Computational Modeling of the Basal Ganglia – Functional Pathways and Reinforcement Learning

Pierre Berthet
Computational Modeling of the Basal Ganglia

Functional Pathways and Reinforcement Learning

Pierre Berthet
Abstract

We perceive the environment via sensor arrays and interact with it through motor outputs. The work of this thesis concerns how the brain selects actions given the information about the perceived state of the world and how it learns and adapts these selections to changes in this environment. This learning is believed to depend on the outcome of the performed actions in the form of reward and punishment. Reinforcement learning theories suggest that an action will be more or less likely to be selected if the outcome has been better or worse than expected. A group of subcortical structures, the basal ganglia (BG), is critically involved in both the selection and the reward prediction.

We developed and investigated two computational models of the BG. They feature the two pathways prominently involved in action selection, one promoting, D1, and the other suppressing, D2, specific actions, along with a reward prediction pathway. The first model is an abstract implementation that enabled us to test how various configurations of its pathways impacted the performance in several reinforcement learning and conditioning tasks. The second one was a more biologically plausible version with spiking neurons, and it allowed us to simulate lesions in the different pathways and to assess how they affect learning and selection. Both models implemented a Bayesian-Hebbian learning rule, which computes the weights between two units based on the probability of their activations and co-activations. Additionally the learning rate depended on the reward prediction error, the difference between the actual reward and its predicted value, as has been reported in biology and linked to the phasic release of dopamine in the brain.

We observed that the evolution of the weights and the performance of the models resembled qualitatively experimental data. There was no unique best way to configure the abstract BG model to handle well all the learning paradigms tested. This indicates that an agent could dynamically configure its action selection mode, mainly by including or not the reward prediction values in the selection process.

With the spiking model it was possible to better relate results of our simulations to available biological data. We present hypotheses on possible biological substrates for the reward prediction pathway. We base these on the functional requirements for successful learning and on an analysis of the experimental data. We further simulate a loss of dopaminergic neurons similar to that re-
ported in Parkinson’s disease. Our results challenge the prevailing hypothesis about the cause of the associated motor symptoms. We suggest that these are dominated by an impairment of the D1 pathway, while the D2 pathway seems to remain functional.
Sammanfattning

Vi uppfattar omgivningen via våra sinnen och interagerar med den genom vårt motoriska beteende. Denna avhandling behandlar hur hjärnan väljer handlningar utifrån information om det upplevda tillståndet i omvärlden och hur den lär sig att anpassa dessa val till förändringar i den omgivande miljön. Detta lärande antas bero på resultatet av de genomförda åtgärderna i form av belönning och bestraffning. Teorier om förstärkninginlärnings (”reinforcement learning”) hävdar att en åtgärd kommer att vara mer eller mindre sannolik att väljas om resultatet tidigare har varit bättre respektive sämre än väntat. En grupp av subkortikala strukturer, de basala ganglierna (BG), är kritiskt involverad i både urval och förutsägelse om belönning.

Vi utvecklade och undersökte två beräkningsmodeller av BG. Dessa har två nervbanor som är kritiskt involverade i val av beteende, en som specifikt befrämjar (D1) och den andra som undertrycker (D2) beteende, tillsammans med en som är viktig för förutsägelse av belönning. Den första modellen är en abstrakt implementation som gjorde det möjligt för oss att testa hur olika konfiguritioner av dessa nervbanor påverkade resultatet i förstärkninginlärning och betingningsuppgifter. Den andra var en mer biologiskt detaljerad version med spikande nervceller, vilken tillåt oss att simulera skador i olika nervbanor samt att bedöma hur dessa påverkar inlärning och beteendeval. Båda modellerna implementerades med en Bayes-Hebbiansk inlärningsregel som beräknar vikter mellan två enheter baserade på sannolikheten för deras aktiveringar och coaktiveringar. Dessutom modulerades inlärningshastigheten beroende på felet i förutsägelse av belönning, dvs skillnaden mellan den verkliga belönningen och dess förutsagda värde, vilket har kopplats till fasisk frisättning av dopamin i hjärnan.


Med den spikande modellen var det möjligt att bättre relatera resultaten av våra simuleringar till biologiska data. Vi presenterar hypoteser om möjliga biologiska substrat för mekanismerna för belöningsförutsägelse. Vi baserar
dessa på dels på de funktionella kraven på framgångsrik inlärning och dels på en analys av experimentella data. Vi simulerar också effekterna av förlust av dopaminerga neuroner liknande den som rapporteras i Parkinsons sjukdom. Våra resultat utmanar den rådande hypotesen om orsaken bakom de associerade motoriska symtomen. Vi föreslår att effekterna domineras av en nedreglering av D1 inflytandet, medan D2 effekten verkar vara intakt.
Acknowledgements

I would like to first thank my supervisor Anders, who always had time to discuss all the aspects of the works during these years, for the feedback and support. I also wish to thank Jeanette for providing valuable insights and new angles to my research. I appreciated the basal ganglia meetings we had with Micke, Iman, and Örjan who furthermore, and along with Erik, took care to explain all the aspects of the Ph.D. studies. To Tony, Arvind and all the faculty members at CB, you are sources of inspiration.

I additionally give special thank to Pawel, always rising spirits with infallible support, to Bernhard, who showed me many coding tips and was a great travel companion, along with Phil, who helped me with the NEST module. Simon, Mr Lundqvist, Marcus, Cristina, Henrike, David, Peter, Pradeep and everyone else at CB and SBI, thank you for the stimulating and enjoyable atmosphere.

I am deeply thankful toward the European Union and the Marie Curie Initial Training Network program, for the great opportunity it has been to be part of this interdisciplinary training and wonderful experience. It was truly a privilege, and I look forward to meeting again everyone involved, Vasilis, Tahir, Radoslav, Mina, Marc-Olivier, Marco, Love, Johannes, Jing, Javier, Giacomo, Gerald, Friedemann, Filippo, Diego, Carlos, Alejandro, Ajith and Björn. Similar acknowledgements to the organizers and participants of the various summer schools I attended, I have been lucky to study with dedicated teachers and cheerful fellow students, and I hope we will come across each other again. I especially tip my headphones to the IBRO for actively, and generously, fostering and promoting collaborations between future scientists from all over the world.

Harriett, Linda and the administrative staff at KTH and SU, you have my sincere gratitude for all the help and support you provided. This work would not have been possible without PDC and their supercomputers Lindgren, Milner and Beskow.

I am immensely grateful to Laure, whose support and magic have made this journey into science, Sweden, and life, such a fantastic experience. I also want to thank my parents for having backed me all those years. Salutations and esteem to all my friends who keep me going forward, and special honors to the Savoyards.
To my parents, family, and friends
Contents

Abstract iv

Sammanfattning vi

Acknowledgements viii

List of Papers xiv

List of Figures xv

Abbreviations xvi

1 Introduction 17
  1.1 Aim of the thesis .................................. 18
  1.2 Thesis overview .................................. 18

2 Theoretical Framework 19
  2.1 Connectionism / Artificial Neural Networks ............ 19
  2.2 Learning ........................................... 20
    2.2.1 Supervised Learning ............................ 21
    2.2.2 Reinforcement Learning ........................ 21
    2.2.3 Unsupervised Learning .......................... 22
    2.2.4 Associative Learning ............................ 22
      2.2.4.1 Classical conditioning ...................... 22
      2.2.4.2 Operant conditioning ....................... 23
      2.2.4.3 Conditioning phenomena .................... 23
    2.2.5 Standard models of learning ..................... 24
      2.2.5.1 Backpropagation ............................ 24
      2.2.5.2 Hebb’s rule ................................. 25
  2.3 Models of Reinforcement Learning ..................... 25
    2.3.1 Model-based and model-free ..................... 26
    2.3.2 Rescorla-Wagner ............................... 26
List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I:  **Action selection performance of a reconfigurable basal ganglia inspired model with Hebbian-Bayesian Go-NoGo connectivity**
DOI: [10.3389/fnbeh.2012.00065](https://doi.org/10.3389/fnbeh.2012.00065)
My contributions were to implement the model, to design and perform the experiments, to analyze the data and to write the paper.

PAPER II:  **Optogenetic Stimulation in a Computational Model of the Basal Ganglia Biases Action Selection and Reward Prediction Error**
DOI: [10.1371/journal.pone.0090578](https://doi.org/10.1371/journal.pone.0090578)
My contributions were to conceive, to design and to perform the experiments, to analyze the data and to write the paper.

PAPER III:  **Functional relevance of different basal ganglia pathways investigated in a spiking model with reward dependent plasticity**
Berthet, P., Lindahl, M., Tully, P. J., Hellgren-Kotaleski, J. & Lansner, A. *Submitted*
My contributions were to implement and to tune the model, to conceive, to design and to perform the experiments, to analyze the data and to write the paper.

Reprints were made with permission from the publishers.
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Reinforcement learning models</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Actor-Critic architecture</td>
<td>30</td>
</tr>
<tr>
<td>2.3</td>
<td>Eligibility trace</td>
<td>32</td>
</tr>
<tr>
<td>3.1</td>
<td>Simplified representation of the principal connections of BG</td>
<td>45</td>
</tr>
<tr>
<td>3.2</td>
<td>Activity of dopaminergic neurons in a conditioning task</td>
<td>62</td>
</tr>
<tr>
<td>4.1</td>
<td>Evolution of BCPNN traces in a toy example</td>
<td>92</td>
</tr>
<tr>
<td>4.2</td>
<td>Schematic representation of a possible mapping onto biology of the BCPNN weights</td>
<td>93</td>
</tr>
<tr>
<td>5.1</td>
<td>Schematic representation of the model with respect to biology</td>
<td>99</td>
</tr>
<tr>
<td>5.2</td>
<td>Schematic representation of the RP setup</td>
<td>102</td>
</tr>
<tr>
<td>5.3</td>
<td>Raster plot of the spiking network during a change of block</td>
<td>105</td>
</tr>
<tr>
<td>5.4</td>
<td>Weights and performance of the abstract and spiking models</td>
<td>107</td>
</tr>
<tr>
<td>5.5</td>
<td>Box plot of the mean success ratio of various conditions of the spiking model</td>
<td>112</td>
</tr>
<tr>
<td>5.6</td>
<td>Evolution of the weights and selection of the habit pathway</td>
<td>122</td>
</tr>
<tr>
<td>5.7</td>
<td>Raster plot of the network in a delayed reward learning task</td>
<td>124</td>
</tr>
<tr>
<td>5.8</td>
<td>Performance of the model in a delayed reward learning task</td>
<td>126</td>
</tr>
<tr>
<td>5.9</td>
<td>Evolution of the weights in the D1, D2 and RP pathways in simulations of PD degeneration of dopaminergic neurons</td>
<td>129</td>
</tr>
</tbody>
</table>
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>attention deficit hyperactivity disorder</td>
</tr>
<tr>
<td>AI</td>
<td>artificial intelligence</td>
</tr>
<tr>
<td>BCM</td>
<td>Bienenstock, Cooper &amp; Munro</td>
</tr>
<tr>
<td>BCPNN</td>
<td>Bayesian confidence propagation neural network</td>
</tr>
<tr>
<td>BG</td>
<td>basal ganglia</td>
</tr>
<tr>
<td>CR</td>
<td>conditioned response</td>
</tr>
<tr>
<td>CS</td>
<td>conditioned stimulus</td>
</tr>
<tr>
<td>DBS</td>
<td>deep brain stimulation</td>
</tr>
<tr>
<td>EEG</td>
<td>electroencephalogram</td>
</tr>
<tr>
<td>EWMA</td>
<td>exponentially weighted moving averages</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional magnetic resonance imaging</td>
</tr>
<tr>
<td>GP</td>
<td>globus pallidus</td>
</tr>
<tr>
<td>GPe</td>
<td>external globus pallidus</td>
</tr>
<tr>
<td>GPI</td>
<td>internal globus pallidus</td>
</tr>
<tr>
<td>HMF</td>
<td>hybrid multiscale facility</td>
</tr>
<tr>
<td>LH</td>
<td>lateral habenula</td>
</tr>
<tr>
<td>LIF</td>
<td>leaky integrate and fire</td>
</tr>
<tr>
<td>LTD</td>
<td>long-term depression</td>
</tr>
<tr>
<td>LTP</td>
<td>long-term potentiation</td>
</tr>
<tr>
<td>MAP</td>
<td>maximum a posteriori</td>
</tr>
<tr>
<td>MEG</td>
<td>magnetoencephalography</td>
</tr>
<tr>
<td>MSN</td>
<td>medium spiny neuron</td>
</tr>
<tr>
<td>NEST</td>
<td>neural simulation tool</td>
</tr>
<tr>
<td>OFC</td>
<td>orbitofrontal cortex</td>
</tr>
<tr>
<td>PD</td>
<td>Parkinson’s disease</td>
</tr>
<tr>
<td>PFC</td>
<td>prefrontal cortex</td>
</tr>
<tr>
<td>PyNN</td>
<td>python neural networks</td>
</tr>
<tr>
<td>RP</td>
<td>reward prediction</td>
</tr>
<tr>
<td>RPE</td>
<td>reward prediction error</td>
</tr>
<tr>
<td>SARSA</td>
<td>state-action-reward-state-action</td>
</tr>
<tr>
<td>SN</td>
<td>substantia nigra</td>
</tr>
<tr>
<td>SNc</td>
<td>substantia nigra pars compacta</td>
</tr>
<tr>
<td>SNr</td>
<td>substantia nigra pars reticula</td>
</tr>
<tr>
<td>STDP</td>
<td>spike-timing-dependent plasticity</td>
</tr>
<tr>
<td>STN</td>
<td>subthalamic nucleus</td>
</tr>
<tr>
<td>TD</td>
<td>temporal difference</td>
</tr>
<tr>
<td>UR</td>
<td>unconditioned response</td>
</tr>
<tr>
<td>US</td>
<td>unconditioned stimulus</td>
</tr>
<tr>
<td>VTA</td>
<td>ventral tegmental area</td>
</tr>
<tr>
<td>WTA</td>
<td>winner takes all</td>
</tr>
</tbody>
</table>
1. Introduction

Understanding the brain is a challenge like no other, with possible major impacts on society and human identity. True artificial intelligence (AI) could indeed render any human work unnecessary. Computational neuroscience has a great role to play in the realisation of such AI. It is a challenge as well for society and ethics, as it needs to be decided how this knowledge can be used, and how far the society is willing to change, in the face of the coming discoveries.

The brain, and especially the human brain, is a complex machinery, with a very large number of interdependent dynamical systems, at different levels. It is suggested that the evolutionary reason for the brain is to produce adaptable and complex movements. The human brain, and its disproportionately large neocortex, seems to be capable of much more, but some argue that it all comes down to enriching the information that is going to be used to perform a movement, be it planning, fine tuning, estimating the outcome, or even building a better representation of the world in order to compute the best motor response possible in that environment. One thing seems certain, there is only through motor output that one can interact on and influence the environment (Wolpert et al., 1995). It has furthermore been proposed that the functional role of consciousness could revolve around integration for action (Merker, 2007).

If the brain is not yet understood at a degree which enables the simulation of such a true AI, the crude functional roles of subparts of this complex system appear however to be roughly fathomed. The development of new techniques of investigation and of data collection expands the biological description of such systems, but may fall short of offering non-trivial explanations or relevant hypotheses about these observations. This is precisely where computational modeling can be used, and it shows the need for some theoretical framework, or top-down approaches, to be able to build a model of all the available data. The interdependent relationship between theoretical works and experimental observations is the engine which will carry us on the path towards a better understanding of the brain and to its artificial implementation.

Our perception of the world rarely matches physical reality. This can boils down to the fact that our percepts can be biased towards our prior beliefs, which are based on our previous experience. Learning can thus be considered as the update of the beliefs with new data, e.g. the outcome of an event.

The basal ganglia (BG) are a group of subcortical structures which inte-
grate the multi-modal sensory inputs with cognitive and motor signals in order to select the action deemed the most appropriate. The relevance of the action is believed to be based on the expected outcome of this action. The BG are found in all vertebrates and are supposed to learn to select the actions through reinforcement learning, that is that the system receives information about the outcome of the action performed, i.e. a positive or a negative reward, and this action will be more or less likely to be selected again in the similar situation if the outcome has been better or worse than expected, respectively.

1.1 Aim of the thesis

In this thesis, we will present and use a model of learning which basically relies on the idea that neurons and neural networks can compute the probabilities of these events and their relations. The aim of the thesis is to investigate if action selection can be learned based on computation derived from Bayesian probabilities through the development of computational models of the BG which use reinforcement learning. Another goal is to study the functional contributions of the different pathways and characteristics of the BG on the learning performance via simulated lesions in our models, and notably a neurodegeneration observed in patients with Parkinson’s disease. This should lead to the development of new hypotheses about the biological substrate for such computations and also to better determine the reason behind the architecture and dynamics of these pathways.

1.2 Thesis overview

As it is often the case, it is by trying to build a copy of a complex system that one unravels requirements that might have been overlooked by a pure descriptive approach, that one stumbles upon various ways to obtain a similar result, and is thus faced with the task to assess these different options and their consequences. This thesis does not pretend to cover all aspects of learning, of the neuroscience of the basal ganglia, or of all the related models. The first part should provide the necessary background to understand the work presented, and some more to grasp the future and related challenges. We will start by presenting the theoretical framework used to model learning. We will then consider the biology involved and will detail the various levels and structures at work and their interdependencies. The framework for our model and learning rule will then be introduced, as well as some relevant other computational models. We will then describe our models and discuss some of our main results, and we will comment on the relation between these different results.
2. Theoretical Framework

Marr (1982) described different levels of analysis in order to understand information processing systems: 1. the computational level, what does the system do, and why? 2. the algorithmic or representational level, how does it do it? 3. the physical implementation level, what biological substrates are involved? In the 2010 edition of the book, Poggio added a fourth level, above the others: learning, in order to build intelligent systems. We will briefly present the theory of connectionism and artificial neural networks before giving an overview of what is one of these networks most interesting features: learning, and its different procedures.

2.1 Connectionism / Artificial Neural Networks

Connectionism is a theory commonly considered to have been formulated by Donald O. Hebb in 1949, where complex cognitive functions are supposed to emerge from the interconnection of simple units performing locally basic computations (Hebb, 1949) (but see Walker (1992) for some history about this theory). This idea still serves today as the foundation of most of computational theories and models of the brain and of its capabilities. Hebb also suggested another seminal theory, about the development of the strength of the connections between neurons, which would depend on their activity (see section 2.2.5.2 for a description). These interconnected units form what is called an (artificial) neural network and the architecture, as well as the unit properties, can vary depending on the model. In this work, the units represent neurons, or functional populations of neurons, and we will thus use the same term to refer to biological and computational neurons. Similarly, the connections represent synapses and learning occurs as the modification of these synapses, or changes of their weights. The activation function of a neuron can be linear or nonlinear, and among the latter, it can be made to mimic a biological neuron response. This function determines the activation of the neuron given its input, which is the sum of the activity received through all its incoming connections, in a given time window, time interval, or step. The response of the network can be obtained by accessing the activation levels of the neurons considered, as read out.
The Perceptron is one of the first artificial neural network, and is based on the visual system (Rosenblatt [1958]). It is a single layer network where the input, representing visual information for example, is fed to all neurons of the output layer, possibly coding for a motor response, through the weights of the connections. Depending on the strength of the connections, where a synaptic weight has to be defined, different inputs trigger different output result. The flow of information is here unidirectional: feedforward. Furthermore, this network can learn, that is, the strength of the connections can be updated to modify the output for a given input. The delta rule is one of the commonly used learning rule, where the goal is to minimize the error of the output in comparison with a provided target output, through gradient descent (section 2.2.5.1). A logical evolution is the multilayer Perceptron, which is also feedforward and has all-to-all connections between layers. This evolution is able to learn to distinguish non-linearly separable data, unlike the regular Perceptron. The propagation of the signal occurs in parallel via the different units and connections. Moreover, the use of the delta-rule can be refined as it is now possible to apply it not only to the weights of the output layer, but also to those of the previous layers, through backpropagation.

Other types of neural network can additionally contain recurrent, or feedback, connections. Of primary interest is the Hopfield network, which is defined by symmetric connections between units (Hopfield [1982]). It is guaranteed to converge to a local minimum, an attractor, and is quite robust to input distortions and node deletions. It is a distributed model of an associative memory so that, once a pattern has been trained, the presentation of a sub-sample should lead to the complete recollection of the original pattern, a process called pattern completion. Different learning rules can be used to store information in Hopfield networks, the Hebbian learning rule being one of them, where associations are being learned based on the inherent statistics of the input (section 2.2.5.2). A Boltzmann machine is also a recurrent neural network, and is similar to a Hopfield network with the difference that its units are stochastic, thus offering an alternative interpretation of the dynamics of network models (see Dayan & Abbott [2001] for a formal description of the various network models). Of relevance in this work are spiking neural networks, which we use in the works presented here, where the timing of the occurrence of the events plays a critical role in the computations.

2.2 Learning

Learning is the capacity to take into account information in order to create new, or to modify existing, constructions about the world. It has been observed from the simplest organism to the most evolved ones (Squire, 1987). The ability to
remember causal relationship between events, and to be able to predict and to anticipate an effect, procures an obvious advantage in all species (Bi & Poo, 2001). Learning is usually gradual and is shaped by previous knowledge. It can be of several types and can occur through different processes. We will detail the most relevant forms of learning in this section, as well as models developed to account for them. Different structures of the brain have been associated with these learning methods. Cerebellum is supposed to take advantage of supervised learning, whereas the cortex would mostly use unsupervised learning, and reinforcement learning would be the main process occurring in the BG (Doya, 1999, 2000a).

2.2.1 Supervised Learning

In supervised learning, the algorithm is given a set of training examples along with the correct result for each of them. The goal is for the classifier to, at least, learn and remember the correct answers even when the supervisor has been removed, but more importantly, to be able to generalise to unseen data. It can be thought of as a teacher and a student, first the teacher guides the student through some example exercises all the way to the solution. Then the student has to get the correct answer without the help of the teacher. Finally, the goal is for the student to find solutions to unseen problems. The efficacy of this method relies obviously on the quality of the training samples and on the presence of noise in the testing set, notably for neural networks. Multilayer Perceptrons use supervised learning through a backpropagation of the error.

2.2.2 Reinforcement Learning

In this type of learning, which can be seen as intermediate between supervised and unsupervised, the algorithm is never given the correct response. Instead, it only receives a feedback, i.e. information that it has produced a good or a bad solution. With that knowledge only, it has, through trial and error, to discover the optimal solution, that is, the one that will maximize the positive reward given the environment (Kaelbling et al., 1996). It is assumed that the value of the reward impacts the weights update accordingly. A reinforcement learning algorithm is often encapsulated in an ‘agent’ and the setting the agent finds itself in is called the environment of the agent. The agent can perceive the environment through sensors and has a range of possibilities to act on it. A refinement of this approach is to implement a reward prediction (RP) system, which enables the agent to compute the expected reward in order to compare it to the actual reward. Using this difference between the expected reward value and the actual reward, the reward prediction error (RPE), instead of the raw reward signal improves both the performance in learning, especially when
the reward schedule changes, and the stability of the system in a stationary environment. Indeed, once the reward is fully predicted, $RPE = 0$, there is no need to undergo any modifications.

2.2.3 Unsupervised Learning

Here, it is up to the algorithm to find a valuable output, as only inputs are provided, without any correct example, or additional help or information about what is expected. It thus needs to find any significant structure in the data in order to return useful or relevant information. This obviously assumes that there is some statistical structure in the data. The network self-organizes itself based on its connections, dynamics, plasticity rule and the input. The goal of this method is usually to produce a representation of the input in a reduced dimensionality. As such, unsupervised learning can share common properties with associative learning.

2.2.4 Associative Learning

As we have mentioned, associative learning extracts the statistical properties of the data. There are two forms: classical and operant. However, these two types of conditioning, and especially the operant one, can be viewed as leaning on the reinforcement learning approach as unconditioned stimuli might represent a reward by themselves, or can even be plain reinforcers. Both methods produce a conditioned response (CR), similar to the unconditioned response (UR), to occur at the time of a conditioned stimulus (CS), in anticipation of the unconditioned stimulus (US) to come. The difference is that the delivery of the US does not depend on the actions of the subject in classical conditioning, whereas it does in operant conditioning. The basic idea is that an event becomes a predictor of a subsequent event, resulting from the learning due to the repetition of the pattern.

2.2.4.1 Classical conditioning

In 1927, Pavlov reported his observations on dogs. He noticed they began to salivate abundantly, as when food was given to them, upon hearing the footsteps of the assistant bringing the food, but still unseen to the animals (Pavlov, 1927). He had the idea that some things are not learned, such as dogs salivating when they see food. Thus, an US, here food, always triggers an UR, salivation. In his experiment, Pavlov associated the sound of a bell with the food delivery. By itself, ringing the bell, the CS, is a neutral stimulus and does not elicit any particular response. However, after a few associations, the dogs were salivating at the sound of the bell. This response to the bell is thus called the
CR, a proof that the food is expected even though it has not yet been delivered. Pavlov noted that if the CS was presented too long before the US, then the CR was not acquired. Thus, temporal contiguity is required but is not sufficient. The proportion of the CS-US presentations over those of the CS only is also of importance (Rescorla, 1988), as we will detail it further in section 2.2.4.3. It should also be noted that the conditioning could depend on the discrepancy between the perceived environment and the subject’s *a priori* representation of that state (Rescorla & Wagner, 1972), which should thus be taken into consideration. This form of conditioning became the basis of Behaviorism, a theory already developed by Watson, focusing on the observable processes rather than on the underlying modifications (Watson, 1913).

### 2.2.4.2 Operant conditioning

Skinner extended further the idea to modification of voluntary behavior, integrating reinforcers in the learning into what is known as operant conditioning (Skinner, 1938). He believed that a reinforcement is fundamental to have a specific behavior repeated. The value of the reinforcement drives the subject to associate the positive or negative outcome with specific behaviors. It was first studied by Thorndike, who noticed that behaviors followed by pleasant rewards tend to be repeated whereas those producing obnoxious outcomes are less likely to be repeated (Thorndike (1998), originally published in 1898). This led to the publication of his law of effect.

This type of learning has also been called trial and error, or generate and test, as the association is made between a sensory information and an action. Based on this work, it has been suggested that internally generated goals, when reached, could produce the same effect as an externally delivered positive reinforcments.

### 2.2.4.3 Conditioning phenomena

Conditioning effects depend on timing, reward value, statistics of associations of individual stimuli and on the internal representations of the subject. We here detail some of the most studied methods and phenomena, and most are valid for both types of conditioning, the US being a reward in operant conditioning. It should be noted first that the standard temporal presentations are noted as: delay, the US is given later but still while the CS is present; trace, the occurrence of the CS and of the US does not overlap; and simultaneous, both CS and US are delivered at the same time. Furthermore, multiple trials are usually needed for the learning to be complete. The inter-stimulus interval (ISI) between the onset of the CS and of the US is critical for the conditioning performance.
We have already seen in section 2.2.4.1 how the standard acquisition is made. Let us note that the acquisition is faster and stronger as the intensity of the US increases (Seager et al., 2003). Secondary (n-ary) conditioning occurs when the CR of a CS is used to pair a new CS to that response. Conditioning can occur when the US is delivered at regular intervals, both spatial and temporal. If the CS is given after the US, the CR tends to be inhibitory. Extinction happens when the CS is not followed by the US, up to the point when the CS does not produce the CR. This phenomenon is also remarked when neither the US and the CS are presented for a long time. Recovery is the following re-apparition of the CR. It can be achieved by re-associating the CS with the US, and in that case reacquisition is usually faster than the initial acquisition (Pavlov, 1927), or by reinstatement of the US even without the CS being presented. It can also result from spontaneous recovery. Without either the CS or US being presented since the complete extinction, there is an interval of time where a presentation of the CS will trigger, spontaneously, the CR. The mechanisms of this phenomenon are still unclear (Rescorla, 2004). Blocking is the absence of conditioning of a second stimulus, CS2, when presented simultaneously with a previously conditioned CS1. It has been suggested that the predictive power of a CS1 prevents any association to be made with a CS2. Latent inhibition refers to the slowed down acquisition of the CR when the CS has been previously exposed without the US. It has been suggested to result from a failure in the retrieval of the normal CS-US association, or in its formation (Schmajuk et al., 1996).

2.2.5 Standard models of learning

We will define a method commonly used for training neural networks, the backpropagation algorithm, which we have mentioned in the previous section, along with another widely implemented theory of synaptic plasticity (section 3.2), Hebb’s rule. We will detail more specifically reinforcement learning models in the next section (2.3).

