T, z_T\} \tag{4.16}$$

We will calculate the expected value of the relevant observables (such as value or entropy) over a sampled path from these two sets of random variables. The

[3] Note that the notational mapping to the model developed in Chapter 3 is $s_t^L \rightarrow s_t$ and $s_t^X \rightarrow z_t$.

dynamics of the environment that generates these paths are given by a transition tensor $T(s_t, a_t, s_{t+1})$ where we consider that the passive action probabilities $p(a_t)$ for the environment dynamics are uniform and independent from the transition i.e. $T(s_t, a_t, s_{t+1}) \approx T(s_t, s_{t+1})p(a_t)$, so the passive action dynamics are mostly random. Given an already defined optimal policy π^*, this allows us to define the affordance function as a marginalization on actions of the action preferences A, i.e.

Definition (Affordance function): For a set of states $s_t \in \mathcal{S}$, a converged total action preference function[4] $A^T = A$ and a constant β (we assume $\beta = 1$ for simplicity) we define the affordance function,

$$g(s_t, s_{t+1}) = \log \left[\beta \sum_a T(s_t, s_{t+1}) A(s_{t+1}, a) \right]. \tag{4.17}$$

The intuition for this definition goes as follows; we want a function that represents the future utility of a state. In order to know where that future utility lies we use the transition dynamics of the environment T to fetch us the future states s_{t+1} that are available for a given present state s_t.

However, this is not enough because we want to know not only the future accessible states, but those which have been useful for the agent in past experience. Given that the agent can record that experience in its advantage function A, we can perform a marginalization over the actions a for each future state s_{t+1} that is available for a current state s_t, leading to the weighted sum of transitions.

Internal Model

Besides the passive dynamics of the environment we also define an internal model for the agent, given by $\hat{T}(s_t, s_{t+1}) \in S^{N \times N}$ and learned through the following TD iterative algorithm (Glascher et al., 2010; Gardner et al., 2018)

$$\hat{T}(s_t, s_{t+1}) \leftarrow \hat{T}(s_t, s_{t+1}) + \eta[\mathbb{1}(s_{t+1} = s') - \hat{T}(s_t, s_{t+1})] \tag{4.18}$$

where η is a learning rate and $\mathbb{1}(s_{t+1} = s')$ is the unitary count function that is 1 every time a transition from a specific state s_t to s_{t+1} occurs. Given that we're associating this state representation to a more medial region of the striatum (see Section 3.4.2), we're also leaning on a previously state hypothesis by Bornstein

[4]Choice made in order not to overload notation.

and Daw (2011) regarding a model-free/model-based divide between dorsal and medial striatal circuits (see also Section 1.3).

Given the base state representation s^X defined in Chapter 3, we can learn the transition matrix $\hat{T}$ by running the agent, with a learned optimal policy through the environment. The resulting matrix can be seen in Figure 4.5.

The matrix consists of two diagonal sets of transitions, as for each pre 2nd tone state (states 0-40 in Figure 4.6) there's a chance that a second tone can occur at that time point, which will lead the agent to a post 2nd tone state (states 40-80). One can also observe the structure of these transitions by looking each individual probability trace, shown in Figure 4.6.

Perceptual Policy

For the derivations presented in the sections conscerning the application of this method to the IDT, we'll assume that the dimensionality of the observed and latent variables are the same i.e. $n_s = n_z$.

To define the actual *perceptual policy* conditional distribution over the set of latent future states $z \in \mathcal{Z}$ and observed states $s \in \mathcal{S}$, $\rho(z|s_t)$, we will anchor ourselves on the environment model $\hat{T}$ defined in Equation 4.18 and perform a weighted distributed representation over each one of the rows with a state-dependent **filtering function** $\phi(z) \in \mathbb{R}^{n_z}$.

Although the internal model matrix $\hat{T}$ was previously defined as a a function of $\hat{T}(s_t, s_{t+1})$, we will now use z as a pointer to a latent future state s_{t+1}, given that these two spaces share the same dimensionality. The future latent state z is defined as such due to the fact that it will only be a concrete state after being sampled from the perceptual policy ρ, which is the filtered internal model matrix $\hat{T}$ that we will derive in the following sections.

We define the filtering function ϕ as a free variational function which will be optimized in order to reflect the affordance structure of the state-action space. This function is a pivotal component of the derivation that allows us to integrate our computational affordance hypothesis into the RL-as-inference framework (Levine, 2018). We achieve this by performing the the element-wise product between $\hat{T}$ and ϕ i.e.

Definition: (Perceptual Policy) The perceptual policy distribution is defined as a filtering of the internal model $\hat{T}$ by the function ϕ via

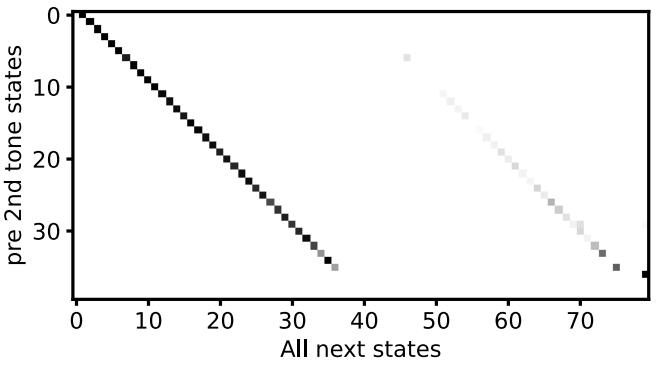


Figure 4.5: Transition matrix for the Interval Discrimination Task, learned with an optimal policy $\pi^*(a|s)$, darker colors are higher transition probability.

$$\rho_\phi(z|s) \approx \hat{T}(s,z) \odot e^{\phi(z)} \tag{4.19}$$

which, can be written also as a joint probability

$$p_\phi(s_t, z) = \hat{T}(s_t, z) e^{\phi(z)} \tag{4.20}$$

this latter equation will be used to compute the optimization needed to obtain the correct form of ϕ.

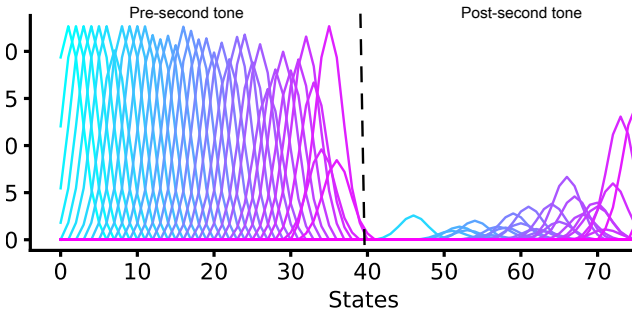


Figure 4.6: Traces of the probability distributions present in $\hat{T}(s_t, s_{t+1})$. Blue traces correspond to earlier states, purple traces to later ones, each corresponding to a row in the transition matrix $\hat{T}$

Compression in the non-isomorphic case

Given that conceptualisation of affordances lies under a principle of compression, we'll briefly outline how this could be implemented in this framework i.e. we'll briefly look at the non-isomorphic case $n_{\hat{s}} < n_z$, where $\hat{s} \in \hat{S}$ is a compressed version of the state space S. Given, that the internal model $\hat{T}$ contains all of the state-transition information available in the environment, we can use this matrix and compress the underlying transition structure for future states. In any given environment, not all states will be available for future transitions, so it's likely that the pattern of transitions from s to a future state z can be encoded into a lower dimensional space.

As we're dealing with probability transition matrices which are non-negative, one of the most straightforward approaches to factorising them is to use Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999, 2001). In NMF, the primary objective is to decompose a given non-negative matrix $\hat{T}$ into two non-negative matrices W and H, such that $\hat{T} \approx WH$. The matrices W and H are often interpreted as the basis and coefficient matrices respectively, underpinning the data represented by $\hat{T}$. The formulation of NMF is inherently tied to the optimization problem where one seeks to minimize the discrepancy between $\hat{T}$ and the product WH, subject to the non-negativity constraints, concretely

$$\min_{W,H} \|\hat{T} - WH\|_F^2 \quad \text{s.t.} \quad W \geq 0, H \geq 0 \tag{4.21}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This optimization is typically non-convex and non-analytic, and is tackled through iterative methods that alternately update W and H, gradually improving the approximation WH to $\hat{T}$. A widely adopted approach for solving this optimization problem is through the use of update rules derived from the Karush-Kuhn-Tucker (KKT) conditions for the non-negativity constraints (Lee and Seung, 2001).

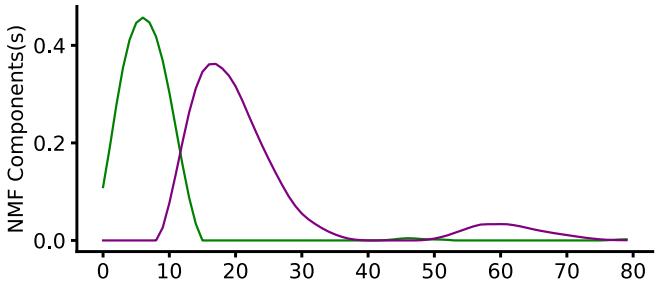


Figure 4.7: First two non-negative components of the basis matrix W for the transition matrix $\hat{T}$.

For our particular case in the Interval Discrimination Task we can fully describe the transition structure of the observed states s with two components that span the short and long interval transitions as one can see in Figure 4.7. These components contain a bimodal structure which spans both pre (before state 40) and post-second tone states (past state 40), summarizing the structure that's visible in the traces presented in Figure 4.6.

This architecture is equivalent to defining a two-dimensional encoding of S, which for our particular construction would be equivalent to defining a bottleneck with the transition structure $S \rightarrow \hat{S} \rightarrow Z$, where the initial compression of S would be performed through the NMF method.

Given this compression of S one could then adapt the perceptual policy definition so as to be based of the basis matrix W, concretely

$$\rho_{\phi}(z|\hat{s}) \approx W(\hat{s}, z) \odot e^{\phi(z)} \tag{4.22}$$

This serves as a basic example of how compression on the observed states could be performed. In the following sections we will explore how affordances could be implemented in this framework so as to generate a new distribution of future latent states Z.

4.3.1 Defining a novel loss function

In the most general case, solving an MDP should entail obtaining both the optimal behavioral policy alongside the perceptual policy. In the maximum entropy paradigm, finding the optimal behavioral policy implies maximizing the entropy of the distribution π_{θ} to explore more effectively the full state space.

The global loss function for this problem will be defined through two sub-losses, $J_q(\theta)$ and $J_p(\phi)$ for the **behavioral** and **perceptual** policies, respectively. The optimization of $J_q(\theta)$ can be done through a multitude of ways such as using soft actor-critic (Haarnoja et al., 2018), DQN (Mnih et al., 2015) or specifically the multi-agent model presented in Chapter 3 (Cruz et al., 2022)[5]. We take inspiration from the framework developed by Levine (2018), where the behavioral policy loss is written as

$$J_q(\theta) = -\mathbb{E}_{\tau_s \sim \pi}[\sum_{t=0}^{T} r(s_t, a_t) + L(q_{\theta})] \tag{4.23}$$

where $r(s, a)$ is the optimality function and $L(q_{\theta})$ is the description length (Rissanen, 1978; Grünwald, 2007) associated with event (s_t, a_t) under the variational

[5]The most consistent solution to this loss function would be methods that are entropy regularized as well, but we relax this condition in order to include the method developed in Chapter 3

distribution $q_\theta(s_t, a_t)$, which we'll try to optimise so as to match the optimal policy π^* (Levine, 2018). It's explicit form is given by

$$L(q_\theta) = -\log q_\theta(s_t, a_t). \tag{4.24}$$

We will expand this initial loss function by adding a novel term relative to the latent states z over the perceptual policy ρ defined via the joint p_ϕ, considering now that the optimality function is the affordance function g defined earlier. The optimisation also takes into account that for the perceptual policy we want the opposite goal to be optimised in what regards entropy i.e. it should be minimised instead of maximised, concretely

$$J(\theta, \phi) = J_q(\theta) - J_p(\phi) \tag{4.25}$$

where

$$J_p(\phi) = -\mathbb{E}_{\tau_z \sim \rho}[\sum_{t=0}^{T} g(s_t, z_t) + L(p_\phi)] \tag{4.26}$$

where the full dependencies of J_p are given by $J_p(\phi, s_t, z_t)$ which we shorten to just $J_p(\phi)$ for easier readability. The future latent state z_t is defined as $z_t = s_{t+1}$ and the term $L(p_\phi)$ refers to the description length (Grünwald, 2007) of event (s_t, z_t) over the joint perceptual policy p_ϕ given by

$$L(p_\phi) = -\log p_\phi(s_t, z_t). \tag{4.27}$$

Note that the entropy for a given future latent state and state pair (z_t, s_t) is defined as [6]

$$H(p_\phi) = -\sum_{t=0}^{T} p_\phi(s_t, z_t) \log p_\phi(s_t, z_t) = -\sum_{t=0}^{T} p_\phi(s_t, z_t) L(p_\phi) \tag{4.28}$$

where we've inserted the definition of the description length of the event (s_t, z_t) under p_ϕ in the latter expression.

[6] A similar definition can used for the variational approximation of the optimal policy q_θ

Entropy minimisation of the perceptual policy

The intuition behind this optimisation goal goes as follows; in order to correctly evaluate value over a given latent state, the relevant state identities must remain constant. As we're sampling from a distribution, the degree to which the future latents z_t remain constant for a given s_t depends on how entropic the joint p_ϕ is. A more entropic distribution p_ϕ would generate more variability in terms of which future latent states z_t would be associated with a given observed state s_t.

Due to the fact that we're trying to perform credit assigment to these future states z_t, if they are not stable enough then this would not be possible, as the corresponding value representations would not be stable as well. In order to correctly evaluate the value functions associated with the latent representation z_t we should minimise the entropy of the corresponding sub-loss function J_p, which will lead to a joint distribution p_ϕ with low entropy as well.

How can we map these intuitions regarding entropy to the corresponding corticostriatal loops that would perform these computations? Although there are no direct measures of the distribution of entropy in state representations across cortex, we can link the underlying reasoning to anatomy by following a line of reasoning that takes into account another observable, the cortical temporal constants referred in Section 1.2.2.

Having a lower entropy in the perceptual policy distribution means that the latent future states that will be selected are more stable; this also implies that the underlying neural temporal constants of the neurons keeping these representations should also be higher. Higher temporal constants have been associated with higher-order associative and prefrontal areas in the cortex (Murray et al., 2014), which neatly map to the corresponding higher order corticostriatal circtuis that we're associating with the X-agent, specifically mapping into the dorsomedial striatum (Alexander and Crutcher, 1990; Haber and Knutson, 2010).

Final expression of the perceptual policy

We now have all the required ingredients to derive the final expression of our perceptual policy loss function. Expanding the point-estimate term over p_ϕ in Equation 4.26 gives

$$J_p(\phi) = -\mathbb{E}_{\tau_z \sim \rho}[\sum_{t=0}^{T} g(s_t, z_t) - \log p_\phi(s_t, z_t)] \tag{4.29}$$

and using the definitions for both the affordance function g and the joint perceptual policy p_ϕ the complete expression is given by

$$J_p(\phi) = -\mathbb{E}_{\tau_z \sim \rho}[\sum_{t=0}^{T} \log \sum_a T(s_t, z_t) A(z, a) - \log \hat{T}(s_t, z_t) e^{\phi(z_t)}]. \tag{4.30}$$

We also make the assumption that the model learned by the agent is close enough to the actual environment probability transition matrix i.e. $\hat{T}$ should converge to T, thus we can simplify the equation above by cancelling $\hat{T}$ and T, concretely

$$J_p(\phi) = -\mathbb{E}_{\tau_z \sim \rho}[\sum_{t=0}^{T} \log \frac{\cancel{T(s_t, z_t)} \sum_a A(z_t, a)}{\cancel{\hat{T}(s_t, z_t)} e^{\phi(z_t)}}]. \tag{4.31}$$

With this we now have the final expression for the perceptual policy loss

$$J_p(\phi) = -\mathbb{E}_{\tau_z \sim \rho}[\sum_{t=0}^{T} \phi(z_t) - \log \sum_a A(z_t, a)] \tag{4.32}$$

which we will try to solve in order to obtain the optimal filtering ϕ that will be applied later on to the internal model M in order to obtain the perceptual policy distribution.