2.2.5.1 Backpropagation

The delta-rule is a learning rule used to modify the weights of the inputs onto a single layer artificial neural network. The goal is to minimise the error of the output of the network through gradient descent. A simplified form of the weight $w_{ij}$ can be formalised as:

$$\Delta w_{ij} = \alpha(t_j - y_i)x_i$$

(2.1)

where $t_j$ is the target output, $y_i$ the actual output, $x_i$ is the $i$th input and $\alpha$ the learning rate.
Backpropagation is a generalisation of the delta-rule for multi-layered feed-forward networks. Its name comes from the backward propagation of the error through the network once the output has been generated. Multiplying this error value with the input activation gives the gradient of the weight used for the update.

2.2.5.2 Hebb’s rule

Donald O. Hebb postulated in his book *The Organization of Behavior* in 1949 (Hebb, 1949) that when a neuron is active jointly with another one to which it is connected to, then the strength of this connection increases (see Brown & Milner (2003) for some perspectives). This became known as Hebb’s rule. Furthermore, he speculated that this dynamic could lead to the formation of neuronal assemblies, reflecting the evolution of the temporal relation between neurons.

Hebb’s rule is at the basis of associative learning and connectionism and is an unsupervised process. For example, in a stimulus response association training, neurons coding for these events will develop strong connections between them. Subsequent activations of those responsive to the stimulus will be sufficient to trigger activity of those coding for the response. He suggested these modifications as basis of a memory formation and storage process. Hebb’s rule has been generalised to consider the case where a failure in the presynaptic neuron to elicit a response in the postsynaptic a neuron would cause a decrease in the connection strength between these two neurons. This would enable to represent the correlation of activity between neurons, and to be one of the basis of computational capabilities of neural networks. Hebb’s theory has been experimentally confirmed many years later (Kelso et al., 1986) (and see Brown et al. (1990) and Bi & Poo (2001) for reviews).

2.3 Models of Reinforcement Learning

We will present the various ways to look at the modeling approach and then detail some of the relevant, basic models of learning (fig. 2.1). Central to most of them is the prediction error, that is, the discrepancy between the estimated outcome and its actual value. There have been several shifts of paradigms over time, form Behaviorism, to Cognitivism and to Constructivism; theories are proposed and predictions can be made. Confronted with experimental evidences, there might be a need for an evolution, for an update of these hypotheses. Which is exactly how learning is modeled today.
2.3.1 Model-based and model-free

Model-based and model-free strategies share the same primary goal, that is to maximize the reward by improving a behavioural policy. Model-free ones only assert and keep track of the associated reward for each state. This is computationally efficient but statistically suboptimal. A model-based approach builds a representation of the world and keeps track of the different states and of the associated actions that lead to them. If one considers the current state as the root node, then the search for the action to be selected is equivalent to a tree search, and is therefore more computationally taxing than a model-free approach, but it also yields better statistics. It is also commonly believed to be related to the kind of representations the brain performs, especially when it comes to high level cognition. However, there are suggestions that both types could be implemented in the brain using parallel circuits (Balleine et al., 2008; Daw et al., 2005; Rangel et al., 2008). The following formalisations are all model-free approaches.

2.3.2 Rescorla-Wagner

The Rescorla-Wagner rule, derived from the delta-rule, describes the change in associative strength between the CS and the US during learning (Rescorla
This conditioning does not rely on the contiguity of the two stimuli but mainly on the prediction of the co-occurrence. The variation of the associative strength $\Delta V_A$ resulting from a single presentation of the US is limited by the sum of associative weights of all the CS present during the trial $V_{all}$.

$$\Delta V_A = \alpha A \beta_1 (\lambda_1 - V_{all})$$  \hspace{1cm} (2.2)

Here $\lambda_1$ is the maximum conditioning strength the US can produce, representing the upper limit of the learning. $\alpha$ and $\beta$ represent the intensity, or salience, of the CS and US, respectively, and are bounded in $[0,1]$. So, in a trial, the difference between $\lambda_1$ and $V_{all}$ is treated as an error and the correction occurs by changing the associative strength $V_A$ accordingly. This model has a great heuristic value and is able to account for some conditioning effects such as blocking. Moreover, its straightforward use of the RPE made it very popular and triggered interests which have thus led to the developments of new models. It however fails to reproduce experimental results of latent inhibition, spontaneous recovery and when a novel stimulus is paired with a conditioned inhibitor.

### 2.3.3 Action value algorithms

In this type of models, the aim of the agent is to learn, for each state, the action maximizing the reward. For every states exist subsequent states that can be reached by means of actions. The value function estimates the sum of the reward to come for a state, or a state-action pairing, instead of learning directly the policy. A policy is a rule that the agent obeys to in selecting the action given the state it is in, and can be of different strictness. A simplistic policy could be to always select the action associated with the highest value, and is called a greedy policy. Other commonly used policies are the $\varepsilon$-greedy and softmax selections. The $\varepsilon$-greedy policy selects the action with the highest value in a proportion of $1 - \varepsilon$ of the time, and a random action is selected in a proportion $\varepsilon$. The softmax selection basically draws the action from a distribution which depends on the action values (see section 2.3.3.3 for a formal description). Markov decision processes are similar in that the selection depends on the current action value and not on the path of states and actions taken to reach the current state, i.e. all relevant information for future computations is self-contained in the state value. $V^\pi(s)$ represents the value of state $s$ under policy $\pi$. Thus, $Q^\pi(s,a)$ is the value of taking action $a$ in state $s$ under policy $\pi$. This is referred to as the action value, or $Q$-value. The initial values are set by the designer, and attributing high initial values, known as ‘optimistic initial conditions’, biases the agent towards exploration. We will present three approaches that can be used to learn these action values.
2.3.3.1 Temporal difference learning

The temporal difference (TD) method brings to classical reinforcement learning models the prediction of future rewards. It offers a conceptual change from the evaluation and modification of associative strength to the computation of predicted reward. In this model, the RPE is represented by the TD error (see O’Doherty et al. (2003) for relations between the brain activity and TD model computations). The TD method computes bootstrapped estimates of the sum of future rewards, which vary as a function of the difference between successive predictions (Sutton, 1988). As the goal is to predict the future reward, an insight of this method is to use the future prediction as training signal for current or past predictions, instead of the reward itself. Learning still occurs at each time steps, where the prediction made at the previous step is updated depending on the prediction at the current step, with the temporal component being the computation of the difference of the estimation of $V^{\pi}(s)$ at time $t$ and $t+1$, defined by:

\[
V^{\pi}_{t+1}(s_t) = V^{\pi}_t(s_t) + \alpha \delta_{t+1}
\]

\[
\delta_{t+1} = R_{t+1} + \gamma V^{\pi}_t(s_{t+1}) - V^{\pi}_t(s_t)
\]

with $\alpha > 0$ and where $\delta$ is also called the effective reinforcement signal, or TD error. A change in the values occurs only if there is some discrepancy between the temporal estimations, otherwise, the values are left untouched.

2.3.3.2 Q-learning

Q-learning learns about the optimal policy regardless of the policy that is being followed and is also a TD, model-free, reinforcement learning algorithm. Q-learning is thus called off-policy, meaning it will learn anyway the optimal policy no matter what the agent does, as long as it explores the field of possibilities. If the environment is stochastic, Q-learning will need to sample each action in every states an infinite number of times to fully average out the noise, but in many cases the optimal policy is learned long before the action values are highly accurate. If the environment is deterministic, it is optimal to set the learning rate equal to 1. It is common to set an exploration-exploitation rule, such as the previously mentioned softmax or $\varepsilon$-greedy, to optimize the balance between exploration and exploitation. The update of the $Q$-values is defined by:
\[ Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \times \delta_{t+1} \]  

where \( R_{t+1} \) is the reward obtained after having performed action \( a_t \) from state \( s_t \) and with the learning rate \( 0 < \alpha_t(s_t, a_t) \leq 1 \) (it can be the same for all the pairs). The learning rate determines how fast the new information impacts the stored values. The discount factor \( \gamma \) sets the importance given to future rewards, where a value close to 1 makes the agent strive for long-term large reward, while 0 blinds it to the future and makes it consider only the current reward.

A common variation of \( Q \)-learning is the SARSA algorithm, standing for State-Action-Reward-State-Action. The difference lies in the way the future reward is found, and thus in the value used for the update. In \( Q \)-learning, it is the value of the most rewarding possible action in the new state, whereas in SARSA, it is the actual value of the action taken in the new state. Therefore, SARSA integrates the policy by which the agent selects its actions. Its mathematical formulation is thus the same as equation 2.5 but its error signal is defined as:

\[ \delta_{t+1} = R_{t+1} + \gamma \max_a Q_{t+1}(s_{t+1}, a) - Q_t(s_t, a_t) \]  

2.3.3.3 Actor-Critic

In the previous methods, the policy might depend on the value function. Actor-critic approaches, derived from TD learning, explicitly feature two separate representations, schematised in fig. 2.2, of the policy, the actor, and of the value function, the critic [Barto et al., 1983; Sutton & Barto, 1998]. Typically, the critic is a state-value function. Once the actor has selected an action, the critic evaluates the new state and determines if things are better or worse than expected. This evaluation is the TD error, as defined in equation 2.4. If the TD error is positive, it means that more reward than expected was obtained, and the action selected in the corresponding state should see its probability to be selected again, given the same state, increased. A common method for this is by using a softmax method in the selection:

\[ \pi_t(s, a) = Pr(a_t = a \mid s_t = s) = \frac{e^{p(s, a)/i}}{\sum_b e^{p(s, b)/i}} \]  

29
Figure 2.2: Representation of the Actor-Critic architecture. The environment informs both the Critic and the Actor about the current state. The Actor selects an action based on its policy whereas the Critic emits an estimation of the expected return, based on the value function of the current state. The Critic is informed about the outcome of the action performed on the environment, that, it knows the actual reward obtained. It computes the difference with its estimation, the TD error. This signal is then used by both the Actor and the Critic to update the relevant values. Figure adapted from Sutton & Barto (1998).

where \( t \) is the temperature parameter. For low value, the probability of selecting the action with the highest value tends to 1, where the probabilities of the different actions tends to be equal for high temperature. \( p(s,a) \) is the preference value at time \( t \) of an action \( a \) when in state \( s \). The increase, or decrease, mentioned above, can thus be implemented by updating \( p(s_t,a_t) \) with the error signal, but other ways have been described. Both the critic and the actor get updated based on the error signal, the critic adjusts its prediction to the obtained reward. This method has been widely used in models of the brain and of the BG especially, and is commonly integrated in artificial neural networks (Cohen & Frank, 2009; Joel et al., 2002).

2.3.3.4 Eligibility trace

A common enhancement of TD learning models is the implementation of an eligibility trace, which can increase their performance, mostly by speeding up learning. One way to look at it, is to consider it as a temporary trace, or record, of the occurrence of an event, such as being in a state or the selected action (see fig. 2.3 for a visual representation). It marks the event as eligible for learning, and restricts the update to those states and actions, once the error signal is delivered (Sutton & Barto, 1998). These types of algorithm are denoted TD(\( \lambda \)),
because of their use of the $\lambda$ parameter determining how fast the traces decay. It does not limit the update to the previous prediction, but to the recent earlier ones as well, acting similarly to a short-term memory. Therefore, it speeds up learning by updating the values of all the event which traces overlap with the reward, effectively acting as an $n$-steps prediction, instead of updating only the 1-step previous value. The traces decay during this interval and learning is slower as the interval increases, until the overlap between the eligible values and the reward disappears, preventing any learning. This is similar to the conditioning phenomena described in section 2.2.4.3 where temporal contiguity is required. This implementation is a primary mechanism of the temporal credit assignment in TD learning. This relates to the situation when the reward is delivered only at the end of a sequence of actions. How to determine which action(s) has(ve) been relevant to attain the reward? From a theoretical point of view, eligibility traces bridge the gap to Monte Carlo methods, where the values for each state or each state-action pair is updated only based on the final reward, and not on the estimated or actual values of the neighbouring states. The eligibility trace for state $s$ at time $t$ is noted $e_t(s)$ and its evolution at each step can be described as:

$$e_t(s) = \begin{cases} \gamma \lambda e_{t-1}(s) & \text{if } s \neq s_t; \\ \gamma \lambda e_{t-1}(s) + 1 & \text{if } s = s_t, \end{cases}$$

for all non-terminal states $s$. This example refers to the accumulating trace, but a replacing trace has also been suggested. $\lambda$ reflects the forward, theoretical view, of the model, where it exponentially discounts past gradients. Implementation of the eligibility trace to the TD model detailed in equation 2.3 gives:

$$V^\pi(s_{t+1}) = V^\pi(s_t) + \alpha \delta_{t+1} e_{t+1}(s_t)$$  \hspace{1cm} (2.9)

We will now see in the next chapter the mechanisms of learning in biology, and the structures involved. We will try to draw the links between the ideas described here and the biological data before detailing the methods used in the presented works.
Figure 2.3: Eligibility trace. Punctual events, bottom line (dashed bars), see their associated eligibility traces increase with each occurrence, top line. However, a decay, here exponential, drives them back with time to a value of zero (dashed horizontal line).
3. Biological background

3.1 Neuroscience overview

It is estimated that the human skull packs $10^{11}$ neurons, and that each of them makes an average of $10^4$ synapses. Additionally, glial cells are found in an even larger number as they provide support and functional modulation to neurons. There are indications that these cells might also play a role in information processing (Barres, 2008). Moreover, there is also a rich system of blood vessels and ventricles irrigating the brain. The human brain consumes 20-25% of the total energy used by the whole body, for approximately 2% of the total mass.

We will first introduce neuronal properties as well as the organisation of the brain into circuits and regions and their functional roles. We will also describe how learning occurs in such systems. We will then focus on the BG by detailing some experimental data relevant to this work. Detailed sub-cellar mechanisms are not in the focus of this work and we will cover only basics of neuronal dynamics, as our interest resides mostly in population level computations (see Kandel et al. (2000) for a general overview of neuroscience). Therefore, we will not only present biological data, but will also include the associated hypotheses and theories about the functional implications of these experimental observations; hypotheses which might have been produced in computational neuroscience works.

3.1.1 Neuronal and synaptic properties

The general idea driving the study of the brain, of its structure, connections, and dynamics, is that information is represented in, and processed by, networks of neurons. The brain holds a massively parallelized architecture, with processes interoperating at various levels. As the elementary pieces of these networks, neurons are cells that communicate with each other via electrical and chemical signals. They receive information from other neurons via receptors, and can themselves send a signal to different neurons through axon terminals. Basically, the dendritic tree makes the receptive part of a neuron. Information converge onto the soma and an electric impulse, an action potential, might be triggered and sent through the axon, which can be several decimeters long.
The synapses at the axon terminal are the true outputs of neurons. Synapses can release, and thus propagate to other nearby neurons, the neural message, via the release of synaptic vesicles containing neurotransmitters. This is true for most connections, but electrical signals can also be directly transmitted to other neurons via gap junctions. The vast majority of the neurons are believed to always send the same signal, that is, to deliver a unique neurotransmitter. It must be noted that a few occurrences of what has been called bilingual neurons have been reported, where these neurons are able to release both glutamate and GABA (Root et al., 2014; Uchida, 2014). These two neurotransmitters are the most commonly found in the mammalian brain, and are, respectively, of excitatory and inhibitory types. However, a neuron can receive different neurotransmitters, from different neurons. A neurotransmitter has to find a compatible receptor on the postsynaptic neuron in order to trigger a response, which is a cascade of subcellular events. In turn, these events could affect the generation of action potentials of the postsynaptic neuron but could also lead to change in the synaptic efficacy and have other neuromodulatory effects. Synaptic plasticity is a critical process in learning and memory. It represents the ability for synapses to strengthen or weaken their relative transmission efficacy, and is usually dependent on the activity of the pre- and postsynaptic neurons, but can also be modulated by neuromodulators (section 3.2).

A common abstraction is to consider that neurons can be either active, i.e. sending signals, or inactive, i.e. silent. The signal can be considered as unidirectional, and this allows to specify the presynaptic neuron as the sender, and the postsynaptic neuron as the receiver. The activity is defined as the number of times a neurons sends an action potential per second and can range from very low firing rates, e.g. 0.1 Hz, to over 100 Hz. Essentially, a neuron emits an action potential, also called a spike, if the difference in electric potential between its interior and the exterior surpasses a certain threshold. The changes in this membrane potential are triggered by the inputs the neuron receives: excitatory, inhibitory and modulatory. When the threshold is passed, a spike will rush through the axon to deliver information to the postsynaptic neurons.

Neurons exhibit a great diversity of types and of firing patterns, from single spike to bursts with down and up states. They also display a wide range of morphological differences. As we have mentioned, the signal they send can be of excitatory or inhibitory nature, increasing and decreasing, respectively, the likelihood of the postsynaptic neuron to fire. But this signal can also modulate the effect that other neurotransmitters have, via the release of neuromodulators such as dopamine, and this can even produce long term effects. Returning to the fact that neurons typically receive different types of connections from different other types of neurons, along with the interdependency of the neurotransmitters, of which there are more than 100, the resulting dynamics can be
extremely complex.

The shape and the activation property of a neuron also play a critical role in its functionality. Projection neurons usually display a small dendritic tree and a comparably long single axon, and a similar polarity is usually observed among neighbours. However, interneurons commonly exhibit dense connectivity on both inputs and outputs, in a relatively restricted volume. The activation function represents the relationship between a neuron’s internal state and its spike activity.

Neurons can have different response properties. For example it has been shown that the firing rates of neurons in the hippocampus depend on the spatial localisation of the animal, with only a restricted number of neurons showing a high activity for a specific area of the field (Moser et al., 2008; O’Keefe & Dostrovsky, 1971) or to the orientation, or speed, of a visual stimulus in the primary visual cortex (Priebe & Ferster, 2012). We will detail a computational model of neurons in section 4.1.2.

Even if there is a lot of support to the doctrine considering the neuron as the structural and functional unit of the nervous system, there is now the belief that ensemble of neurons, rather than individual ones, can form physiological units, which can produce emergent functional properties (Yuste, 2015). Non-linear selection, known as ‘winner-takes-all’ (WTA), through competition of populations sharing interneurons can be achieved. Inhibitory interneurons, which can be involved via feed-forward, feed-back, or lateral connections, can also help to reduce firing activity and to filter the signal, notably for temporal precision.

Noise, as a random intrinsic electrical fluctuations within neural networks, is now considered to have a purpose and to serve computations instead of being a simple byproduct of neuronal activity, e.g. stochastic molecular processes (Allen & Stevens, 1994), thermal noise, ion channel noise (see Faisal et al. (2008) and McDonnell et al. (2011) for reviews). It can, for example, affect perceptual threshold or add variability in the responses (Soltoggio & Stanley, 2012).

3.1.2 Neural coding

Neural coding deals with the representation of information and its transmission through networks, in terms of neural activity. In sensory layers, it mostly has to do with the way a change of the environment, e.g. a new light source or a sound, impacts the activity of the neural receptors, and how this information is transmitted to downstream processing stages. It assumes that changes in the spike activity of neurons should be correlated with modifications of the stimulus. The complementary mechanism is neural decoding, which aims to find what is going on in the world from neuronal spiking patterns.
The focus of many studies has been on the spike activity of neurons, spike trains, as the underlying code of the representations. The neural response is defined by the probability distribution of spike times as a function of the stimulus. Furthermore, at the single neuron level, the spike generation might be independent of all the other spikes in the train. However, correlations between spike times are also believed to convey information. Similarly, at the population level, the independent neuron hypothesis makes the analysis of population coding easier than if correlations between neurons have some informative value. It requires to determine if the correlations between neurons bring any additional information about a stimulus compared to what individual firing patterns already provide. Two possible strategies have been described regarding the coding of the information by neurons. The first one considers the spike code as time series of all-or-none events, such that only the mean spiking rate carries meaningful information, e.g. the firing rate of a neuron would increase with the intensity of the stimulus. The second one is based on the temporal profile of the spike trains, where the precise spike timing is believed to carry information (see (Dayan & Abbott, 2001) for detailed information on neural coding).

Neurons are believed to encode both analog- and digital-like outputs, such as for example, speed and a direction-switch, respectively (Li et al., 2014b). Furthermore, synaptic diversity has been shown to allow for temporal coding of correlated multisensory inputs by a single neuron (Chabrol et al., 2015). This is believed to improve the sensory representation and to facilitate pattern separation. It remains to determine what temporal resolution is required, as precise timing could rely on the number of spikes within some time window, thereby bridging the gap between the two coding processes. Additionally, evidences of first-spike times coding have been reported in the human somatosensory system and could account for quick behavioral responses believed to be too rapid to be relying on firing rates estimate over time (VanRullen et al., 2005).

The quality of a code is valued on its robustness against error and on the resource requirements for its transmission and computation (Mink et al., 1981). In terms of transmission of information, two possible mechanisms to propagate spiking activity in neuronal networks have been described: asynchronous high activity and synchronous low activity, producing oscillations (Hahn et al., 2014; Kumar et al., 2010).

In summary, the various coding schemes here detailed could be all present in the brain, from the early stage sensory layers to higher associative levels.
3.1.3 Local and distributed representations

Information can be represented by a specific and exclusive unit, be it a population or even a neuron, coding for a single concept, and would thus be called local. The grandmother cell representation is an example of this, where studies have shown that different neurons responded exclusively to single, specific, stimuli or concepts, such as one’s grandmother or (and) famous people (Quiroga, 2012; Quiroga et al., 2005). This implies that memory storage is there limited by the number of available units, and that without redundancy, a loss of the cell coding for the object would result in a specific loss of its representation. The pending alternative is that neurons are part of a distributed representation, a cell assembly, and are therefore involved in the coding of multiple representations. According to recent findings, the brain uses sparse distributed representations to process information, which makes it resilient to single neuron failure and allows for pattern-completion (see Bowers (2009) for a review).

3.1.4 Organization and functions of the mammalian brain

We here present a general view of relevant systems, and will focus more on the BG in further sections. Progress and discoveries in this domain are tied with pathophysiology and other investigations of cerebral injuries, linking functions with biological substrate. For example, patients with damages to the left inferior frontal gyrus may suffer from aphasia. Another example is the patient HM, who suffered a bilateral removal of his hippocampi and some of his medial temporal lobes, and had extremely severe anterograde amnesia (Squire, 2009).

The mammalian brain is made of several structures, spatially localised and functionally different. It features two hemispheres, each one in charge of the sensory inputs and motor outputs of the contralateral hemibody. They are densely interconnected and the congregation made by these axons crossing sides forms a tract called corpus callosum. Furthermore, the largest constituent of the human brain is the cortex, but a number of subcortical structures are also critical to all levels of cognition and behavior, such as: amygdala, thalamus, hypothalamus, cerebellum and BG.

The neocortex is formed of several lobes, anatomically delimited by large sulci. This gross classification also carries some functional relevance, with the occipital lobe mostly involved in visual processing, the parietal lobe in sensorimotor computations, the temporal lobe in memory and in the association of inputs from the other lobes and of language, and the frontal lobe in advanced cognitive processes, e.g. attention, working memory, planning, and reward-related functions. The neocortex is not critical for survival, and is in-
deed absent in non-mammals. It has phylogenetically evolved from the pallium, a three-layered structure present in birds and reptiles. Its size has seen a relatively recent increase to make up to two thirds of the human brain, thus stressing the evolutionary advantage it provided. The cortex features an horizontal layered structure, with each of these six layers having a distribution of neuronal cell types and connections. This structure also presents lateral, or recurrent, connectivity, which has been shown to support bi-stability and oscillatory activity (Shu et al. 2003).

A modular structure has additionally been described. Cortical columns, hypercolumns and minicolumns consist of tightly inter-connected neurons, and had traditionally been more seen as functional units than anatomical columns, even though this architecture is found throughout cortex, supporting a non-random connectivity (Perin et al. 2011; Song et al. 2005; Yoshimura et al. 2005). Furthermore, minicolumns have been shown to share similar inputs and to be highly co-active (Mountcastle 1997; Yoshimura et al. 2005). They typically consist of a few tens of interconnected neurons, coding for a similar input or attribute. Hypercolumns comprise several minicolumns sharing a common feedback inhibition. According to a popular theory, they cause a competition among these minicolumns in order to force only one of them to be active at a time. However, there are also abundant long range interactions between minicolumn across different hypercolumns. Columns are embedded within distributed networks (see Rockland 2010 for some details on macrocolumns).

Returning to the fact that specific brain structures can be critically involved in certain functions, let us mention that hippocampus is critical in both short- and long-term memory and spatial navigation, where place cells and spatial maps have been described (see Best et al. 2001; Moser et al. 2008 for a review). Also, amygdala is central in emotions and fear (Cardinal et al. 2002) while hypothalamus regulates homeostasis and overall mood. Cerebellum plays an important role in motor control, especially in timing, fine tuning of movements and automatic or unconscious executions and is believed to benefit from supervised learning (Doya 1999 2000a). Thalamus serves as a hub dispatching sensory and motor signals to and from cortex, as well as controlling subcortico-cortical communications. Finally, the BG are connected to all the aforementioned structures and are believed to be essential in action selection and reinforcement learning. We will detail the anatomical and functional properties of BG in section 3.3.
3.1.5 Experimental techniques

Experimental data do not yet provide exhaustive information about the brain. However, there is already an immense amount of data about specific components and at various temporal and spatial scales. Apart from studying dead tissues, recording the electrical activity of neurons or of populations of neurons is the most accessible signal to obtain, be it through invasive recording electrodes or with non-invasive electroencephalogram (EEG) sensors. Depending on the object of interest, experimentalists can choose from various methods to investigate and decode the neural activity. Commonly, these techniques can be classified along a two dimensional representation: temporal and spatial resolution of the recordings.

In order to gain knowledge on a function of a neural object, for example neurotransmitters, neuronal type or population, or region, it is useful to actively modify the dynamic of the system, in order to observe the response to the controlled perturbation. Stimulation electrodes can, through the delivery of electric current, trigger neurons in the vicinity to fire. Pharmacology can be used to target neurons’ receptors, channels, transporters and enzymes and is thus mostly of interest in the study of neurotransmitters actions, which can impact from local cells to large areas (Raiteri, 2006). Similarly, the studies of brain lesions also provides knowledge about the underlying functional organisation. Studies now usually involve advanced technology where both the anatomy and some dynamics can be recorded. We will here mention some of the most common ones, which are relevant for the next sections.

Electron microscopy and calcium imaging capabilities have improved with two-photon techniques and confocal image analysis. For larger scales, light microscopy and patch-clamp enable to record and to dynamically affect a neuron with electrical stimuli. At the population level, field potentials and multi-electrode arrays can be used. A recent technique has gained a lot of fame, as it enables to selectively activate or inhibit populations of neurons through the genetic modification of their membranes in order to feature light sensitive ion channels, so called optogenetics. The activity of the neurons can thereby be controlled with flashes of light, contrary to electrodes which indiscriminately affect all the cells in the nearby volume, and both can be performed in vivo (see Lenz & Lobo (2013) for a review of studies investigating the BG mechanisms with optogenetics, and Gerfen et al. (2013) for a characterisation of transgenic mice used for the study of neuro-anatomical pathways between the BG and the cerebral cortex). Studies of the connectome, that is of the physical wiring of the nervous system (Chuhma et al. [2011] Oh et al. [2014]), developmental biology (Rakic, 2002) and techniques such as CLARITY, enabling to see through tissues (Lerner et al., 2015), all have the potential to bring extremely
valuable knowledge. At the systems and brain area level, functional magnetic resonance imaging (fMRI) probably is the most common but its temporal resolution is in the order of a second. Similarly, magnetoencephalography (MEG), can map brain activity. Finally, behavioural assessments give indications on the functional impact of controlled modifications to the subject. Relatively defined areas of the brain can be non-invasively activated or inhibited via transcranial magnetic stimulation. A common experimental setup in humans is still EEG acquisition, capturing the electrical signals from the brain from electrodes placed on the skin of the head of patients. All these techniques are used mostly to record the activity in response to a perturbation, and each can tell us a part of the picture, as they all work at different temporal and spatial scales, and can be very invasive. A large part of the data mentioned in this work come from studies using one or more of: electron microscopy imaging, patch clamp and electrode recordings, transgenic and optogenetics techniques, behavioural works and pharmacology.