4.3.2 Solving for the perceptual policy

We can now optimize the expression obtained above in regards to $\phi(z)$ by performing gradient descent. First we need to impose a normalisation constraint on p_ϕ so that it remains a distribution, i.e., for a given observed state s_t the joint p_ϕ should be normalised, i.e. for a given path over the pairs (s_t, z_t)

$$\sum_{t=0}^{T} p_\phi(s_t, z_t) = 1. \tag{4.33}$$

As we're calculating an expected value over a path of future latent states z_t, given an observed state s_t, the expectation $\mathbb{E}_{\tau_z \sim \rho}$ turns into

$$J_p(\phi) = -\sum_{t=0}^{T} p_\phi(s_t, z_t)[\phi(z_t) - \log \sum_a A(z_t, a)] + \lambda(\sum_{t=0}^{T} p_\phi(s_t, z_t) - 1) \quad (4.34)$$

where λ is the lagrangian constraint over the normalisation condition. Calculating the minima over the filtering function ϕ for a given latent z gives us

$$\frac{\partial J(\phi)}{\partial \phi(z)} = 0 \Rightarrow p_\phi(s_t, z)[\phi(z) - \log \sum_a A(z, a) + \lambda] = 0 \quad (4.35)$$

as now only the terms relative to the chosen z survive the derivative. This leads us to a closed form expression for $\phi(z)$

$$\phi(z) = \log \sum_a A(z, a) - \lambda \quad (4.36)$$

the normalisation constant λ can be obtained through the constraint condition. For a given s_t we sum over all possible z

$$\sum_z p_\phi(s_t, z) = 1 \Rightarrow \sum_z \hat{T}(s_t, z) e^{\phi(z)} = 1 \quad (4.37)$$

which can be simplified into

$$\sum_z \hat{T}(s_t, z) e^{\log \sum_a A(z,a) - \lambda} = 1 \quad (4.38)$$

thus giving the λ for a specific state s_t

$$\lambda(s_t) = \log \sum_z \sum_a \hat{T}(s_t, z) A(z, a). \quad (4.39)$$

Limit case for the normalisation constant

We now have an unfortunate dependency on s_t, however, given that we're performing the sum over all future states z and all future actions a, with the internal model $\hat{T}$ being a normalised transition matrix ($\sum_z \hat{T}(s_t, z) = 1$) we can define an upper bound constant K which is dependent on the maximum value that the action preferences can take for a given action a, $\max_z A(z, a)$. One can then show that $|\lambda(s_t)| \leq |K|$ as for a given state s_t we have the following inequality

$$\left| \log \sum_z \hat{T}(s_t, z) \sum_a A(z, a) \right| < \left| \log \max_z n_s \sum_a A(z, a) \right| \tag{4.40}$$

where n_s is the total number of states. This allows us to use the normalisation term as a constant instead of it being dependent on an observed state s_t. The final equation for the filtering vector $\phi(z)$ is then

$$\phi(z) = \log \left[\frac{1}{K} \sum_a A(z, a) \right] \tag{4.41}$$

which can now be calculated in a straightforward manner for a given future state z.

4.4 Linking with the multi-agent model

In the specific case where the optimal policy is learned via the DLS agent action preference (see the definition of the multi-agent model in Chapter 3) A^L, we can write the filtering vector for agent X as

$$\phi^X(z) = \log \left[\frac{1}{K} \sum_a A^L(z, a) \right]. \tag{4.42}$$

As we can already see from this equation, we have constructed a hierarchy amongst the agents X and L. The state representation of agent X is now fully dependent on the action preferences of the L agent. These action preferences are the filter which will select, from the learned transition model $\hat{T}^L$ of agent L, which states z_t the new agent should pay attention at each s_t (see the MDP diagram in Figure 4.3).

The filtering function for agent X - ϕ^X - is thus given by the expression

$$\phi^X(z) = \log\left[\sum_a \sum_p \omega_{p,L} A^{p,L}(z,a)\right] - \log K \approx \log\left[\sum_a \sum_p \omega_{p,L} A^{p,L}(z,a)\right] \quad (4.43)$$

so the final perceptual policy expression is given by a sampling from the filtered transition model, i.e.

$$z \sim \rho^X(z|s_t) \sim \hat{T}^L(s_t,z) \odot \sum_a \sum_p \omega_{p,L} A^{p,L}(z,a). \quad (4.44)$$

For a given state state observed by the agent L - s_t the perceptual policy (a vector of size n_z) will select which future states the agent should pay attention to as containing useful affordances in the environment. These selected states (z which in the notation of Chapter 3 will turn into s_t^X), will serve as the new state representation on which a new set of action preferences A^X will be learned, thus creating a hierarchical state structure anchored on previously learned action preferences. We will see in the following section how this representation has the right expressive power to model the experimental data shown in Chapter 2 with the implementation achieved in 3.

4.4.1 Filtering mechanism via the perceptual policy

As we calculate the perceptual policy by performing the computation in Equation 4.44, we obtain a filtering of states that only allows for post-second states to occur as potential transitions for a given state s_t. This can be seen, through the resulting matrix, shown in Figure 4.8, where the previous existing transitions pre-second tone (before state 40) have now a very low probability (values close to zero are painted white), while only the post-second tone transitions remain as potential states that can be accessed by the perceptual policy.

This filtering occurs due to the nature of the active suppression of motion that the agents need to perform and which increases the overall activity of the indirect pathway action preferences $A^{L,I}$. The high values of the indirect pathway push the total action preferences of the $L-$agent A^L to zero in the pre-second tone region. This then leads to a negative mask for the pre-second tone states and a positive mask probability for the post second tone states. This effect is

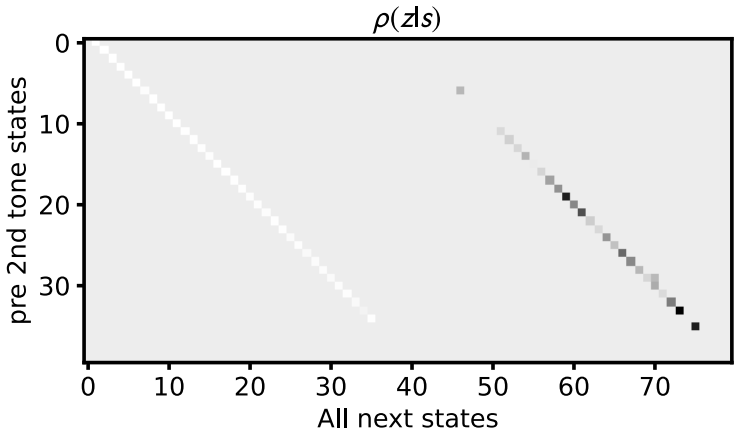


Figure 4.8: Traces of the probability distributions obtained from the optimal solution of Equation 4.22. The white diagonal on the leftmost side of the matrix represents zero probability states, with all transitions now happening to post-second tone states (past state 40).

in line with our initial conception of *affordances* as a potential mechanism to generate new state representations, as only the states on which an action is actually performed (remember that high indirect pathway values lead to a HOLD action to be executed) become relevant as a potential learning substrate for future reward.

The mechanism here presented can also be thought of as a form of preemptive planning, or pre-planning, where a new representation emerges that already contains the correct behavior that should be performed in an previously experienced environment; in the case of the timing task, choose the short noseport in the first half of the delay period and the long noseport in the second half. This creates an overlap in both non-rewarded states (as breaking fixations incur in a negative reward) and positively rewarded states, the implication being that the agents use a learned model of the world to generate a drive for action that is anchored on a prediction of future reward.

The perceptual policy now creates a more compact state representation, which do not contain stimulus information. As was shown in 3, for the multi-agent model, the X-agent can base its action preference learning on this alternative state representation, leading to the emergence of the breaking fixation behavior which was absent in the single agent model.

4.4.2 An anatomical hypothesis for the computation of the perceptual policy

We hypothesize that the mechanism described above can be one of many potential candidates for the formation of these representations. Disambiguation in this case will only be possible if one actually perturbed upstream cortical areas, which might have both the thalamocortical action preference information and the model of the environment that the agent might learn from experience.

Following the anatomy, we hypothesize a potential circuit that could allow for the generation of a more testable hypothesis. Although the usual striatal hierarchy along the striatum goes from medial to lateral and ventral regions of the striatum (Haber and Knutson, 2010), which to some extent would imply that the perturbation would be inverse of the one we're proposing, we're anchoring the hypothesis of this mechanism on the intracortical networks (as described in Section 1.2.2) and their bidirectional connectivity amongst both sensorimotor and prefrontal areas.

The bidirectional connections amongst sensorimotor and associative cortex circuits might provide the necessary bottom-up information transfer of state representations with high-granularity, currently assumed to lie in dorsolateral circuits to those more abstract and less granular representations that might exist in dorsomedial and ventral circuits. Specifically, the sensorimotor cluster made consisting of the secondary somatomotor (MOs) and the primary sensorimotor areas (SSs) connect directly to the cluster of associative sub-regions made up of the dorsal anterior cingulate area (ACAd) and the medial and ventrolateral regions of the orbitofrontal cortex (ORBm, ORBvl), the former being a constituent region of the prefrontal cortex which, as was been mentioned before, has been associated with higher order abstractions (Zingg et al., 2014)[7].

As described in section 1.2.2, these circuits shape a set of parallel striatocortical circuits(Hunnicutt et al., 2016; Hintiryan et al., 2016; Foster et al., 2021) which can be mapped to the same circuits that we associated with each one of the agents in the model developed in Chapter 3. Specifically, the sensorimotor cluster, connecting to the dorsolateral caudoputamen (CP.dl) [8] is associated with the dorsolateral circuit and *L-agent* and conversely the dorsomedial caudoputamen (CP.dm) is associated with the *X-agent*.

The recurrent connectivity within the associative cluster makes it hard to disentangle a more specific circuit, so understanding information flow would require highly region-specific ablations in order to probe the effect of optogenetic inhibitions. However, more global perturbations might allow to continue shaping this intuition, i.e. inhibition of all sub-regions of the associative cortex mentioned could potentially block the shaping of the abstract representations which lead to the creation of the drive signal in the dorsomedial agent described in Chapter 3. This perturbation could have the effect of decreasing the amount of breaking fixations that mice do during the delay period. However, one can also imagine that these drive signals are essential for solving the task itself, as in the end they contain the necessary task structure information that allows for the optimal behavior to emerge.

Specific inhibition of more local subnetworks, such as the one consisting of the MOs, ACAd and ORBm might provide some information as to how this information itself is transferred amongst regions. We propose that the filtering process presented in Section 4.4 lies within these cortical bidirectional sub-circuits and as such, perturbation of the connectivity between them might have an equivalent

[7]Region nomenclature is taken from Zingg et al. (2014)
[8]Region nomenclature adapted from Foster et al. (2021)

effect to the direct cortical inhibition proposed above. For a visual presentation of this hypothesis, please refer to the full circuit diagram in Figure 4.9, adapted from (Zingg et al., 2014; Foster et al., 2021; Hintiryan et al., 2016; Hunnicutt et al., 2016).

4.5 Discussion

The question of what state representation animals use to solve any kind of task is a fundamental and unsolved one. Most attempts rely on a lot of modeling assumptions about what's relevant for the animal. At the level of sensorimotor circuits, these assumptions are usually reasonable, even if laden with a tint of antropomorphism.

The most salient events of the task, things such as visual or auditory stimuli and movement triggers are useful anchors that to some extent have been well studied in terms of their assumed presence in the animals perceptive and neural response field. However, when one starts dealing with more abstract representations, ones which are less connected with the immediate experience of the organism, the problem shifts to what processes might be responsible for creating those abstract representations. Here, the processes are not directly anchored to the stimuli but probably in a secondary fashion, as the relevant elements that guide their creation can be assumed of a higher level as well, such as securing a future reward in an uncertain environment. This entails some notion of the future states of that environment and consequently and consequently a probabilistic model of what states are more likely to occur given current information.

In Chapter 3 we developed a multi-agent model that was able to express a set of experimental measures of mice performing a temporal interval discrimination task. The requirements of this dataset ruled out a single agent reinforcement learning model due to the lack of expressivity regarding the breaking fixation behavior and the dynamics of both direct and indirect pathway activities, expressed as action-preferences (see Section 3.3). In order to increase the expressivity of this model, a new agent was introduced that contained a more abstract state representation of the task (Section 3.4).

However, this representation lacked a formal justification as to what process might lead to its formation. One potential mechanism that could allow the shaping of such a distribution could be an Information Bottleneck process. The intuition behind this idea is that abstraction might imply some amount of compression.

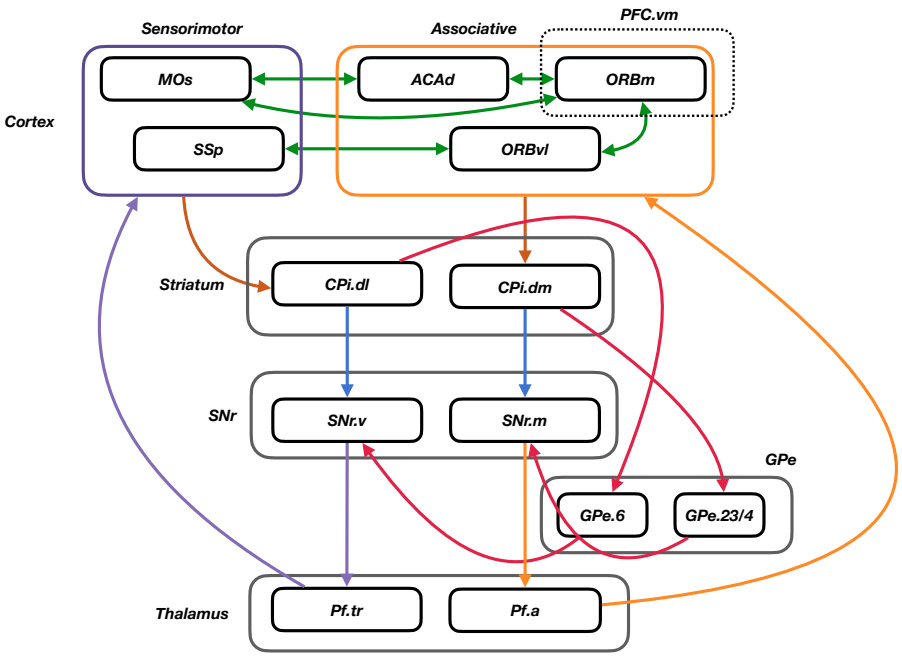


Figure 4.9: Full diagram of the circuit potentially responsible for the bottom-up filtering process that leads to the generation of a more abstract state representation defined in section 3.4.1. Nomenclature adapted from Foster et al. (2021); Zingg et al. (2014).

In the case of the timing task, one can think that all states before the decision boundary are "left-action states" and all the states post-decision boundary are "right-action states". This is equivalent to treating each collection of states as a single entity, which could be encoded in some form of perceptual policy, that takes the direct sensorimotor representation and associates the corresponding "class" to which the current state belongs to.

The Information Bottleneck method is a potential approach to finding the correct policies for such a set of state representations. We explored, through a toy example, what could be a potential adaptation of this method to the Interval Discriminatio Task in Section 4.2. As expected, the method gives us the correct policies for the latent policies (that would be equivalent to the more abstract representations) but in order to do so it relies on the specification of how many such states should exist. In a sense, the solution is already given in the choice, which as a representation scheme works but not as a model for what might be happening in the animal.

We proposed an alternative process, also based on information theory principles inspired by the work done in Levine (2018), which assumes that the agent is following a multi-objective loss function. This objective, expressed in Equation 4.26, assumes that the organism is trying to simultaneously obtain the behavioral policy that maximizes reward, whilst simultaneously trying to reduce the entropy of its perceptual policy, which represents an encoding of observed low-level states to a set of higher-level future latent states.

The optimal perceptual policy is obtained by optimising a mask over the low-level state representation, thus only making relevant the states that were acted upon. This connection between the filtering of states and their associated actions is assumed to be a potential formal description of the concept of affordances, put forward by James Gibson, which might open up a new set of theoretically relevant processes that lead to the generation of state representation hierarchies. By taking the transition model learned by the agent and filtering it with the action-preferences of the low-level action-preferences it turns out that one obtains the same states as the ones assumed in the model developed in Chapter 3.

Although we've only explored this process in the context of the IDT, one can assume that the overall concept of filtering might be useful in other task circumstances. For instance, when in spatial environments which contain regions that might be deemed as dangerous for an agent or organism, this filtering might be a quicker way of learning the necessary inhibitory signals that would allow the animal to avoid those regions of space. The same could be applied to learning va-

lences for particular classes of stimuli which, though slightly different, might have the same value in a specific context; for example, the value of a card suit in a game might be the same for the whole suit, although all the cards in a suit are different.

As we shall see in the final chapter of this thesis, this idea might shape the beginning of a new formalism that can have some expressive and predictive power for corticostriatal circuits, but for now it remains as an already useful formalism that can lead to the specific anatomical hypothesis regarding how corticostriatal circuits generate task representations in the domain of the interval discrimination task.

Chapter 5

General Discussion

The functional role of the basal ganglia and its connections with other brain regions has been a topic of research and debate for centuries. From high-level philosophical theories of it being the nucleus from which volition might emerge, to complex theoretical and biophysical models of its activity, this elusive brain region still contains a deeply complex set of inner workings that have required multiple scientific areas to be joined together in order to build a shared understanding of these circuits.

As we've seen in Chapter 1, this century-wide conversation uniting neurobiology and computer science has led to one of the most fascinating conceptual developments regarding brain function, that of the brain computing *Reward Prediction Errors* as a way to learn the value of states and actions in an environment. From this concept, a multitude of other ideas have emerged, anchored on the basic principles that guide the optimality conditions present within *Reinforcement Learning*. Added to this, questions about the role of the basal ganglia in movement disorders moved in parallel to this narrative, allowing for a rich tapestry of hypothesis and results (Mink, 1996) to be interwoven into the fabric of knowledge we have today. Its seemingly paradoxical role in simultaneously allowing for suppression and production of actions opened up a line of inquiry that is still open to this day, with one circuit emerging as an essential element; the direct and indirect pathways of the basal ganglia.

In order to better understand this circuit, recordings were made from the dorsolateral striatum of mice performing a temporal interval discrimination task requiring voluntary suppression of movement. One crucial aspect of the behavior of these animals was the pattern of breaking fixations and consequent choices. As was shown in Chapter 2, it's intriguing that even though mice present a subopti-

mal behavior by breaking fixation before the occurrence of the second tone, their subsequent choices match what would be the "correct" choice if a second tone *had* occurred. This behavior generated a bimodal distribution of breaking fixations that also mimicked the actual structure of the task, hinting at a potential latent plan that could serve as the driving force behind this behavior i.e. the mice, even failing to perform some trials, had some knowledge of *what would be the correct choice* at every time point.

Due to co-activation, on movement onset, of dorsolateral striatum direct and indirect pathways having been previously observed, (Cui et al., 2013; Tecuapetla et al., 2014; Markowitz et al., 2018) this region served as the optimal candidate to study what could be the underlying dynamics of motor suppression and expression. Indeed, co-activation of these pathways was observed during movement-related epochs, specifically during trial initiation and choice reporting at one of the side ports. Notably, during periods of active immobility, sustained activity was observed in the indirect pathway while activity in the direct pathway remained comparatively lower. Furthermore, optogenetic silencing of dorsolateral striatum iMSNs also demonstrated that this preferential engagement was causally implicated in action suppression.

Modeling this task, in particular, the suboptimality of their behavior in conjunction with a dynamic pattern of activity during the *absence of choice,* is not a trivial task. If one takes into account that an extremely basic *Q-Learning* or simple actor-critic agent can solve the task with perfect accuracy, one needs to ask where the inefficiencies might be coming from. Initial culprits might be the inaccurate estimates of time (Gibbon, 1977; Gouveâ et al., 2015; Motiwala et al., 2020) that mice might perform, but that wouldn't by itself generate the breaking fixation behavior, as badly estimating an interval wouldn't directly lead to not hearing the second tone stimuli or fully hallucinating a second tone at the incorrect timepoint in the delay period.

Alongside these observations, functional asymmetry was detected in the effects of optogenetic inhibition. Silencing dorsolateral striatum iMSN activity impaired the animals' capacity to suppress contra-lateral actions during active immobility periods but did not affect the vigor of these choices. In contrast, inhibition of dorsolateral striatum dMSN activity did not impact the likelihood of premature behavior but increased the time required for animals to report a choice post-fixation break, aligning with prior studies that implicate the basal ganglia in action vigor regulation (Turner and Desmurget, 2010; Panigrahi et al., 2015; Dudman and Krakauer, 2016).

In Chapter 3, we've exposed the inner workings of a model that contains not only biophysical detail but also behavioral expressivity. We've also generated a hypothesis in Chapter 4 regarding a potential mechanism of abstraction that allows for higher-order state representations to be formed through an attention mechanism, inspired by James Gibson's concept of affordances and Information Theory. We'll quickly recap how this framework fits in the overall literature and what contributions might still be possible in future work.

5.1 Multiplicity of agents

5.1.1 Computational role of the direct and indirect pathways

As we've seen in Chapter 1, initial research posited functionally opposing roles for the basal ganglia's two major pathways, confirmed by later experiments which showed that these pathways differentially influence aspects of behavior such as locomotion, motor sequence production, reinforcement, and decision-making (Kravitz et al., 2010, 2012; Roseberry et al., 2016; Tecuapetla et al., 2016; Sippy et al., 2015; Tai et al., 2012; Yttri and Dudman, 2016). However, concurrent activation of both pathways is often observed during movement, contradicting this traditional view (Cui et al., 2013; Tecuapetla et al., 2014; Markowitz et al., 2018; Cox and Witten, 2019).

A hypothesis reconciling this contradiction suggests that normal behavior involves promoting one motor plan while suppressing potentially competing ones (Mink and Thach, 1993). Dopaminergic modulation of the two largest classes of medium spiny neurons supports this framework, with dMSNs and iMSNs responding to positive and negative reward prediction errors, respectively (Collins and Frank, 2014; Gurney et al., 2015).

Studies on action selection have often overlooked the critical role of action suppression. This neglect may arise because an action is overtly observable, while successful suppression remains latent. Additionally, standard operant behaviors in laboratory settings do not always require explicit suppressive mechanisms. To effectively study action suppression, experimental paradigms must be designed to reveal the process's hallmarks (e.g., failure to remain immobile when required, as in the dynamic reward contingencies of specific tasks).

Experimental evidence also indicates the importance of suppression in decision-making, as shown in saccade experiments, where inhibiting iMSNs impaired monkeys' decision-making abilities (Amita and Hikosaka, 2019; Hikosaka et al., 2019; Kim et al., 2017). Traditional models, which define a single value for each action incremented or decremented by dopamine-encoded reward prediction errors, are limited in explaining dopamine's modulatory effects on incentive and asymmetric learning outcomes (Frank et al., 2007; Jocham et al., 2011; McClure et al., 2003; Berridge, 2012; Salamone et al., 2005).

Alternative models like those focusing on vigor are also insufficient, as they don't fully consider the impacts of dopamine depletion or differential effects on choice valence (Niv et al., 2007). The Opponent Actor Learning (OpAL) model by Collins and Frank (2014) seems to offer a more nuanced approach by recognizing the differential encoding of dopamine in distinct striatal neuronal populations and their projections to different basal ganglia output nuclei (Kravitz et al., 2010) however, the underlying plasticity rules do not take into account the non-linearities and negative baselines inherent in the dendritic volume changes observed experimentally (Iino et al., 2020; Gurney et al., 2015), as well as not allowing for the co-activation of the pathways due to the mutually exclusive nature of their update equations.

Even though the OpAL model effectively captures dopamine's influence on both learning and performance and the effects of optogenetic perturbations, showing that specific action values can be modulated through D1 and D2 stimulation (Tai et al., 2012), its direct adaptation does not by itself allow for the modeling of the *absence of movement*. Even taking into account other approaches present in the robotics literature that separate the processing of reward and punishment, (Navarro-Guerrero et al., 2017; Elfwing and Seymour, 2017) similar issues emerge as the specificity of the observed D1/D2 dynamics are not present if one implements directly the underlying learning rules.

Given this, we approached the problem of modeling this data by adopting a set of architectural motifs present in the literature and inserting the missing components that in principle should allow for the expression of the behavioral and neural profiles present in the experimental data.

What was accomplished in modeling terms was a multi-agent configuration, where the role of these two pathways can now more directly be understood within the context of a multi-level hierarchical configuration of state representations. The roles of these two populations of direct and indirect pathway neurons are now not only seen through a lens of behavioral control via direct suppression or activation of

motor programs but also as components of a more complex configuration of parallel circuit motifs that are in direct competition with each other for expression.

5.1.2 Further predictions of the multi-agent corticostriatal model

The decision to utilize a multi-agent architecture for our model was informed by two critical observations. First, both physiological data and manipulative experiments indicated that the DLS primarily functions to inhibit rather than facilitate action. Second, a single-agent architecture with a state-space representation incorporating elapsed time and second tone failed to generate broken fixations and the underlying dynamics present in the direct and indirect pathway activities.

Specifically, action values remained consistently low during the delay period once the agent was aware of the second tone, resulting in no action initiation. This is a natural consequence of the negative RPEs that the agent received every time an action was chosen during the delay period, increasing the indirect pathway action preferences $A^{L,I}(s_t, a_t)$ of the DLS agent and setting them above the direct pathway action preferences $A^{L,D}(s_t, a_t)$, triggering the HOLD action during the entirety of the delay period. The convergence to this combination of action preferences is a valid solution to the task, as there is no natural incentive within the reward structure of the task for fixations to be broken, hence no positive RPEs are received by the agent during the delay period and consequently the direct pathway, which in the single-agent case is the only driving force for action execution, does not accumulate positive energy in any of the available actions.

Given that even experimentally, DLS dMSNs did inhibition did not produce an expected decrease in breaking fixations (under the assumption that the drive to move would be coming from this population), the drive to break fixations has to come from another circuit. Looking into the parallel nature of the corticostriatal-thalamocortical loops that were described in Section 1.2.3, one potential candidate was assumed to be a higher order circuit, potentially in the DMS region of the striatum, which due to its anterior projections to prefrontal regions could contain a more abstract state representation of the task. Given that these parallel circuits show a canonical architecture, being composed of direct and indirect pathways and projecting to specific regions of the sub-nuclei of the basal ganglia (STN, GPe, GPi as shown in Section 1.2), the choice of a similar architecture for the new agent was natural.

This new agent, of which we were initially agnostic as to the brain region that it would represent, would also contain similar direct and indirect pathway action-preferences, sharing the action space but having a distinct state representation. We thus introduced a modified state-space representation that lacked information about the second tone. This assumption was the result of a loss function minimization (see Section 4.26) that took an internal model of the environment and learned an *attention mask* over the states with the highest *affordances* present in that environment. This allowed for a higher task-abstraction level to emerge on which a new set of action preferences could be learned, enabling the generation of a driving function that ended up being responsible for the broken fixations. Follow-up experiments corroborated that the DMS is integral to the neural circuit implementing this second agent, sustaining the drive to act during the delay interval. In Section 5.2 we'll dive into a bit more detail as to the intuitions behind the conceptualization of such a *perceptual policy* and also attempt to contextualize this concept in the general machine learning literature where it can potentially lead to fruitful developments.

It is important to mention that the underlying model is not the only solution to this problem, but instead consists of a model class. If instead of adding an X-agent with two pathways, one adds an agent with a single positively enforcing pathway, the same results emerge and the breaking fixation is conserved. In the limit case, a model with two agents, one with just a DLS indirect pathway action preference and a DMS agent with just the direct pathway agent would be the model with the minimum number of transfer function parameters that could fully reproduce the existing set of results.

This puts forward two important questions; what is then the role of the DLS dMSN and conversely what is the role of the DMS iMSN populations? One potential hypothesis is that given the particular state representations that each particular population inherits from its cortical afferents (Zingg et al., 2014; Haber and Knutson, 2010), and the fact that the dMSN populations are only active during action execution (a result which is expressed in the full action-preferences) the only missing element is the functional role of the DMS indirect pathway activity, which according to our model is not active during action execution (see Figure B.1). In this model, the indirect pathway activity is then contextually dependent on the representation, and as the more abstract representation does not "care" about actually executing actions but instead in trying to predict *what would be the future action* it serves no role. However, this disambiguation would only be possible if actual recordings were done in this brain region, the hypothesis remaining as such

until contrary evidence.

5.1.3 Multi-agent models of the brain

The conceptualization of different neural circuits as individual agents is not new, although the specifics of the proposed architectures are less granular, overlaying wider sets of corticostriatal circuits. One such proposal by Bornstein and Daw (2011), associates different regions of the striatum to specific reinforcement learning model types; DMS with model-based algorithms and DLS with model-free, both working as parallel circuits performing computations that use different amounts of information regarding state-transition dynamics of the environment. This distinction is made to allow for the simultaneous expression of TD-learning-based predictions in the DLS and the task-related representations that seem to coexist in DMS circuits (Yin et al., 2004b, 2005).

In the case of the model proposed in this work, the model-based component exists as a substrate for state abstraction and not for directly calculating value, as the corresponding model-based circuits are assumed to exist in the cortical afferents of the striatal domain where the higher-level agent exists. Previous models suggest that the model-based component was also part of the DMS circuit, implying a different computation altogether. Although there are anatomical differences in D2 populations in DMS in relation to DLS (the latter with greater concentration) (Yin and Costa, 2009), this is probably not sufficient to assume an entirely different computation. In our model, we propose that the underlying value computation is the same but the state-representation on which the canonical striatal circuit operates is different, resulting in the different types of signals observed experimentally (Kimchi and Laubach, 2009a,b).

It's still unclear what is the differentiating factor between the state representations allocated to each agent. However, there's some evidence that these processes might be the consequence of predictive processes. Given that DMS circuits are downward projections of prefrontal and other associative cortices, it might be the case that these higher-order circuits' objective is to use more abstract representations in order to predict future states of the world (Russek et al., 2017). Some approaches that have taken this route use the Successor Representation (SR) as the main algorithmic process for future state prediction. Given that prediction involves some notion of the probabilities of certain events, having such a model in terms of a discounted state transition matrix might be one potential strategy that allows these circuits to operate (Momennejad et al., 2017). Our choice of

defining an internal model $\hat{T}$ is somewhat inspired by this idea, but given that the successor is not a normalized transition probability matrix, we've chosen to use an approach that already provides those probabilities (Glascher et al., 2010; Gardner et al., 2018) without requiring an extra computation. One disadvantage of this approach is that there's no associated time constant parameter for the underlying predictions, something that might be of value in order to integrate a set of experimental results that show the existence of differing temporal horizons across the striatum (Wei et al., 2021). However, the extension of this model to one where this is taken into account is quite straightforward via the substitution of the transition matrix by a normalized SR matrix.

5.2 Multiplicity of ümwelts

> *"If we further consider that a subject is related to the same or to different objects by several functional cycles, we shall gain insight into the first principle of Ümwelt theory: all animals, from the simplest to the most complex, are fitted into their unique worlds with equal completeness. A simple world corresponds to a simple animal, a well-articulated world to a complex one."* — Jakob von Uexküll, A Stroll Through the Worlds of Animals and Men, 1934

One way in which a system could be able to understand an environment where the amount of information present is above the storage capacity of its memory system would be to split that information into chunks and create a modular representation of that same environment. Dividing in modules allows for parallel processing, a strategy that not only scales computational power but also allows for a distributed correction system. If one module fails at its job, other modules might come into play to keep the system running.

The mirror of this for a biological organism is its own representation of the world. Facing the same problem as the system above, it's not too far-fetched to think that similar strategies might have been employed in order to allow for an organism to reach evolutionary fitness within its niche ecological realm. Given this hypothesis, what mechanisms might have emerged in order to tackle this issue?

One potential way in which this problem might be solved is by allowing for a hierarchy of representations to be made which, according to a given context, are chosen to allow for the correct *behavioral policies* to be chosen. Given that

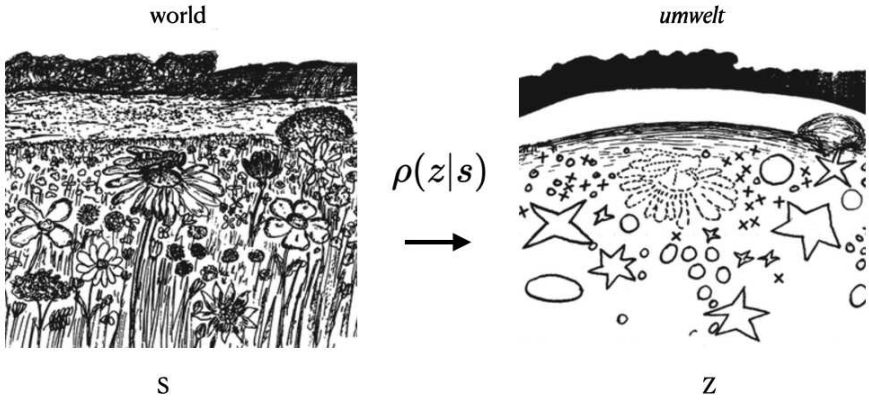


Figure 5.1: Emergence of a perceptual policy $\rho(z|s)$ in the niche ecology of a bee in its ecological niche, adapted from Von Uexkull (1934)

these representations live in more abstract representational spaces, they should affect mostly the perceptual relationship between the encoded percepts and the observed states of the world that the organism faces, creating a mapping $\rho(z|s)$ between the observed states s and the encoded latent states z. This *perceptual policy* can now be seen as a hierarchy-level dependent window into some aspects of the organism's environment, which given that there are multiple levels present due to the necessary decomposition and modularization of the underlying encoding process, shape a myriad of windows living at these different levels of the abstraction hierarchy.

These windows must allow for a simultaneous decomposition of spatial and temporal perceptual hierarchies; as an example, a bee moving in an area of dense forestation should have the appropriate filters to avoid attempting the pollination of flowers that will not contain pollen. This shapes a visual *affordance* map which selects for primary visual characteristics of the relevant plants, allowing for efficient sampling of the space. Conversely, temporal hierarchies are also highly relevant, as they are also able to shape the corresponding affordance of objects with temporal dynamics, e.g. an apple that will decompose and become non-nutritious as time passes. Given the different timescales on which specific dynamics occur, the nervous system needs to keep track of all the evolutionarily relevant representa-

tions that span across different timescales, requiring thus a gradient of temporal representations to be defined across the same hierarchies of abstraction that were mentioned earlier.

If one now thinks of the brain as a whole, a factorization of the functional relevance of specific brain areas can now be defined under a set of abstraction domains. For instance, in the sensorimotor realm, a division into low sensory level representations, such as the ones present in V1 might feed information not only into more abstract visual areas but also into representations with some episodic structure, which might feature hippocampal projections as their afferent (Pennartz et al., 2011). This can, in mammals and specifically in humans, be turned into representations that have a compositional and semantic property, shaping the formation of language which might lead itself to the formation of even higher order conceptual objects, containing abstract semantic connections (Fuster, 2001).

On the motor side, the same trajectory can be drawn, but downwards, with the sharing of the conceptual realm with the sensory level, feeding into the formation of plans and actions which then lead into the specific motor programs that will be executed by the relevant motor systems. In all of these levels, the relevant spatio-temporal discretizations are also employed, both on the sensory and the motor level as in order to keep higher order representations, which emerge out of the sustained activity of the lower order areas and which compress this activity, should have longer temporal constants.