3.2 Plasticity

Neurons in the living brain can adapt in response to experience. It is believed to be not only critical during development but also for adaptation and learning during adulthood (Chalupa et al., 2011). It was suggested at the end of the 19th century that memories might be stored as an anatomical changes in the strength and numbers of neuronal connections (Ramón Y Cajal, 1894). Plasticity corresponds to all the mechanisms that can lead to adaptation, on a time scale ranging from milliseconds to years (see Citri & Malenka (2008) and Tetzlaff et al. (2012) for reviews). Synaptic plasticity refers to the different aspect of use-dependent synaptic modifications (Kreutz & Sala, 2012). These modifications change the functional properties of the synapse. Synaptic plasticity is commonly considered as the major process, but other intrinsic modifications can also alter the integration of postsynaptic potentials. Disturbances in plasticity result in diseases affecting cognitives functions and behaviours, such as in addiction, compulsive behaviours, learning and memory troubles e.g. in Alzheimer’s disease, or even psychiatric disorders (Selkoe, 2002) (see section 3.4 and Berretta et al. (2008) for some overview of diseases related to dysfunctional synaptic plasticity).

Learning in biological systems occurs through the modification of the efficacy of the synaptic transmission of a signal between the pre- and postsynaptic neurons. We have already mentioned Hebb’s rule which bases the changes on the correlation and causality of the activity of the neurons (section 2.2.5.2).

There are data suggesting that a backpropagation of spikes occurs, mostly in the dendritic part of the neuron. This is supposed to carry the information
about the activity of the postsynaptic neuron back to the presynaptic one, in order to enable adaptation and changes to happen at the presynaptic part of the synapse.

We will here present a brief overview of the mechanisms of plasticity. A more specific description of the synaptic plasticity in the BG with a focus on dopamine will be done in section 3.3.4 (see Gerfen & Surmeier (2011); Tritsch & Sabatini (2012) for a review).

3.2.1 Short-term plasticity

Short-term plasticity, usually referred to as short-term adaptation describes the mechanisms with which the response of a postsynaptic neuron evolves, mostly in a monotonic fashion, depending on the activity of presynaptic neurons. Critically, this adaptation does not last in time, and the neuron exhibits a generic response again after some time, on a timescale of some milliseconds to a few minutes. The underlying mechanisms are believed to be synaptic processes, such as depletion or augmentation of releasable vesicles of neurotransmitters, and the resulting effects can be an increased or decreased synaptic transmission efficacy. Functionally, it is supposed to enable, for instance, the encoding of a signal with a larger dynamic range (Pfister et al., 2010) and locking to the stimulus (Brody & Korngreen, 2013). In the BG, both short-term depression and short-term facilitation have been suggested to play an important role in the activity of the network (Lindahl et al., 2013). Additionally, dopamine has been shown to be involved in the modulation of short-term synaptic plasticity at striatal inhibitory synapses (Tecuapetla et al., 2007).

3.2.2 Long-term plasticity

As opposed to short-term adaptation, the timescale of the effects of long-term plasticity ranges from minutes to years. This is how more permanent memories are supposed to be formed, by synaptic modifications that can last for an extended time, but which result from a few seconds of neuronal activity (Nabavi et al., 2014). These modifications can, as for the short-term ones, lead to a strengthening or a weakening of the synapses, long-term potentiation (LTP) and long-term depression (LTD), respectively. Evidence of experience-dependent plasticity have been detailed, where boutons and dendritic spines appear and disappear, along with synapse formation and elimination (see Holtmaat & Svoboda (2009) for a review). Hebbian and anti-Hebbian plasticity have notably been reported at cortico-striatal synapses, and further seems to be dopamine-dependent (Fino et al., 2005, 2008; Pawlak & Kerr, 2008; Reynolds & Wickens, 2002; Shen et al., 2008). Recently, GABA has also been shown
to be able to reverse the polarity of synaptic plasticity from Hebbian to anti-Hebbian learning, and thereby LTP to LTD, and vice versa (Paille et al., 2013).

3.2.3 Homeostasis

Homeostasis is the property of keeping a constant internal balance independently of external changes. It could be obtained through modification of intrinsic excitability, synaptic potentiation, depression, depletion or augmentation (Song et al., 2000). Such mechanisms would prevent the neuron, and the network (Maffei & Fontanini, 2009), to fall outside their dynamic range (see Turrigiano, 1999) for a review), which could otherwise result from, for example, the positive feedback loop associated with Hebbian learning. These long term changes in synaptic strength require coordination with multiple synapses, and possibly with other neuronal properties, to maintain the stability and functionality of neural circuits (Burrone & Murthy, 2003; Watt & Desai, 2010; Zenke et al., 2013). Homeostatic mechanisms have notably been reported in neurons in the striatum (Azdad et al., 2009; Fino et al., 2005) (see section 3.3.4 for more details).

3.2.4 STDP

The temporal information of a spike event has been suggested to be the mechanism underlying the causality relation, implying an asymmetrical form of Hebbian learning. Basically, two patterns can occur: either a presynaptic spike happens consistently before the postsynaptic one, thereby leading to LTP of the involved synapses, or conversely, it occurs after the postsynaptic spike, triggering LTD. Such learning rule is called spike timing dependent plasticity (STDP). This type of learning has been experimentally observed, and the temporal resolution of the pre- and postsynaptic spike pairing was shown to be in the range of a few milliseconds, and to vary across synapses and brain regions (Bell et al., 1997; Bi & Poo, 1998; Dan & Poo, 2004; Markram et al., 1997; Sjöström et al., 2001). This precision allows the use of the learning rule in temporal coding paradigms (spike timing patterns). STDP processes association between spikes in an unsupervised and temporally asymmetric Hebbian way (for a review see Caporale & Dan, 2008 and Feldman, 2012). Repeated activation of the presynaptic neuron followed by the firing of the postsynaptic one within a short temporal scale, milliseconds, leads to LTP of the synapse, whereas the contrary triggers LTD. However, other factors were shown to be also involved in the modulation of the synaptic plasticity, such as the rate (Sjöström et al., 2001), the dendritic location (Froemke et al., 2005), the cell and receptor types (Shen et al., 2008; Surmeier et al., 2007) and the presence of neuromodulators (Pawlak & Kerr, 2008; Reynolds & Wickens, 2002).
Bidirectional long-term plasticity has, for example, been observed on both types of striatal interneurons, albeit with a cell-specificity (Fino et al., 2008).

### 3.3 Basal ganglia

**Contraria sunt complementa.**
- Niels Bohr

The BG, which comprise several subcortical (forebrain) structures and pathways, are found in all vertebrates and are evolutionarily well preserved through this family (Pombal et al., 1997; Reiner et al., 1998; Stephenson-Jones et al., 2011). We will here outline the structural and functional aspects of the BG. We will start with the striatum, which is the entry nucleus of the BG and we will then continue with the description of the downstream pathways and nuclei. We will then focus on the dopaminergic nuclei and the feedback they provide to the BG. We will then depict the major role that dopamine plays in the BG and in learning. Even if we do not go into great details, we will mention, in the following section, more than what is required to grasp the computational models of this thesis in order to provide bases for developments.

The BG are commonly considered to be critical in action selection (DeLong, 1990; Graybiel, 1995, 2005; Mink, 1996; Redgrave et al., 1999). Specifically, it has been suggested that they could have evolved as a centralised selection device, specialised to resolve conflicts over access to limited motor and cognitive resources (Redgrave et al., 1999). As this supposes, they might not be restricted to motor command, but as most of the experimental studies predominantly consider only motor responses, we will from here on primarily refer to the motor control aspects. Dysfunctions in the BG can cause a number of symptoms, and are involved in several diseases such as Parkinson’s and Huntington’s diseases, but also obsessive-compulsive disorders (see section 3.4 and Kreitzer & Malenka, 2008 for an overview of the BG functions and dysfunctions).

The classical view of the BG is that two parallel pathways, here noted D1 or direct and D2 or indirect, exert opposing influences on, promoting or inhibiting respectively, motor outputs (Albin et al., 1989; Alexander & Crutcher, 1990; DeLong, 1990; Freeze et al., 2013; Nambu, 2008) (fig. 3.1). The striatum is believed to code action values (Kim et al., 2009; Lau & Glimcher, 2008; Samejima et al., 2005), and to thereby act as the interface between the state related information it gets as inputs and the motor command that should be selected upon this information. Critically, the cortico-striatal synapses are plastic. They are the substrate of the learning of the action selection and enable the
BG to adapt to changes in the environment and in the reward obtention.

A functional topology has been described in the BG: projections from the direct and indirect pathways, originating from populations of neurons coding for a similar action, converge to populations in the internal globus pallidus (GPi) and in the substantia nigra pars reticula (SNr) coding for that same action \cite{Alexander et al. 1986}. The GPi/SNr are the output of the BG and they gate the motor output through their connections to the thalamus, cortex, and brainstem. Unlike the cortex, which features mostly excitatory glutamatergic projection neurons, the BG rely mainly on inhibitory GABAergic projection neurons. An additional pathway from the striatum to the dopaminergic nuclei is supposed to be involved in the prediction of the reward. The feedback information from these dopaminergic nuclei to the striatum and cortex is believed to be an important factor in the synaptic plasticity.

In order to offer some self-consistency within the following sections, we felt that some overlap in the information concerning the connectivity was unavoidable, as outputs from a nucleus could be the inputs of another one.

### 3.3.1 Striatum

The striatum is the entry structure of the BG and can be divided in the caudate nucleus and the putamen, forming the dorsal part, and the nucleus accumbens as the ventral part. The striatum receives glutamatergic inputs from the cortex, the thalamus and the limbic system \cite{Bolam et al. 2000, Doig et al. 2010, Gerfen 1992, McGeorge & Faull 1989, Parent 1990, Smith et al. 2011}, dopaminergic inputs from neurons in the substantia nigra pars compacta (SNc) and in the ventral tegmental area (VTA) \cite{Ilango et al. 2014, Joel & Weiner 2000}, along with serotonergic and noradrenergic inputs from the raphe and pedonculopontine nuclei \cite{Wall et al. 2013}.

The striatum is mostly composed (95%) of GABAergic medium spiny neurons (MSNs) along with different types of inhibitory interneurons \cite{Chuhma et al. 2011, Kemp & Powell 1971}. MSNs are believed to express either D1 or D2 dopamine receptors, equally distributed throughout the striatum \cite{Gerfen 1992}. A strong lateral to medial gradient of D2 receptors has however been noticed in rodents \cite{Joyce et al. 1986}. Inhibitory connections from these MSNs target either the output nuclei of the BG: the GPi and the SNr, making the direct pathway (also called D1 pathway), or the external globus pallidus (GPe) which in turns sends inhibitory connections to the subthalamic nucleus (STN), itself projecting excitatory connections to the same output nuclei as the direct pathway, the GPi/SNr, making the indirect pathway (D2 pathway) \cite{Chuhma et al. 2011, Gerfen et al. 1990, Parent & Hazrati 1995, Surmeier et al. 1997}. However, MSNs expressing both receptors have been reported
Figure 3.1: Simplified diagram of the principal connections of the basal ganglia. They consist of the striatum, GPe, GPi, SNr, SNC and STN. Prominent projections are represented, while some of the connections described in the text are missing, for clarity.

PFC = prefrontal cortex, D1 / D2 = D1 / D2 medium spiny neurons, GPe = external globus pallidus, GPi = internal globus pallidus, STN = subthalamic nucleus, SNr = substantia nigra pars reticula, SNC = substantia nigra pars compacta, LH = lateral habenula, VTA = ventral tegmental area, RMTn = rostromedial tegmental nucleus.
(Nicola et al., 2000; Surmeier et al., 1992) and a study has suggested that virtually all striatal projection neurons could contain dopamine receptors of both classes (Aizman et al., 2000). Still, it was discussed that the contradictory evidence from other studies (Gertler et al., 2008; Jaber et al., 1996; Surmeier et al., 1997) could be reconciled with this result by assuming that MSNs of the direct pathway contain a low level of D2 receptors compared to the D1 one, and vice versa for those of the indirect pathway (Aizman et al., 2000).

A significant feature of MSNs is that their response has been shown to code for actions. Specific optogenetic stimulations of D1 and D2 MSNs, in vivo, lead to motor initiation and suppression respectively (Freeze et al., 2013; Kravitz et al., 2010, 2012; Lenz & Lobo, 2013; Tai et al., 2012). Similar observations have been made through genetic deletion of key protein expression in either D1 or D2 MSNs (Bateup et al., 2010; Durieux et al., 2009). Ablation of D2 neurons in the ventral striatum of mice does not lead to enhanced motor activity but increases the preference to conditioned stimuli, thereby suggesting a role for D2 MSNs in the reinforcement process (Durieux et al., 2009). Stimulation of D2 MSNs does not seem to evoke a motor response, but instead suppresses or decreases motor activity (Kravitz et al., 2010, 2012; Sippy et al., 2015; Tai et al., 2012), thereby supporting the idea that the resulting global or possibly focal inhibition of the GPi/SNr is not enough to trigger a motor response.

Further distinctions of the MSNs can be made, based on their embryology and morphology, their gene expression and furthermore on the topography of their projections (Johnston et al., 1990; Joyce et al., 1986; Nakamura et al., 2009). Striosomal compartments, ‘islands’ embryologically older and also termed patches in rodents, can notably be distinguished from the matrix ‘sea’ by their differential expression of most neurotransmitter-related molecules in the striatum (Crittenden & Graybiel, 2011; Graybiel, 1990; Johnston et al., 1990; Nakamura et al., 2009; Newman et al., 2015). The ratio of 15% patch to 85% matrix area is relatively constant in rats, rhesus monkeys and humans, despite a 19-fold increase in the total striatal area from rat to human (Johnston et al., 1990).

Neurons in the striatum do not form a layered or columnar structure, unlike in the neocortex. However, the observed modular organisation could carry a similar function (Graybiel & Ragsdale, 1978). The presence of different types of inhibitory neurons and the modular organisation described fulfill similar functional requirements to those of the columnar organization (Chuhma et al., 2011; Szydlowski et al., 2013; Taverna et al., 2004, 2008). Dendrites from striosomal neurons span the striosome compartment, whereas those of matrix neurons do not cross the patch/matrix border (Fujiyama et al., 2011). We will now focus on the inputs to the MSNs, with respect to the various, and
mixed, classification we have just described.

3.3.1.1 Inputs

Cortical and thalamic terminals form synapses with both direct and indirect pathway MSNs and, critically, are plastic (Doig et al., 2010) (see 3.3.4 for information on the synaptic plasticity at these synapses). These projections are furthermore topographically organized and target both the patch and matrix compartments (Crittenden & Graybiel, 2011; Gerfen, 1992; Graybiel et al., 1987; Wiesendanger et al., 2004). Striosomes receive inputs from the limbic system, the orbitofrontal and prefrontal cortex (OFC and PFC, respectively), and from the insula (Crittenden & Graybiel, 2011; Eblen & Graybiel, 1995; Friedman et al., 2015; Graybiel, 2008). Additionally, it has been reported that the striosomes might be specifically avoided by sensori-motor projections, which thus seem to only target the matrisomes (Crittenden & Graybiel, 2011; Flaherty & Graybiell, 1993; Fujiyama et al., 2011). The latter thus receive inputs from the motor cortex, somato-sensory areas and the parietal lobe, an organisation which has been reported in rats, cats and monkeys (Flaherty & Graybiell, 1993; Gerfen, 1984; Malach & Graybiel, 1986). Nevertheless, sub-regions of the striatum, containing both striosomes and matrisomes have been reported to receive inputs from related cortical regions (Kincaid & Wilson, 1996). The two compartments can also be differentiated based on the layer of the cerebral cortex from which they receive the most inputs: layer II/III-Va for the matrix, and from deeper layer, Vb and VI, for striosomes (Gerfen, 1989, 1992; Kincaid & Wilson, 1996). MSNs of both the direct and indirect pathway have been shown to respond to bilateral multimodal sensory information (Cui et al., 2013a; Hikosaka et al., 1989). D1 MSNs seems to get stronger inputs from sensory cortex, whereas D2 MSNs receive stronger inputs from motor cortex (Wall et al., 2013). However, D2 MSNs seem to require less input strength than D1 ones to become active. For the minimal input with which the latter would emit their first spike, the D2 MSNs would fire at 15 Hz (Gertler et al., 2008). MSNs seem to have a spontaneous low firing rate in vivo (Berke et al., 2004), and require a strong correlated input to activate (Nisenbaum & Wilson, 1995). Direct pathway MSNs are believed to receive more glutamatergic inputs than those of the indirect pathway (Gerfen & Surmeier, 2011; Gertler et al., 2008). Beside, D1 MSNs in the matrix receive more synaptic terminals from intra-telencephalically projecting cortical neurons than from pyramidal tract projecting neurons, a pattern that is reverted for D2 MSNs (Deng et al., 2015; Reiner et al., 2010). Matrisomes receive three times more thalamic inputs than striosomes, further supporting the differential processing of the inputs through the BG pathways (Fujiyama et al., 2006).
Moreover, a delay of 6 to 7 ms between cortical responses and MSNs ones, and anatomical tracing data, suggest an absence of thalamic engagement in the initial MSNs response (Reig & Silberberg, 2014). Also, the inhibition component in the MSNs response followed the excitation by a few milliseconds, indicating the involvement of intra-striatal interneurons, possibly recruited by the same cortical excitatory input.

3.3.1.2 Ventro-dorsal distinction

An additional organisation of the striatum commonly considered is to distinguish the ventral part from the dorsal one (see Humphries & Prescott (2010) and Balleine et al. (2007) for a review). In relation to the actor-critic architecture, the activity in the ventral striatum of monkeys has been linked with the RP (Schultz et al., 1992), thereby possibly endorsing the role of the critic (van der Meer & Redish, 2011). Furthermore, activity in the medial striatum has been suggested to be task-related, and that it can bias animals to collect rewards (Kimchi & Laubach, 2009). Similarly to striosomes, the ventral striatum has been associated with the processing of inputs from limbic, associative and orbitofrontal areas. MSNs in the nucleus accumbens seem to preferentially target non-dopaminergic neurons in the VTA, some of which project back to the nucleus accumbens (Xia et al., 2011). However, MSNs in the dorsal striatum have been shown to project to GABAergic neurons in the SNr but not to dopaminergic neurons in the SNc (Chuhma et al. 2011). The ventral part of the putamen and the nucleus accumbens (part of the ventral striatal complex) receive input from the limbic cortex (Albin et al., 1989; Alexander & Crutcher, 1990). All these regions are believed to influence the RPE computation in the VTA, through the ventral striatum (O’Doherty et al., 2006). Further stressing the ventro-dorsal distinction, dopaminergic neurons from the SNc have been reported to co-release GABA onto dorsal striatum MSNs of both the direct and indirect pathways (Tritsch et al., 2012) whereas co-release of glutamate was observed onto projections neurons in the nucleus accumbens, part of the ventral striatum (Stuber et al., 2010; Tecuapetla et al., 2010). This classic distinction has been challenged to some extent, Voorn et al. (2004) suggesting a dorsolateral and ventromedial organisation instead, based on behavioural and anatomical data.

In a study where they recorded simultaneously in the ventral and dorsal striatum in rats during a two-armed bandit task, Kim et al. (2009) noted signals related to the upcoming choice only in the latter. Once the action had been initiated, they observed signals conveying the choice, its action-value and its outcome by neurons in both structures. This agrees with the hypothesis that the striatum codes action values and can update these representations based on
the outcome. In a similar study, but this time with recordings in the dorsomedial and dorsolateral striatum, Thorn et al. (2010) found that ensemble activity pattern increased as training progressed and was correlated with performance, whereas the associative (dorsolateral) striatum exhibited an increase of its activity in a first time before fading out as training progressed. Furthermore, it did not correlate with behavioural aspects and the authors suggested that it could be important for performance monitoring and could mean a modulation of the access of dorsolateral sensorimotor loops to the control of action by the dorso-medial loops.

A slight deviation from the actor-critic framework has been discussed by Atallah et al. (2007) where the dorsal striatum would be primordial for performance but not for learning, and would thereby represent the actor, whereas the ventral striatum would be critical for both learning and performance, here embodying the role of a director. However, this ventro-dorsal distinction has lost some interest in the past few years and may need to be revisited, as experimental evidence seemed to blur the demarcation (Nieh et al. 2013), to the benefit of the other distinction: the patch / matrix one. Two studies by Kravitz et al. (2010, 2012) suggested that, through optogenetic activation, neurons in the ventral striatum may have similar functions to those in the dorsal striatum.

3.3.1.3 Reward prediction

In studies using fMRI, the striatum and the OFC exhibited responses related to the TD error in reinforcement learning tasks with various reward probabilities (Berns et al. 2001, 2003). However, as it is believed that the RPE is actually encoded by dopaminergic neurons (section 3.3.4), it may be that the signal in the striatum has more to do with the RP than the actual RPE. Indeed, the similarity in the response of striatal and dopaminergic neurons, with both exhibiting RPE coding in rats (Oyama et al. 2010), could be linked with the idea that the striatum holds the values of both the current and the previous states/actions, such that the TD can be computed (Morita et al. 2012). Besides, no relation was denoted between the activity of these striatal neurons and the licking movements during the conditioned stimulus presentation. This goes to support a prediction specific pathway, decoupled from the actual action selection (Oyama et al. 2010). Only the dopaminergic neurons showed negative RPE coding, i.e. a decrease of activity in response to an omission of an expected reward, where only positive coding RPE was observed for striatal neurons. This indicates a possible dual pathway organisation, similar to the D1 and D2 matrisomal MSNs, in the prediction, where one pathway would be involved in the positive RP, and the other one for negative RP. These pathways would target dopaminergic neurons and GABAergic interneurons,
respectively, in the dopaminergic nuclei (section 3.3.3). Moreover, findings of inverse RPE coding in GPi neurons in monkeys (Hong & Hikosaka 2008) can back the idea that the RP can be used in the selection process. Beside, striosomes have been found to project directly to the SNc, with some collaterals to the GP, whereas matrisomes send GABAergic connections to neurons in the GPe and in the GPi/SNr (Fujiyama et al., 2011; Gerfen et al., 2013; Gerfen, 1984; Lévesque & Parent, 2005; Watabe-Uchida et al., 2012). Thus, it has been suggested that connections from striosomes to dopaminergic neurons could convey RPs in the same way matrisomes signal action values (Amemori et al., 2011) and that they could be the source of reward related information, part of a distinct reward evaluation circuit (Stephenson-Jones et al., 2013).

With respect to an actor-critic architecture, the role of the actor has therefore been associated with matrisomes who can impact the selection through the D1 and D2 pathways, and the critic would be played by striosomes who can exert control on the dopaminergic neurons, and would thus be involved in computing the RPE (Fujiyama et al., 2011; Houk et al., 1995). What has been commonly reported is the plasticity onto MSNs D1 and D2, however there was no indication about the neurons being part of a striosome or a matrisome. The classical D1 and D2 pathways originate from matrisomes. To our knowledge, there is no experimental data assessing the projections from striosomes with respect to the dopamine receptor type these MSNs express, but a circuitry similar to a direct and indirect pathway has been described (Fujiyama et al., 2011, 2015). The direct pathway would correspond to the direct projection from striosomes, possibly from MSNs expressing the dopamine D1 receptor, onto dopaminergic neurons, and the indirect pathway would represent the circuit from striosomes, possibly from MSNs expressing the D2 receptor type, to the GP and onto dopaminergic nuclei. However, in a review, Humphries & Prescott (2010) suggested that patches of the dorsal striatum are made of MSNs expressing both D1 and D2 receptors whereas they considered the striosomes in the ventral striatum to be predominantly of the D1 type. MSNs in matrisomes belong either to the direct or indirect pathway, even if some limited overlap has been reported, notably with evidences that D1 MSNs also project to the GPe (Fujiyama et al., 2011; Kawaguchi, 1997; Lévesque & Parent, 2005). Striosomal MSNs for their part, target specifically the SNc and the VTA (Fujiyama et al., 2011; Watabe-Uchida et al., 2012), and the presence of some axon collaterals to the GP have been described (Fujiyama et al., 2011). All this calls for further investigations of these possible RP-related loops.

As dopaminergic neurons project back to striosomes, the specific role of these connections still has to be determined but it seems reasonable to assume a similar role as with the matrix, that is of modulating the synaptic plasticity at the level of striosomes. Due to the topology of the inputs received by the
striosomes, it was considered that they could be involved in processing cognitive and emotion-related functions in the computation of the prediction of the reward. Indeed, this has been experimentally supported in a cost-benefit conflict task with rats, where optogenetic inhibition of the prelimbic cortex, which preferentially projects to striosomes, leads to the disinhibition of the latter and to a shift of behavioural choices towards high cost high reward options. Excitation of the same region resulted in a bias of the selection away from that choice (Friedman et al. 2015).

3.3.1.4 Interneurons

There are several distinct interneuron types in the striatum, and even if we do not implement these populations specifically in this work, we will briefly describe their roles in modulating striatal activity (for a review, see Silberberg & Bolam (2015) and Tepper & Bolam (2004)). As opposed to fast spiking GABAergic interneurons in the cortex, those in the striatum lack the reciprocity with their target neurons, thereby providing only feedforward inhibition (Gittis et al. 2010; Planert et al. 2010). MSNs are believed to interact also through these striatal interneurons (Graybiel et al. 1994). The presence of monosynaptic GABAergic connections between MSNs has however been reported, mostly between MSNs featuring the same dopamine receptor type but also from the D2 MSNs to D1 MSNs (Taverna et al. 2004, 2008). Regarding interneurons, both cholinergic and GABAergic types have been reported (Kreitzer 2009; Tepper & Bolam 2004; Tepper et al. 2010). They receive various excitatory, mostly from the cortex and thalamus, and modulatory inputs and contribute to determining whether MSNs, D1 and D2, fire or not (Aosaki et al. 1998; Bracci et al. 2002; Kawaguchi 1997; Koós & Tepper, 1999; Tepper et al. 2008). In particular, the activation of thalamo-striatal axons induced a burst of cholinergic interneurons, itself triggering a transient, presynaptic suppression of the cortical inputs onto MSNs, thereby providing a gating role to thalamus (Ding et al. 2010).

MSNs seems to fluctuate between an hyperpolarised down-state and a depolarised up-state, possibly constrained by the local circuitry (Nicola et al. 2000). It has been reported that application of dopamine could excite the fast spiking interneurons, as cocaine does, thereby exerting a major inhibitory influence on MSNs (Bracci et al. 2002). Feedforward inhibition occurs mostly via fast spiking interneurons, involved in the information processing during action selection (Gage et al. 2010; Kawaguchi 1993; Kawaguchi et al. 1995; Szydlowski et al. 2013), and via tonically active ones, believed to be involved in coding the motivational value and probability of the reward (Apicella 2007; Apicella et al. 2011; Wilson et al. 1990; Yamada et al. 2004). These fast
spiking interneurons target both D1 and D2 MSNs and can deliver strong and homogeneous inhibition (Planert et al. 2010). In fact, inhibitory synaptic potential from single GABAergic interneurons have been shown to be powerful enough to silence a large number of projection neurons simultaneously (Koós & Tepper 1999). Inhibition across interneuron populations has also been reported (Friedman et al. 2015). Striatal cholinergic interneurons are even able to trigger GABA release from dopaminergic terminals, in order to inhibit MSNs (Nelson et al. 2014). Additionally, synchronous activity of these cholinergic interneurons can directly cause release of dopamine in the striatum, bypassing the need for dopaminergic neurons activity (Threlfell et al. 2012). Acetylcholine depolarises fast spiking interneurons but attenuate the GABAergic inhibition of MSNs (Koós & Tepper 2002). Furthermore, optogenetic activation of cholinergic axons has led to disynaptic inhibition of MSNs through two circuits with different kinetics (English et al. 2011). Indeed, dendrites and axons of interneurons have been shown to cross compartmental borders, notably between matrisomes and striosomes, in what could be a limbic control of the striosome on the behaviours driven by sensorimotor information processing matrix (Crittenden & Graybiel 2011).

3.3.2 Other nuclei and functional pathways

As we have mentioned previously, the BG are believed to be critical in action selection, and it is thought the direct and indirect pathways compete to control the selection by releasing inhibition on the structures downstream, the thalamus and brainstem, themselves responsible for relaying and sending the actual motor command (Grillner & Robertson 2015; Grillner et al. 2005; Mink, 1996). However, the presence of relatively segregated parallel BG-thalamo-cortical loops are also believed to exhibit some integration and competition between them (Graybiel et al. 1994; Haber 2003; Lehéricy et al. 2004; Menéndez et al. 1999) (additionally, see Sesack & Grace 2010 for details on the microcircuitry of the BG). As we have seen, the striatal GABAergic D1 and D2 MSNs in the matrix are the source of the direct and indirect pathways. The D1 MSNs inhibit the GPi/SNr whereas the D2 MSNs inhibit the GPe, which in turn sends inhibitory connections to the STN. The latter then exhibit feedback excitatory connections with the GPe, and feedforward ones to the GPi/SNr. An additional pathway, involved in reward related information processing, stems from MSNs in striosomes, which project to the dopaminergic nuclei.