For each of these "cuts" within the potential hierarchy/heterarchical and spatiotemporal abstraction levels, the hypothesis we're putting forward is that one can define a specific sub-agent, with its own representation and reward function and its own domain-specific set of tasks to perform in conjunction with other inter and intra-level agents. Each agent representing a particular canonical circuit responsible for computing within a given hierarchy level, and similarly to the way in which organism has an ümwelt derived from its ecological niche, would have an equivalent ümwelt for its own inner-world information "abstraction ecology", which would be consistent with the over-arching mental ecology that would sustain the *whole* of perception for the underlying organism. Although this opens up a fractal of ümwelts, it's still a decomposable one as each component remains within the scope of the three principles that seem to organize the architecture of nervous systems: *modularity, hierarchy, and heterarchy.*

The role of entropy in this framework can be explained succinctly as follows; as one moves up the abstraction ladder, the more stable these representations should be and as such, their respective *perceptual policy* entropy should be lower as well. If

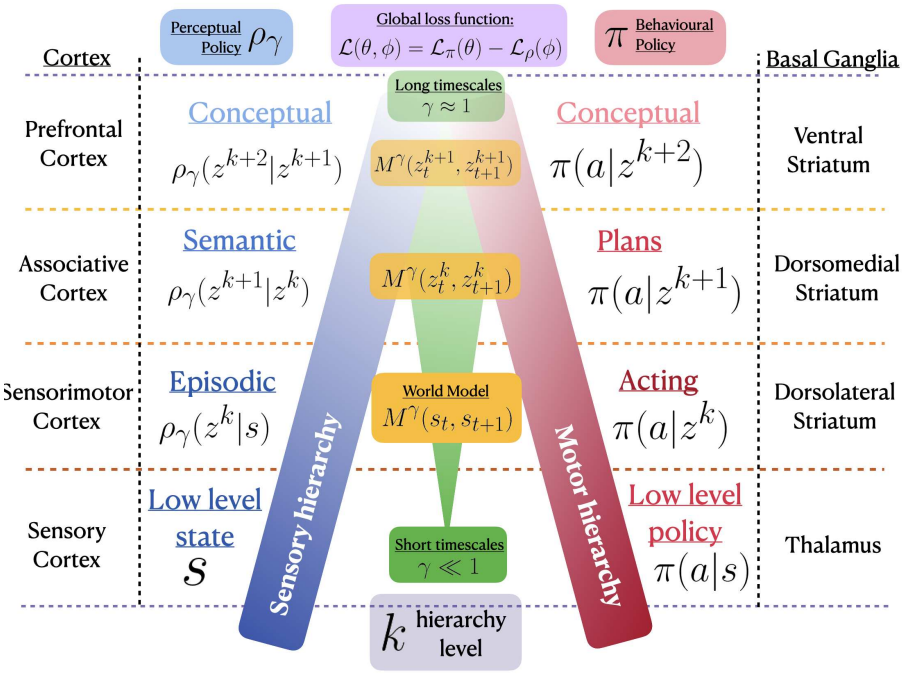


Figure 5.2: Schematic of the underlying framework of this thesis with the different hierarchical levels and their corresponding corticostriatal loops as horizontal cuts through the hierarchy.

not then the high-level representations which anchor the lower-level policies would not be stable enough to allow learning (Keshmiri, 2020). This can be further justified by the spectrum of time horizons that seem to exist across the striatum, with bigger time constants associated with higher-order corticostriatal loops and lower time constants lower-order ones (Wei et al., 2021). The bigger time constants can be associated with lower entropy as the underlying representations span over a wider set of state transitions in the environment, changing less over time, akin to how in the interval discrimination task one could cluster SHORT and LONG states during each half of the delay period. As these states are constant throughout these distinct periods, their relative entropy would be smaller in relation to the actual temporal dimension has been abstracted into two larger regions.

This conceptualization of the brain's functional architecture, inspired by the conjunction of both Von Uexküll and Gibson's frameworks, might open a fruitful ground out of which one might be able to derive more model classes that try to take into account all the underlying processes that were taken as assumptions in this thesis. The processes that generate the state representations are still of a tabular form instead of being derived from a dynamic process. The state representations themselves still only have a unidirectional communication channel between hierarchies and finally, the development of the right mathematical framework that is able to describe all of this in a generalizable form that can hold any number of abstraction levels.

It is with this in mind that we will dive into what could be potential approaches, within machine learning, that could open up a pathway for such a framework to be developed.

5.2.1 Affordances as a mechanism of abstraction

The idea of encoding a latent representation of the world in a compressed form is not novel. As was shown in Section 4.2, the classical Information Bottleneck method could be a potential alternative in the case where the dimensionality of the problem is pre-defined. Other approaches with a similar flavor to the one used in this thesis (see Chapter 4) are based on a state-enconding scheme (Ghugare et al., 2023), which aims at compressing state-space information within a maximum entropy minimization process similar to the one used to obtain the state representation of the DMS agent. This approach involves defining a compressing encoder from the high-dimensional state observations to some latent state, a predictive model of future latent states, and a policy over the latent states. This con-

figuration assumes a single agent which operates only under the low-dimensional latent representations derived from high-dimensional observations.

The equivalent affordance mechanism to our proposed mechanics for state compression can be somewhat compared to the latent prediction term which tries to guess which latent state should occur given an action, which is similar to the final construction obtained by sampling from the perceptual policy ρ in order to calculate the optimal policy π. In our case, the optimality condition is given by the sampled states being the next state, which shapes a sort of temporal compression, or anticipation. Potential work could be done in trying to understand how these two approaches could be joined together into a more global framework.

Other approaches, such as the one proposed by Khetarpal et al. (2020) encode the definition of affordances in terms of what it would mean for an agent to carry out that action "successfully". The authors define a notion of *affordances* based on another notion; that of *intent*, a map of states to desired states after executing a certain action. Affordances are then the subset of actions that lead to the predefined intents, which could be moving into a position that's closer to a reward or, for example, choosing a short or long port after a second tone is heard in our timing task. Although the mechanics that implement these affordances are different, the guiding principle is still the same; compressing information about the environment by reducing the available state-action pairs to those that have been previously useful to the task, which in a way also serves as a mask over the state-action space leading to a speeding up of the learning of value functions in MDPs. This idea has also been used in robotics applications as part of the learning process through image segmentation, which leads to associating specific affordances to objects present in the segments or by calculating *action scores* which represent the likelihood of a given action being performed successfully for a given object (Yang et al., 2023).

Affordance computational models in vision and robotics are often tackled through object-action recognition tasks. Nguyen et al. (2017) employs a dual-network approach with CNNs for object detection and pixel-wise affordance labeling, while Do et al. (2018) integrates affordance and object detection in a single CNN, trained end-to-end. Chuang et al. (2018) utilize GNNs for affordance reasoning and RNNs for action explanation from an egocentric viewpoint. A criticism of these methods(Hämäläinen et al., 2019; Kjellström et al., 2011) is their reliance on feature extraction and supervised learning without world model exploration, hindering adaptation to environmental interactions.

Recent works pivot towards RL for affordance discovery through world interac-

tion. Nagarajan and Grauman (2020) discusses an RL agent learning affordances by executing predefined high-level actions within a 3D environment, while Grabner et al. (2011) approach affordance detection by simulating human actions and assessing success likelihood. These models, however, are constrained by their predefined action sets and lack real-world action learning capabilities.

Other theories posit that there's a direct link between Affordances and the underlying reinforcement learning mechanisms, where actions are tried and optimized based on the reinforcement signals received. Unsuccessful actions are negatively reinforced, which could potentially be linked to a functional role for the D2 striatal populations, which prompt the avoidance of such motions (Chater, 2009) or avoid movement to be performed (Cruz et al., 2022; Kravitz et al., 2012).

Efficient affordance learning involves a balance between the exploration of new actions and the exploitation of known successful ones, a process central to reinforcement learning and described as the exploration/exploitation dilemma (Sutton et al., 1998). Affordance perception also involves generalization, allowing for recognition and interaction with novel objects based on similar experiences. Two mechanisms are proposed for this generalization: policy generalization through perceptual feature similarity and model-based reinforcement learning, involving motion planning and prediction (Gershman and Daw, 2017). Affordance learning in humans might also be enhanced by categorization, aiding in faster learning and facilitating social communication and acculturation (Kjellström et al., 2011; Cisek, 2007).

All the aforementioned ways in which affordances can be implemented showcase the flexibility of the concept as a theoretical tool for both understanding the underlying representations of neural activity in animals and providing potentially useful frameworks to develop novel algorithmic strategies for robotics. In this work, we presented a new implementation of this concept, adapted to the specific needs of the actor-critic framework being used to model the direct and indirect pathway activities, but it might also be the case that this implementation can prove useful for other applications and in the process generate new approaches and intuitions for building more efficient reinforcement learning models.

5.2.2 Hierarchies in minds and machines

As was mentioned earlier, we can categorize a set of principles that seem to guide the formation and underlying architecture of nervous systems; *modularity, hierarchy, and heterarchy.* The concept of hierarchy has been largely explored in the

context of reinforcement learning models, with approaches that utilize principles anchored on multi-agent models with some degree of heterarchy or intra-level information transfer.

One of the most fundamental problems where having hierarchies helps is in decomposing very high-dimensional state spaces. In the case of a complex task for example, defining levels that contain sub-tasks that further contain sub-sequences of actions might lead to a more stable learning process as the information required to solve each level decreases. Abstracting the complexity of the environment into tangible representations allows agents and organisms to have a better chance of developing the necessary policies to navigate efficiently in a given environment.

One possible implementation is through clustering states with correlated outcomes, thus reducing the total number of states that the agent needs to learn about. This clustering can be done either in the spatial or temporal domain, leading to the formation of temporal or spatial abstractions. *Hierarchical Reinforcement Learning* (HRL) provides a useful framing to include a wide range of model classes that could potentially lead to useful prediction of neural correlates for abstractions.

A classic implementation of the Hierarchical Reinforcement Learning idea is the *Options Framework*. This architecture allows agents to not only select from basic actions but also from higher-level strategies, or options, each with its own policy, initiation set, termination conditions, and associated pseudo-reward functions. These options represent temporally extended sequences of actions, simplifying decision-making by abstracting over many steps that would otherwise need to be considered individually. As an example, in the temporal discrimination task modeled in this thesis, two options could be defined; doing the left action in the first half of the delay period and the right action in the second action. The Information Bottleneck example in Section 4.2 provides a similar construction to what would be obtained via the options framework.

HRL can be also implemented in a model-based and model-free way, where the underlying sub-routines might have (or not) an underlying predictive model of the environment. One strategy for obtaining multi-level temporal abstractions that consolidate can be seen in the work by Momennejad and Howard (2018) who introduced a multi-scale model of predictive representations that addresses limitations in existing models that utilize the successor representation. By calculating the SR with different temporal discounts, one is able to obtain representations that are relevant at different time-scales, facilitating the compression of goals and sub-goals into their respective temporal hierarchy. This idea could allow the construction of options at different timescales, an idea that could also be explored

within the context of the model developed in Chapter 4.

Traditional SR models require computationally intensive operations to determine the sequence of future states from a given starting point. These operations involve non-linear manipulations of large SR matrices, which can be prohibitive in terms of computational resources. Momennejad and Howard (2018)'s model circumvents this issue by utilizing the derivatives of multiple SRs, each abstracted at different temporal scales. This method allows for the inference of both the distance and sequential order of successor states, thus enabling the reconstruction of entire expected future trajectories.

Besides the algorithmic strategies it is also important to consider if hierarchies are observable in different organisms. Multiple results seem to suggest that hierarchies do in fact play a role in the shaping of both behavior and neural activity. Botvinick (2012) suggested that only a few extensions, with plausible neural correlates, are needed to implement HRL, such as a mechanism for representing active subroutines, likely linked to the dorsolateral prefrontal cortex, with other work supporting the notion of the brain segmenting behavior into discrete units, fitting well within the HRL framework.

In the neural domain, research has identified specific brain regions associated with the hierarchical organization of behavior. The dorsolateral prefrontal cortex, for instance, has been implicated in representing active subroutines and task sets, as highlighted by studies such as those conducted by Koechlin and Summerfield (2007), which show that this region is involved in tracking the hierarchical structure of tasks. Furthermore, the ventral striatum has been shown to generate distinct reward prediction error signals at different levels of the task hierarchy, as illustrated by Diuk et al. (2013), providing evidence that it is sensitive to both overarching goals and immediate subgoals.

Moreover, the dorsal anterior cingulate cortex has been proposed as a key player in the selection amongst temporally abstract actions. Holroyd and Yeung (2012) hypothesized that this region might trigger activity in the dorsolateral prefrontal cortex to guide the execution of selected actions, thus serving a crucial function in HRL's multi-level decision-making architecture. This is consistent with the role of the cingulate in regulating overall levels of task engagement and the selection of actions within a hierarchical framework.

Another perspective is offered by Ito and Doya (2009), who suggests that HRL mechanisms are distributed along a ventromedial-to-dorsolateral axis within the striatum, comprising parallel, hierarchically interrelated learning modules. They

postulate that different regions of the striatum—ventral, dorsomedial, and dorsolateral—are responsible for guiding action selection at varying levels of temporal granularity, from high-level goals to fine-grained movements.

One of the key behavioral correlates of HRL is the ability to break down tasks into manageable subgoals, a feature that has been extensively studied in the context of problem-solving and skill acquisition. For instance, Thaler et al. (2013) demonstrated that when individuals are faced with complex tasks, they tend to organize their actions hierarchically by formulating intermediate goals that guide subsequent behavior. This aligns with the concept of options in HRL, where high-level actions (options) are composed of multiple lower-level actions.

Further behavioral evidence comes from studies on task-switching and multitasking, which show that people tend to structure their behavior in a way that reduces the cognitive load associated with switching between different tasks. Badre (2008) showed that hierarchical structures could predict the patterns of task-switch costs, indicating that humans naturally employ a hierarchical organization to streamline the execution of multiple tasks.

The chunking of actions into sequences, which can be rapidly executed with little conscious oversight, is another behavioral hallmark of HRL. Studies on habit formation, such as those by Graybiel (2008), have revealed that with repetition, sequences of actions become chunked into cohesive units that can be triggered by contextual cues, suggesting a behavioral mechanism for the options framework within HRL. However, one fundamental problem with this framework is that options do not tell us how to build the underlying representations - there's a priori knowledge of how to obtain things and given the anatomical and experimental evidence organisms seem to bootstrap of the lower level representation and create higher-level ones. *Our proposed framework addresses this issue by allowing for these high-level representations to be anchored on lower level ones via a information minimisation principle.*

Finally, and going back to the proposed model structure outlined in 4, one could potentially extend the model to also consider multiple time-scales and further representational levels, as outlined in Figure 5.2. As the structure of the *perceptual policy* $\rho(z|s)$ is dependent on a lower level representation, one could potentially stack these representations indefinitely, with the condition that the lowest level representation has such a big number of states that would make that hierarchical decomposition useful. By stacking the perceptual policies in different levels k, one could build a set of functions $\rho(z^k|z^{k-1})$ that would represent the latent state representation k built from the lower level $k-1$. Added to this, and given that the

perceptual policy definition is dependent on a model of the world dependent on a specific temporal discount factor γ, so as to constrain that representation at a given timescale $M\gamma(z_t^k, z_{t+1}^k)$. Potential mechanisms to learn these world representations could be adapted from the Successor Representation Dayan (1993); Stachenfeld et al. (2017); Momennejad et al. (2017).

This would give then a set of perceptual policies $\rho^\gamma(z^k|z^{k-1})$ and environment models $M^\gamma(z_t^k, z_{t+1}^k)$ that would potentially shape the learning of the value functions for a given level k and temporal discount γ. This extension of the model will be the next step in the development of this framework, as well as its application to tasks designed specifically to prune out hierarchical structures out of behavioral and neural signals.

5.3 Closing remarks

In this thesis, we have delved into a mechanistic understanding of the basal ganglia, examining the functional interplay between the direct and indirect pathways, particularly during action suppression and co-activation during movement. These insights might be pivotal, given the basal ganglia's involvement in disorders characterized by impaired inhibitory control across both motor and cognitive domains, in allowing for potential new treatments of these disorders.

We employed a multi-agent model based on reinforcement learning, capturing the neural and behavioral profiles within the dorsolateral and dorsomedial striatum. This model simulates the behavior effects of inhibiting specific striatal populations and predicts activity profiles within different striatal regions. The model posits each agent within a distinct state representation that drives reinforcement learning through different reward prediction errors as the animal interacts with its environment. This approach integrates hypotheses regarding the parallel circuits' functions, such as arbitration between control types in response to task demands.

Enhancing this multi-agent model, we incorporate state representation abstraction via affordances in a hierarchical setting, a theoretical progression from simple sensorimotor to complex, abstract representations linked to frontal cortical areas. This abstraction allows the model to predict actions and manage conflicting behavioral drives, aligning with traditional basal ganglia circuit functions and contemporary reinforcement learning control frameworks. We propose that the abstraction process, which could also potentially be implemented by an adaptation of a Variational Inference process over latent states, compresses state representations to form

more general classes of action states, aligning with the general conceptualization of affordances. This hierarchy of state representations could reflect the varying degrees of abstraction in the corticostriatal projectome, from specific sensorimotor events to more generalized and compressed environmental representations.

Although the proposed framework has only been applied to the temporal discrimination task presented in this thesis, future work will imply the application of this framework to other tasks and also allow the development of a more general algorithm structure. Throughout this thesis, we've focused on applying the basic architectural principles of the corticostriatal circuits into shaping the models used to predict behavior and neural activity. Hopefully, this process can be done in the direction of improving existing Multi-agent Hierarchical Reinforcement Learning frameworks and allowing for more efficient applications of these principles, as well as to further contribute to the progress of understanding how artificial intelligence models could also be shaped by brain-inspired architectures.

Unexpectedly, the shape of the underlying thesis being proposed has taken a somewhat fractal structure. We've proposed that the concept of Ümwelt can be extended beyond individual organisms and be applied to specific neural circuits. Thus, it's possible that the brain can be seen not only as a single information processing entity but as the collection of a myriad of such entities operating in concert and using different views of the world in order to shape a greater picture that can be useful for the organism. In a sense, it seems that the Roman depiction of Janus can also be folded within, with multi-faced entities that cooperate in order to form a cohesive interaction with the world hopefully leading to a graceful passage through the gateway of time.

References

A. Abdi, N. Mallet, F. Mohamed, A. Sharott, P. Dodson, K. Nakamura, and P. Magill. Prototypic and arkypallidal neurons in the dopamine-intact external globus pallidus. *J. Neurosci.*, 35(17):6667–6688, April 2015.

R. D. Adams, L. van Bogaert, and H. Vander Eecken. Striato-nigral degeneration. *Journal of Neuropathology & Experimental Neurology*, 23(4):584–608, 1964.

J. P. Aggleton, A. Pralus, A. J. Nelson, and M. Hornberger. Thalamic pathology and memory loss in early alzheimer's disease: moving the focus from the medial temporal lobe to papez circuit. *Brain*, 139(7):1877–1890, 2016.

R. L. Albin, A. B. Young, and J. B. Penney. The functional anatomy of basal ganglia disorders. *Trends Neurosci.*, 12(10):366–375, October 1989.

G. E. Alexander and M. D. Crutcher. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in neurosciences*, 13 (7):266–271, 1990.

G. E. Alexander, M. R. DeLong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.*, 9:357–381, 1986.

M. Alexander and J. Wickens. Analysis of striatal dynamics: the existence of two modes of behaviour. *J. Theor. Biol.*, 163:413–438, 1993.

H. Amita and O. Hikosaka. Indirect pathway from caudate tail mediates rejection of bad objects in periphery. *Sci Adv*, 5(8):eaaw9297, Aug 2019.

E. C. Anderson et al. The relationship between uncertainty and affect. *Frontiers in psychology*, 10:2504, 2019.

A. Aransay, C. Rodríguez-López, M. García-Amado, F. Clascá, and L. Prensa. Longrange projection neurons of the mouse ventral tegmental area: a single-cell axon tracing analysis. *Front. Neuroanat.*, 9:1–24, 2015. doi: 10.3389/fnana.2015.00059.

M. Azpilcueta Navarro. Commentaries on usury. In *Las Siete Partidas, Volume 5*, pages 211–229. University of Pennsylvania Press, 1966.

D. Badre. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200, 2008.

I. I. I. Baird and L. C. Advantage updating. 1993.

L. C. Baird. Advantage updating. Technical report, Wright Laboratory Technical Report WL-TR-93-1146, available from the Defense Technical information Center, Cameron Station, Alexandria, VA 22304-6145, 1993.

B. Balleine, M. Delgado, and O. Hikosaka. The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, 27:8161–8165, 2007.

R. S. Bapi and K. Doya. A sequence learning architecture based on cortico-basal ganglionic loops and reinforcement learning. In *ICONIP'98*, Kita-Kyushu, Japan, 1998.

I. Bar-Gad, G. Havazelet-Heimer, E. Ruppin, and H. Bergman. Reinforcement driven dimensionality reductions; a model for information processing in the basal ganglia. *Journal of Basic and Clinical Physiological and Pharmacology*, 11:305–320, 2000.

A. G. Barto. Adaptive critics and the basal ganglia. *Models of Information Processing in the Basal Ganglia*, 1995.

J. Beeler, M. Frank, J. McDaid, and E. Alexander. A role for dopamine-mediated learning in the pathophysiology and treatment of parkinson's disease. *Cell Reports*, 2:1747–1761, 2012.

J. A. Beeler, N. Daw, C. R. M. Frazier, and X. Zhuang. Tonic dopamine modulates exploitation of reward learning. *Front. Behav. Neurosci.*, 4:170, November 2010. doi: 10.3389/fnbeh.2010.00170.

D. Beiser and J. Houk. Model of cortical-basal ganglia ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology*, 79: 3168–3188, 1998.

D. G. Beiser, S. E. Hua, and J. C. Houk. Network models of the basal ganglia. *Curr. Opin. Neurobiol.*, 7(2):185–190, Apr. 1997.

R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

G. S. Berns and T. J. Sejnowski. A computational model of how the basal ganglia produce sequences. *J Cognit Neurosci*, 10:108–121, 1998.

K. Berridge and M. Kringelbach. Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Curr. Opin. Neurobiol.*, 23(3):294–303, 2013. doi: 10.1016/j.conb.2013.01.017.

K. Berridge and T. Robinson. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Review*, 28: 309–369, 1998.

K. C. Berridge. From prediction error to incentive salience: Mesolimbic computation of reward motivation. *European Journal of Neuroscience*, 35(7):1124–1143, 2012. doi: 10.1111/j.1460-9568.2012.07990.x.

Å. Bertler and E. Rosengren. Occurrence and distribution of dopamine in brain and other tissues. *Experientia*, 15:10–11, 1959.

A. Björklund and S. B. Dunnett. Fifty years of dopamine research. *Trends in Neurosciences*, 30(5):185 – 187, 2007. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2007.03.004.

J. Bogenpohl, S. Ritter, R. Hall, K. Thompson, B. Mercer, and Y. Smith. Adenosine a2a receptor in the monkey basal ganglia: ultrastructural localization and colocalization with the metabotropic glutamate receptor 5 in the striatum. *The Journal of comparative neurology*, 520(3):570–589, 2012.

A. M. Bornstein and N. D. Daw. Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, 21(3):374-80.(3):374–380, June 2011.

M. M. Botvinick. Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6):956–962, 2012.

S. L. Bressler and V. Menon. Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences*, 14(6):277–290, 2010.

J. Brown, D. Bullock, and S. Grossberg. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J Neurosci*, 19(23):10502–10511, 1999. doi: 10.1523/JNEUROSCI.19-23-10502.1999.

G. W. Bruyn, G. T. Bots, and A. Staal. Familial bilateral vascular calcification in the central nervous system. *Psychiatr Neurol Neurochir*, 67:342–376, 1964.

D. J. Bucci. Posterior parietal cortex: An interface between attention and learning? *Neurobiology of Learning and Memory*, 91(2):114–120, 2009.

K. F. Burdach. *Vom baue und leben des gehirns*. Dyk, Leipzig, 1819–1826.

A. Carlsson, M. Lindqvist, and T. Magnusson. 3,4-dihydroxyphenylalanine and 5-hydroxytryptophan as reserpine antagonists. *Nature*, 180(4596):1200, nov 1957.

M. B. Carpenter. *Core Text of Neuroanatomy*. Williams & Wilkins, 2nd edition, 1981.

M. B. Carpenter, R. Fraser, and W. R. Mehler. The cells of origin and course of the cortico-pallidal projection in the monkey. *Journal of Comparative Neurology*, 132(2):223–243, 1968a.

M. B. Carpenter, R. A. R. Fraser, and W. R. Mehler. The organization of the pallidosubthalamic fibers in the monkey. *Brain Research*, 11:522–599, 1968b.

M. Cazorla, F. D. de Carvalho, M. O. Chohan, M. Shegda, N. Chuhma, S. Rayport, S. E. Ahmari, H. Moore, C. Kellendonk, J. A. Gingrich, et al. Dopamine d2 receptors regulate the anatomical and functional balance of basal ganglia circuitry. *Neuron*, 81(1):153–164, 2014.

V. S. Chakravarthy, D. Joseph, and R. S. Bapi. What do the basal ganglia do? a modeling perspective. *Biol Cybern*, 103(3):237–253, 2010. doi: 10.1007/s00422-010-0401-y.

J.-M. Charcot. *Leçons sur les maladies du système nerveux, Nucleus nervosen, vol. 2*. Delahaye, Paris, 1877.

J.-M. Charcot. *Leçons du Mardi: Policlinique de la Salpêtrière, 1887–1888*. Bureaux du Progrès Médical, Paris, 1888. Lesson of June 12, 1888.

N. Chater. Rational and mechanistic perspectives on reinforcement learning. *Cognition*, 113(3):350–364, 2009. doi: 10.1016/j.cognition.2008.06.014. URL https://doi.org/10.1016/j.cognition.2008.06.014.

C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler. Learning to act properly: Predicting and explaining affordances from images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 975–983. IEEE, 2018. doi: 10.1109/CVPR.2018.00108. URL https://doi.org/10.1109/CVPR.2018.00108.

P. Cisek. Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1585–1599, Sep 2007.

A. Collins and M. Frank. Opponent actor learning (opal): modeling interactive efects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, 121(3):337–366, July 2014.

L. S. Colzato, J. Pratt, and B. Hommel. Dopaminergic control of attentional flexibility: Inhibition of return is associated with the dopamine transporter gene (dat1). *Frontiers in Human Neuroscience*, 4:1–6, jan 2010. doi: 10.3389/fnhum.2010.00053. URL http://dx.doi.org/10.3389/fnhum.2010.00053.

L. S. Colzato, H. A. Slagter, M. de Rover, and B. Hommel. Dopamine and the management of attentional resources: Genetic markers of striatal d2 dopamine predict individual differences in the attentional blink. *Journal of Cognitive Neuroscience*, 23(11):3576–3585, nov 2011. doi: 10.1162/jocn_a_00049. URL http://doi.org/10.1162/jocn_a_00049.

C. I. Connolly and J. B. Burns. *A state-space striatal model*, pages 163–183. MIT Press, 1995.

J. L. Contreras-Vidal and W. Schultz. A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience*, 6(3):191–214, 1999.

J. L. Contreras-Vidal and G. E. Stelmach. A neural model of basal ganglia-thalamocortical relations in normal and parkinsonian movement. *Biol Cybern*, 73(5):467–476, 1995.

J. Cox and I. B. Witten. Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.*, 20(8):482–494, August 2019.

A. D. Craig. How do you feel–now? the anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1):59–70, 2009.

B. F. Cruz, G. Guiomar, S. Soares, A. Motiwala, C. K. Machens, and J. J. Paton. Action suppression reveals opponent parallel control via striatal circuits. *Nature*, 607(7919):521–526, jul 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-04894-9. URL `https://www.nature.com/articles/s41586-022-04894-9`.

G. Cui, S. B. Jun, X. Jin, M. D. Pham, S. S. Vogel, D. M. Lovinger, and R. M. Costa. Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, 494(7436):238–242, Feb. 2013. ISSN 0028-0836. doi: 10.1038/nature11884. URL `https://www.nature.com/articles/nature11884`.

K. D'Ardenne, N. Eshel, J. Luka, A. Lenartowicz, L. Nystrom, and J. Cohen. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proc. Natl. Acad. Sci. Unit. States Am.*, 109:19900–19909, 2012. doi: 10.1073/pnas.1116727109/-/DCSupplemental.

N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, 8(12):1704–1711, dec 2005.

P. Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

P. Dayan and K. C. Berridge. Model-based and model-free pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, 14:473–492, 2014.

P. Dayan and N. D. Daw. Decision theory, reinforcement learning, and the brain. *Cognition, Affect, & Behavior Neuroscience*, 8(4):429–453, Dec 2008.

P. Dayan, Y. Niv, B. Seymour, and N. Daw. The misbehavior of value and the discipline of the will. *Neural Netw.*, 19(8):1153–1160, Oct 2006.

M. DeLong, M. Crutcher, and A. Georgopoulos. Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. *Journal of Neuroscience*, 3(8):1599–1606, 1983.

M. R. DeLong. Activity of basal ganglia neurons during movement. *Brain research*, 40(1):127–135, 1972.

M. R. DeLong, M. D. Crutcher, and A. P. Georgopoulos. Primate globus pallidus and subthalamic nucleus: functional organization. *Journal of neurophysiology*, 53(2):530–543, 1985.

J. M. Deniau and G. Chevalier. Disinhibition as a basic process in the expression of striatal functions. ii. the striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus. *Brain Res.*, 334(2):227–233, May 1985.

D. Denny-Brown and N. Yanagisawa. The role of the basal ganglia in the initiation of movement. *Res. Publ. Assoc. Res. Nerv. Ment. Dis.*, 55:115–149, 1976.

G. Di Chiara. Nucleus accumbens shell and core dopamine: differential role in behavior and addiction. *Behavioural Brain Research*, 137:75–114, 2002.

C. Diuk, K. Tsai, J. Wallis, M. Botvinick, and Y. Niv. Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, 33(13):5797–5805, 2013.

T.-T. Do, A. Nguyen, and I. Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5882–5889. IEEE, 2018. doi: 10.1109/ICRA.2018.8460902. URL `https://doi.org/10.1109/ICRA.2018.8460902`.

P. Dominey. Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics*, 73:265–274, 1995.

H. Dorfman and S. Gershman. Controllability governs the balance between pavlovian and instrumental action selection. *Nat. Commun.*, 10(1):5826, 2019.

K. Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw*, 12(7-8):961–974, 1999. doi: 10.1016/s0893-6080(99)00046-5.

J. T. Dudman and J. W. Krakauer. The basal ganglia: from motor commands to the control of vigor. *Current Opinion in Neurobiology*, 37(158-166), April 2016.

H. Ehringer and O. Hornykiewicz. Verteilung von noradrenalin und dopamin (3-hydroxytyramin) im gehirn des menschen und ihr verhalten bei erkrankungen des extrapyramidalen systems. *Klin. Wochenschr*, 38:1236–1239, 1960.

S. Elfwing and B. Seymour. Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the maxpain algorithm. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2017.

E. V. Evarts. A technique for recording activity of subcortical neurons in moving animals. *Journal of Applied Physiology*, 24(5):584–586, 1968.

B. J. Everitt and T. W. Robbins. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neuroscience*, 8(11): 1481–1489, 2005.

M. S. Fanselow and H.-W. Dong. Are the dorsal and ventral hippocampus functionally distinct structures? *Neuron*, 65(1):7–19, 2010.

J. Féger, P. Robeldo, and N. Renwart. The subthalamic nucleus: new data, new questions. In *The Basal Ganglia III*. Plenum Press, 1991.

C. E. Feierstein, M. C. Quirk, N. Uchida, D. L. Sosulski, and Z. F. Mainen. Representation of spatial goals in rat orbitofrontal cortex. *Neuron*, 51(4):495–507, 2006.

I. Fisher. *The Theory of Interest*. Macmillan, 1930.

P. Foldiak. Forming sparse representations by local anti-hebbian learning. *Biological Cybernetics*, 64:165–170, 1990.

K. A. Ford and S. Everling. Neural activity in primate caudate nucleus associated with proand antisaccades. *J. Neurophysiol.*, 102(4):2334–2341, October 2009.

N. N. Foster, J. Barry, L. Korobkova, L. Garcia, L. Gao, M. Becerra, and H.-W. Dong. The mouse cortico-basal ganglia-thalamic network. *Nature*, 598(7879): 188–194, Oct 2021.

M. Foucault. *The Order of Things: An Archaeology of the Human Sciences*. Pantheon, New York, 1970.

C. François, G. Percheron, and J. Yelnik. Localization of nigrostriatal, nigrothalamic and nigrotectal neurons in ventricular coordinates in macaques. *Neuroscience*, 13:61–76, 1984.

M. J. Frank, A. A. Moustafa, H. M. Haughey, T. Curran, and K. E. Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, 104(41):16311–16316, 2007.

B. S. Freeze, A. V. Kravitz, N. Hammack, J. D. Berke, and A. C. Kreitzer. Control of basal ganglia output by direct and indirect pathway projection neurons. *J. Neurosci.*, 33(47):18531–18539, November 2013.

Y. Fu, G. Paxinos, C. Watson, and G. Halliday. The substantia nigra and ventral tegmental dopaminergic neurons from development to degeneration. *J. Chem. Neuroanat.*, 76:98–107, 2016. doi: 10.1016/j.jchemneu.2016.02.001.

T. Fukai and S. Tanaka. A simple neural network exhibiting selective - activation of neuronal ensembles: from winner-take-all to winners-share-all. *Neural Computation*, 9:197–238, 1997.

J. M. Fuster. The prefrontal cortex–an update: time is of the essence. *Neuron*, 30 (2):319–333, 2001. doi: 10.1016/s0896-6273(01)00285-9.

A. Galashov, S. Jayakumar, L. Hasenclever, D. Tirumala, J. Schwarz, G. Desjardins, W. M. Czarnecki, Y. W. Teh, R. Pascanu, and N. Heess. Information asymmetry in KL-regularized RL. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1lqMn05Ym`.

M. P. H. Gardner, G. Schoenbaum, and S. J. Gershman. Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, 285(1891): 20181645, November 2018. doi: 10.1098/rspb.2018.1645.