It has been hypothesised that the parallel loops could represent action channels (Redgrave et al. 1999), an idea which can be supported by the reported consistent functional topology throughout the BG (Alexander et al. 1986; Romanelli et al. 2005). These loops could furthermore enable processed infor-
formation to be used as inputs back to the striatum, to either fine tune the selection or to enable sequence of actions. Their reentrant feature might also allow for model-based learning and thus for improved predictions. It might enable the system to compute in which state it would probably end up given the current state and the possible actions, and might use the expected reward associated in the evaluation of the choices. In a recursive way, this could be a part of the action selection.

In comparison to the striatum, the STN is smaller in volume and in number of cells, by a ratio of 1/60 to 1/100, the GPe by a ratio of 1/12 and the GPi by 1/20 (Yelnik, 2002). This decrease of neural tissue when flowing from cerebral cortex to the downstream nuclei of the BG could suggest a convergence of the information. This has led to the hypothesis that the BG could operate a compression of cortical information controlled by reinforcement learning (Bar-Gad et al., 2003). The GPi/SNr have been suggested to exhibit a somatotopically organized structure, where the GPi would be involved in axial and limb movements and the SNr would be associated with head and eye movements (Gerfen & Surmeier, 2011). The GPi and SNr receive similar inputs and the latter has been considered as a caudo-medial extension of the former.

The cascade of disinhibition ends in the thalamus, targeted by the inhibitory connections from the output nuclei of the BG (Chevalier & Deniau, 1990; Penney & Young, 1981; Ueki, 1983). The functional organisation observed in the BG seems to be preserved in the downstream projections to thalamus and brainstem (Mana & Chevalier, 2001; Mengual et al., 1999). Neurons in the GPi/SNr serve as an inhibitory gate of motor output, and can be modulated by the BG pathways (Freeze et al., 2013; Nambu, 2008). Indeed, activation of D1 MSNs will inhibit the tonically active GABAergic SNr neurons, which project to brainstem motor regions, thereby disinhibiting them and thus promoting motor output (Freeze et al., 2013; Grillner et al., 2005; Sippy et al., 2015). Additionally, SNr neurons also project to thalamus, which could be involved in the control of task performance and part of a suggested feedback loop through striato-thalamo-cortical circuits. Neurons in the GPi, GPe, STN and SNr show spontaneous activity without any synaptic input (Atherton & Bevan, 2005; Delong et al., 1984; Freeze et al., 2013; Gernert et al., 2004).

The indirect pathway plays an opposite role to the direct pathway in the selection (Kravitz et al., 2010, 2012; Tai et al., 2012). However, the details and specificity of its contribution are still unknown. It has been hypothesised that it could operate a surround-inhibition of competing actions in the GPi/SNr during a motor response (Mink, 1996). This seems to assume a particular somatotopic organisation of the action representations in the various structures of the BG, where actions mechanistically similar would be neighbours, an organisation that has been reported (Alexander et al., 1986, Romanelli et al., 2011).
Concerning the GPe, there are limited knowledge about its functional impact. It is mostly considered as a relay in the indirect pathway between D2 MSNs and STN neurons. However, due to its reciprocal connections with the latter, several studies have suggested a role of oscillatory pacemaker to this GPe-STN pair (Bevan et al., 2002; Plenz & Kital, 1999). Sato et al. (2000) reported connections from GPe neurons targeting mostly the STN and the SNr, but also the GPi and the striatum, suggesting its involvement on most structures of the BG.

The STN is supposed to be involved in behavioural switching, that is, allowing a new behaviour to be undertaken by terminating the current action (Hikosaka & Isoda, 2010). It has indeed been suggested that the BG could enable the automatic execution of sequences generated in cortical areas (Jin et al., 2014). The STN also receives input from the cortex, the thalamus and the GPe (Temel et al., 2005) and projects to the GPi, to form the so called hyperdirect pathway, and back to the GPe. It also has disynaptic projections to the cerebellar cortex (Bostan et al., 2010), which projects back to the striatum, supposedly carrying motor and non-motor information and considered to be involved in delay learning conditioning (Bostan et al., 2013). It is commonly believed to act as a global suppression pathway (Frank, 2006; Nambu, 2004). It could be especially useful following a change in the environment requiring critical action selection (Nambu, 2004; Nambu et al., 2002). Moreover, experimental and computational studies in rats have underlined the role of the STN for global action suppression in case of conflicting input cues (Baladron & Hamker, 2015; Baunez et al., 2001). It has also been suggested that a high STN activity could increase reaction time. Recently, artificial inhibition of the basal forebrain has also been shown to be sufficient to induce a rapid behavioural stopping in rats, when applied before a go response initiation (Mayse et al., 2015). This neural mechanism, outside of the fronto-BG circuit, adds further support to a more specific inhibitory role of the indirect pathway of the BG.

The output nuclei of the BG project to thalamus which in turn projects back to the striatum (Kimura et al., 2004; McHaffie et al., 2005). Furthermore, connections from motor thalamus, which is co-active with voluntary and passive movements of discrete body arts (Nambu, 2011), have been shown to exhibit a high degree of specificity in topographical projections to the striatum (Mengual et al., 1999). The motor circuit within the striato-pallidal system is believed to receive a delayed read-out of cortical motor activity (Marsden & Obeso, 1994). The thalamus is commonly considered as the source of efference copy signals for the BG (Fee, 2012). This efference copy, also called corollary discharge, enables the removal of the agent self-produced modifications on the environ-
ment from its sensory perception of the world, in order to evaluate what has really changed (Blakemore et al., 2000). The functional role of the efference copy in the BG has also been suggested to contribute to the establishment of the eligibility traces, in that it cleans the noise from all the firing patterns of the striatal cells during the selection (Lisman, 2014). Indeed, as the actual selection seems to happen downstream of the striatum, possibly in the GPi/SNr or even in the thalamus, the upstream structures are not aware of which action has been selected and performed (Fee, 2014; Redgrave & Gurney, 2006; Schroll & Hamker, 2013). It has been observed that the activity of MSNs following the selection is more driven by that choice than by its pre-selection activity (Lau & Glimcher, 2008; Lisman, 2014). Nambu (2004) hypothesised that the efference copy could help the indirect pathway to ensure the activation of only the selected command.

The functional organisation of the BG has led to different interpretations, but they however all relate to some extent to action selection and reward related processes (Chakravarthy et al., 2010) (see Schroll & Hamker (2013) for more theories). It has been suggested that motor inputs and contextual inputs could be the source of two subcircuits, one involved in the classical action selection and the other one in extending the repertoire of possible actions (Sarvestani et al., 2011). Besides, Bar-Gad & Bergman (2001) hypothesised that reinforcement signals and local competitive rules in the BG could serve to reduce the dimensionality of the sparse cortical information in order to increase the correlation of the neuronal activity, thereby increasing the efficacies of feedforward and lateral inhibitory connections.

The BG have also been linked with sequences generations, supposing that subnetworks such as the GPe-STN reciprocal interactions could provide traces of previous activity to effectively work as a short term memory (Beiser & Houk, 1998; Berns & Sejnowski, 1998; Dominey et al., 1995). There are some evidence that the BG could indeed encode concatenation of action sequences (Jin & Costa, 2010; Jin et al., 2014; Wörgötter & Port, 2005). Related to that is the representation of time, which is, to the least, pivotal to the BG, and might possibly be implemented within its circuits (Gershman et al., 2014; Jin et al., 2009; Mauk & Buonomano, 2004) (see section 3.3.4 for an example, where the inhibitory effect of the predicted reward matches temporally the actual reward delivery). It remains to be understood how this is computed. Hypotheses mostly rely on assemblies of neurons, tuned to different time intervals through oscillations (Matell & Meck, 2004), or on the integration of ramping activity (Rivest et al., 2010; Simen et al., 2011) or even on dynamics of recurrent networks (Buonomano & Laje, 2010).
3.3.3 Dopaminergic nuclei

The dopaminergic neurons are found in two nuclei, the VTA and the SNc. Although they might not be officially classified as part of the BG, they share a lot of connections with them, and especially with the striatum, and are critical to their functions through the release of the dopamine neuromodulator. They project and release dopamine (but not only, see below, and for details on dopamine and its effects, see section 3.3.4) to different targets, both cortical and subcortical, and receive inputs from different areas (Lammel et al., 2011, 2012, 2014; Watabe-Uchida et al., 2012). The reward expectation and actual reward signals could be provided by the pedunculopontine nucleus, the amygdala, the hypothalamus, or the OFC (Matsumoto & Hikosaka, 2007; Nieh et al., 2015; Okada et al., 2009). These nuclei also have GABAergic projection neurons, targeting the striatum and PFC (Carr & Sesack, 2000).

The VTA and the SNc also exhibit different electrophysiological properties and molecular features (Roeppe 2013). A distinction in the firing patterns of VTA versus SNc dopaminergic neurons in response to aversive events has been revealed (Matsumoto & Hikosaka, 2009b). Additionally, a similar distinction holds also when based on the targeted areas of the projections from these dopaminergic neurons. The ventral striatum receives inputs from the VTA and the ventromedial SNc, which are also regions inhibited by aversive stimuli, and the dorsal striatum receives inputs from the dorsolateral SNc, themselves excited by the same stimuli (Lerner et al., 2015; Lynd-Balta & Haber, 1994; Matsumoto & Hikosaka, 2009b). In rats, cats and primates, the dorsal SNc seems to target mostly the matrix, while the ventral part projects to striosomes (Gerfen et al., 1987; Jimenez-Castellanos & Graybiel, 1989; Langer & Graybiel, 1989; Prensa & Parent, 2001). Matsuda et al. (2009) however found only a preferential innervation of matrisomes and striosomes by the dorsal and ventral SNc, respectively. A recent study reported a possible division of dopaminergic neurons suggesting distinct information streams: one targeting the dorsomedial striatum and the other one the dorsolateral striatum (Lerner et al., 2015). The authors additionally noted reciprocally connections from these striatal regions to the same dopaminergic neurons.

The dopaminergic neurons project to a wide range of areas, and of interest here, to a very wide striatal area, innervating both the striosome and matrix compartments (Matsuda et al., 2009). Furthermore, the dopamine signal seems to be widely broadcast, that is, a single dopaminergic neuron in the SNc exert a strong influence on a large number of striatal neurons (Matsuda et al., 2009; Threlfell & Cragg, 2011). Additionally, the dopaminergic neurons in the SNc affect the GP and the STN, collaterally from the axons targeting the striatum (Prensa & Parent, 2001). Surprising observations revealed that dopaminer-
gic neurons projecting to the striatum could co-release glutamate or GABA, possibly following a ventrodorsal distinction of the striatum (Nieh et al., 2013; Stuber et al., 2010; Tecuapetla et al., 2010; Tritsch et al., 2012). This expanded repertoire of influence of the dopaminergic neurons has still to be functionally asserted. It can not account for the whole modification of MSNs’ excitability, as the experiments which reported such effects did so with the diffusion of dopamine agonists.

Concerning the inputs to the dopaminergic neurons, it has been shown that SNc neurons receive excitatory inputs from somato-sensory and motor cortices, and from the STN, whereas neurons in the VTA receive inputs from the lateral habenula (LH) and the rostromedial tegmental nucleus. However, the most abundant source of inputs to these dopaminergic nuclei is found in the BG, and especially in the striatum and the pallidum, and are mostly ipsilateral (Lerner et al., 2015; Watabe-Uchida et al., 2012). Dopaminergic neurons receive monosynaptic inhibitory inputs from striosomes and the SNr, and polysynaptic disinhibitory inputs from the GP through the SNr (Fujiyama et al., 2011; Tepper & Lee, 2007). In the SNc, at least 30% of the synaptic inputs to the dopaminergic neurons are glutamatergic, and 40-70% are GABAergic (Henny et al., 2012). It has recently been reported that frontal cortical neurons project to the VTA and the SNc, and that activity in the presynaptic population results in an elevation of the dopamine level in the striatum (Kim et al., 2015). Lateral hypothalamus also sends both excitatory and inhibitory connections to the VTA, which projects back to the former (Nieh et al., 2015). Another study reported connections from paraventricular, in addition to lateral, hypothalamus, and from dorsal raphe to dopaminergic and GABAergic neurons in the VTA, all supposed to be implicated in reward-related behaviours (Beier et al., 2015). Moreover, Beier et al. (2015) have described an anterior cortex-VTA-lateral nucleus accumbens circuit, detailed as reinforcing given that excitation of anterior cortex leads to dopamine release in the lateral nucleus accumbens. This suggests a top-down control of midbrain dopaminergic neurons, bypassing the striatum. Their study also showed direct connections from neurons in the nucleus accumbens to dopaminergic and GABAergic neurons in the VTA, possibly also involved in a top-down control of the reinforcement signal. Chuhma et al. (2011) further reported that GABAergic interneurons in the substantia nigra (SN) seemed to be preferentially targeted by MSNs in the striatum, with a few direct connections from MSNs to the dopaminergic neurons in the SNc. This is also a support of a dual control of the dopaminergic neuron activity by striatal neurons, capable of triggering both decreases and increases in the activity of the targeted midbrain neurons.

The VTA receives inputs from widely distributed brain areas (see Sesack & Grace (2010) for a review), but of particular interest are the PFC-VTA and
striato-VTA loops. Within the VTA, GABAergic neurons, possibly interneurons, preferentially target dendrites of dopaminergic neurons (Omelchenko & Sesack, 2009). It was further noted that these GABAergic neurons also receive local inputs, suggesting a complex circuitry within the VTA. Optogenetic activation of these neurons reduced the excitability of the dopaminergic neurons (van Zessen et al., 2012). Dopaminergic neurons in the VTA seem to get phasically excited by reward and reward predicting events, in a manner consistent with RPE coding, whereas GABAergic neuron activity appears to be determined predominantly by reward predicting stimuli (Cohen et al., 2012). Pan et al. (2013) have described anticorrelated activity of inhibitory neurons, seemingly projection neurons from the SNr, with midbrain dopaminergic neurons, with the former firing before the onset of the phasic dopaminergic response. Additionally, the authors showed that optogenetic activation of SNr projection neurons was sufficient to suppress firing of dopaminergic neurons. These results, tend to provide support to the hypothesised role of these interneurons in the inhibition of the phasic dopaminergic signal, a critical mechanism for extinction and for the RPE computation. This agrees with the indirect pathway from striosomes to dopaminergic neurons via the GP reported in rats by Fujiyama et al. (2011).

The activity of the dopaminergic neurons has been reported to decrease with an unexpected loss of a predicted reward or with a negative reward (e.g. painful stimulus) (Matsumoto & Hikosaka, 2007). However, Matsumoto & Hikosaka (2009b) and Tan et al. (2012) have observed both inhibition and excitation of dopaminergic neurons by aversive stimuli, and that these dopaminergic neurons were anatomically distinct. The inhibition by an aversive stimulus might be transmitted from the LH (Ji & Shepard, 2007; Lammel et al., 2012; Matsumoto & Hikosaka, 2007; Stephenson-Jones et al., 2013), which has been shown to be excited (inhibited) by CS predicting aversive (positive) events (Matsumoto & Hikosaka, 2009a) (see Hikosaka & Isoda, 2010 for a review). However, as most of neurons in LH are prominently glutamatergic, they must target inhibitory neurons as relay. Such relay has been found in rats and primates, where GABAergic neurons in the rostromedial tegmental nucleus receive excitatory inputs from the LH, but also exhibit a negative RPE and project to the dopaminergic neurons somata (Hong & Hikosaka, 2011). Optogenetic activation of fibers from the LH to the rostromedial tegmental nucleus promoted behavioural avoidance in mice, mimicking the effect of an aversive stimulus (Stamatakis & Stuber, 2012). It has indeed been reported that LH sends inhibitory projections to the SNc and the VTA (Ji & Shepard, 2007; Quina et al., 2014), and itself receives connections from a subset of GPi neurons, linked with reward related signaling (Hong & Hikosaka, 2008). Furthermore, reward probability and uncertainty about the timing of the delivery
do not seem to be the only factors which extend the duration of the phasic response of dopaminergic neurons (Hollerman & Schultz, 1998; Matsumoto & Hikosaka, 2007; Oyama et al., 2010). The intrinsic value of an outcome or event might be coded by the limbic system and especially the amygdala, even though plasticity in the amygdalo-striatal pathway has been reported (Namburi et al., 2015). Interestingly, no significant change in the response profile of the dopaminergic neurons during the reward phase of a trial has been shown in relation to variation in the motivation (Oyama et al., 2010). The authors suggested that this could be due to adaptation to the increased reward value, by the prediction system. An other possibility for the coding of the motivation could be in the tonic activity of the dopaminergic neurons (Hamid et al., 2015). Let us note that the tonic activity of striatal neurons has also been linked with the encoding of motivational information (Apicella, 2002; Yamada et al., 2004).

Returning to the phasic activation of the dopaminergic neurons, their optogenetic activation slows extinction and can drive new learning of antecedent clues during associative blocking procedures (Steinberg et al., 2013). The authors argued that this could support the idea that RPE signaling by dopaminergic neurons is causally related to cue-reward learning.

About the RPE computation, Stephenson-Jones et al. (2013) suggested that the striosomes might be in a position where they could control directly the dopaminergic neurons in the midbrain, and indirectly via the LH (Hong & Hikosaka, 2011). Synapses onto dopaminergic neurons in the VTA exhibit synaptic plasticity, which seems to be dopamine-dependent (Bonci & Malenka, 1999; Lüscher & Malenka, 2011; Thomas et al., 2000). Notably, cocaine exposure leads to LTP in midbrain dopaminergic neurons (Liu et al., 2005; Ungless et al., 2001), whereas amphetamine inhibits it through D2 receptors activation (Jones et al., 2000). Jones et al. (2000) further reported bidirectional synaptic modifications on VTA dopaminergic neurons. These results could support the idea that the striosomal-dopaminergic neuron pathway would be the substrate of the RPE computation, acting as the critic and learning to predict the reward. However, there is a need for a more systematic testing of the variables at play, similarly to the studies on cortico-striatal plasticity (section 3.3.4), for instance on the dependence on the activation of AMPA and NMDA receptors.

3.3.4 Dopamine and synaptic plasticity

*Dopaminergic neurons [...] cannot say much to the rest of the brain but what they say must be widely heard.*

- Glimcher (2011)

Dopamine is one of the, at least, 100 neurotransmitters present in the brain.
It affects its targets primarily via two classes of receptors: D1 and D2. As we have seen, it is predominantly released by two nuclei, the VTA and the SNc, and we will here consider only these sources. Extracellular dopamine concentration is critical for modulating plasticity, notably at cortico-striatal synapses (Calabresi et al., 2000; Pawlak & Kerr, 2008; Reynolds & Wickens, 2002; Reynolds et al., 2001; Shen et al., 2008; Surmeier et al., 2007; Wickens et al., 2003a; Yagishita et al., 2014). Elsewhere, dopamine has also been reported to modulate STDP as much as to turn it into its symmetrical, anti-Hebbian, shape, in PFC (Ruan et al., 2014). In lateral PFC, the blocking of D1 and D2 receptors impairs the learning of new stimulus-response associations and cognitive flexibility but not the memory of pre-existing associations (Puig et al., 2014).

Importantly, the phasic activity of dopaminergic neurons has been shown to encode the difference between the current reward, or expected reward, and a weighed average of the history of the outcome, and is similar to the TD error of reinforcement learning models (Bayer & Glimcher, 2005; Dayan & Balleine, 2002; Glimcher, 2011; Hamid et al., 2015; Hollerman & Schultz, 1998; Montague et al., 1996; Schultz, 2007a; Schultz et al., 1993, 1997, 1998; Suri, 2002; Suri et al., 2001) (also see Wörgötter & Porr (2005) for a review of TD models and their relation to biological mechanisms). It has to be noted that any reward predicting event can be used in place of an actual reward, as it acts as a temporal transfer of the reward (section 2.3.3.1). Therefore, a reward predicting event generates a transient dopamine increase, whereas the actual reward delivery will not, as it is fully predicted (Oyama et al., 2010; Schultz et al., 1997). The same inhibition of the phasic change could also occur for the reward predicting event, as it itself gets predicted by an earlier event, similarly to n-ary conditioning (section 2.2.4.3).

At the delivery of a reward, if it has not been expected, a phasic increase of the dopamine level occurs (fig. 3.2). However, if it has been expected and obtained, no significant variation of the dopamine level is observed. Furthermore, if the expected reward is omitted, the firing rate of the dopaminergic neurons decreases, at the time when the reward is expected. Eventually, the dopamine signal also adapts to the omission (Pan et al., 2013). Basically, dopaminergic phasic signals are believed to inform on whether recent actions have led to a reward, in order to update the decision policy towards rewarded actions (Hamid et al., 2015; Stopper et al., 2014). It is therefore a surprising outcome that is critical for learning rather than a predicted one, similarly to a concept in information theory (section 4.3.3).

The activity of dopaminergic neurons has been shown to be modulated according to the value of the upcoming action (Morris et al., 2006). An increase in the firing rate of the dopaminergic neurons triggers a transient surge in extracellular dopamine concentration sufficient for behavioural condition-
ing (Lavin et al., 2005; Tsai et al., 2009). Improved learnings from negative outcomes have been observed in human subjects following a transient decrease in the dopamine level (Cox et al., 2015). Similarly, optogenetic inhibition of dopaminergic neurons mimics a negative RPE in a classical conditioning setup (Chang et al., 2015) and causes avoidance of the associated action in an operant conditioning task (Hamid et al., 2015). Moreover, mice which could not produce dopamine demonstrated less interactions toward a sucrose delivering tube than controls (Cannon & Palmiter, 2003). Such mice became hypoactive and let themselves die of starvation even though food was readily available. Restoration of dopamine production in the nucleus accumbens reinstated exploratory behaviours while in the caudate putamen it restored feeding and nest

ing (Szczypka et al., 2001). This further supports the idea that dopamine plays a role in goal directed behaviours and not specifically in reward processes.

In the BG, the pre- and postsynaptic activity as well as the dopamine level are the obvious factors for synaptic modification, for which the standard STDP learning rule can not account for (see section [3.2.4] for detail on STDP and section [4.2.1] for a presentation of some models of synaptic plasticity). However, some evolutions have been offered to this rule (Farries & Fairhall, 2007; Frémaux et al., 2010; Gurney et al., 2015; Izhikevich, 2007; Legenstein et al., 2008). Additionally, dopamine-dependent synapses in the striatum of mice with NMDA receptor knock out exhibit almost no LTP (Dang et al., 2006), which adds support to the idea that multiple factors are involved in the plasticity. Moreover, GABAergic signaling has been shown to be able to control the polarity of STDP at cortico-striatal synapses: the blockade of postsynaptic GABA receptors reverses the temporal order of plasticity for both D1 and D2 MSNs (Paille et al., 2013). Therefore, taking into account the diversity of striatal neurons seems necessary to unfold the complex spike timing dependences observed at cortico-striatal synapses (Beretta et al., 2008; Fino & Venance, 2010).

The definitive account of the synaptic change at cortico-striatal synapses has yet to be reported, due to the complexity of recording and controlling the four (at least) interdependent variables: pre- and postsynaptic spike timing, dopamine receptor type and dopamine level. The precise timing required by STDP might be important (Shen et al., 2008) (see Fino & Venance, 2010 for a review of studies which found STDP in the striatum), but other studies only remarked a looser form of learning, based on correlated activity of the pre- and postsynaptic neurons (Yagishita et al., 2014) (see Pawlak et al., 2010) for a review of studies reporting non-standard STDP in the striatum). The dopamine-dependent plasticity of cortico-striatal synapses is believed to be bidirectional: both LTP and LTD have been observed, although there is no clear consensus on the factors involved (Fino et al., 2005; Reynolds & Wickens, 2002; Shen,
Peaks of dopamine are believed to lead to LTP of cortico-striatal synapses of D1 MSNs whereas dips would lead to LTP of cortico-striatal synapses of D2 MSNs (Nair et al. 2015). Shen et al. (2008) have described LTP in D1 MSNs and LTD in D2 MSNs with phasic increase of dopamine, and vice versa with phasic decrease of dopamine. However, Fino et al. (2005) have reported anti-Hebbian STDP at cortico-striatal synapses in rats in vivo, without detailing the dopamine receptor type and without varying the dopamine level. These results are opposite to those of Pawlak & Kerr (2008), who found standard STDP in rat brain preparations. They furthermore noted that blocking the dopamine receptors prevented both LTP and LTD. Additionally, local influences in the striatum have been shown to impact the release of dopamine by affecting directly axon terminals (Ca-chope & Cheer 2014; Nelson et al. 2014), synchronous activity of striatal cholinergic interneurons is sufficient to provoke dopamine release by activation of nicotinic receptors on axons of dopaminergic neuron (Threlfell et al., 2012). The functional role of this mechanism remains to be elucidated, but the authors suggested that it could go beyond the processes of action selection. The synaptic plasticity of thalamo-striatal connections has been suggested to display higher plasticity in patch (synapses on dendritic spines) than in matrix (dendritic shafts) (Fujiyama et al. 2006).
A problem persists however, since it has yet to be determined how the activity on such short timescale as STDP interacts with the long-term feedback that usually occurs, corresponding to the temporal credit assignment problem (Berke & Hyman, 2000; Houk et al., 1995).

It has been reported that, in monkeys, dopaminergic neurons can encode a context-dependent prediction error (Nakahara et al., 2004). Furthermore, the activity of the dopaminergic neurons has also been linked with aversion, saliency, novelty and uncertainty signaling (Bromberg-Martin et al., 2010; Schultz, 2007a,b). The RPE has been suggested to reflect the integration from different reward dimensions, e.g. amount, type and risk, into a common subjective scale of value (Lak et al., 2014). This indicates that the dopaminergic signal carries rich and complex information about reinforcement.

Dopamine seems to affect the response properties of striatal neurons (Calabresi et al., 2000; Shen et al., 2008; Surmeier et al., 2007). Indeed, dopamine depletion has been shown to increase the excitability of MSNs, by enhancing the responsiveness of the remaining synapses, compensating for the observed decrease of the number of glutamatergic synapses (Azdad et al., 2009; Day et al., 2006; West & Grace, 2002) (see section 3.4.1 for more details on the phenomena, and Gerfen & Surmeier (2011) for a review on the effects of dopamine on projection systems in the BG). In rats, the use of dopamine agonists and antagonists causes a decrease and an increase, respectively, in reaction time (Leventhal et al., 2014). In the dorso-medial striatum, the ratio of AMPA/NMDA has been reported to increase in D1 MSNs and decreased in D2 MSNs after learning (Shan et al., 2014).

The dopamine phasic activity commonly described occurs however, with a very short latency following the presentation of a reward (Schultz et al., 1997; Tobler et al., 2005). This short latency, < 100 ms, is believed to be too fast to enable the necessary processes of the RPE computation, as it is even faster than a saccadic eye movement. It has therefore been suggested to account for novelty salience, where an unpredicted event should attract attention, and should most likely trigger a response (Berridge, 2007; Redgrave et al., 2013). Indeed, the superior colliculus features a direct projection to the SN and is sensitive to the location of luminance changes (Comoli et al., 2003; McHaffie et al., 2006). The information rich phasic dopamine signal observed could thereby result from the overtraining of the task. Climbing fibers have also recently been shown to display an activity similar to the TD error during eyeblink conditioning in mice (Ohmae & Medina, 2015). The dopaminergic signal might thus not be the only way to inform about the RPE. Neurons in the OFC and in amygdala also exhibit responses to reward events such as reward predicting instructions and predictions, or to internally generated reward goals (Hernádi et al., 2015; Schultz et al., 2000).
The dopaminergic signal could act selectively in a form of stimulus-reward learning, in which incentive salience is conferred to reward cues and is subject to individual differences (Flagel et al., 2011). The accurate valuation of the stimulus and comparison with its predicted value would require foveation and processing of the sensory information, possibly involving the PFC for estimating the value of future actions (O’Reilly & Frank, 2006; Potts et al., 2006). Redgrave & Gurney (2006) have suggested that this short latency signal could be more of a ‘sensory prediction error’ than a RPE. This short latency dopamine release could have a similar role to the reward-related dopamine phasic changes, that is to to link the action temporally close to the novel stimulus (Redgrave & Gurney, 2006). It would indeed avoid the contamination of the memory of the possible responsible events with those directly related to the attention shifting actions, and would thereby ease the difficult credit assignment problem (Izhikevich, 2007; Redgrave et al., 2008). Another, non exclusive hypothesis is that the short latency burst could drive the system to take an action, in response to the novel stimulus. This would rely on the change in the intrinsic excitability of the MSNs caused by the variation of the dopamine level. It has also been remarked that the phasic activity of SN dopaminergic neurons could signal the initiation and termination of specific action sequences (Jin & Costa, 2010).