C. R. Gerfen. Synaptic organization of the striatum. *Journal of Electron Microscopy Technique*, 10(3):265–281, 1988.

C. R. Gerfen. The neostriatal mosaic: multiple levels of compartmental organization. *Trends in neurosciences*, 15(4):133–139, 1992.

S. J. Gershman and N. D. Daw. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68:101–128, 2017. doi: 10.1146/annurev-psych-122414-033625. URL `https://doi.org/10.1146/annurev-psych-122414-033625`.

R. Ghugare, H. Bharadhwaj, B. Eysenbach, S. Levine, and R. Salakhutdinov. Simplifying model-based rl: Learning representations, latent-space models, and policies with one objective, 2023.

J. Gibbon. Scalar expectancy theory and weber's law in animal timing. *Psychol. Rev.*, 84(3):279–325, May 1977.

J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979. Print.

A. Gillies and G. Arbuthnott. Computational models of the basal ganglia. *Movement Disorders*, 15(5):762–770, 2000.

R. Giufrida, G. Li Volsi, G. Maugeri, and V. Perciavalle. Influences of pyramidal tract on the subthalamic nucleus in the cat. *Neuroscience Letters*, 54:231–235, 1985.

J. Glascher, N. Daw, P. Dayan, and J. P. O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, May 2010. doi: 10.1016/j.neuron.2010.04.016.

T. S. Gouveâ, T. Monteiro, A. Motiwala, S. Soares, C. K. Machens, and J. J. Paton. Striatal dynamics explain duration judgments. Technical report, Max Planck Institute for Biological Cybernetics, jun 2015.

H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR 2011*, pages 1529–1536. IEEE, 2011. doi: 10.1109/CVPR.2011.5995327. URL `https://doi.org/10.1109/CVPR.2011.5995327`.

D. Graeber. *Debt: The First 5 000 Years*. Melville House, Brooklyn, N.Y, 2011.

A. Graybiel, T. Aosaki, A. Flaherty, and M. Kimura. The basal ganglia and adaptive motor control. *Science*, 265:1826–1831, 1994.

A. M. Graybiel. Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, 31:359–387, 2008.

S. Grillner, J. Hellgren, A. Menard, K. Saitoh, and M. Wikstrom. Mechanisms for selection of basic motor programs-roles for the striatum and pallidum. *Trends Neurosci.*, 28(7):364–370, July 2005.

H. J. Groenewegen, C. I. Wright, A. V. Beijer, and P. Voorn. Convergence and segregation of ventral striatal inputs and outputs. *Annals of the New York Academy of Sciences*, 877(1):49–63, 1999.

I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska. A survey of Actor-Critic reinforcement learning: Standard and natural policy gradients. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, 42(6):1291–1307, Nov. 2012.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

K. Gurney, T. J. Prescott, and P. Redgrave. The basal ganglia viewed as an action selection device. In *International conference on artificial neural networks*, pages 1033–1038, 1998.

K. Gurney, T. J. Prescott, J. R. Wickens, and P. Redgrave. Computational models of the basal ganglia: from robots to membranes. *Trends in Neurosciences*, 27 (8), 2004.

K. N. Gurney, M. D. Humphries, and P. Redgrave. A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS Biol.*, 13(1):e1002034, January 2015.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

S. Haber, E. Lynd, C. Klein, and H. Groenewegen. Topographic organization of the ventral striatal efferent projections in the rhesus monkey: An anterograde tracing study. *Journal of Comparative Neurology*, 293(2):282–298, 1990.

S. N. Haber and B. Knutson. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35(1):4–26, 2010.

S. N. Haber, J. L. Fudge, and N. R. McFarland. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.*, 20(6):2369–2382, March 2000.

G. Halliday, S. Reyes, and K. Double. *Substantia Nigra, Ventral Tegmental Area, and Retrorubral Fields*. Elsevier, third edition, 2012. doi: 1016/B978-0-12-374236-0.10013-6.

C. Hammond and J. Yelnik. Intracellular labelling of rat subthalamic neurones with horseradish peroxidase: computer analysis of dendrites and characterization of axon arborization. *Neuroscience*, 8:781–790, 1983.

T. Hauser, E. Eldar, and R. Dolan. Separate mesocortical and mesolimbic pathways encode effort and reward learning signals. *Proc. Natl. Acad. Sci. Unit. States Am.*, 114(35):E7395–E7404, 2017. doi: 10.1073/pnas.1705643114.

L. N. Hazrati, A. Parent, S. Mitchell, and S. N. Haber. Evidence for interconnections between the two segments of the globus pallidus in primates: a pha-l anterograde tracing study. *Brain research*, 533(2):171–175, 1990.

S. V. Hegarty, A. M. Sullivan, and G. W. O'Keeffe. Midbrain dopaminergic neurons: a review of the molecular circuitry that regulates their development. *Developmental Biology*, 379(2):123–138, 2013. doi: 10.1016/j.ydbio.2013.04.014.

L. Heimer, D. Zahm, L. Churchill, P. Kalivas, and C. Wohltmann. Specificity in the projection patterns of accumbal core and shell in the rat. *Neuroscience*, 76: 689–707, 1997.

O. Hikosaka and R. Wurtz. Visual and oculomotor functions of monkey substantia nigra pars reticulata. iii. memory-contingent visual and saccade responses. *Journal of Neurophysiology*, 49(5):1268–1284, 1983.

O. Hikosaka and R. H. Wurtz. Modification of saccadic eye movements by gaba-related substances. ii. effects of muscimol in monkey substantia nigra pars reticulata. *J. Neurophysiol.*, 53(1):292–308, 1985.

O. Hikosaka, Y. Takikawa, and R. Kawagoe. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol. Rev.*, 80(3):953–978, July 2000.

O. Hikosaka, H. F. Kim, H. Amita, M. Yasuda, M. Isoda, Y. Tachibana, and A. Yoshida. Direct and indirect pathways for choosing objects and actions. *Eur. J. Neurosci.*, 49(5):637–645, March 2019.

H. Hintiryan, N. N. Foster, I. Bowman, M. Bay, M. Y. Song, L. Gou, and H.-W. Dong. The mouse cortico-striatal projectome. *Nat. Neurosci.*, 19(8):1100–1114, August 2016.

C. B. Holroyd and N. Yeung. Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, 16(2):122–128, 2012.

J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia might generate and use neural signals that predict reinforcement. In J. Houk, J. Davis, and D. Beiser, editors, *Workshop paper*, volume 13, pages 249–270. MIT Press, 1995a.

J. C. Houk, J. L. Davis, and D. G. Beiser. *Models of information processing in the basal ganglia*. MIT Press, Cambridge, 1995b.

I. Huerta-Ocampo, J. Mena-Segovia, and J. P. Bolam. Convergence of cortical and thalamic input to direct and indirect pathway medium spiny neurons in the striatum. *Brain Structure and Function*, 219(5):1787–1800, 2014.

B. J. Hunnicutt, B. C. Jongbloets, W. T. Birdsong, K. J. Gertz, H. Zhong, and T. Mao. A comprehensive excitatory input map of the striatum reveals novel functional organization. *eLife*, 5(vember2016):1–32, Nov. 2016.

J. R. Hunt. Progressive atrophy of the globus pallidus. *Brain*, 40(1-4):58–148, 1917.

A. Hämäläinen, K. Arndt, A. Ghadirzadeh, and V. Kyrki. Affordance learning for end-to-end visuomotor robot control. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1781–1788. IEEE, 2019. doi: 10.1109/IROS40897.2019.8968596. URL `https://doi.org/10.1109/IROS40897.2019.8968596`.

Y. Iino, T. Sawada, K. Yamaguchi, M. Tajiri, S. Ishii, H. Kasai, and S. Yagishita. Dopamine d2 receptors in discrimination learning and spine enlargement. *Nature*, 579(7800):555–560, March 2020.

I. Ilinsky and K. Kultas-Ilinsky. Sagittal cytoarchitectonic maps of the macaca mulatta thalamus with a revised nomenclature of the motor-related nuclei validated by observations on their connectivity. *Journal of Comparative Neurology*, 262(3):331–364, 1987.

M. Ito and K. Doya. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, 29(31):9861–9874, 2009.

P. N. Izzo and J. P. Bolam. Cholinergic synaptic input to different parts of spiny striatonigral neurons in the rat. *Journal of Comparative Neurology*, 269(2):219–234, 1988.

D. Jaeger, H. Kita, and C. J. Wilson. Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum. *J. Neurophysiol.*, 72(5): 2555–2558, November 1994.

G. Jocham, T. A. Klein, and M. Ullsperger. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of Neuroscience*, 31(5):1606–1613, 2011. doi: 10.1523/JNEUROSCI.3904-10.2011.

D. Joel and I. Weiner. The connections of the primate subthalamic nucleus: indirect pathways and the open-interconnected scheme of basal ganglia-thalamocortical circuitry. *Brain research reviews*, 23(1-2):62–78, 1997.

D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw*, 15(4):535–547, June 2002.

J. M. Kemp and T. P. Powell. The site of termination of afferent fibres in the caudate nucleus. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 262(845):413–427, 1971.

S. Keshmiri. Entropy and the brain: An overview. *Entropy (Basel)*, 22(9): 917, 2020. doi: 10.3390/e22090917. Published online 2020 Aug 21. PMCID: PMC7597158. PMID: 33286686.

M. Keuken, P. Bazin, L. Crown, J. Hootsmans, A. Laufer, C. Müller-Axt, E. Putten, and J. Van Der. Quantifying inter-individual anatomical variability in the subcortex using 7 t structural mri. *Neuroimage*, 94:40–46, 2014. doi: 10.1016/j.neuroimage.2014.03.032.

K. Khetarpal, Z. Ahmed, G. Comanici, D. Abel, and D. Precup. What can i do here? a theory of affordances in reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020. URL `https://arxiv.org/abs/2009.11694`. Updated: May 05, 2023.

H. F. Kim, H. Amita, and O. Hikosaka. Indirect pathway of caudal basal ganglia for rejection of valueless visual objects. *Neuron*, 94(4):920930.e3, May 2017.

E. Y. Kimchi and M. Laubach. The dorsomedial striatum reflects response bias during learning. *The Journal of Neuroscience*, 29(47):14891–14902, 2009a. doi: 10.1523/JNEUROSCI.4060-09.2009.

E. Y. Kimchi and M. Laubach. Dynamic encoding of action selection by the medial striatum. *Journal of Neuroscience*, 29(10):3148–3159, 2009b. doi: 10.1523/JNEUROSCI.5206-08.2009.

A. E. Kincaid, J. B. Penney, A. B. Young, and S. W. Newman. The globus pallidus receives a projection from the parafascicular nucleus in the rat. *Brain research*, 553(2):18–26, 1991.

A. E. Kincaid, T. Zheng, and C. J. Wilson. Connectivity and convergence of single corticostriatal axons. *J. Neurosci.*, 18(12):4722–4731, June 1998.

H. Kita and T. Kita. Cortical stimulation evokes abnormal responses in the dopamine-depleted rat basal ganglia. *J. Neurosci.*, 31(28):10311–10322, July 2011.

H. Kita, C. Ohye, M. Kimura, and McKenzie. *The basal ganglia V*, pages 77–94. Plenum Press, 1996.

H. Kita, H. Tokuno, and A. Nambu. Monkey globus pallidus external segment neurons projecting to the neostriatum. *Neuroreport*, 10(7):1467–1472, May 1999.

S. T. Kitai and J.-M. Deniau. Cortical inputs to the subthalamus: intracellular analysis. *Brain Research*, 214:411–415, 1981.

H. Kitano, I. Tanibuchi, and K. Jinnai. Neuronal activity in the motor area of the substantia nigra pars reticulata in monkeys during postural adjustments under different behavioral contexts. *Neuroscience Research*, 31(2):99–112, 1998.

H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. doi: 10.1016/j.cviu.2010.08.002. URL `https://doi.org/10.1016/j.cviu.2010.08.002`.

E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235, 2007.

A. V. Kravitz, B. S. Freeze, P. R. L. Parker, K. Kay, M. T. Thwin, K. Deisseroth, and A. C. Kreitzer. Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature*, 466(7306):622–626, jul 2010.

A. V. Kravitz, L. D. Tye, and A. C. Kreitzer. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.*, 15(6):816–818, June 2012.

R. Krebs, B. Schott, and E. Düzel. Personality traits are differentially associated with patterns of reward and novelty processing in the human substantia nigra/ventral tegmental area. *Biol. Psychiatry*, 65(2):103–110, 2009. doi: 10.1016/j.biopsych.2008.08.019.

R. Krebs, D. Heipertz, H. Schuetze, and E. Duzel. Novelty increases the mesolimbic functional connectivity of the substantia nigra/ventral tegmental area (sn/vta) during reward anticipation: evidence from high-resolution fmri. *Neuroimage*, 58, 2011. doi: 10.1016/j.neuroimage.2011.06.038.

A. C. Kreitzer and R. C. Malenka. Striatal plasticity and basal ganglia circuit function. *Neuron*, 60(4):543–554, November 2008.

S. W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

S. Kung and K. Diamantars. Ieee international conference on acoustics, speech, and signal processing. volume 2, pages 861–864, 1990.

L. Lai and S. J. Gershman. Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pages 195–232. Elsevier, 2021.

S. R. Lapper and J. P. Bolam. Input from the frontal cortex and the parafascicular nucleus to cholinergic interneurons in the dorsal striatum of the rat. *Neuroscience*, 51(3):533–545, 1992.

B. Lau and P. W. Glimcher. Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.*, 27(52):14502–14514, December 2007.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *CoRR*, abs/1805.00909, 2018.

R. Lowe and T. Ziemke. Exploring the relationship of reward and punishment in reinforcement learning. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 2013.

J. MacInnes, K. Dickerson, N. Chen, and R. Adcock. Cognitive neurostimulation: learning to volitionally sustain ventral tegmental area activation. *Neuron*, 89(6): 1331–1342, 2016. doi: 10.1016/j.neuron.2016.02.002.

D. Majid, W. Cai, J. Corey-Bloom, and A. Aron. Proactive selective response suppression is implemented via the basal ganglia. *J. Neurosci.*, 33(33):13259–13269, aug 2013.

G. Marini, L. Pianca, and G. Tredici. Descending projections arising from the parafascicular nucleus in rats. trajectory of fibers, projection pattern and mapping of terminations. *Somatosensory & Motor Research*, 16(3):207–222, 1999.

J. E. Markowitz, W. F. Gillis, C. C. Beron, S. Q. Neufeld, K. Robertson, N. D. Bhagat, and S. R. Datta. The striatum organizes 3d behavior via moment-to-moment action selection. *Cell*, 174(1):44–58, 2018.

D. Marr. *Vision: A Computational Approach*. Freeman Co, 1982.

D. Marr and T. Poggio. From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488, 1977.

K. J. Mastro, R. S. Bouchard, H. A. K. Holt, and A. H. Gittis. Transgenic mouse lines subdivide external segment of the globus pallidus (gpe) neurons and reveal distinct gpe output pathways. *The Journal of Neuroscience*, 34(6):2087–2099, 2014.

H. R. Maturana and F. J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy of Science. D. Reidel, Dordrecht, Holland, 1980.

S. M. McClure, N. D. Daw, and P. Read Montague. A computational substrate for incentive salience. *Trends in Neurosciences*, 26(8):423–428, 2003.

D. A. McRitchie, C. D. Hardman, and G. M. Halliday. Cytoarchitectural distribution of calcium binding proteins in midbrain dopaminergic regions of rats and humans. *Journal of Comparative Neurology*, jan 1996. doi: 10.1002/(SICI)1096-9861(19960101)364:1⟨121::AID-CNE11. URL `https://doi.org/10.1002/(SICI)1096_9861(19960101)364:1<121::AID-CNE11`.

J. W. Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.*, 50(4):381–425, November 1996.

J. W. Mink and W. T. Thach. Basal ganglia motor control. ii. late pallidal timing relative to movement onset and inconsistent pallidal coding of movement parameters. *J. Neurophysiol.*, 65(2):301–329, February 1991.

J. W. Mink and W. T. Thach. Basal ganglia intrinsic circuits and their role in behavior. *Curr. Opin. Neurobiol.*, 3(6):950–957, December 1993.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

G. Mogenson, D. Jones, and C. Yim. From motivation to action: functional interface between the limbic system and the motor system. *Progress in neurobiology*, 14(2-3):69–97, 1980.

I. Momennejad and M. W. Howard. Predicting the future with multi-scale successor representations. *bioRxiv*, page 449470, 2018. doi: 10.1101/449470.

I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, 2017. doi: 10.1038/s41562-017-0180-8.

P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.*, 16(5): 1936–1947, March 1996.

M. Morales and E. Margolis. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nat. Rev. Neurosci.*, 2017. doi: 10.1038/nrn.2016.165.

A. Motiwala, S. Soares, B. Atallah, J. Paton, and C. Machens. Dopamine responses reveal efficient coding of cognitive variables. *arXiv preprint arXiv:2005.05841*, 2020.

A. Motiwala, S. Soares, B. V. Atallah, J. J. Paton, and C. K. Machens. Efficient coding of cognitive variables underlies dopamine response and choice behavior. *Nature Neuroscience*, 25:738–748, 2022.

M. Muenter and J. Whisnant. Basal ganglia calcification, hypoparathyroidism, and extrapyramidal motor manifestations. *Neurology*, 18(11):1075–1083, 1968.

J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, and X.-J. Wang. A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.*, 17(12):1661–1663, December 2014.

T. Nagarajan and K. Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A. Nambu. A new dynamic model of the cortico-basal ganglia loop. *Progress in Brain Research*, 143:461–466, 2004.

A. Nambu. Somatotopic organization of the primate basal ganglia. *Frontiers in Neuroanatomy*, 5:26, 2011.

H. J. W. Nauta and M. Cole. Efferent projections of the subthalamic nucleus: an autoradiographic study in monkey and cat. *Journal of Comparative Neurology*, 180:1–16, 1978.

N. Navarro-Guerrero, R. J. Lowe, and S. Wermter. Improving robot motor learning with negatively valenced reinforcement signals. *Frontiers in neurorobotics*, 11: 10, 2017.

A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. doi: 10.1109/IROS.2017.8206484. URL `https://doi.org/10.1109/IROS.2017.8206484`.

Y. Niv, N. D. Daw, D. Joel, and P. Dayan. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520, Apr 2007.

J. O'Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454, April 2004.

R. A. Ohrenstein. The socio-economic time dimension in talmudic literature. *International Journal of Social Economics*, 23(4/5/6):357–369, 1996. doi: 10.1108/03068299610121912.

E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

H. Okada, H. Yamakawa, and T. Omori. Two dimensional evaluation reinforcement learning. In *Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence: 6th International Work-Conference on Artificial and Natural*

Neural Networks, IWANN 2001 Granada, Spain, June 13–15, 2001 Proceedings, Part 1 6. Springer Berlin Heidelberg, 2001.

J. Olszewski and D. Baxter. *Cytoarchitecture of the Human Brain Stem*. S. Karger, Basel, 1954.

D. Oorschot. Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: A stereological study using the cavalieri and optical disector methods. *Journal of Comparative Neurology*, 366:580–599, 1996.

D. E. Oorschot, N. Lin, B. H. Cooper, J. N. Reynolds, H. Sun, and J. R. Wickens. Synaptic connectivity between rat striatal spiny projection neurons in vivo: Unexpected multiple somatic innervation in the context of overall sparse dendritic connectivity. *Basal Ganglia*, 2(1):43–52, 2002.

B. Panigrahi, K. Martin, Y. Li, A. Graves, A. Vollmer, L. Olson, and J. Dudman. Dopamine is required for the neural representation and control of movement vigor. *Cell*, 162(6):1418–1430, September 2015.

A. Parent and L. N. Hazrati. Functional anatomy of the basal ganglia. ii. the place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Res. Brain Res. Rev.*, 20(1):128–154, january 1995. doi: 10.1002/mds.10138. URL `https://doi.org/10.1002/mds.10138`.

J. G. Parker, J. D. Marshall, B. Ahanonu, Y.-W. Wu, T. H. Kim, B. F. Grewe, and M. J. Schnitzer. Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. *Nature*, 557(7704):177–182, May 2018.

J. Parkinson. *An Essay on the Shaking Palsy*. Whittingham and Rowland for Sherwood, Neely and Jones, London, 1817.

I. P. Pavlov. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, 1927.

C. M. A. Pennartz et al. The hippocampal–striatal axis in learning, prediction and goal-directed behavior. *Trends in Neurosciences*, 34(10):548–559, 2011. et al.

G. R. Penny, C. J. Wilson, and S. T. Kitai. Relationship of the axonal and dendritic geometry of spiny projection neurons to the compartmental organization of the neostriatum. *Journal of Comparative Neurology*, 251(3):323–341, 1986.

G. Percheron, J. Yelnik, and C. François. A golgi analysis of the primate globus pallidus. iii. spatial organization of the striatopallidal complex. *Journal of Comparative Neurology*, 227:214–227, 1984.

G. Percheron, C. Francois, J. Yelnik, G. Fenelon, and B. Talbi. *The basal ganglia related system of primates: Defnition, description and informational analysis*, pages 3–20. Springer US, Boston, MA, 1994.

M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan, and C. D. Frith. Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. *Nature*, 442(7106):1042–1045, 2006.

M. Pignatelli and A. Bonci. Role of dopamine neurons in reward and aversion: A synaptic plasticity perspective. *Neuron*, 86(5):1145–1157, oct 2015. doi: 10.1016/j.neuron.2015.04.015. URL `https://doi.org/10.1016/j.neuron.2015.04.015`.

T. Prescott, F. Montes González, K. Gurney, M. Humphries, and P. Redgrave. A robot model of the basal ganglia: behavior and intrinsic processing. *Neural Netw.*, 19(1):31–61, jan 2006.

J. A. Rafols and C. A. Fox. The neurons in the primate subthalamic nucleus: a golgi and electron microscopic study. *Journal of Comparative Neurology*, 168: 75–112, 1976.

S. Ramon y Cajal. *Histologie du Systeme Nerveux de l'Homme et des Vertebres*, volume 2. Maloine, 1899.

R. Ranaldi. Dopamine and reward seeking: the role of ventral tegmental area. *Rev. Neurosci.*, 25(5):621–630, 2014. doi: 10.1515/revneuro-2014-0019. URL `http://doi.org/10.1515/revneuro-2014-0019`.

P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023, 1999.

P. Redgrave, T. Prescott, and K. Gurney. Is the short-latency dopamine response too short to signal reward error? *Trends in Neuroscience*, 22(4):146–151, 1999a.

J. Reynolds, B. Hyland, and J. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413(6851):67–70, Sep 2001.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

T. Robbins and B. Everitt. Neurobehavioural mechanisms of reward and motivation. *Current Opinion in Neurobiology*, 6:228 – 236, 1996.

M. R. Roesch, D. J. Calu, and G. Schoenbaum. Dopamine neurons encode the better option in rats deciding between diferently delayed or sized rewards. *Nat. Neurosci.*, 10(12):1615–1624, December 2007.

D. H. Root, H.-L. Wang, B. Liu, D. J. Barker, L. Mód, P. Szocsics, and M. Morales. Glutamate neurons are intermixed with midbrain dopamine neurons in nonhuman primates and humans. *Sci. Rep.*, 6(1):30615, 2016. doi: 10.1038/srep30615.

T. K. Roseberry, A. M. Lee, A. L. Lalive, L. Wilbrecht, A. Bonci, and A. C. Kreitzer. Cell-type-specifc control of brainstem locomotor circuits by basal ganglia. *Cell*, 164(3):526–537, January 2016.

D. L. Rosin, A. Robeya, R. L. Woodard, P. G. Guyenet, and J. Linden. Immunohistochemical localization of adenosine a2a receptors in the rat central nervous system. *Journal of Comparative Neurology*, 401(2):163–186, 1998.

E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, and N. D. Daw. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13(9):e1005768, 2017. doi: 10.1371/journal.pcbi.1005768. eCollection 2017 Sep.

S. J. Russo and E. J. Nestler. The brain reward circuitry in mood disorders. *Nature Reviews Neuroscience*, 14(9):609–625, 2013.

A. F. Sadikot, A. Parent, and C. François. Efferent connections of the centromedian and parafascicular thalamic nuclei in the squirrel monkey: A pha-l study of subcortical projections. *Journal of Comparative Neurology*, 315(1):137–159, 1992.

J. D. Salamone, M. Correa, S. M. Mingote, and S. M. Weber. Beyond the reward hypothesis: Alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology*, 5:34–41, 2005. doi: 10.1016/j.coph.2004.09.004.

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

I. Sano et al. Distribution of catechol compounds in human brain. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Enzymological Subjects*, 32(3): 586–587, 1959.

F. Sato, P. Lavallée, M. Lévesque, and A. Parent. Singleaxon tracing study of neurons of the external segment of the globus pallidus in primate. *J. Comp. Neurol.*, 417(1):17–31, 2000.

E. Schneidman, N. Slonim, N. Tishby, R. d. van Steveninck, and W. Bialek. Analyzing neural codes using the information bottleneck method. *Advances in Neural Information Processing Systems, NIPS*, 2001.

T. Schonberg, N. D. Daw, D. Joel, and J. P. O'Doherty. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, 27(47):12860–12867, 2007.

W. Schultz. The primate basal ganglia between the intention and outcome of action. 1995.

W. Schultz. Updating dopamine reward signals. *Current Opinion in Neurobiology*, 23(2):229–238, 2013. doi: 10.1016/j.conb.2012.11.012.

W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.*, 13(3):900–913, March 1993.

W. Schultz, P. Dayan, and R. P. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, Mar 1997.

R. Schwarting and J. Huston. The unilateral 6-hydroxydopamine lesion model in behavioral brain research. analysis of functional defcits, recovery and treatments. *Prog. Neurobiol.*, 50(2-3):275–331, Oct 1996.

E. Shink, M. D. Bevan, J. P. Bolam, and Y. Smith. The subthalamic nucleus and the external globus pallidus: Two tightly interconnected structures that control the output of the basal ganglia in the monkey. *Neuroscience*, 73(2):335–357, 1996.

T. Sippy, D. Lapray, S. Crochet, and C. Petersen. Celltype-specifc sensorimotor processing in striatal projection neurons during goal-directed behavior. *Neuron*, 88(2):298–305, October 2015.

B. F. Skinner. The behavior of organisms: an experimental analysis. 457, 1938.

K. Smith, K. Berridge, and J. Aldridge. Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *Proc. Natl. Acad. Sci. U. S. A*, 108(27):E255–E264, 2011. doi: 10.1073/pnas.1101920108.

Y. Smith and A. Parent. Differential connections of caudate nucleus and putamen in the squirrel monkey (saimiri sciureus). *Neuroscience*, 18(2):347–371, 1986.

Y. Smith, L.-N. Hazrati, and A. Parent. Efferent projections of the subthalamic nucleus in the squirrel monkey as studied by the pha-l anterograde tracing method. *Journal of Comparative Neurology*, 294:306–323, 1990.

Y. Smith, M. Bevan, E. Shink, and J. Bolam. Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience*, 86(2):353–387, sep 1998.

S. Soares, B. V. Atallah, and J. J. Paton. Midbrain dopamine neurons control judgment of time. *Science*, 354(6317):1273–1277, December 2016.

O. Sporns. *Networks of the Brain*. MIT press, 2010.

K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. The hippocampus as a predictive map. *Nat. Neurosci.*, 20(11):1643–1653, Nov 2017.

H. Steiner and K. Y. Tseng. *Handbook of basal ganglia structure and function*. Academic Press, 2016.

R. Suri and W. Schultz. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, 121:350–354, 1998.

R. Suri, J. Bargas, and M. Arbib. Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103(1):65–85, 2001. doi: 10.1016/s0306-4522(00)00554-6.

R. E. Suri and W. Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3): 871–890, 1999. doi: 10.1016/s0306-4522(98)00697-6.

R. Sutton, A. Barto, et al. *Introduction to reinforcement learning*. MIT press Cambridge, 1998.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.

R. S. Sutton, A. G. Barto, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, (5):834–846, 1983.

L. Swanson. The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.*, 9(1-6):321–353, 1982. doi: 10.1016/0361-9230(82)90145-9.

L. W. Swanson and M. Bota. Foundational model of structural connectivity in the nervous system with a schema for wiring diagrams, connectome, and basic plan architecture. *Proceedings of the National Academy of Sciences*, 107(48): 20610–20617, 2010.

L.-H. Tai, A. M. Lee, N. Benavidez, A. Bonci, and L. Wilbrecht. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.*, 15(9):1281–1289, September 2012. doi: 10.1038/nn.3188.

P. Tano, P. Dayan, and A. Pouget. A local temporal difference code for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

F. Tecuapetla, S. Matias, G. Dugue, Z. Mainen, and R. Costa. Balanced activity in basal ganglia projection pathways is critical for contraversive movements. *Nat. Commun.*, 5:4315, July 2014.

F. Tecuapetla, X. Jin, S. Lima, and R. Costa. Complementary contributions of striatal projection pathways to action initiation and execution. *Cell*, 166(3): 703–715, July 2016.

D. Thaler, S.-C. Chen, P. D. Nixon, C. E. Stern, and R. E. Passingham. The functions of the medial premotor cortex. ii. the timing and selection of learned movements. *Experimental Brain Research*, 95(2):206–222, 2013.

N. Tishby and D. Polani. Information theory of decisions and actions. In *Perception-action cycle: Models, architectures, and hardware*, pages 601–636. Springer New York, New York, NY, 2010.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

R. Turner and M. Desmurget. Basal ganglia contributions to motor control: a vigorous tutor. *Current Opinion in Neurobiology*, 20(6):704–716, December 2010.

L. van Bogaert. Contribution clinique et anatomique à l'etude de la paralysie agitante juvenile primitive. *Revue Neurologique*, 2:315–326, 1930.

S. D. Vann, J. P. Aggleton, and E. A. Maguire. What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, 10(11):792–802, 2009.

A. Vesalius. *De humani corporis fabrica, libri septem.* Oporinus, Basel, 1543.

J. Von Uexkull. *A foray into the worlds of animals and humans: with a theory of meaning.* U of Minnesota Press, 1934.

N. R. Wall, M. De La Parra, E. M. Callaway, and A. C. Kreitzer. Differential innervation of direct- and indirect-pathway striatal projection neurons. *Neuron*, 79(2):347–360, July 2013.

M. Watabe-Uchida, L. Zhu, S. K. Ogawa, A. Vamanrao, and N. Uchida. Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, 74(5): 858–873, June 2012. doi: 10.1016/j.neuron.2012.03.017.

M. Watanabe and D. P. Munoz. Presetting basal ganglia for volitional actions. *J. Neurosci.*, 30(30):10144–10157, July 2010.

C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

W. Wei, A. Mohebi, and J. D. Berke. A spectrum of time horizons for dopamine signals. *bioRxiv*, 2021. doi: 10.1101/2021.10.31.466705.

A. P. Weible. Remembering to attend: the anterior cingulate cortex and remote memory. *Behavioral Brain Research*, 245:63–75, 2013.

A. C. L. Westphal. Über eine dem bilde der cerebrospinalen grauen degeneration ähnliche erkrankung des centralen nervensystems ohne anatomischen befund, nebst einigen bemerkungen über paradoxe contraction. *Archiv für Psychiatrie und Nervenkrankheiten*, 14(1):87–95, 767–773, 1883.

T. Wichmann and M. Kliem. Neuronal activity in the primate substantia nigra pars reticulata during the performance of simple and memory-guided elbow movements. *Journal of Neurophysiology*, 91(2):815–827, 2004.

T. Wichmann, H. Bergman, P. A. Starr, T. Subramanian, R. L. Watts, and M. R. DeLong. Comparison of mptp-induced changes in spontaneous neuronal discharge in the internal pallidal segment and in the substantia nigra pars reticulata in primates. *Exp. Brain Res.*, 125(4):397–409, April 1999.

J. Wickens, R. Kotter, and M. Alexander. Effects of local connectivity on striatal function: stimulation and analysis of a model. *Synapse*, 20:281–298, 1995.

J. Wickens et al. Two dynamic modes of striatal function under dopaminergic–cholinergic control: Simulation and analysis of a model. *Synapse*, 8:1–12, 1991.

J. R. Wickens. Basal ganglia: structure and computations. *Network: computation in neural systems*, 8(4):R77–R109, 1997.

M. S. Williams and P. S. Goldman-Rakic. Widespread origin of the primate mesofrontal dopamine system. *Cerebr. Cortex*, 8(4):321–345, 1998. doi: 10.1093/cercor/8.4.321.

T. Willis. *Cerebri anatome cui accessit nervorum descriptio et usus*. Flesher, Martyn & Allestry, London, 1664.

C. J. Wilson and P. M. Groves. Fine structure and synaptic connections of the common spiny neuron of the rat neostriatum: A study employing intracellular injection of horseradish peroxidase. *Journal of Comparative Neurology*, 194(3): 599–615, 1980.

B. D. Winters, L. M. Saksida, and T. J. Bussey. Object recognition memory: neurobiological mechanisms of encoding, consolidation and retrieval. *Neuroscience & Biobehavioral Reviews*, 32(5):1055–1070, 2008.

Y. Yamada, K. Yamamura, and M. Inoue. Neural mechanisms of oral somatosensory function: a review of the literature and implications for future research. *Journal of oral biosciences*, 47(1):1–11, 2005.

H. Yang, J. W. de Jong, Y. E. Tak, J. Peck, H. S. Bateup, and S. Lammel. Nucleus accumbens subnuclei regulate motivated behavior via direct inhibition

and disinhibition of vta dopamine subpopulations. *Neuron*, 97(2):434–449, 2018. doi: 10.1016/j.neuron.2017.12.022.

X. Yang, Z. Ji, J. Wu, and Y.-K. Lai. Recent advances of deep robotic affordance learning: A reinforcement learning perspective. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. doi: 10.1109/tcds.2023.3277288. In press.

T. Yashukawa, T. Kita, Y. Xue, and H. Kita. Rat intralaminar thalamic nuclei projections to the globus pallidus: A biotinylated dextran amine anterograde tracing study. *Journal of Comparative Neurology*, 471(2):153–167, 2004.

J. Yelnik, C. Francois, and D. Tand. Proceedings if the third congress of european neuroscience society. 1997.

H. Yin, B. Knowlton, and B. Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19:181–189, 2004a.

H. H. Yin and R. M. Costa. Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience*, 12(3):333–342, 2009. doi: 10.1038/nn.2261.