Other plastic connections have been described in relation to the BG, notably synapses from neurons in the nucleus accumbens onto VTA dopaminergic neurons, where drug evoked plasticity has furthermore been reported (Lüscher & Malenka, 2011; Ungless et al., 2001, 2004) (see Berretta et al. (2008) for an overview regarding synaptic plasticity in the BG). The knowledge of the effect of dopamine on the SNr, the GPi, and the GPe is somewhat limited, but there are some indications that it could also modulate synaptic plasticity (Cossette et al., 1999). In the BG, the D1 and D2 dopamine receptors can also be found in the substantia nigra (Levey et al., 1993).

3.3.5 Habit learning and addiction

Addictive and habitual behaviours could both depend on the same processes, where the former would be the extreme result of an otherwise ordinary mechanism. One of the main effects of drug addiction is the inhibition of GABAergic neurons in the VTA (Lüscher & Malenka, 2011) and see Hyman et al. (2006) for a review). Recently, there has been a frantic increase in the list of behaviours that could develop into an addiction problem. The culprit of all these claims is the dopaminergic signal and is usually framed by fMRI studies showing a similar pattern of activity in locations associated with dopamine in the brain when doing activity X than when injecting heroine, therefore X must be
considered as a drug.

At the biological level, the striatum, along with its dopaminergic innervations, seems to play a significant role in drug-dependence, with an involvement progressing from ventral to dorsal parts (see Gerdeman et al. (2003), Everitt & Robbins (2005) and Lüscher & Malenka (2011) for a review). Action encoding has been suggested to undergo a dynamic reorganisation in the dorsal striatum as habit learning develops (Jog et al. 1999). The addiction and compulsive behaviour could result from a loss of effect of the phasic dopamine change on the plasticity, or even of the inversion of LTD into LTP, further strengthening the compulsive response (Nishioku et al. 1999). Parvaz et al. (2015) have found an impaired activity in human cocaine addicts relative to negative RPEs, and suggested this could be the reason for continued drug use even after conviction. However it does not address if this impaired neural response is the cause or the consequence of cocaine addiction. Synaptic plasticity at synapses onto dopaminergic neurons in the VTA has been shown to be massively impacted by cocaine exposure, inversing the learning rule, and having long lasting effects (Lüscher & Malenka 2011). Similarly, habit development has been prevented by dopamine deafferentation in the dorsolateral striatum (Paure et al. 2005).

Seger & Spiering (2011) have defined five common, but not universal, features of habit learning: it is inflexible, slow or incremental, unconscious, automatic, and insensitive to reinforcer devaluation (see Yin & Knowlton (2006) for a review on the role of the BG in habit formation). Lesion and pharmacological studies have been helpful to characterise the role of structures and pathways. For example, inactivation of the infra-limbic PFC leads to an absence of the response associated with the devaluated reward (Coutureau & Killcross, 2003). Inhibition of the ventral striatum once a habit is formed does not impact the occurrence of this habit, but is however critical for its acquisition. The performance rely on the dorsal striatum, and even with its inhibition during acquisition, the removal of this inhibition leads to immediate similar performance as those of control rats (Atallah et al. 2007). The authors therefore suggested a division between the ventral and dorsal striatum in habit formation, assigning the role of the actor to the dorsal striatum, and of the critic to the ventral striatum. More specifically, lesions of the dorsolateral striatum have been shown to prevent habit formation in rats, even with extensive training (Yin et al. 2004).

It has therefore been suggested that habit formation proceeds from a gradual loss of the dopamine dependence of synapses related to neurons in the dorsolateral striatum. However, ventral striatum abilities would remain unaffected. To put that into perspective, let also note that high-frequency stimulations of the nucleus accumbens have been reported to not significantly affect the activity of the VTA or its discharge pattern but to cause a potent decrease in the number of dopaminergic neurons firing spontaneously in the SNc (Sesia et al.)
Such stimulations have been suggested to contribute to the disruption of pathological habit formation by affecting the SNC-dorsal striatum projections. In a study considering the patch/matrix specificity, Canales (2005) showed a silencing of the matrisomes following drug exposure. He suggested that this inactivation of the matrix, and thereby the reliance on the striosome based pathways only, could represent neural end-points of the shift from goal-directed selection to conditioned habitual behaviour.

Synchronised activity from the BG has been proposed as a mechanism teaching cortical pathways in order to establish a habit outside the BG (Atallah et al., 2007). Additionally, Howe et al. (2011) reported a change in the oscillatory frequencies in the ventromedial striatum with the development of learning, possibly suggesting a more effective inhibition of MSNs and thus a more fixed pattern of activity in the output of the BG. Burguiere et al. (2013) observed impaired fast spiking neuron striatal microcircuits, of which the compensation was possible through focused optogenetic activation of the lateral OFC, thereby restoring the inhibition of compulsive behaviours. These results could suggest the development of a dysfunctional control of the inhibition in the striatum in habit formation and compulsive or addictive behaviours. The balance of inhibition between MSNs and interneurons is most likely involved in regulating a dynamic decision transition threshold (Bahuguna et al., 2015).

Habit formation mechanisms have also received lights from studies on bird songs. In song birds, vocal experimentation and learning require a BG-related circuit, the anterior forebrain pathway (Olveczky et al., 2005). A lesion of this circuit prevents the proper development of songs but has only a few restricted effects in adults (Scharff & Nottebohm, 1991). This structure has also been shown to be necessary for modifying the spectral, but not the temporal structure of bird song (Ali et al., 2013). This further suggests a temporal representation external to the BG. Blocking the output of this anterior forebrain pathway prevented the improvement of performance normally observed in song learning in Bengalese finches. However, the unblocking of this pathway contribution after training leads to immediate excellent behaviours. Moreover, inhibiting the activity in the output nucleus of the anterior forebrain pathway during learning not only prevents any improvement of the performance during training but also completely blocks the learning.

It has been suggested that the BG could monitor the results and outcomes of behavioural variations engaged by other brain regions and can direct these brains regions towards valued behaviours (Atallah et al., 2007; Charlesworth...
such pathways outside of the BG have been hypothesised to enable faster information transfer thanks to fewer synaptic contacts than through the BG, and thus reduced reaction times and automatic functioning (Ashby et al., 2007). It has to be mentioned that the cerebellum has been linked with habit formation, but seemingly only for motor skill learning (Callu et al., 2007; Hikosaka et al., 2002; Salmon & Butters, 1995).

The mechanisms leading to the desensitisation of the synaptic plasticity to the RPE are still not understood and there is a need for new theories and models (see Smith & Graybiel (2014) for a review of suggestions). For example, it has been suggested that addictive behaviours could result from the irruption or intrusion of a model-free representation in a selection that should rely on model-based knowledge, such that the construction of the world and its tenants are not taken into account anymore by the habit system (see Dolan & Dayan (2013) and Graybiel (2008) for a review). A general assumption concerning habit formation and addiction is the development of reward-independent circuit, possibly outside the BG.

3.4 BG-related diseases

As the BG are massively involved in action selection, associated dysfunctions most commonly lead to motor troubles. Attention Deficit Hyperactivity Disorder (ADHD), Tourette’s syndrome, dystonia, Parkinson’s disease (PD) and some psychiatric disorders have been related to some extent with the BG. Imbalances between the neuronal activity of the two main pathways have been described as underlying the motor symptoms (Kreitzer & Malenka, 2008). Additionally, some movement disorders have been suggested to result from dysfunctional GABAergic microcircuits (interneurons) (see Gittis & Kreitzer (2012) for a review). Computational models of reinforcement learning have recently been used to assess the role of disturbance of the dopaminergic system in the diseases mentioned, but also in addiction or in schizophrenia (for a review, see Maia & Frank (2011)).

3.4.1 Parkinson’s disease

PD is a debilitating neurodegenerative disease, the second most common after Alzheimer’s disease. It is mostly idiopathic, with weak genetic correlations and a relatively large number of environmental causes. It has a prevalence of 1% and 4% in the population over, respectively, 60 and 80 years old (see de Lau & Breteler (2006) for a review). Patients suffering from Parkinsonian symptoms show a variety of troubles, usually starting with motor related afflictions, such as tremor, dyskinesia, postural instability and rigidity. Evolution is
generally gradual with time and psychiatric disturbances usually occur at later stages, with depression and dementia, along with general cognitive difficulties, as the most commonly observed.

The biological hallmark of PD is a degeneration of dopaminergic neurons in the SNc which is believed to cause imbalance in the pathways of BG (Albin et al., 1989; DeLong, 1990). Indeed, one of the most efficient treatment of the motor symptoms is the administration of dopamine agonists and of L-DOPA, which is a precursor of dopamine, noradrenaline and adrenaline (all part of catecholamines). L-DOPA can cross the blood-brain barrier and can therefore increase the dopamine concentration through enzymatic processing. It must be noted that patients treated with L-DOPA at too high doses can experience psychotic episodes and eventually develop dyskinesia (Iravani et al., 2012). The SNc is not the only dopaminergic nucleus affected in PD, it has been shown that the VTA also degenerates in the disease, although less than the SNc (Alberico et al., 2015). Motor symptoms are thought to occur not before a loss of 50 to 70% of the dopaminergic neurons (Obeso et al., 2004, 2008).

As a result of the dopaminergic neurons depletion, different consequences have been reported at the cellular level. However, a general increase in neural activity has been observed, affecting mostly D2 neurons, which seem to become oversensitive (Bamford et al., 2004). D1 MSNs have been depicted to become supersensitive to D1 receptor activation, even though the actual number of these receptors decreases (Gerfen, 2003). In rats, after acute dopamine depletion, MSNs and cholinergic interneurons become more excitable whereas GABAergic interneurons become less excitable (Fino et al., 2007). Moreover, hypoactivity in the GPe has been described. This, in turns, results in a higher activity in the STN, thus increasing the inhibitory output of the GPi and the SNr (see Obeso et al. (2008) for a review). However, it is still unclear how the increased firing of the STN does not rescue the activity in the GPe, as it has reciprocal connections to it. A direct modulatory effect of dopamine onto the STN has been described, where a decrease of the neuromodulator causes an increase in the activity of the nucleus (Kreiss et al., 1996).

Dysfunctions in the indirect pathway are thought to be critical in PD. Synaptic plasticity has recently been the focus of several studies and is considered as a possible source for the imbalance of the dynamics in the BG. Manipulations of LTD of synapses of D2 MSNs via the administration of D2 receptor agonists have been linked with reduced motor deficits (Kreitzer & Malenka, 2007). Bilateral activation of D2 MSNs using optogenetic stimulation has been shown to elicit a Parkinsonian state, with bradykinesia, freezing and reduced locomotion in mice (Kravitz et al., 2010). Within the striatum, the strength of recurrent connections between MSNs has been shown to be greatly reduced in PD models (López-Huerta et al., 2013; Taverna et al., 2008). The
Dopamine loss has been shown to facilitate LTD in D1 MSNs but to favor LTP in D2 MSNs (Schroll et al., 2014; Shen et al., 2008). Loss of dopamine-dependent synaptic plasticity has been reported on subthalamo-nigral synapses in experimental parkinsonism in rats (Dupuis et al., 2013). The authors hypothesized that the resulting loss of LTD could facilitate the spreading of pathological activity to the SNr. Recently, it has been suggested that, in PD conditions, structural plasticity does not correlate with functional plasticity in the motor area M1 (Guo et al., 2015). Pallidotomy, which is the surgical destruction of the GPi, leads to improvements of dyskinesia (Lozano et al., 1995) but is also associated with adverse events such as contralateral weakness, visual field defects and new learning impairments (Trépanier et al., 2000).

Deep Brain Stimulation (DBS) is another neurosurgical procedure aiming to offer therapeutic benefits to patients. It furthermore provides the possibility of reversibility and adaptability compared to pallidotomy. It consists in the implantation of electrodes deep in the brain, most commonly onto the STN, delivering high frequency electric pulses continuously (100 Hz and more) (Benabid, 2003). It has proven to be very effective to stop tremor, chronic pain and dystonia notably (see Kringelbach et al., 2007 for a review). However, it is still unclear as to how these high frequency stimulations of the STN lead to a relief of the symptoms (see Chiken & Nambu, 2014 for a review of hypotheses). As previously mentioned, in the PD state, the STN is already more active than normally. It thus seems counter intuitive to benefit from stimulating it even more. Furthermore, high frequency stimulation of the STN is not linked with any effect on the dopamine level (Hilker et al., 2003). It has been suggested that such high frequencies would actually induce neural silencing, which would be in line with the effects observed. The classical view suggesting an increase of D2 pathway inhibition has recently been challenged by studies showing that lesions of the thalamus or of the GPe do not cause bradykinesia or akinesia, suggesting a highly organized network effect (Obeso et al., 2000). Following studies shifted their focus on to neural dynamics such as oscillations. Beta-band synchronous oscillations in the dorsolateral region of the STN in patients with PD have been described, as well as increased synchrony. The latter has been suggested to result from the increased inhibitory contribution from the striatum onto the GPe caused by over activity of the indirect pathway observed in PD (Kumar et al., 2011). More recently, studies have brought the attention on the involvement of abnormal synaptic plasticity as a cause of motor and cognitive dysfunctions (Iravani et al., 2012). Computational models have here been critical to test the associated hypotheses (Schroll et al., 2014). It has furthermore been suggested that the imbalance could instead lie between the matrix and striosome compartments and could be the cause of disorders in the BG, such as PD motor troubles (Crittenden & Graybiel, 2011).
3.4.2 Huntington’s disease

Huntington’s disease is also a neurodegenerative disease but is of genetic origin. The evolution of the disease impacts negatively motor coordination and cognitive abilities (see Walker & Walker (2007) for a review). Life expectancy is around 20 years after the onset of the first symptoms. It is the most common genetic cause of abnormal involuntary writhing movements, chorea. The phenotype also includes dystonia, incoordination, cognitive decline and behavioural difficulties. There is no cure for the disease even though some of the mechanisms are well described. Striatum is the area of the brain which is the most damaged by the disease, but its effects spread beyond this nucleus during the later stages. MSNs projecting to the GP die because of protein aggregation, forming inclusion bodies within the cell, growing enough to stop neuronal functions such as the transport of neurotransmitters. It has been suggested that chorea and motor disorders could result from this cell loss, possibly decreasing the D2 pathway contribution, or causing an imbalance between the D1 and D2 pathway, leading to the disinhibition of unwanted motor outputs (Albin, 1995; DeLong, 1990). In mice, a specific near-complete deletion of the vesicular inhibitory amino-acid transport, responsible of the storage of GABA and glycine in synaptic vesicles, in matrisomal MSNs, leads to an elevated locomotor activity, muscle rigidity and other symptoms similar to those of Huntington’s disease, thereby suggesting a link of this disease with a dysfunctional GABAergic signaling of MSNs (Reinius et al., 2015).

3.4.3 Schizophrenia

Schizophrenia is a mental disorder affecting 0.7% of the population (Saha et al., 2005). Its etiology links it with both genetic and environmental factors, even though the diagnostic still relies on the symptoms described by the patient, the signs observed by the clinician and the history of the disorders. The most characteristic features are the positive symptoms of hallucinations and delusions. The negative symptoms commonly consist in a reduced social engagement, a dysfunctional working memory, a loss of motivation (avolition) and a reduced speech output (alogia). This condition puts a lot of strain on the patient and its relatives.

Early in childhood, e.g. at age four, future schizophrenia patients show normal cognitive functioning. The cognitive impairment is an early emerging gradual progressive growth, where a substantial part of the neuropsychological decline is already observed before the onset of the disease (Meier et al., 2014; Woodberry et al., 2008). A puzzling fact about schizophrenia is that no one with congenital or early blindness has been diagnosed with this mental disease (see Silverstein et al., 2013) for discussions on the mechanisms that may be
Recently, an hypothesis involving dopamine has received both attention and support from experiments. Amphetamines, which increase the level of dopamine, are believed to worsen the psychotic symptoms of the patients, as do optogenetic stimulations of frontal cortex neurons projecting to the VTA and the SNc in mice (Kim et al., 2015). It was suggested that one role of dopamine is to render salient a stimulus, in order to focus on its relation with the environment. Psychosis would thus be a result of a dysfunctional dopamine system, adding saliency to common and insignificant events or perceptions. Delusions would be the consequence of the attempt to make sense of these false beliefs (Kapur, 2003). Reward associated dysregulation could be compensated by recruitment of prefrontal areas, in an effort to reduce the dissonance between beliefs and sensory integration. Experiments have indeed found abnormal RPE responses in the midbrain, striatum, and limbic system in patients with psychosis (Murray et al., 2008). Additionally, the reaction times of patients with schizophrenia have been shown to be faster than controls to neutral stimuli. It has been suggested that patients fail to make a distinction between events that are motivationally salient and neutral ones. Deluded patients reach their conclusion with significantly less evidence than controls (Garety et al., 1991) and show abnormal bias in the face of disconfirmatory evidence (Warman, 2008). This indicates that patients develop false beliefs or fail to draw the correct conclusions from the available information. Positive symptoms could result from the wrong integration of new evidence causing false prediction errors to be propagated through a Bayesian system (see Fletcher & Frith (2009) for a review). It has been reported that patients with psychosis show a tendency to update their reward values faster than controls and exhibit a lower degree of choice perseverance (Li et al., 2014a). It has also been reported that delusional ideation is associated with a perceptual instability and a stronger belief-induced bias on reported perception. Furthermore, an enhanced connectivity between the OFC and the cortex has been described in patients (Schmack et al., 2013). Moreover, the genes involved in dendritic plasticity have been associated with psychiatric disorders and with schizophrenia specifically (Tsuboi et al., 2015).

The positive symptoms of the disease have also been suggested to result from a failure related to the efference copy (Pynn & DeSouza, 2013). An inability to recognize the efference copy signal as one’s own could lead to believe that actions result from an external agent. This would support the hypothesis that an efficient RP system might prevent abnormal strengthening of the connections in the D1 and D2 pathway. This would, in turn, expose the associations to be less prominent. Furthermore, the processing of potential future rewards, through stronger activation of the ventral striatum and the right
anterior insula, might be abnormal in patients (Wotruba et al., 2014). There is so far no cure to the disease, and the main treatment is symptomatic, through antipsychotic medications.
4. Methods and models

*The most important questions of life are indeed, for the most part, really only problems of probability.*
- Pierre Simon Laplace, Théorie analytique des probabilités, 1812

4.1 Computational modeling

Computational modeling is the representation of a model in mathematical and algorithmic formalism. It requires the selection of the relevant features for the model along with their adequate mathematical descriptions and parameters. Numerical tools can then be used to study the equations when analytical means can not easily determine the dynamics of the system. The quantitative results of computational models enable direct comparisons with real world data, thereby providing a way to validate or falsify the model and its predictions. Computational models in neuroscience provide the ability to integrate biological data from different experiments, possibly from different scales. Furthermore, they can serve to test hypotheses and to make predictions. Another advantage is the possibility to access any parameter or variable value in the model, a feature which is, as we have mentioned previously (section 3.1.5) still problematic in experimental studies. A model is valued mainly on the testable predictions it can offer. Additionally, trying to integrate the available data in order to re-create and simulate the brain is a great way to realise what relevant information is missing and this generates opportunities for new hypothesis and can thus guide experimental investigations. An increase in our knowledge and comprehension of the brain’s intricate mechanisms therefore results from this interaction between modeling and biology, whereas purely descriptive approaches might fail to grasp what is important and what is required to understand this very complex system. Moreover, the capability to simulate neurological diseases, psychiatric disorders and lesions is a great asset of computational neuroscience.

The level of detail can range from molecular and subcellular mechanisms to abstract models of whole brains. However, in the current state of knowledge, considering higher cognitive functions seems to require higher levels of abstraction. Indeed, the amount of integrable data increases, as do the number
of gaps where data is lacking. This is related to the approach taken towards modelisation. Is the model build form the integration of detailed piece of data or is it shaped by a theory of the mechanisms of the process studied (section 4.1.1)? For a particular system, Ockham’s razor asks for the selection the simplest model that can faithfully characterise the behavior of the system, and, at the same time be simple enough to keep a high explanatory power. Finally, the computational cost of a model has to be considered, as even with the computational capacity of the current supercomputers, there is no realistic way to run large-scale simulations with a high level of detail in a decent amount of time. However, the development of exascale computing and of neuromorphic hardware could very soon provide executions close to real time, and possibly even faster (section 4.1.3).

Computational neuroscience is a critical tool in the study of the brain. With a more complete understanding of the brain will come great societal evolutions.

4.1.1 Top-down and bottom-up

A top-down approach starts with an abstract hypothesis or theory on how a complex event or function could be algorithmically computed. The formulation of the model then has to respect the intended computational elements. On the contrary, the bottom-up approach aims to get all the computations possible from the elements of a structure in order to capture the function of the system. In computational biology, the bottom-up approach relates to the study of the emerging properties from the implementation of a complete biological description of all the experimentally characterised components of a system. A pure bottom-up description is uncommon and, in practise, what differs is the extent to which the modeller takes into account one or the other approach to constrain the model. For example, in this work, which started as a top-down approach, we used biological data in order to constrain our model when a computational functionality could be implemented in various ways. Both considerations can be applied depending on the level of organisation, i.e. any levels, except maybe the highest and the lowest, can be set indifferently as the reference layer for any of these two approaches. It also defines the effect, constraining or biasing, a higher level area has on low level sensory motor processing, in biology, and conversely, low level processes constrain what kind of computations can emerge from the biology.

4.1.2 Simulations

Simulation is the main tool in computational studies of the brain. It requires the selection of a scale of detail, which depends on the question addressed, and the definition of a model, describing the neural system at that scale. Along
with the simulation itself, there is a need to record and store the data produced, in order to run analyses before the visualisation of the results.

4.1.2.1 Neuron model

We implemented what is probably the most common biological neuron model: the leaky integrate and fire (LIF) neuron. Only sub-threshold dynamics are formally described, i.e. when the membrane potential \( V_m \) is below the spike threshold \( V_{th} \). A spike is artificially emitted every time \( V_m \geq V_{th} \), at which point \( V_m \) is reset to its reset potential \( V_{res} \) and held there for a time \( t_{ref} \), simulating a refractory period. The evolution of the membrane potential is defined as:

\[
C_m \frac{dV_m}{dt} = -g_L(V_m - V_L) + I(t)
\]  
(4.1)

where \( C_m \) is the membrane capacitance, \( g_L \) and \( V_L \) are the leak conductance and the leak potential, or reversal potential, respectively. \( I(t) \) is the input current to the cell, which is related to the number of spikes sent by pre-synaptic neurons and their associated synaptic weights.

The simplicity of this model allows for large scale modeling, but at the same time restricts the diversity of dynamical features of biological neurons that it can reproduce. The main reason that drove us to select this neuron model was the compatibility issues with the neuromorphic hardware specifications of the European project “BrainScaleS” European project (section 4.1.3).

4.1.2.2 NEST

The Neural Simulation Tool (NEST) is a software package used to run relatively large scale simulations of spiking point neurons, and such simulation has been described as an electrophysiological experiment taking place inside the computer (Gewaltig & Diesmann, 2007). It includes several synapse and neuron models and can be extended with new models, functions and libraries. This simulator focuses on the dynamics, size and structure of neural systems rather than on the exact morphology of individual neurons. It can represent spikes in continuous time and mechanisms such as plasticity and learning can be implemented. NEST can be run in parallel on multiprocessor computers and computer clusters to increase the memory and to speed up the simulation. A python interface has been developed, enabling the integration of the many extension modules for scientific computing available (Eppler et al., 2008).

We can also mention NEURON, which enables detailed multicompartmen- tal neuronal models (Hines & Carnevale, 2001), Nengo, which focuses more on large-scale simulations and high-level cognitive functions (Bekolay et al., 2014), and Brian, also a spiking neural networks simulator with the aim of
making the writing of simulation code as easy as possible (Goodman & Brette, 2009). PyNN, python neural networks (Davison et al., 2008), is a simulator-independent language, where the code written once can be run directly on any simulators supported, e.g. NEST, NEURON or Brian.

4.1.3 Neuromorphic Hardware

The human brain, despite draining almost one fourth of the total energy consumption of the body, is a highly energy efficient system. Indeed, it requires only 10-20 watts, compared to the several thousands devoured by a supercomputer, nowhere near capable of simulating in real time a detailed model of the brain, or to plainly be able to perform complex cognitive tasks (see Lansner & Diesmann (2012) for a review on challenges associated with spiking neural networks and large-scale simulations). Reverse engineering is one way and one aim of neuromorphic computing.

Comprehending the brain as a sum of electrical components, it is fair to imagine that analog circuits could reproduce at least some of its dynamics and behaviors (Mead, 1990). Additionally, the development of new architectures could provide novel approaches for real-world problems, where conventional computer architecture can struggle to achieve good results. Recently, mixed analog and digital as well as pure digital chips are also included in the “neuromorphic” designation. The aim is still to understand how individual neurons, circuits and systems can enable complex computations, and how information is represented. Such was the goal of the European projects FACETS-ITN and BrainScaleS, in which we were involved. The focus was also on the interaction of multiple temporal and spatial scales. The resulting mixed signal hardware, based on custom design analog circuits emulating neurons and synapses, can run much faster than real-time, up to 10000 times (Schemmel et al., 2008, 2010). Similarly to classical cluster computers, the bottleneck is the memory access in neuromorphic hardware. Be it for reading or writing updated weights, or to commonly read state information of the system, these operations are costly in term of time but also with regards to the code implementation.

Plasticity takes a significant toll on the simulation time on standard hardware architecture and neuromorphic solutions represent an opportunity to considerably reduce that cost in the simulations. Unfortunately, due to portability troubles, they have not been as helpful, yet, as one could have hoped for in the development of models featuring plasticity or non-conventional implementations. However, the current neuromorphic solutions, analog and digital, have proven to be efficient on non-plastic, well developed models.

This lack of easily available plasticity modules acts as if yet another level of constrains is put on the modeling part, this time from the hardware, as one
has to adapt to the restricted possibilities of implementation. Hopefully, these issues will be overcome in the near future, with the continuation of the Brain-ScaleS’ hardware production into the Human Brain Project. The development of memristor technology has recently attracted lots of interests for their capabilities to mimic synaptic adaptation and plasticity, and could become an additional component in future neuromorphic systems (Prezioso et al., 2015). In some cases, due to hardware limitations, it is simply impossible to port a feature of the model, such as the learning rule here used on the current version of the hybrid multiscale facility (HMF), in others, it requires relatively important development for the feature to be implemented in the software package of the hardware, such as needed with the purely digital SpiNNaker hardware. However, a decrease of the memory and computational footprints of the learning rule we use (section 4.3.4) has brought closer real-time simulation of a cortex model in high performance computing (Vogginger et al., 2015). Furthermore, the hardware implementation of this learning rule in dedicated VLSI could also unload computational burden on classical supercomputer architectures (Farahini et al., 2014).

4.2 Relevant computational models

We will here give an overview of the most influential models of synaptic plasticity and of the BG before presenting the framework of our work and our models, of which a more exhaustive description of the latter can be found in the papers included in the thesis. We will here focus especially on the spiking version of our model. Since plasticity is at the core of our model, we will first detail specifically some of the most common learning rules used in computational neuroscience.

4.2.1 Models of synaptic plasticity

The Hebbian learning principle was formulated on theoretical grounds, and biological evidence came years later (section 2.2.5.2). In rate-based models, the rates of firing of the pre- and postsynaptic neurons are measured over time and determine the sign and amplitude of the plasticity. Among these models, we can mention the one proposed by Oja (1982) and the BCM by Bienenstock et al. (1982). In the former, Hebbian plasticity additionally features a decay term and neurons can compute a principal component of its input stream. About BCM, a sliding threshold controlling the switch from LTD to LTP prevents the problem of positive feedback.