H. H. Yin, B. J. Knowlton, and B. W. Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.*, 19(1):181–189, January 2004b.

H. H. Yin, S. B. Ostlund, B. J. Knowlton, and B. W. Balleine. The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.*, 22(2): 513–523, July 2005.

H. H. Yin, B. J. Knowlton, and B. W. Balleine. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behav. Brain Res.*, 166(2):189–196, 2006.

E. A. Yttri and J. T. Dudman. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature*, 533(7603):402–406, May 2016.

T.-S. Zhang, A. N. Placzek, and J. A. Dani. In vitro identification and electrophysiological characterization of dopamine neurons in the ventral tegmental area. *Neuropharmacology*, 59(6):431–436, jun 2010. doi: 10.1016/j.neuropharm.2010.06.004. URL `http://doi.org/10.1016/j.neuropharm.2010.06.004`.

Y. Zhang, K.-H. Larcher, B. Misic, and A. Dagher. Anatomical and functional organization of the human substantia nigra and its connections. *eLife*, 6:1–23, 2017. doi: 10.7554/eLife.26653.

B. Zingg, H. Hintiryan, L. Gou, M. Y. Song, M. Bay, M. S. Bienkowski, N. N. Foster, S. Yamashita, I. Bowman, A. W. Toga, and H.-W. Dong. Neural networks of the mouse neocortex. *Cell*, 156(5):1096–1111, 2014. doi: 10.1016/j.cell.2014.02.023.

Appendix A

IB Simulations for various initial conditions

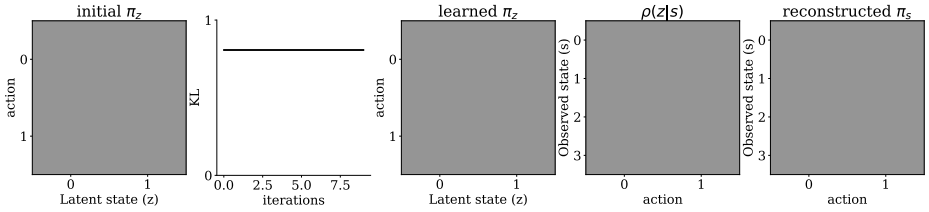


Figure A.1:

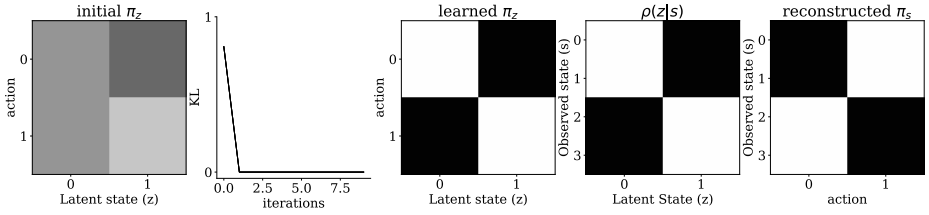


Figure A.2:

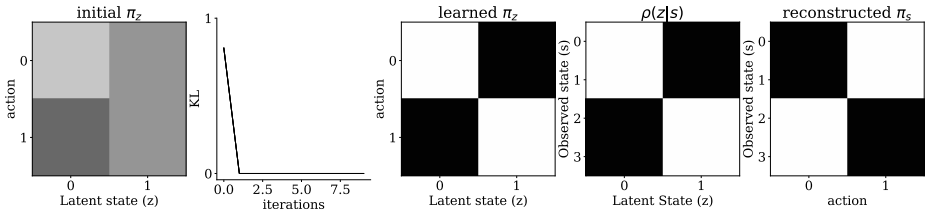


Figure A.3:

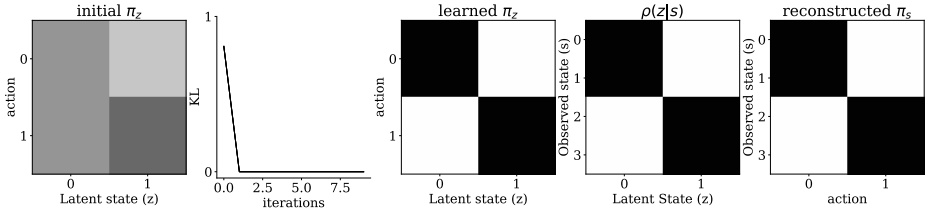


Figure A.4:

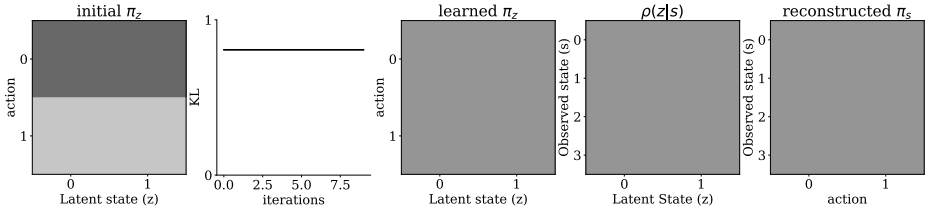


Figure A.5:

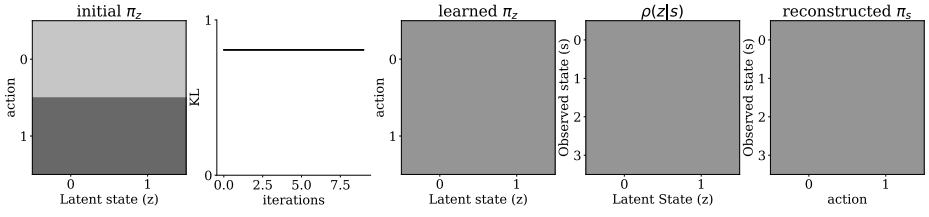


Figure A.6:

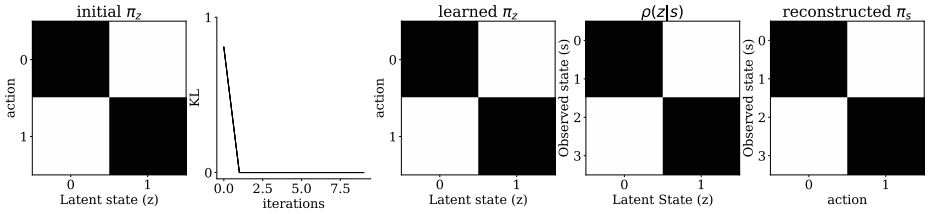


Figure A.7:

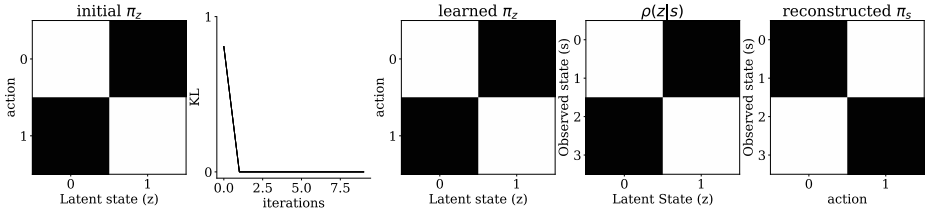


Figure A.8:

Appendix B

Multi-agent model simulation details

All simulations were performed on an 80 dimensional state-space. Each trial consisted on a trajectory with a pre-specified dwell time sampled from a gaussian distribution with mean $\mu_t = 67ms$ and variance $\sigma_t = 20ms$, sampled from a predetermined set of 30 dwell times that covered the interval between $t_{max} = 100ms$ and 65 ms. Convergence was assessed by verifying that all value functions and action preferences showed a normalized difference between epochs under 0.01 for more than 50,000 trials, with a total number of 250 000 trials.

The training set for the second tone times (in seconds) was given by $T_{train} = [0.1, 0.3, 0.5, 0.6, 0.8, 1.05, 1.26, 1.38, 1.62, 1.74, 1.95, 2.2, 2.4, 2.5]$ with the decision boundary being set at 1.5s. The learning rates were all initialized at 20 and decreased for each state proportionally to the number of visits to each state. Rewards for correct and incorrect/broken fixation trials were set at 10 and -5, respectively.

The transfer function $f^p(\delta_t)$ parameters were defined as $[b_0^p, b_1^p, b_2^p, b_3^p]$. The values for the direct pathway actor were set at $[-0.7, 8, 3.7, 1], [-0.7, 3.7, 8, -1]$ for the indirect and $[-4, 8.5, 0.45, 1]$ for the third actor. The weights for each actor $\omega_D = -\omega_I$ were all equal to 1. The discount parameter for each critic γ was 0.98 and the temperature of the softmax function β for the policy was 1.5 for both training and testing. In order to simulate the effects of pathway specific optogenetic inhibition, we generated control and manipulated datasets (1.5k trials) drawing single trial intervals from a set identical to the one used to test the animals ($T_{test}[0.6, 1.05, 1.26, 1.38, 1.62, 1.74, 1.95, 2.4]$).

Algorithm 1 Opponent Multi-Agent Actor-Critic Algorithm

Define $\{\beta, \gamma, \sigma, \rho, k_i^p, \tau_i\}$
Initialize states $\{s_t^{D/I}, s_t^{\chi}\}$
Zero initialization of critics $\{V^{D/I}(s^{D/I}), V^{\chi}(s^{\chi})\}$
Zero initialization of action preferences $\{A^{D/I}(s^{D/I}, a), A^{\chi}(s^{\chi}, a), A^{T}(s^{D/I}, s^{\chi}, a)\}$
for *episode* **do**
    Sample dwell time $t_d \sim \mathcal{N}(\mu_t, \sigma_t)$
    for *step* in *episode* **do**
        Update Learning rates:
        $\alpha(s_t^{D/I}) \leftarrow \alpha(s_t^{D/I}) + (\mu(s^D/I) + 1)^{-1}$
        $\alpha(s_t^{\chi}) \leftarrow \alpha(s_t^{\chi}) + (\mu(s^{\chi}) + 1)^{-1}$
        Choose action and get next state:
        Check if HOLD action should be triggered:
        if $A^T(s_t^{D/I}, s_t^{\chi}, a) < 0 \quad \forall \quad a = \{SHORT, LONG\}$ **then**
            $A^T(s_t^{D/I}, s_t^{\chi} HOLD) = 0$
        else
            $A^T(s_t^{D/I}, s_t^{\chi}, HOLD) = -\infty$
        end if
        Sample new action $a_t \sim \pi(a|s^{D/I}, s^{\chi})$
        Observe $\{s_{t+1}^{D/I}, s_{t+1}^{\chi}, r_t\}$ from $\{K^{D/I}(t, \tau_i, t_d), K^{\chi}(t, t_d)\}$
        Update Critic:
        $\delta_t^{D/I} = r_t + \gamma V(s_{t+1}^{D/I}) - V(s_t^{D/I})$
        $\delta_t^{\chi} = r_t + \gamma V(s_{t+1}^{\chi}) - V(s_t^{\chi})$
        $V(s_t^{D/I}) \leftarrow V(s_t^{D/I}) + \alpha(s_t^{D/I})\delta_t^{D/I}$
        $V(s_t^{\chi}) \leftarrow V(s_t^{\chi}) + \alpha(s_t^{\chi})\delta_t^{\chi}$
        Update Action Preferences:
        $A^p(s_t^D, a_t) \leftarrow A^p(s_t^D, a_t) + \alpha(s^{D/I})[f^D(\delta_t^{D/I})$
        $A^p(s_t^I, a_t) \leftarrow A^p(s_t^I, a_t) + \alpha(s^{D/I}) f^I(\delta_t^{D/I})$
        $A^p(s_t^{\chi}, a_t) \leftarrow A^p(s_t^{\chi}, a_t) + \alpha(s^{\chi}) f^{\chi}(\delta_t^{\chi})$
        $A^T(s^{D/I}, s^{\chi}, a) = \omega_D A^p(s^{D/I}, a) + \omega_I A^p(s^{D/I}, a) + \omega_{\chi} A^p(s^{\chi}, a)$
    end for
end for

B.1 Full set of simulation metrics

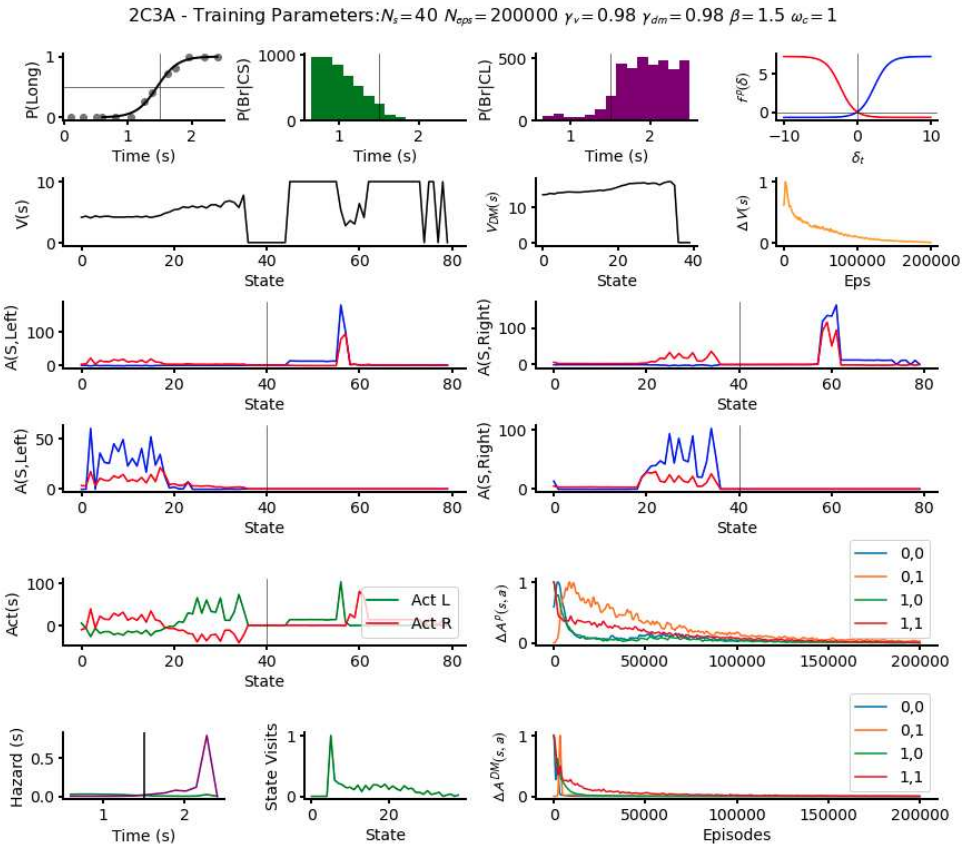


Figure B.1: Full set of simulation metrics and outputs from the two pathway L, X model.

Appendix C

Maximum entropy Reinforcement Learning

We start off by going through the derivation of the *Variational Inference* (VI) loss function and adapting the formalism to having also an *Affordance Representation* (AR), which we will potentially use to compress relevant state information as agents move through a particular Markov Decision Process (MDP).

We start by defining a path in the space of state-action pairs $s_t \in \mathcal{S}, a_t \in \mathcal{A}$ by

$$\tau_s = \{s_0, a_0, ..., s_T, a_T\} \tag{C.1}$$

We can then express the probability of a trajectory τ_s as

$$p(\tau_s) = p(s_0) \prod_{t=0}^{T} p(a_t) T(s_{t+1}|s_t, a_t) \tag{C.2}$$

i.e. we hypothesisze that in the case of a random trajectory through this space, the action probabilities are uniform, as they're not dependent in any policy π. This allows us to extract $p(a_t)$ as a constant C^T from the product above and define $p(\tau_s)$ as

$$p(\tau_s) \propto p(s_0) \prod_{t=0}^{T} T(s_{t+1}|s_t, a_t) \tag{C.3}$$

However, just by this definition we don't have any idea if the ongoing path is optimal or not. In order to have this information we need to constraint the states in the trajectory with a new binary variable o_t which will filter the optimal states i.e. a state-action pair at time t will be optimal if $o_t = 1$.

We can also express the probability of observing a sequence of observations $o_{1:T}$ given a trajectory τ_s as:

$$p(o_{1:T} = 1|\tau_s) = f(s_t, a_t) \tag{C.4}$$

We can compute the posterior probability of a trajectory given an observed sequence of observations using Bayes' rule. We can express the posterior probability of a trajectory τ_s given the observed sequence $o_{1:T} = 1$ as proportional to the product of their probabilities by

$$p(\tau_s|o_{1:T} = 1) \propto p(o_{1:T} = 1|\tau_s)p(\tau_s) \tag{C.5}$$

We can define a policy $\pi_\theta(\tau_s)$ as:

$$\pi_\theta(\tau_s) = p(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t)p(s_{t+1}|s_t, a_t) \tag{C.6}$$

We can then minimize the Kullback-Leibler (KL) divergence between the policy and the posterior as:

$$\min_\theta D_{KL}(\pi_\theta(\tau_s)||p(\tau_s|o_{1:T})) = \mathbb{E}_{\tau_s \sim \pi}[\sum_{t=0}^{T} f(s_t, a_t)] + \mathcal{H}(\pi_\theta) \tag{C.7}$$

This gives us the well known *Maximum Entropy* formalism of *Reinforcement Learning* where in the variational optimisatio of the underlying policy π_θ we also concurrently maximize the underlying entropy of that policy, guaranteeing that exploration is maintained for a given path τ_s.