Hebbian learning and STDP are now the most common form of learning rule implemented, especially in spiking networks, as they are both supported
by experimental data. Network stability, self-stabilisation of the weights and rate, is achieved in STDP when the LTD side dominates over the LTP one (Song et al., 2000). Variations and extensions of these rules have been offered in order to account for more complex integration and new experimental data (see Gerstner & Kistler (2002b) for STDP related descriptions and Gerstner & Kistler (2002a) for mathematical formulations of Hebbian learning). For example, in the reinforcement learning framework, a third term has been added in the STDP learning rule in order to provide information guiding the learning based on the reward value or on the RPE, similar to the effect of dopamine (section 3.3.4). Such a learning rule, commonly denoted reward-modulated STDP, has shown the need for a critic part to predict the reward for each task and stimulus separately (Frémaux et al., 2010). Furthermore, the implementation of TD learning in spiking networks through the use of reward-modulated STDP has displayed good result in various tasks, such as in a simulated Morris watermaze and cartpole problems (Frémaux et al. 2013). Alternatively, a more direct implementation of the TD learning rule has been done without STDP but assumed uncommon biological dynamics such as a temporally restricted synaptic plasticity (Potjans et al., 2009). Reward-based and correlation-based methods are similar in open-loop conditions but notably different in the closed-loop condition, which is the normal setup of behavioral feedback (Wörgötter & Porr, 2005). The reward, in classical TD learning, enters the update in an additive way, whereas a multiplicative correlation drives STDP. However, TD(λ) methods make use of eligibility traces, which enter the algorithm in a multiplicative way. Rate based learning rules dismiss the temporal ordering of the spikes critical in STDP but use the average rate over a temporal window.

The online update rule of STDP can be implemented by assuming that each spike leaves a trace which decays exponentially in absence of spike. The weight is updated by an amount based on the value of the trace of the presynaptic spike when the postsynaptic spike occurs, and by an amount based on the value of the trace of the postsynaptic spike when a pre-synaptic spike happens. Another common addition to the standard STDP are upper and lower bounds for the weights. It results from the pairwise interaction in STDP that at high enough firing rates, pre- before postsynaptic spikes can be picked up as post- before presynaptic pairs, leading to depression of the synapse, which is not observed in biology (Sjöström et al., 2001). Assuming a triplet interaction, a combination of a pre- and two postsynaptic spikes with different time constants, instead of the pairwise one, can avoid such depression. Different phases of synaptic plasticity have been suggested and a mathematical model details three steps: the setting of synaptic tags, a trigger process for protein synthesis, and a slow transition to synaptic consolidation (Clopath et al., 2008). This enables the retention of the strength of the synapses over longer time scales.
Most other models will not update the weights as long as no new induction is elicited, they would however decay, meaning that information can not be read out without suffering modifications (Feldman [2012]).

In an attempt to take into account the bidirectional and dopamine receptor type dependent plasticity reported at cortico-striatal synapses, the STDP learning rule was modified with a non-standard form of the plasticity function, derived from in-vitro data, and implemented an eligibility trace. The resulting rule, denoted E-STDP, showed good similarities with experimental data on plasticity and learning at cortico-striatal synapses (Gurney et al. [2015]).

We can also mention the hypothesis of the first-spike times, where synaptic plasticity is tuned to the earliest arriving spikes, through STDP (Guyonneau et al. [2005]).

Even if some biological mechanisms have been a posteriori presented as explanation of the processes taking places during the synaptic plasticity, it is not the focus of these models, which are phenomenological. However, it should here be briefly noted that such biophysical models investigate the biochemical and physiological operations involved in the expression of synaptic plasticity. Most of these models rely on calcium flux dynamics and NMDA receptors to induce plasticity, both LTP and LTD. They can reproduce various STDP curves (Froemke et al. [2005]; Shouval et al. [2002]).

4.2.2 Computational models of the BG

Simulations of reward learning experiments have already been done, but a majority were based on mean rate neurons (Frank [2005]; Gurney et al. [2001a]; O’Reilly & Frank [2006]; Schroll et al. [2014]). More recently, models of the BG implementing spiking neurons were described (Baladron & Hamker [2015]; Stewart et al. [2012]) (see Delong & Wichmann [2009] for an overview of the development of the models). We will not go into the details of these models, several good review papers have been published and should be considered for further descriptions (Cohen & Frank [2009]; Doya [2007]; Gillies & Arbuthnott, 2000; Helie et al. [2013]; Joel et al. [2002]; Moustafa et al. [2014]; Samson et al., 2010; Schroll & Hamker [2013]; Wörgötter & Port [2005]). The architecture of the networks ranges from high level abstract description to more detailed construction, with some even focusing only on the dynamic of MSNs (Bahuguna et al. [2015]; Beiser & Houk, 1998; Humphries et al. [2009]), and implementing a fourth factor in the learning rule: the dopamine receptor type of the MSNs (Gurney et al. [2015]). The majority now seems to include the D1, D2 and RP pathways with several nuclei (Chersi et al. [2013]; Collins & Frank [2014]; Frank, 2006; Jitsev et al. [2012]; Stewart et al. [2012]), but models with plastic connections in all these three pathways are quite uncommon. The first spik-
ing network of the BG implemented the actor-critic TD learning, combining local plasticity rules with a global signal (Potjans et al. 2009). Their model showed similar performance compared to the discrete time, standard, TD algorithm but could not be mapped onto the BG. Baladron & Hamker (2015) presented a model with plasticity in three pathways: direct, indirect and hyperdirect pathway, using Izhikevich’s learning rule (Izhikevich 2007). They did not implement the RP but instead fed the learning rule with the raw reward value, depending only on the success or not of the trial. As a result, they showed that the direct pathway selected an action by disinhibiting the thalamus, that the indirect pathway was responsible for avoiding repeated mistakes and that the hyperdirect pathway could delay the response until a correct decision is made.

Another common implementation is to model the inputs to the BG or the raw reward signal as external Poisson processes or constant activity driving the input layer or the reward population (Stewart et al. 2012). However, some models feature the pre-motor cortex and thalamus (Chersi et al. 2013; Frank, 2006), or the OFC in decision making (Frank & Claus 2006). Moreover, it has been suggested that the cortical layer Va could represent the current state whereas the layer Vb would code for the previous one, offering a straightforward correspondence with the TD error computation (Morita et al. 2012). Chersi et al. (2013) have proposed a multilayer model of spiking neurons of the cortico-BG-thalamo-cortical circuit organised in parallel loops, with STDP and eligibility traces. One of the main result of the model is its ability to suppress an habitual behavior to engage into a goal-directed action. In a similar approach, Doya et al. (2002) have described a multiple model-based reinforcement learning architecture, composed of modules with a state prediction model and a reinforcement learning controller. The learning rates are weighted by a responsibility signal, similar to the contextual relevance or responsibility of modules, hypothesised by Amemori et al. (2011) to be determined by errors in prediction of environmental features and furthermore assigned to striosomes. The modules specialised for different domains in the state space, enabling a good performance in the pendulum swing task.

In his model, Frank (2006) implements a model of the BG based on a combination of error-driven and Hebbian learning, where a difference in pre- and postsynaptic activity is computed between a selection and a reward phase. Additionally, the dopamine level impacts the activity of the MSNs. Previously, Frank (2005) had assumed an RPE sign-dependent cortico-striatal synaptic plasticity. Connections onto D1 MSNs would be potentiated with increased levels of dopamine, whereas those onto D2 MSNs would be for levels below baseline. An interesting development of reinforcement learning has been to include the reaction time in the reward rate (Bogacz & Larsen 2011) as an
incentive to make a choice as fast as possible.

Models also provide the possibility to simulate diseases and lesions, with PD being the prominent subject of interest, stressing the relation with biology (Frank, 2005; Frank et al., 2004), see Maia & Frank (2011), Delong & Wichmann (2009) and Wiecki & Frank (2010) for a review.

Conditioning phenomena have been a major focus of early models, based on Hebbian learning rules, enabling relatively straightforward sequence learning (Berns & Sejnowski, 1998). Models based on TD learning became widely used to explain what has become a major focus of attention: the phasic dopaminergic signal described by Schultz et al. (1997). This followed the hypothesis that the dopaminergic neurons activity, the plasticity of cortico-striatal connections, the TD signal, and the striatum organisation are all linked (Berns et al., 2001; Houk et al., 1995; Montague et al., 1996; Suri, 2002; Suri et al., 2001; Suri & Schultz, 2001). In particular, Houk et al. (1995), suggested a modular organisation of the striosomes as specific reward predictors. However, they assumed a sub-compartmentalisation of the striosomes based on reciprocal and moreover specific connections from the dopaminergic neurons. The actor-critic framework was nevertheless a popular way to bridge the gap between theoretical TD models and experimental data (see section 2.3.3.3 and Joel et al. (2002), for a review). Such models are able to mimic the dopamine signal during various conditioning paradigms, such as the ramping of activity before an expected reward and the burst at the occurrence of the CS (Niv et al., 2005). Also based on TD learning, the function of the multiple cortico-striato-nigro-thalamo-cortical loops has been studied in another model, suggesting a spread of limbic and cognitive information to motor loops by spiral connections (Haruno & Kawato, 2006). This could be supported by the topological connections to striosomes and matrisomes (section 3.3.1). O’Reilly & Frank (2006) presented a model derived from the TD algorithm where the actor would be trained by the critic and could control motor output, and where dopaminergic projections to PFC could gate the updating of the working memory. The decision to maintain and to update working memory representations is based on the reward-related information provided by the dopamine signal.

One of the first and most influential computational model of the BG credits them with the role of action selection (Gurney et al., 2001a,b), following the author’s earlier hypothesis that the BG deal with the allocation of restricted resources to competing systems (Redgrave et al., 1999). It assumes and implements a global disinhibition of GPi/SNr by activity in STN and a focal inhibition sent from striatal connections, effectively inhibiting all populations, or promoting a specific population, respectively. It can be noted that the model already implements LIF spiking neurons. However, the model does not include learning, dopamine only impacts the activation of the MSNs. Therefore, the
authors limited their discussion to the way the selection is affected by various modifications such as an increased or decreased level of dopamine, or by varying the inhibition / excitation rate in the model. A contrasting view has been offered, attributing a different role to the BG: to compress cortical information according to a reinforcement signal (Bar-Gad et al., 2000, 2003).

More recently but adapted from Gurney et al. (2001a), Stewart et al. (2012) detailed a model of the BG with spiking neurons and using the delta rule for learning. The spike patterns observed in the model show good similarity with those of experimental recordings in the ventral striatum of rats, in a two-armed bandit task. The authors suggested that an implementation similar to Potjans et al. (2009) of slowly decaying Q-values, enabling the inclusion of the estimate of the following state’s value, would allow their model to learn to associate states that occur sequentially in time. They furthermore mentioned the need to incorporate the ability to make precise temporal predictions. Such a feature is not the main focus of this thesis but is nonetheless a critical property in reinforcement learning. Dopamine has been suggested to be a prominent component in such timing mechanisms (Daw et al., 2006; Doya, 2000b; Gershman et al., 2014; Rivest et al., 2010; Schultz, 2007a; Suri & Schultz, 1998). Coincidence detection of oscillatory processes has notably been suggested as a possible candidate (Buhusi & Meck, 2005; Matell & Meck, 2004).

As a further development from their 2009 model, Potjans et al. (2011) implemented a more realistic dopaminergic population, in which firing rate coded for the RPE. The authors suggested that learning from negative rewards is impaired compared to the classical TD algorithm. They however made some assumptions regarding the connectivity of the direct and indirect pathways for the actor, which do not seem to be supported by experimental data. This model was further extended to enable learning from negative reward by segregating the impact of the level of dopamine on the plasticity of the D1 and D2 MSNs (Jitsev et al., 2012). They also implemented a direct and indirect pathway, based on the dopamine receptor type expressed by the MSNs, in both the critic and the actor.

The implementation of both synaptic plasticity and dopamine-dependent sensitivity of MSNs has led models to capture the impact of the neuromodulator on the plasticity, the choice incentive and the reaction time (Collins & Frank, 2014). In their model, the weights in the two pathways are updated symmetrically depending on the sign of the RPE.

4.3 Bayesian framework

The learning rule is the mathematical theory at the basis of the modifications that can occur in artificial neural networks in order to improve their output and
performance, usually through repetition and by updating the weights and bias values of the connections (see section 2.2 for some overview). We will first introduce the concepts leading to the formalisation of our learning rule.

4.3.1 Probability

Probability is a measure of the likelihood of an event to occur. It is bounded between zero and one, where one means absolute certainty and zero impossibility. The classical view of probability is the frequentist view. Here, there is no information prior to the model specification, thus, phenomena are either known or unknown. Repeatability is key to this approach, as the probability is the ratio of having a particular outcome out of all the possible outcomes, and would thus converge to a number between zero and one as the number of observations increase. It focuses on $P(X|Y)$, that is, the probability of the data $X$, given the hypothesis $Y$. The hypothesis is usually the null hypothesis, and is fixed.

The alternative consideration is the “Bayesian” view, where probabilities are seen as a measure of the belief about the predicted outcome of an event (Doya et al., 2007). Conversely to the frequentist approach, unknown quantities are here treated probabilistically and it is the data that is fixed. The hypothesis can be true or false, with some probability between 0 and 1. Therefore, the Bayesian view is centered on $P(Y|X)$, the probability of the hypothesis given the data.

4.3.2 Bayesian inference

There are two sources of information in Bayesian theory, data and prior knowledge. Data is the current sensory evidence about the environment. Prior knowledge relies on the memory of past similar experiences. These two sources are combined in a optimal way to generate a prediction about the outcome of a considered event. The imperfection of our own sensory system would thus call for such inference. It has been shown that Bayesian inference is what human do when they learn new movements and skills, in a model-based schema (Körding & Wolpert, 2004; Wolpert et al., 1995), similar to statistical approaches (Griffiths & Tenenbaum, 2006). The brain could thus represent the world in terms of probabilities, where prior knowledge of a distribution of events could be combined with sensory evidence to update the posterior distribution (Friston, 2005; Yang & Shadlen, 2007). The use of Bayesian inference in the study of the brain is relatively recent, and has been applied with the idea that it could describe how the brain deals with uncertainty. This uncertainty is inherent to the world and to the noise from sensory inputs. Furthermore, it has been shown
that Bayesian probabilities can be coded by artificial neural networks and spiking neurons (Boerlin et al., 2013; Buesing et al., 2011; Deneve, 2008a; Doya et al., 2007). The finality is assumed to be the posterior probability $P(Y|X)$, which is the probability of an hypothesis, or belief, given the evidence. Bayes’ theorem derives it as a consequence of a prior probability $P(Y)$ and a likelihood function $P(X|Y)$:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \quad (4.2)$$

where $P(X|Y)$ is the likelihood, that is, the prediction that the sensory evidence observed can occur given the hypothesis engaged. $P(Y)$ is the prior probability, that is the probability, often deduced from experience, that this hypothesis is true before integrating new sensory evidence. It relates to the distribution of the possible states of the world. The sum of all the hypotheses under consideration should be one. $P(X)$, the evidence, is often called the marginal likelihood (and is non-zero) and it acts as a normalising factor. It is the same for all the hypotheses considered. It can thus be noted that if we have uniform priors, then the posterior probability relies on the likelihood. Conversely, if the likelihood for all the sensory evidence is the same, the estimate of the posterior probability is made based on the priors. Generally however, the posterior probability depends both on the prior and on the likelihood. Bayes’ theorem states how a belief of a certain hypothesis $P(X)$ should be updated, based on how well the evidences were predicted from the hypothesis $P(X|Y)$.

Additionally, the relationship between the joint and the conditional probabilities can be noted:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \quad (4.3)$$

If the variables $X$ and $Y$ are independent, then their joint probability is reduced to the product of two probabilities:

$$P(X,Y) = P(X)P(Y) \quad (4.4)$$

and we thus obtain:

$$P(X|Y) = P(X), \quad P(Y|X) = P(Y) \quad (4.5)$$

The variables are otherwise said to be dependent if this condition is not met.

The ability to predict is a marker of an internally generated representation of the world, based on the experience a subject has gained from previous associations of events. If the probability of the co-occurrence of two stimuli is greater than predicted by the probability of their isolated occurrence, it might be valuable to form a memory of such relation. Updating this internal record of
probability with each perception of a stimulus enables the representation of the world to steer away from confusion and disjointed events. Prior occurrences of elements of a perceived stimuli association have to be considered in order to infer if they form truly an association or of they are only random events. This details how prior knowledge can impact the sensory experience. Furthermore, events that are fully predictable are ignored, as they do not trigger a modification of the belief system. However, this can lead to problem when strong beliefs can lead to overlooking informative evidence. Ultimately, abnormality in perception can arise from an unadapted representation of the world, and vice versa (see section 3.4.3 for a description on how this can be relevant in psychiatric diseases such as schizophrenia).

Predictions are not only made about when events may occur, but also about the relation they share. Letting aside the normalising denominator $P(X)$ in eq. 4.2, the posterior probability $P(Y|X)$ is proportional to the product of the likelihood $P(X|Y)$ with the prior probability $P(Y)$. The hypothesis about the model that maximises the posterior probability is called the maximum a posteriori (MAP) estimate. An interesting way of using the posterior probability is to feed it as the prior probability in the following step. For example, if the world changes as independent sensory observations $x = (x_1, x_2, \ldots, x_t)$ are made, having knowledge about this evolution, a transition probability $P(Y_t|Y_{t-1})$, makes it possible to set the new prior at time $t$ as the posterior at time $t - 1$ multiplied by this transition probability:

$$P(Y_t|x_1, \ldots, x_{t-1}) \propto P(Y_t|Y_{t-1})P(Y_{t-1}|x_1, \ldots, x_{t-1})$$ (4.6)

The following gives the sequence of the Bayesian estimation:

$$P(Y_t|x_1, \ldots, x_t) \propto P(x_t|Y_t)P(Y_t|Y_{t-1})P(Y_{t-1}|x_1, \ldots, x_{t-1})$$ (4.7)

This recursive estimation is called the Bayes filter, which is a widely used probabilistic approach for estimating an unknown probability density function recursively over time.

We will detail next how these probabilistic representations can be used to quantify information.

4.3.3 Shannon’s information

As we have mentioned, the probability of occurrence is linked with how predictable an event is. The observation of a particular value $x$ of a random variable $X$ with probability $P(X)$ defines how informative this observation is. Indeed, if $P(X = x)$ is high, there is not much of a surprise to it, however, if $P(X = x)$ is low, then observing it is informative. This can be related to the
RPE dependent plasticity of reinforcement learning models (section 2.2.2) and to the plasticity observed at cortico-striatal synapses (section 3.3.4). In order to quantify the information of the event $X = x$, it is convenient to take the logarithm of the inverse of the probability:

$$
\log \frac{1}{P(X = x)} = - \log P(X = x) \quad (4.8)
$$

Thus, for a fully predicted outcome, $P(X = x) = 1$, information is zero and increases as $P(X = x)$ becomes smaller. Additionally, taking the logarithm enables us to measure the information of two independent events $x$ and $y$, with joint probability $P(x, y)$, by the sum of each event:

$$
\log \frac{1}{P(x, y)} = \log \frac{1}{P(x)P(y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)} \quad (4.9)
$$

It can be noted here that the unit of information resulting from the use of a binary logarithm is called a **bit**.

An additional way of quantifying the information is through the measure of the amount of uncertainty associated with the value of $X$, also called the entropy of $X$, $H(X)$:

$$
H(X) = - \sum_{x \in X} P(x) \log_2 P(x) \quad (4.10)
$$

Therefore, for deterministic conditions, i.e. $P(X = x) = 1$ and $P(X \neq x) = 0$, entropy is zero, $H(X) = 0$, and is maximal, $\log N$, for a uniform distribution over $N$ values. The mutual information $I(X; Y)$ given by the sensory input $Y$ about the world state $X$ can thus be determined by the difference between the entropy of the *a priori* distribution of the inputs, $H(X)$, and the entropy of the *a posteriori* distribution of the world state given the stimulus, averaged over all the repetitive presentations of the same stimulus, $H(X|Y)$, also called the conditional entropy:

$$
I(X; Y) = H(X) - H(X|Y) \quad (4.11)
$$

This details the source of the variability observed in the entropy. Further description of entropy and information for spike trains can be found in [Dayan & Abbott, 2001; Doya *et al.*, 2007; Rieke *et al.*, 1997].

### 4.3.4 BCPNN

Summing up on the previous sections, the use of a Bayesian approach brings several benefits in the quest to understand the brain. First, it predicts how an ideal perceptual system would combine prior knowledge with sensory information. Secondly, algorithmic formulation of the Bayesian estimations can
provide interpretations on the functional computations of neural systems, e.g. neural decoding. Thirdly, it offers a method to optimally decode neural data, which can be noisy, given certain nature of signal and noise (Doya et al., 2007).

The Bayesian confidence propagation neural network (BCPNN) is an artificial neural network implementing a Hebbian learning rule inspired from a naïve Bayes classifier, and was developed in analogy with the Hopfield model (Lansner & Ekeberg, 1989). Probabilities or confidences of observing certain events are represented by the activity of the corresponding units. Critically, the synaptic weight between two neurons represents the statistics of their activations and co-activations, and the BCPNN learning rule describes how these weights are computed.

The standard BCPNN learning rule is a Hebbian unsupervised, associative, two-factor learning rule. It has been successfully implemented in a variety of tasks such as associative and working memory (Lansner, 2009; Meli & Lansner, 2013; Sandberg et al., 2000), memory consolidation (Fiebig & Lansner, 2014), attentional blink (Lundqvist et al., 2010; Silverstein & Lansner, 2011), olfaction modeling (Kaplan & Lansner, 2014) and reinforcement learning (Johansson et al., 2003). Obviously, in a reinforcement learning paradigm there is a need for an additional component that can alter and bias these probabilities in order for the system to adapt to the appropriate behavior. This has been done through the implementation of the learning rate \( \kappa \). It initially resulted from the need to have some basic control on the plasticity window, and allowed to turn the plasticity on, \( \kappa \neq 0 \), and off, \( \kappa = 0 \). However, it also brought the capacity for the model to take into account a reinforcement signal.

In the model presented in this work, we focus on the connections between state coding populations and action coding populations, and their weights represent the probability that, given that the system is in a specific state, a considered action will be the one selected. Therefore, by coupling the learning rate to a set reward mapping, it becomes possible to bias the plasticity towards the desirable behaviour. The learning rule thus evolved from a two-factor to a three-factor learning rule. \( \kappa \) is a global parameter that affects all the BCPNN weights, and can take any non-negative value. It controls the impact of the recent correlations on the weights.

The goal for the model is to, basically, learn to select the action that leads to the reward. Hence, if the selected action is not rewarded, one may wants to decrease the probability of doing it again under the same circumstances. Conversely, if some reward is obtained, the weight should be increased. These plastic weights correspond to the cortico-striatal synapses in the BG and our model features a D1 and a D2 pathway, in charge of promoting and suppressing, respectively, the selection of an action. Additionally, a similar process occurs in the RP pathway, corresponding to the strioso-dopaminergic nu-
clei pathway described in biology, to compute the confidence with which a specific action, given a specific state, will lead to a reward.

We will first describe the standard BCPNN before going into the specifics of our model of the BG. The BCPNN learning rule can be implemented with spiking or non-spiking neural networks. We will here focus on the spike based BCPNN, implemented in our most detailed model, but the general principles are similar to the more abstract implementation (see Holst (1997); Lansner & Ekeberg (1989) for a specific description of the non-spiking version). Results form the spike-based learning rule support the view that neurons can represent information in the form of probability distribution (Tully et al., 2014). The spiking BCPNN learning rule differs from STDP in that the precise order of firing of the pre- and postsynaptic neurons does not necessarily impact the sign of the synaptic changes. The plasticity depends more on the correlated activity over a defined time interval, where correlated activity would increase the connection strength whereas uncorrelated activity would decrease them. Moreover, the use of eligibility traces allow for delayed learning (Tully et al., 2014). This probabilistic approach provides a possible explanation as to how the brain can deal with uncertainty. Bayesian inference is indeed well suited to detail how the incoming information can be integrated with previous knowledge.

The BCPNN has been classically used with a columnar architecture, which is most commonly discussed in the context of the cortex. This modular organisation, i.e. hypercolumns and minicolumns, has been shown to exhibit biological relevance and thus fit functional and anatomical cortical structures, which feature several layers and multiple inhibitory processes (section 3.1.4). This generic model of the cortex considers minicolumns as the smallest functional units, organised in hypercolumns that regulate their activity by feedback inhibition. Hypercolumns are, in turn, organised into cortical areas that build up the cortex. The columnar structure of the network arises naturally from the derivation of a naïve Bayesian classifier, as it requires the representation of mutually exclusive events (see (Holst, 1997; Johansson, 2006; Lansner & Holst, 1996; Sandberg, 2003) for rigorous mathematical investigations).

Usually, the standard BCPNN is described in the context of a classification task. It is used to find the class \( y_j \), e.g. a specific animal, which is the most likely to explain or account for the various attributes or features observed \( x_1 \ldots n \), where \( x_i \) could code for a specific color or shape for example. In the work presented here, different classes represent possible actions, and the aim is to compute the probability of an action \( y_j \) to be selected given the information about the state of the environment. The firing rates of \( n \) pre-synaptic mini-
columns $x_{1...n}$, i.e. $P(x_{1...n})$, provide information about the firing probabilities of neurons in the postsynaptic minicolumn $y_j$, via Bayes’ rule:

$$P(y_j|x_{1...n}) = P(y_j) \frac{P(x_{1...n}|y_j)}{P(x_{1...n})} \tag{4.12}$$

Assuming the $n$ input attributes as independent both conditionally and unconditionally given $y_j$, the probability of the joint outcome $x_{1...n}$ can be defined as a product:

$$P(x_{1...n}) = P(x_1) \cdot P(x_2) \cdots P(x_n) \tag{4.13}$$

and so can the probability of $x_{1...n}$ given each action $y_j$:

$$P(x_{1...n}|y_j) = P(x_1|y_j) \cdot P(x_2|y_j) \cdots P(x_n|y_j) \tag{4.14}$$

With that, Bayes’ rule, eq. 4.12, can be extended to:

$$P(y_j|x_{1...n}) = P(y_j) \prod_{i=1}^{n} \frac{P(x_i|y_j)}{P(x_i)} \tag{4.15}$$

Here, as the denominator is the same for each action $y_j$, the assumption of independent marginals above is trivial.

Taking the logarithm of this gives a sum which can conveniently be implemented in a neural network as it gives a linear expression in the contribution from the attributes:

$$\log P(y_j|x_{1...n}) = \log P(y_j) + \sum_i \log \left( \frac{P(x_i,y_j)}{P(x_i)P(y_j)} \right) \tag{4.16}$$

Hypercolumns are supposed to represent specific attributes, e.g. color or speed, and minicolumns code their specific values, e.g. blue or fast. Furthermore, being comprised of a cluster of minicolumns, the hypercolumns are convenient modules to control the activity among its various minicolumns. $\pi_{xh_i}$ denotes the relative activity representing the confidence in the presence of the attribute value $x_{h_i}$. A normalisation of this activity in each hypercolumn can be obtain through a strict or soft WTA connectivity within hypercolumns.

Therefore, assuming the modular network topology with $H$ hypercolumns, each containing $n_h$ minicolumns, eq. 4.15 can be written as:

$$\log P(y_j|x_{1...n}) = \log P(y_j) + \sum_{h=1}^{H} \log \sum_{i=1}^{n_h} \left( \frac{P(x_{h_i},y_j)}{P(x_{h_i})P(y_j)\pi_{xh_i}} \right) \tag{4.17}$$

However, in our spiking neural network, it is possible to consider a network topology without hypercolumns, or else that it consists of a single hypercolumn for each state and action population. In the latter case, the normalisation...
is guaranteed. However, it has to be a separate mechanism implemented for the action population. Additionally, it has been suggested that the sum of the contributions from $H$ hypercolumns can be approximated by one term, hence removing the $h$ index. This condition considers that it is statistically unlikely that a specific hypercolumn would receive inputs from more than one other hypercolumn (Tully et al., 2014), given the sparse connectivity in conjunction with the number of possible hypercolumns in the cortex and the number of synapses per neuron. Alternatively, a single minicolumn can be made active within a hypercolumn, by enforcing WTA. Therefore, all this leads to the simplification of eq. 4.17 where a single element coding for the dominant feature replaces the sum over all the minicolumns activities. It is possible to define the support value $s_j$ as the sum of all the $N = Hn_h$ contributions for $y_j$ with weights $w_{x_iy_j}$ and bias $\beta_j$:

$$s_j = \beta_j + \sum_{i=1}^{N} \pi_i w_{x_iy_j}, \quad \beta_j = \log P(y_j), \quad w_{x_iy_j} = \log \left( \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right) \quad (4.18)$$

Finally, a normalisation of the support values using an exponential transfer function furthermore returns the corresponding posterior probabilities.

We will now detail how the probabilities in eq. 4.18 are estimated from the activity of spiking neurons. The spike trains are transformed into traces by exponentially weighted moving averages (EWMA). As new information is learned, old memories decay and get gradually replaced by the most recent evidence. A three-staged EWMA is performed on the spike trains of the presynaptic $i$ and postsynaptic $j$ neurons. Two stages are needed in order to obtain the joint activity of the neurons based on the first EWMA. The addition of an intermediary trace enable learning in situations with delayed reinforcement (Tully et al., 2014). The pre- $S_i$ and postsynaptic $S_j$ spike trains are described by summed Dirac delta pulses with respective spike times $t^i$ and $t^j$ where $t^i_{sp}$ ($sp = 1, 2, \cdots$) are firing times of neuron $i$:

$$S_i(t) = \sum_{sp} \delta(t - t^i_{sp}), \quad S_j(t) = \sum_{sp} \delta(t - t^j_{sp}) \quad (4.19)$$

$Z_i$ and $Z_j$, the traces with the fastest dynamics, i.e. their time constants $\tau_{z_i}$ and $\tau_{z_j}$ are the shortest, are defined as exponentially smoothed average of the spike trains:

$$\tau_{z_i} \frac{dZ_i}{dt} = \frac{S_i}{f_{max} \Delta t} - Z_i + \varepsilon, \quad \tau_{z_j} \frac{dZ_j}{dt} = \frac{S_j}{f_{max} \Delta t} - Z_j + \varepsilon \quad (4.20)$$

which low pass filters the pre- and postsynaptic activities with the defined time constants. $f_{max}$ and $\varepsilon$ define the higher and lower bound, respectively, of the
firing rate, representing the maximal certainty and doubt, respectively, in the presence of the feature coded by the associated neural population. Firing rates within that range represent the estimated probability. Each spike event had a duration of $\Delta t = 1$ ms. Based on the $Z$ traces, the $E$ eligibility traces are computed, and a new equation is introduced to take into account the coincident activity of the $Z$ traces:

$$
\tau_e \frac{dE_i}{dt} = Z_i - E_i, \quad \tau_e \frac{dE_j}{dt} = Z_j - E_j, \quad \tau_e \frac{dE_{ij}}{dt} = Z_i Z_j - E_{ij} \quad (4.21)
$$

$\tau_e$ is the time constant for these $E$ traces. Next, the $P$ traces are EWMA of the $E$ traces with the slowest time constant $\tau_p$:

$$
\tau_p \frac{dP_i}{dt} = \kappa(E_i - P_i), \quad \tau_p \frac{dP_j}{dt} = \kappa(E_j - P_j), \quad \tau_p \frac{dP_{ij}}{dt} = \kappa(P_{ij} - E_{ij}) \quad (4.22)
$$

Ultimately, the $P$ traces are used to compute $w_{ij}$ and $\beta_j$, from eq. 4.18:

$$
\beta_j = \log(P_j), \quad w_{ij} = \log\left(\frac{P_{ij}}{P_i P_j}\right) \quad (4.23)
$$

Fig. 4.1 shows the evolution of the traces, weights and bias in a toy example featuring a pre- and a postsynaptic neuron.

The different trace dynamics can be related to biological processes. The fast time constant of $Z$ traces could be linked with rapid $Ca^{2+}$ influx via ion channels. The $E$ and $P$ traces could be accounted by downstream cellular processes depending on the intracellular $Ca^{2+}$ concentration and by gene expression and protein synthesis, respectively. The bias value is implemented in the neuron model as an activity-dependent current, modifying its intrinsic excitability.

A feature of the dual pathway implementation in our model of the BG is the possibility to compute odds ratio, which are often used in decision making. Here it corresponds to the ratio of the probability of selecting one action, that is its D1 pathway promotion, over the probability of this action being avoided, that is the corresponding D2 pathway inhibition. We have not investigated the potential of such computations in this work but log-odds ratio have been mapped to membrane potential of neurons (Buesing et al., 2011) and it has been shown that spiking neurons could estimate log-odds ratio (Deneve, 2008b).

Finally, one of the BCPNN characteristic is that weights can alternate from positive to negative values, and vice versa, depending on the correlation of the activity between the pre- and postsynaptic units. Even if the release of excitatory glutamate and inhibitory GABA from a single axon terminal has been
Figure 4.1: Example of the evolution of the BCPNN traces, weight and bias in a simulation with a pre-, $i$ (blue traces on the left column), and a postsynaptic, $j$ (green traces), neuron. Red traces represent the co-activity of these neurons. $\kappa$ is initially set to one, allowing the development of the $P$ traces and of the weight $w_{ij}$ and bias $\beta_j$. When the traces of the neurons overlap, attesting some correlated activity, $w_{ij}$ increases. At $t = 750$ ms, $\kappa$ is set to zero, effectively freezing the update of the $P$ traces, weight and bias, but not of the $Z$ and $E$ traces.
Figure 4.2: Schematic representation of a possible mapping onto biology of the BCPNN weights, where the weights here represented do not change sign. The minimal configuration requires only plasticity on the weight $w_{\text{pre,post}}$, with spontaneous firing from the inhibitory neurons. Alternatively, it is possible that all the weights could be plastic. In the middle ground of this spectrum, both $w_{\text{pre,post}}$ and $w_{\text{pre,inh}}$ could be plastic and could decrease to zero.

recently reported in biology \cite{Root2014, Uchida2014}, it is still considered as an uncommon property. A suggested option is thus to either consider the weights as not being specific to a neuron, but instead to represent a module involving both excitatory and disynaptic inhibitory connections, or restrict the use of the weights to positive values only. In the former suggestion, positive weights would represent a net excitatory effect of the sum of the excitatory and inhibitory input to the projecting neurons (fig. 4.2). This can be achieved through various setups. One is to assume a spontaneous firing of the inhibitory interneuron, where a negative weight of the BCPNN would imply a lower excitatory component than this baseline inhibition. One caveat of such implementation is that, in order to represent a zero BCPNN weight, the excitatory contribution from the presynaptic neuron has to equal to the inhibition that the postsynaptic neuron receives from the interneuron. Alternatively, one could assume plasticity of both the $w_{\text{pre,post}}$ and $w_{\text{pre,inh}}$. Each of them could decrease to zero, representing the deletion of the synapse or a way to inactivate it. With regard to the BG, this would correspond to cortico-striatal connections, with presynaptic cortical neurons, postsynaptic MSNs and fast-spiking interneurons providing the inhibition.

4.4 Learning paradigms used

We will here describe the tasks on which the models are tested. We will first present a basic reinforcement learning task, successive learnings, which is sim-
ilar in all our models and works, before mentioning related tasks. We will then
detail delayed learning, which we only test in the spiking version of the model.

4.4.1 Common procedures

A simulation typically consists of several blocks, themselves made up of sev-
eral tens of trials. Within a block, the reward mapping is held constant and
there is usually only one action for each state that leads to the reward. The
system has to first find the correct action for each state, but then, as the reward
mapping changes, it has to learn to stop selecting the previous correct action
to then discover the new correct one. For example, in the spiking model which
comprises three states and three actions (section 5.1), a reward is obtained
when the model, in state $i$, selects the action $j$ that verifies:

$$((i + b) \mod a) \equiv j$$

where $a$ is the number of actions, here $a = 3$, and $b$ is the block number.

Based on the activity of the output layer, a selection is made through a
softmax function. The outcome is computed with the equation above, eq. 4.24,
and the difference between this value and the expected reward, the RPE, is
conveyed to all dopamine-dependent synapses.

The same protocol is applied for reversal learning but there $b$ is replaced
by $b \mod 2$. In effect, the reward mapping alternates between the two same
configurations.

Reacquisition after extinction consists of a first acquisition of a reward
mapping, that is during the first block, while no reward at all is given during
the second block. The length of this block can vary in order to assess if the
length of the extinction has a effect on the reacquisition. Thus, in the third
block, rewards are again delivered and the reward mapping is the same as in
the first block.

In probabilistic reward learning, or multi-armed bandit tasks, the reward is
not always delivered for correct trials. Furthermore, several actions per state
can lead to a reward but with different probabilities. This experimental setup is
often used with animals, in conjunction with a block-dependent reward map-
ning.

4.4.2 Delayed reward

The delay refers to the time between the activity of the neural population cod-
ing for the action that has led to the obtention of the reward, and the delivery
of the reward. Spikes are by nature very short temporal events and thus a
memory of the activity is required. A mechanism that allows for learning in
this situation is also involved in secondary conditioning. Through the use of
temporal traces, as mentioned in the previous section (4.3.4), it is possible to
enable the overlap of the traces of the state and of the action with the reward
signal, e.g. the phasic change of the dopaminergic neuron activity. There is
however a chance, scaling with the delay, that in the interval, other neural pop-
ulations become active and would thus also have their traces overlapping with
the reward. The correct attribution is therefore complex, and is called the distal
reward, or temporal credit assignment problem (see section 2.3.3.4 and Izhike-
vich (2007)). BCPNN learning can deal with delays by keeping a memory of
the activity in the $E$ traces. Associations can be made until their relative traces
have decayed. Similarly, the $Z$ traces of the pre- and postsynaptic neurons need
to overlap in order for the $E_{ij}$ trace to pick up the joint activity. As the traces
decay with time (fig. 2.3), the strongest ones are related to the most recent
firing. Therefore, when the reward occurs, all the non-zero traces of the states
and actions temporally close enough to the reward will be linked to it, increas-
ing the probability of selecting these same actions in the given states, with the
closest being the most strongly associated. Through trial and error, similarly
to TD learning, the model should eventually find the responsible state-action
pair that causes the reward, after having tested out all the other pairs occurring
in between.
5. Results and Discussion

In this chapter we will present and discuss some of the results of our works, including our models. We will consider specifically the implication of these results with regard to the biological data and hypotheses mentioned in section 3.3 and to other theoretical frameworks. Additionally, we will also introduce results from unpublished works. As a note, in papers I, II and III, the Go, direct and D1 pathways represent in fact the same concept, and this is also the case for the NoGo, indirect and D2 pathways.

5.1 Implementation of our models of the BG

We will here present the models used in the papers, through their common features, while focusing on the more complex implementation of the spiking network. We will now bring together the information previously exposed. The description of the model will thus mix theoretical and biological denominations. Additionally, we will briefly describe what happens in the models during a trial.

5.1.1 Architecture of the models

The models are basically composed of three pathways, with plastic connections using the BCPNN learning rule in each of them. The different states, $N_{states}$, and actions, $N_{actions}$, are coded by different subpopulations of neurons.

In the spiking model, within a subpopulation, all neurons code for the same representation, e.g. the same action or the same state. It is the equivalent of a unit in the abstract model. A population is defined as the union of the subpopulations of the same kind, e.g. all the action subpopulations.

Let us first describe two out these three pathways, the D1 and the D2 ones, they are used for the action selection, as commonly described in biology. They both originate from the striatum, and specifically from the matrisesomes. They differ on the dopamine receptor type that the MSNs, from which they both stem, express: either D1 or D2 (see section 3.3 for a description of the biology). In relation to biology, the striatal MSNs in the D1 and D2 pathways are, in our model, part of the matrix and are involved in the action selection, whereas
striosomal MSNs are involved in the RP.

Activity in the D1 pathway increases the probability of selection of the related action, whereas activity in D2 would prevent it (fig. 5.1). The different structures of the BG are here abstracted to the hypothesised functional level in a two layer network, thereby simplifying the polysynaptic projections of disinhibition (D1 pathway) and of the dis-disinhibition of excitatory neurons (D2 pathway). Here, the abstract model and the spiking network differ, as in the former, the D1 and D2 pathway converge onto the same unit, whereas there exists separate D1 and D2 populations in the latter, whereby all actions are coded twice, once by each D1 and D2 population. In this setup, the convergence occurs in an output layer which preserves the topology of the previous layer, and receives inhibitory and excitatory inputs from the D1 and D2 populations in striatum, respectively. It thereby corresponds to the GPi/SNr.

The third pathway, the RP pathway, computes the expected reward, specifically for each state-action pair. In this pathway, plastic synapses connect the state-action pairs coding subpopulations to the RPE coding population, here the dopaminergic neurons. There are \( N_{\text{states}} \times N_{\text{actions}} \) subpopulations coding for all the possible state-action combinations in the RP pathway. Therefore, each state connects to \( N_{\text{actions}} \) subpopulations, each of them coding for a different action but for the same state. Similarly, the information about the action is sent for each action to the \( N_{\text{states}} \) possible states coded in the RP pathway (fig. 5.2). In the D1 and D2 pathways, a subpopulation represents one action and we associate it to a matrisome, but in the RP pathway it codes for a specific state-action pair. A striosome is here defined as regrouping all the RP subpopulations coding for a common action. The number of striosomes therefore equals the number of actions in the model, \( N_{\text{actions}} \). We have thus assigned the RP pathway to the striosomo-nigral pathway. The RPE is simply the difference between the inhibition received from the RP pathway and the excitation triggered by the reward.

The principle, shared by the abstract model, is to increase the relevant D1 weights, which are the ones from the current state to the selected action, when the obtained reward value is superior to its expected value, \( \text{RPE} > 0 \), and at the same time to also increase the weights from the associated active state-action pair to the RPE coding population in order to adjust the prediction for the next occurrence of the situation. Conversely, the D2 weights should be increased if less reward than expected is delivered, \( \text{RPE} < 0 \), as the weights in the RP pathway should be decreased to improve the estimated reward when in the same conditions. Once the reward is fully expected, no updates take place as the RPE equals zero, preventing any modification of the \( P \) traces, and therefore of the weights and biases.

Our implementation of the BCPNN learning rule in a BG model shares
Figure 5.1: Schematic representation of the model with respect to the biological structures mapped. For example, action ‘A’ is coded twice in the matrisomes, by both MSN D1 and MSN D2 subpopulations. These MSNs both project to a subpopulation also coding for action ‘A’ in the GPi/SNr. The selection is done in this layer, based on the activity of the different action coding subpopulations. In the RP pathway, subpopulations code for specific state-action pairs, such as state ‘1’ and action ‘A’. The RPE is however non specific and affect all the plastic connections.
similarities with TD implementations and Actor-Critic models in reinforce-
ment learning. The purpose of the D1 and D2 pathways is to select an action
based on action values, and these pathways therefore endorse the role of the
actor, while the striosome-dopaminergic nuclei loop computes the RPE, which
is comparable to the TD error, and can therefore be seen as the critic. However,
the difference lies in that the values in our model result from BCPNN compu-
tations and are thus derived from the statistics of activation and co-activation
of the neurons. In that way, the values do not represent estimates of future
rewards, but probabilities of the postsynaptic neuron to be active given that
the presynaptic one is active, for the D1, D2 and RP pathways. In our model,
the value function of the latter relies not only on the state, but also takes into
account the action selected to emit a prediction based on this more precise
information. The actual selection is done through a softmax process on the ac-
tivation in the output layer, the GPi/SNr for the spiking model, and the action
layer in the abstract version.

The RPE, even if it is specific to the particular state-action pairing, can
affect all the dopamine dependent plastic connections. The update of the
weights thus depends on the value and sign of the RPE, but also on the type of
dopamine receptor expressed by the neurons, and on the activity of the pre- and
postsynaptic neurons. The RPE is used for the modification of the strengths of
the connections of the D1, D2 and RP pathways.

Here, $\kappa$ in eq. 4.22 is substituted with the RPE as the learning rate. The
updates of the $P$ traces, used for the computation of the weights, in the different
pathways are dependent on the sign of the RPE and on the dopamine receptor
type expressed by the matrisomes.

Now we will specifically detail the computation of the RPE and its impact
on the updates in the spiking model. The weights of the D1 and D2 pathways
are updated when $RPE > 0$ and $RPE < 0$, respectively. However, the weights
of the striosome-nigral connections are updated irrespectively of the sign of the
RPE. The dopaminergic neurons are driven to fire at baseline level in the
absence of inputs from the RP pathway or from the external reward informa-
tion. At such baseline activity, the resulting RPE is thus set to be equal to zero
(eq. 5.1). The relative lack of data on the biological substrate of the critic, i.e.
the RP pathway, has led us to keep a relatively abstract implementation of this
circuit, with the goal to investigate more its implication in the performance.
The RPE is defined as:

$$RPE = \kappa = (\sigma_{dopa}(\beta_{dopa} + q))^\lambda$$

(5.1)

where $q$ is the filtered trace of the spike activity of the dopaminergic neurons,
hereby acting as a proxy for the dopamine level in the model. $\beta_{dopa}$ acts as a
bias to shift the RPE to zero at the normal tonic activity of the dopaminergic
neurons ($\approx 10$Hz). $q$ evolves following:

$$\tau_q \frac{dq}{dt} = \sum_{sp} \delta(t - t_{sp}^{dopa}) - q$$

(5.2)

where $t_{sp}^{dopa}$ are the spike times of the dopaminergic neurons. $\tau_q$ is the time constant for the trace, $\sigma_{dopa}$ and $\lambda$ are the gain and the power. We used an odd number for $\lambda$ in order to keep the information about the sign in matrices as it is critical to determine the localisation of the plasticity in the pathways, and an even number $\lambda$ for striosomes. These computations are done using the volume transmitter implementation in NEST, which enables the synaptic plasticity to be dependent on a non-local neuromodulatory signal, through the integration of spikes from a population of neurons, here the dopaminergic neurons, connected to it (Potjans et al., 2010).

Depending on the action selected, the trial might lead to a reward. For example, in the spiking model, a correct selection leads to a transient increase of the external stimulation of the dopaminergic neurons, whereas an incorrect one triggers a transient decrease of this stimulation.

Now, in the abstract version of the model, when $RPE<0$ or $RPE>0$, the D1 or D2 weights, respectively, between the units coding for the active state and the selected action are decreased. In order to obtain a decrease between co-active units, or an increase between units without correlated activity, the activation vector of the actions of the relevant pathway is changed to its complementary, effectively activating the otherwise silenced other action coding units, and setting to 0 the unit coding for the selected action. The spiking implementation of the learning rule proves to require non-trivial modifications to be able to compute complementary traces of activation, which would be used instead of the original, ‘real’, traces of activation of the neurons in the situations just mentioned.

Both models also uses an efference copy which increases the firing rate of the D1 and D2 subpopulations coding for the selected action, while the others are inhibited. This occurs as soon as an action gets selected. The efference copy helps the pre- and postsynaptic traces to overlap and also happens in the abstract version where the selection produces binary vectors of the activity in the action layer. At the point when the system knows the state and the action, the expected reward can therefore be computed in the RP system. In the spiking version, a specific striosomal subpopulation gets active and represents the RP for the current state and selected action. This results from their input connectivity (fig. 5.2). Each striosomal subpopulation receives inputs from one state and one action, via the efference copy, in a way that it would fire only if both of its inputs are active at the same time.
Figure 5.2: Schematic representation of the RP setup. Each striosomal subpopulation receives inputs from a single state and from a single action. In order to become active, it needs to have both of its inputs active at the same time, therefore, only one subpopulation at a time can become engaged in the RP. A striosome, here shown by dashed grey lines, comprises all the subpopulations coding for the same action. In italics are the supposed biological substrates of the features specified. Here the toy example considers three colored states where the same three actions can be selected.
5.1.2 Trial timeline

We aimed to reproduce experimental protocols by testing our models on blocks of trials. A trial starts by setting one state subpopulation active (see fig. 5.3 for a representation of the activity of the network during a trial). Through the D1 and D2 pathways, the information is conveyed to all the action coding subpopulations.

Then comes the actual selection, which relies on the activation of the action coding subpopulations at the convergence of the D1 and D2 pathways.

In the spiking network, as connections from D1 MSNs to the output layer, the GPi/SNr, are inhibitory, we first need to invert the softmax distribution, as the action associated with the highest probability of being selected is the one coded by the GPi/SNr subpopulation with the lowest activity.

Once the action has been selected, the efference copy increases the activity of the D1 and D2 MSNs coding for that action, while suppressing the others. Additionally, it sends excitatory inputs to the relevant striosome, where the subpopulation which also receives activation from the current state becomes active.

Conveniently, the output of the active RP subpopulation is delivered to the dopaminergic neurons simultaneously with the outcome of the trial. The difference between this RP and the actual reward value, i.e. the RPE, computed through the phasic firing rate of the dopaminergic neurons or offline in the abstract version, is then available to all the plastic dopamine dependent connections: the cortico-D1 and -D2 matrisomal ones and the striosome-dopaminergic nuclei ones.

There is no delay in the updates of the weights, as soon as the RPE differs from zero, the relevant pathways are affected. In the spiking model, we set a resting time after the reinforcement, where only background activity is fed to the network, in order for the traces of one trial to not overlap with those of the next.

5.1.3 Summary paper I

We used the abstract computational model of the BG to investigate the performance of various configurations of this model in action selection. Specifically, in one of these configurations (Actor+RP), the RP associated with each action was added to the activation value of the actions before the selection. We also tested versions of the model without the D1 or D2 connections, as well as a configuration where the selection was made only based on the RP associated with the possible actions. We first demonstrated that in a simple successive learning task with the standard model, the D2 pathway was responsible for suppressing an action which does not lead to the reward anymore, for
example after a change of the reward mapping, and that this enables the D1 pathway to then promote the correct action. Furthermore, the RP system was able to learn to predict the reward, which led to a RPE of 0, thus stopping all the weight updates. We then experimented with various configurations of different reinforcement learning and conditioning tasks, such as simple acquisition, reacquisition after extinction, successive learnings, reversal learnings and probabilistic reward learning. Basically, we found that the standard (Actor) and the Actor+RP versions showed the best performance in these tasks. The model based only on the RP seemed to be particularly effective during reacquisition. We additionally set up an equivalent task to an experimental study with monkeys [Samejima et al. 2005]. This study looked at how the probability of pressing one or the other lever, left and right, depended on the probability of getting a reward following these choices. The standard version exhibited similar results as the experimental data, whereas the Actor+RP configuration seemed to have better performance for low probability choices. We suggested that an agent could thus adapt on-the-fly its mode of selection depending on the situation.

5.1.4 Summary paper II

Based on the same abstract model as for paper I, we simulated the effect of optogenetic stimulations by transiently increasing the activation of the targeted unit or units, just before the selection. We only used the Actor and the Actor+RP configurations. Specifically, we wanted to first test if the model could reproduce the data from the experimental study of Tai et al. (2012) where they could bias the selection of mice in a probabilistic reward two-choice task, left or right, using optogenetic stimulation of D1 or D2 MSNs in the right or left hemisphere, during decision-making. We were able to calibrate the intensity of the delivered activation to get similar shifts in the selection as experimentally reported. We then tested the two configurations of the models in the probabilistic task. In the Actor+RP configuration we additionally increased, or decreased, the RP associated with the action selected when the targeted MSNs were of the D1 or D2 type, respectively. The biases of the selection approached closely those observed with the mice, that is that the biases differentially impacted the selection depending on the recent reward history. The shifts were the most prominent when there had been inconsistency of the rewards in the recent trials. In a third part, we investigated the impact of a transient increase of the RP only, on the selection. This was sufficient to modify the subsequent selections.
Figure 5.3: Raster plot of the spiking model consisting of 3 states and 3 actions, illustrated over 30 seconds of activity during a change in the reward mapping. In A, a single trial is detailed. In A and B the states (blue), D1 MSN (green), D2 MSN (red) and GPi/SNr (purple) populations, grouped by representation coding, are shown. The indices of the states and actions begin from the top, e.g. neurons with an ID between 160-220 represent the D1 and D2 subpopulations coding for action 1. In C are shown a raster plot and an histogram of the dopaminergic neuron activity. The width of the bins is 10 ms. A trial lasts for 1500 ms and starts with the onset of a new state. The vertical orange dashed line signals a change in the reward mapping. From paper III.
5.1.5 Summary paper III

This implementation of the standard abstract model with spiking neurons also led to develop the network into a more biologically plausible representation of the BG. Its main features are the D1 and D2 pathways, which disinhibits or inhibits, respectively, the output nuclei GPi/SNr where the selection is made, but also the RP pathway with the dopaminergic neuron activity coding for the RPE, which modulate the synaptic plasticity (section 5.1.1). The activity of the network during a trial is in accordance with biological data (fig. 5.3). The model performs well in successive learning tasks, and shows similar dynamics as the abstract implementation. We additionally simulated various lesions of the aforementioned pathways as well as the degeneration of the dopaminergic neurons observed in PD. Among the various conditions, it is notable that a lack of the D2 pathway causes worse performance than without the D1 contributions, and that without the RP pathway, the system still learns quite well but the dynamics of the weight updates are suboptimal. The simulated dopaminergic neuron depletion triggers both modifications of the evolution of the weights and a decrease in the performance which scale with the extent of the depletion. Finally, we suggest that it is the disruption in the D1 pathway that predominantly disrupts the learning performance observed in PD patients.

5.1.6 General comments

The basic architecture is similar between the two versions of the model (for descriptions, please refer to the respective papers, I and III, and to section 5.1.1). A comparison of the evolution of their weights and of their performance during a learning task attests to this similarity (fig. 5.4).

The abstract and spiking versions offer some complementarity. The abstract model allowed us to easily implement the RP in the selection, while the spiking implementation provided the opportunity for more biologically plausible simulations such as the lesioning of pathways or neuronal depletion.

An advantage of using the BCPNN learning over action-value based models of reinforcement learning is that there is no need for the maintenance of the previous action or state value to compute the new value. This occurs naturally through the update of the traces and of the weights.

Going back to the reported modification of the intrinsic excitability of the MSNs by dopamine (section 3.3.4), a convenient way to implement such characteristic in our model would be through the bias term $\beta_j$, which can be seen as an intrinsic excitability variable of the neuron model (Tully et al. 2014). Additionally, the implementation of a dopamine-dependent dynamic threshold, as suggested by Bahuguna et al. (2015), would further help to incorporate reaction times in the model. There are two ways to increase the learning speed
Figure 5.4: Weights and performance of the abstract, left column, and spiking, right column, models in a successive learning task with six blocks. The abstract model consisted of 10 states and 5 action and had blocks of 200 trials, the spiking model of 3 states and 3 actions with blocks of 40 trials. Panels A, B, D and E show the Go/D1 and NoGo/D2 weights from state 1 to all the actions. In C and F, the moving average of successful trials over the last 10 trials is represented. In C, the RPE for the abstract model is displayed in red, and not represented for the spiking version. Black dots represent the average value of the weights. These results come from a single simulation of each model. Left column is from paper I.
in our model. One is to augment the RPE, for example by amplifying the firing rate of the dopaminergic neurons for a reinforcement. The second one is to augment the correlation of the pre- and postsynaptic populations. This could be related to the modification of the intrinsic excitability of MSNs by dopamine.

Considering the model-based and model-free distinction, it seems that performing action selection only through model-based representations might prove to be time-inefficient and computationally taxing. Model-free selection enables quicker and simpler decisions, especially when some habituation has been formed, while the BG still allow to adapt to changes in the environment (Dolan & Dayan, 2013; Doll et al., 2012). It has been hypothesised that both computations could be performed in the striatum, with complementary roles from the dorsolateral and dorsomedial striatum, along with the ventral striatum (Daw et al., 2011; Khamassi & Humphries, 2012). However, in an fMRI study, Doll et al. (2015) reported a correlation between the activation of the putamen and model-free prediction errors whereas it was negatively associated with neural prospections, that is the neural substrate of the states planned to visit, which is linked with model-based behaviours. The underlying biology of these representations is still unclear. The lateral PFC and the intra-parietal sulcus have been described to exhibit the neural signature of a state prediction error, assessing the difference between the state computed from the model-based representation and the observed state (Gläscher et al., 2010). It thus seems that computational models can prove to be useful to suggest and to test hypotheses about these representations.

Outside the scope of this thesis is the decision making process, defining the current state, and which most likely involves cortical interactions similar in nature to the standard BCPNN description with hypercolumns and attractor networks, for which computational models have shown good results in associative learning and pattern completion (Lansner, 2009; Lundqvist et al., 2006; Meli & Lansner, 2013; Sandberg et al., 2000).

The switch from the abstract model to a spiking implementation resulted from the will to increase the biological plausibility of our model, in order for its predictions to better suited to experimental assessment. It can be noted that the learning rule implemented does not rely on very precise spike timings, and that rate coded activity could have be sufficient. However, as mentioned, our goal was not only to show that it could work, but to provide a model that could offer a good match to biological data. It furthermore now allows for interesting developments of the model where precise timings and subthreshold activity are important. We can mention that the spiking BCPNN rule can approach standard STDP shape where the precise spike timings are relevant (Tully et al., 2014). Additionally, one of the main reason to implement our model with
spiking neurons, and to specifically use NEST and PyNN, was to be able to run the model on neuromorphic hardware developed in the European project ‘BrainScaleS’, in which we were involved.

The development of the spiking model confronted us with various challenges. Some of them called for biological data which were usually either conflicting or absent. These steps of the implementation proved to be very useful in determining the information needed, and also offered us the possibility to assess and implement hypotheses about circuitry and functionalities.

Our four factor learning rule can account for a variety of learning conditions, similarly to the spike timing dependent eligibility rule detailed by Gurney et al. (2015). However, it still depends on only one neuromodulator, dopamine, whereas it is known that the BG contains many more, e.g. serotonin, acetylcholine, noradrenaline and noripinephrine (Carli & Invernizzi, 2014; Doya, 2008; Doya et al., 2002).

5.2 Functional pathways

One of the main interest in this thesis was to determine the functional relevance of some of the pathways in the BG. We investigated this by simulating the removal or the silencing of such pathways, as well as of some of other properties, such as spontaneous firing, in reinforcement learning tasks. We also simulated the effect of loss of dopaminergic neurons, as reported in PD. This will be further detailed in section 5.5.

5.2.1 Direct and indirect pathways

The activity of the MSNs in our spiking model shows that both D1 and D2 MSNs fire during the selection phase (fig. 5.3), as it has been reported in mice (Cui et al., 2013b; Tecuapetla et al., 2014). Related to probabilistic learning, as in the multi-armed bandit task, it has been suggested that the arbitration required to respond to a high-conflict decision, where the D1 and D2 MSNs coding for a similar action are both active, could involve interneurons and could explain longer reaction times (Bahuguna et al., 2015). Furthermore, the firing dynamics of the neurons in the output population of our model during selection seem to correspond to data obtained in monkeys during learning. A decrease in activity enhances exploration whereas an augmented activity in the GPi/SNr is associated with restricted selection and increased exploitation (Sheth et al., 2011).

However, in order to improve the biological plausibility, we restricted the weights of the D1 and D2 pathways to positive values, i.e. to be excitatory only. This does not affect dramatically the performance, but it however requires the
implementation of lateral inhibition between subpopulations of MSNs featuring the same dopamine receptor type. This compensates for the loss of inhibition resulting from the absence of negative weights. The implementation of inhibitory interneurons (section 3.3.1) should give even better results as it should offer a better mapping of these negative weights.

Following a change of the reward mapping, the dynamics of the weights exhibit a critical initial involvement of the D2 subpopulation coding for the action which was the correct one for the previous block. This is similar to the observed changes in firing rates in the striatum in rats during learning (Kimchi & Laubach, 2009). Additionally, this relates to the improved ability of monkeys who have the highest D2 receptor availability to reach criterion in a reversal learning task (Groman et al., 2011). This is even more important for the spiking model as with the absence of change in the D1 pathway until the RPE becomes positive, this suppression from the D2 pathway is the only way for the system to switch its selection. Once the correct action has been selected, the weights from the active state subpopulation to the related D1 subpopulation increase as those to the other D1 subpopulations decrease. Eventually, as the RPE diminishes to zero, the amplitude of the change of the weights does so too.

The biases are not represented as they offer little information in this particular setup. With the reward mapping and the cyclic occurrences of the states, which thereby call for a similarly cyclic selection of the actions and of the RP populations, the bias values are similar between subpopulations. In less controlled environments, the biases would differ and would add variability in the selection, by enhancing the selectability of the already most frequently selected actions (section 4.3.4).

In the abstract model (paper I), the removal of the Go pathway often leads to poorer performance than the removal of the NoGo pathway. However, in the spiking version, the results are the opposite, the condition without the D1 pathway (noD1) performs a lot better than the one without the D2 pathway (noD2). As explained in paper III, this results from the absence of complementary activation of the subpopulations in the spiking models, where such complementary activation is implemented in the abstract model. This enables the decrease of the weights between co-active units, such that weights are updated in all the pathways without restriction from the sign of the RPE.

The results from the spiking version can be supported by experimental data. Mice with D2 receptor knock-out (D2R-KO) have been reported to perform significantly worse than D1R-KO mice, both in trial duration and success ratio in a reversal learning task (Kwak et al., 2014). Besides, it has been shown that the number of available D2 receptors in the striatum can be an indication of the performance during the reversal phase of reversal learning, but not of
the initial acquisition or of the retention (Groman et al. 2011).

In the simulations of optogenetic stimulation (paper II), only two actions are represented, going left or right, and are respectively coded by D1 and D2 units in the contralateral side. The simulation of optogenetic activation is represented by an external transient increase of the action value of the targeted action coding units, i.e. the D1 or D2 unit in the left or right hemisphere. We also tested the impact of an artificially augmented RP in the plasticity of all the weights and in the subsequent selections. Due to the left/right lateralisation of the actions, the activation of action units in the D1 or D2 pathway impacts the choice of the side. However, as Tai et al. (2012) stressed in their experimental study, the technique could not elicit an activity in what would be a subpopulation of matrisomal MSNs, if we transpose the terminology used in this thesis to their study. Therefore, the selection bias towards the contralateral side of a stimulation of all D2 MSNs in one hemisphere should not be seen as indicating that D2 MSN activation leads to the selection of an action. Such massive activations probably do not occur under normal conditions in the brain. It has indeed been suggested that activity in the D2 pathway might be required only to inhibit actions that could prevent the selection of the desired output, and not of all the possible other options (Baladron & Hamker 2015). Amemori et al. (2011) supposed that the indirect pathway could be involved in the selection of a specific action module, and the direct pathway would thereby select one action within that module. This relates to the surround-inhibition hypothesis of the indirect pathway suggested by Mink (1996). It would indeed allow for non-competing actions to be performed simultaneously, and would probably be more energy efficient.

Investigating the impact of the spontaneous firing observed in the GPi/SNr was an indirect way to assess the role of the D1 and D2 pathways in the selection. This spontaneous firing seems to enhance the relevance of the D1 pathway (fig. 5.3). In the intact spiking model, high activity in the D1 pathway means a lot of inhibition is sent to the GPi/SNr. It is thus possible for a D1 subpopulation to completely silence its associated subpopulation in the output nuclei. Without spontaneous firing (noSF), as the D1 MSNs project inhibitory connections, there would be no change in the activity of these output nuclei. Thus, the selection would be random, unless this inhibition can have an impact by suppressing the excitation from the D2 pathway. However, by itself, the D1 pathway would not be able to impact the selection without spontaneous firing in the output nuclei. Fig. 5.5 shows that the noSF condition performs as well as the standard intact model in the first part of the blocks, which depends a lot on the D2 pathway, which is unaffected by the noSF condition. Even so, this condition fails to improve as much as the intact one during the second half of the blocks, which relies on the promotion of the correct action by the D1
Figure 5.5: Box plot of the mean success ratio and standard deviation of the examined conditions. In A, the first 20 trials of each block were used whereas in B, the analysis was carried out on the last 20 trials. Data from the last 7 blocks of the PD conditions were used. All differences are significant (p<.0001) unless stated otherwise (ns = non-significant). The horizontal dotted line represents chance level. For all conditions except PD33, PD66, noD2 and no Efference, differences within conditions between the first and last 20 trials are significant. NoSF stands for the condition without spontaneous firing of the GPi/SNr output nuclei and noLI for the condition without lateral inhibition in the striatum. PD16, PD33 and PD66 display the results of the 7 blocks following the deletion of respectively 16%, 33% and 66% of the dopaminergic neurons. From paper III.
pathway.

5.2.2 RP and RPE

We will here first present some results related to the RP pathway and to the RPE. Then we will discuss which role the RP pathway can have in the selection and how it can be mapped to biology.

Both the abstract and the spiking version of the model are able to learn to predict the reward, in a successive learning task. This means that the RPE becomes zero, thus resulting in the freezing of the weights (fig. 5.3 C).

We suggest in paper III that a negative RP, e.g. coding for pain, could be conveyed by striosomal D2 MSNs onto dopaminergic neurons, possibly via LH, making an indirect pathway, whereas the positive RP would be sent from striosomal D1 MSNs onto GABAergic interneurons in the dopaminergic nuclei, representing a direct pathway. This would assume a sort of spontaneous firing of the interneurons in the dopaminergic nuclei, for their inhibition to trigger an increase in the activity of their target dopaminergic neurons. Additionally, we propose that dorso-striatal striosomal MSNs could be involved in the selection, supposedly via the reported projections from striosomes to the GP and the re-entrant cortico-BG-thalamus loops (sections 3.3.1 and 3.3.2).

Functionally, a reward predicting event, a CS, should have two effects in the RP system, in conformity with the TD learning models presented in the first part of this thesis (section 2.3). One is to generate a positive RP, a phasic increase of dopamine, a CR, in order to associate statistically relevant previous events to the eventual reward, US. The second one is an inhibition of the dopaminergic neurons at the time of the expected reward, or reward predicting event. Our model does not account for this, but it could be associated, and implemented, with the previously described direct and indirect pathways stemming from striosomes.

Interestingly, in a study recording the dopamine concentration in the nucleus accumbens in rats performing Go/NoGo tasks, it appeared that a reward predicting phasic dopamine change occurs only at action initiation, and not a the apparition of the reward predicting cue (Syed et al., 2015). This could support our implementation of the state-action dependent RP. Additionally, the authors noted a dip of the dopamine concentration right at the time of the initiation of a wrong action, e.g. going left when the cue indicates a reward on the right lever. This could attest of an absence of the inclusion of RP in the selection, as there seems to be no point in doing an action which the animal knows is linked with a negative RP. However, as we have mentioned, a softmax selection might produce this kind of apparently suboptimal pattern of selection, in order to assert that the other option do not lead to even better outcomes.
Returning to the behavioural effect of the dopaminergic signal, it has been shown that the optogenetic activation of VTA dopaminergic neurons, without any other reward delivery, is sufficient for both the reactivation and the reversal phases of operant behaviour (Adamantidis et al., 2011). However, this activation alone could not drive optical self-stimulation-like behaviour. It is possible that the targeted dopaminergic neurons encode different dimension of the reward signal. There might be some uncoupling of the predictive value, which enables n-ary conditioning, and of the reward signal, which acts as the confirmation of the prediction and is required for the synaptic change triggered by the RP to be effective and to last. Therefore, in the study by Adamantidis et al. (2011), it could be that the optogenetic activation of the dopaminergic neurons is equivalent to the predictive value signal, but as no food is obtained, that is no reward comes to confirm the prediction, the plastic changes are canceled, or not consolidated.

In a previous study which used a similar methods, Tsai et al. (2009) were able to elicit place preference in freely moving mice only via phasic optogenetic stimulation of VTA dopaminergic neurons. Based on the results from our second paper, simulating optogenetic activation, we suggest our model could be able to reproduce this place preference. Indeed, by considering the states as representations of cells in a gridspace, and actions as various movements in that space, driving a dopamine increase at a specific localisation in that gridspace will reinforce the actions that lead to it. In order for our model to show a better match to the data, it would require the use of the eligibility trace. In effect, this stimulation would act as a reward. Moreover, regarding the effect on the behaviours of transient changes, excitation or inhibition, of the dopaminergic neurons activity, the results from recent studies using optogenetics are in accordance with those from our simulations (Chang et al., 2015; Hamid et al., 2015).

There might be two, at least, advantages for a biological system in the use of the RPE instead of the absolute reward value. Firstly, it decreases the energy consumption required for the synaptic plasticity, as updates do not occur once the reward is correctly predicted. Using the absolute reward value would require updates each time the reward is obtained, even after the thousandth time. Secondly, in a condition where plasticity would depend only on the reward value, associated with homeostasis, the modification of a synapse would trigger an opposite effect at the others synapses (Watt & Desai, 2010), and as they would occur every time the reward is obtained, this would lead to a much faster decay of the connections and thus of memories, than what would result from the use of RPE. Our models feature both mechanisms. As the RPE converges to zero, the plasticity is frozen and the system does not spend energy to update the synapses and thus the memories are kept. This, in effect, is one
way to solve the stability-plasticity dilemma. However, it could happen that
the system learns to predict an absence of reward, or even a negative reward,
before it has selected all the possible actions. It would thus keep on selecting
a suboptimal option as the weights are not changed.

One of the results of the simulated optogenetic stimulation on the strio-
osomes is that in this implementation of the RP system, the resulting RPE alone
is not enough to assert if the reward obtained was smaller, or if the negative re-
ward had been stronger, in situations where the expectation has been increased
prior to the delivery of a positive or negative reward, respectively. Through
this, it shows the need for an additional pathway informing about the raw, ab-
solute reward value in order to be able to differentiate such situations.

An initialisation of the weights in the RP pathway to low positive values
means that the system would expect some positive outcome from any actions
performed for the first time. Therefore, the RPE will be smaller if it indeed
leads to a reward delivery than if it leads to a negative reward of the same
absolute value as the positive reward. This optimistic a priori belief has been
suggested to behave like a probabilistic prior, where its influence gets reduced
with increasing experience (Stankevicius et al., 2014). Interestingly, this also
relates to the loss aversion observed in humans, which is the tendency for
losses to have a larger hedonic impact than comparable gains (Rick, 2011).

Reacquisition is the relearning of a previously learned and then extin-
guished association. The reacquisition is faster than the first time. In the
BCPNN, the bias, notably, via its role in the excitability of the neuron, could
account for this phenomenon. As the action is selected, the bias decreases, in-
creasing the excitability of the neuron (section 4.3.4). The phenomenon would
therefore be observed as long as the bias has not decayed back to its initial
value. In paper I, we show that if the extinction phase is relatively short, these
reacquisition effects could also result from the incomplete decay of the relevant
weights and traces. Additionally, the inclusion of the RP in the selection is able
to help the system out of the aforementioned dead ends, where it has learned
to expect a negative reward and thus keep on selecting the wrong action. This
provides support to the usefulness of the RP in the selection.

By allowing the weights from the striosomes to the dopaminergic neurons
to become negative we wanted to capture the same idea as for the matrisomes,
i.e. a putative direct and indirect pathways onto the downstream target, the
dopaminergic neurons. For example, a negative weight in the RP pathway
would represent a negative RP, possibly mapped on striosomal D2 MSN con-
nections to inhibitory interneurons in the SNC, or via the LH or the GP, where
such connections have been reported (Fujiyama et al., 2011) (section 3.3). Due
to the lack of data to clearly establish such pathways, we kept a relatively ab-
stract implementation of this circuit, which is indeed similar to the abstract
The role of striosomes has been suggested to provide the dopaminergic neurons with state-related information, or state-based prediction (Fujiyama et al., 2011). Even though we assume that the RP is based on both the state and the action selected, we have tested the performance of the abstract model where the RP was state only dependent or action only dependent. These setups would be closer to the reinforcement learning models where the RP is based on the average reward obtained, from a specific state. These versions work well when only one action for each state leads to the reward. In the successive learning task, due to the simple reward mapping assigning one specific action for each state to the reward, the three different RP systems exhibit similarly good results. In a multi-armed bandit task however, where several actions for each state can lead to the obtention reward, but with different probabilities, the performance of the state-based only RP deteriorates, as does those of the action-based only RP.

Loops have been described from dorsal and ventral MSNs to dopaminergic neurons via GABAergic neurons in the SNr and in the VTA, respectively (Chuhma et al., 2011; Xia et al., 2011). Activation of these MSNs would thus lead to an increase of the activity of the dopaminergic neurons. This could provide a mechanism to increase the dopamine level in response to an action that has been associated with some positive reward or RP and could therefore be the underlying mechanism for n-ary conditioning. Moreover, monosynaptic connections from MSNs to dopaminergic neurons could thus convey the inhibitory signal of the RP, which would counter-balance either the mentioned disinhibition of the dopaminergic neurons or their excitation, caused by of the delivery of a reward.

In our second paper, an artificial increase of the RP in 94% of the time the model selects the right (not left) side leads to an increase of the left side selection ratio, thereby showing that specific modifications of the RP can impact the selection.

Now let us consider the possible circuitry and characteristics of the RP pathway. Nieh et al. (2015) have reported that the lateral hypothalamus sends both excitatory and inhibitory connections to dopaminergic and GABAergic neurons in the VTA. This describes what could be a possible separate RP system to the one stemming from striosomes. However, if one considers that the RP pathway relies on the striosomal-dopaminergic nuclei circuit, with both a direct and an indirect pathway, as suggested by Fujiyama et al. (2011) and Stephenson-Jones et al. (2013), then the reported connections from the lateral hypothalamus plead for a localisation of the synaptic plasticity in the RP pathway at synapses contacting dopaminergic and, possibly, GABAergic neurons. Indeed, this localisation would allow the presynaptic neurons, strioso-
mal MSNs, to take into account the modifications of the dopaminergic neuron activity triggered by the lateral hypothalamus, and of any other inputs the dopaminergic neurons could receive, and could therefore be in command of the RP. This is in support of the setup we implemented in our spiking model. If that were not the case, that is if multiple, less dependent, circuits were involved in the computation of the RP, it might require a more complex functional system in order to obtain the RPE signal that has been largely documented from experimental data. It is however possible that each of these putative circuits could compute a different aspect of the RP, as some recent studies seem to support (Beier et al., 2015; Lammel et al., 2012; Lerner et al., 2015). It is moreover notable that the lateral hypothalamus receives excitatory inputs from the VTA as feedback. Optogenetic excitation of neurons in the lateral hypothalamus triggered activity in dopaminergic neurons in the VTA, which projects back to neurons in the lateral hypothalamus. Furthermore, the optogenetic activation of lateral hypothalamic neurons led well fed mice to binge eat sugar, even when they had to run across a platform delivering electrical shocks to obtain it (Nieh et al., 2015). As the lateral hypothalamus has been reported to be involved in promoting feeding behaviour, it could be that it is biasing the BG to take action, via the incentive salience of the burst of dopaminergic neurons resulting from the optogenetic activation of the neurons in the lateral hypothalamus. This would confirm that the BG have to integrate a lot of different information but also that their selection can be biased by relatively limited change of their inputs, even of the dopaminergic ones.

The sensory prediction error theory, proposed by Redgrave & Gurney (2006), signals any deviation of the environment from its predicted state, and is thus not specific to a reward signal, as opposed to the RPE. This theory could be accommodated with a single mechanism, as it would account for both the RP signal and for its inhibition, and could thus be mapped on the connections from striosomes to dopaminergic and GABAergic neurons in the VTA and the SNc. An absence of sensory prediction error indicates that the outcome has correctly been predicted. This supposes that the transient increase of dopamine triggered by the change in the environment has been successfully inhibited by the projection from neurons coding for this prediction. The difference between the RPE and the sensory prediction error is that there is no transfer of that prediction to the events preceding the unpredicted sensory event, it only associates the relevant event to inhibit the resulting sensory prediction error through repetition, whereas with the RPE, the RP will be transferred backwards in time, event by event, as in TD learning models. There is no associated increase of the value of the event as it becomes a predictor, that is that it inhibits the burst of dopamine of the unpredicted outcome, which could thus serve for n-ary conditioning. As Redgrave & Gurney (2006) further mentioned, the sensory
prediction error might be adapted in the acquisition of behavioral ‘building blocks’, which could then be used to generate sequences of goal-directed actions. This building process has been linked with activity in the medial PFC, which was hypothesised to have a role in determining the relevance of each step to achieving a goal (Matsumoto & Tanaka, 2004).

It has been suggested that the reentrant cortico-BG-thalamus loops could spread limbic and cognitive information to motor loops (Haruno & Kawato, 2006). As we have seen, the brain has been suggested to combine model-based and model-free computations (Doll et al., 2012; Gläscher et al., 2010). These reentrant loops could be the biological substrate for such mechanism. They could provide the system with the expected next state given the considered action, along with the associated expected reward, in a recursive way, in order to deduce the best action given the current policy or goal. Additionally, it is mainly the striosomes who receive inputs from the limbic system and from the frontal cortex whereas matrisomes are believed to receive mostly sensori-motor information (Crittenden & Graybiel, 2011; Flaherty & Graybiell, 1993). This seems to support the involvement of the RP pathway in the evaluation of the available options, by integrating the former through these loops for the selection. An hybrid model, mixing a model-free and a model-based learner, exhibited a better behavioural fit to human data than did each part alone (Gläscher et al., 2010). A consequence of this hypothesis would be that forcing very quick action selection could prevent the integration of the RP or of the model-based construction in the selection, as it should take more time to integrate this information, putatively through several iteration of the cortico-BG-thalamus loops. The distribution of the selection could thus exhibit a bias towards model-free choices when time available for selection is limited. Related to that, the hyperdirect pathway could provide the global inhibition preventing any selection as long as required by these model-based computations to come up with the directions to follow, as suggested by Baladron & Hamker (2015).

We suggest that our probabilistic model could prove to be well suited to learn such a model of the effects of its actions on the environment, that is by learning that from a certain state, selecting a specific action would lead to a certain new state, and so on. Associating the RP with these computations could enable to plan a strategy in order to achieve a specific goal.

5.2.3 Lateral inhibition

In the spiking model detailed in paper III, we have prevented negative weights in the connections onto the D1 and D2 matrisomal MSNs. Additionally, we have implemented lateral inhibition between matrisomal subpopulations of the
same pathway. Without this lateral inhibition, the baseline firing rate of matri-
somal MSNs would shoot up to 15 Hz, which is a lot higher than what has been
observed in biology and in the version of the model with lateral inhibition (<1 Hz, [Kravitz et al. (2010); Samejima et al. (2005)]). The resulting effect be-
tween allowing negative weights and implementing lateral inhibition between
subpopulations of the same pathway is quite similar, which brings support to
the biological substrate of the BCPNN that we presented in fig. 4.2. Besides
the increased firing rate, the performance is not dramatically decreased without
lateral inhibition (fig. [5.5]). The slight drop in performance seems to occur in
the second part of a block, that is when one action should be the clear winner,
thereby inhibiting the other choices through lateral inhibition. Thus, without
these connections, the competing, irrelevant, actions might get more selected
because of our softmax implementation.

We have not specifically implemented interneurons for the lateral inhibi-
tion but instead the MSNs send direct inhibitory connection to those of other
subpopulations, as observed in rats and mice (Taverna et al., 2004, 2008). The
role of interneurons is of great interest in the control they have on MSN ac-
tivation and therefore on the selection. They could, for example, be used to
control the exploration / exploitation ratio, through the lateral inhibition of the
competing actions. Our current spiking model could provide a solid basis to
further investigate this role.

5.2.4 Efference copy

As can be seen in fig. [5.5], the efference copy has a critical role in the learning
in our spiking model as its absence leads to a failure of the learning. We have
set the efference copy to occur concurrently with the state that triggered the
selected action (see the efference copy activation in fig. [5.7]). This can there-
fore increase the overlap between the pre- and postsynaptic traces. However, a
recent study with rats recently showed that the activity in the ventral striatum
and in the OFC once they realized that the decision they took, to wait or not
for a reward to come, turned out to be the bad one, was similar to the activity
when rats made that decision in the first place (Steiner & Redish, 2014). This
could suggest that a working memory mechanism can, possibly through the ef-
ference copy, re-activates populations that were related to a situation that took
place in the past, further suggesting that the efference copy can be delayed in
regard to the responsible action.

There might be a limit in the processing of sequences of action by the BG.
Indeed, with our current implementation, no information processing appears
to be possible in the striatum during the efference copy. This could have some
implication for sequence learning, like playing musical instruments. We sug-
gest that this means that sequences have to be represented outside of the BG, or have to, through learning and repetitions, become merged as single representations, or behavioural building blocks, in the BG. Furthermore, from a Bayesian point of view, the efference copy carries information that can be used to make a prediction about the expected outcome. The discrepancy between the sensory evidence and the expected state of the world resulting from your action is critical in the updates of the posterior probability (see section 4.3.2 for more details on Bayesian inference).

The segregated parallel loops within the BG could also serve as a way to enable the selection and the execution of different actions at the same time, provided that they are not mutually exclusive, e.g. to raise and lower the same arm. This could also be related to the generation of sequences of actions.

The firing rate of striosomes was fixed in our simulations, but it is unlikely that this is the case in biology. Modulations of the cortical inputs, and possibly of those from the efference copy, onto striosomes could account for motivational drives and for goal-based representations by impacting the RP.

5.3 Habit formation

Ashby et al. (2007) hypothesised that the BG would disengage from the selection as a particular behaviour becomes automatised, that is once it has been reliably learned via reinforcements. Such a shift has been reported in monkeys during the learning of categories, where activity in the striatum is initially a predictor of the classification but then activity in the PFC begins to predict the outcome before the striatum (Antzoulatos & Miller, 2014, 2011). In humans, an fMRI study of a categorisation task showed that automaticity develops after an initial learning phase during which the performance depends on activity in the putamen (Waldschmidt & Ashby, 2011). In a review, Hélie et al. (2015) argued that the BG could act as a general purpose trainer for cortico-cortical connections. This suggests that the main role of the BG is to create habitual behaviours for situations occurring frequently. They could keep control of and adapt to, through the RPE dependent plasticity, a change in the environment, enabling a modification of the responses.

In accordance with the view that habits result from the formation of a dopamine independent pathway outside of the BG, we have set up such connections from the state population to a new, denoted motor, population where subpopulations represent the possible actions, similarly to the GPi/SNr. This pathway could be apprehended as a direct excitatory connection from the cortex to the brainstem, in the schematic representation of our model in fig. 5.1 (although not explicitly represented) and in the one of the BG in fig. 3.1. To that effect, plasticity in this habit pathway is dopamine-independent and
Hebbian. The learning rule used is the standard BCPNN rule, i.e. without implementing the RPE as the learning rate, but with $\kappa$ set to one (eq. 4.22). Additionally, the time constant of its $P$ traces are slower than for the other pathways, $\tau_p = 4000$ (eight times those of the D1 and D2 pathways), but still reasonably fast in order to facilitate the visualisation of the dynamics. The GPi/SNr also project to the motor population in a topographical way, respecting the action mapping. This implementation is simple enough to test the functionality of the circuit, while avoiding the complexity of a scrupulous mapping to biology. Within this motor population, there is also lateral inhibition between subpopulations, as well as excitatory feedback within subpopulations. It should here be noted that the figures shown relative to this habit pathway implementation are the result from unpublished work.

One way to address the problem of the actual selection, which has then to integrate the contribution from the habit pathway, could be to apply the softmax selection not on the GPi/SNr anymore but on the activity of the motor population. The motor population thus receives inhibitory inputs from the GPi/SNr and excitatory ones from the habit pathway. Similarly to the GPi/SNr, external excitation drives neurons in the motor population to fire in absence of other inputs. Thus, as the BG select an action, by inhibiting the relevant subpopulation in the GPi/SNr, the associated motor subpopulation gets disinhibited and can fire. The correlation of this activity with the one of the active state subpopulation can therefore be picked by the BCPNN weights of the habit pathway.

In a first time, the BG act as a teacher to the habit pathway, guiding the selection, as can be seen by the delay in the rise of the weights of the habit pathway compared to those of the D1 pathway in the BG, in fig. 5.6. It is even more noticeable in the average selected action when in state 3. The BG learn quite fast to shift its selection with each new block, whereas the weights and selection of the habit pathway trail behind. This is similar to the results from Charlesworth et al. (2012) and to the hypothesis from Ashby et al. (2007) (and reviewed by Hélie et al. (2015)). In our implementation, the habit pathway does not seem to be affected right from the start of each block. It seems that during the first successful trials in a block, the weak weights from cortical neurons onto MSNs D1 do not elicit a strong enough release of the inhibition from the GPi/SNr to the motor subpopulations to enable the latter to fire sufficiently for the BCPNN weights of the habit pathway to grow. This means that habit formation starts only once the selected action has become salient enough in the GPi/SNr. However, once the habit pathway has learned the correct actions to select, the BG become redundant. This is highlighted in block 5 of the simulation depicted in fig. 5.6 where the cortico-striatal connections onto D1 and D2 MSNs are basically rendered inactive for the whole block, before
Figure 5.6: BG versus habit pathway. Evolution of the weights in the D1, D2 and habit pathways in the spiking model, over 8 blocks of 50 trials, averaged over 20 simulations. The three colour coded lines represent the average weight of the connections from the subpopulation coding for state 3 to the three action coding populations in the D1, D2 (A), and habit pathways (B). The black dots represent the total average of the plotted weights. In (C) are displayed the average action that each the BG and the habit pathway would select, when in state 3. Vertical dashed grey lines denote the start of each new block. In block 5, the gain of the cortico-striatal connections was set to 0, basically inactivating any BG impact on the GPi/SNr.
being restored back. The action the habit pathway promotes for state 3 is thus the same in block 5 as for the previous one, as the BG do not anymore impact the activity in the motor population. It is confirmed by the fact that the average weights from state 3 to the motor subpopulation coding for action 3 in the habit pathway stay elevated during these two blocks. Once the BG can impact again the motor population, in block 6, the weights and the selection of the habit pathway start again to follow the guidance provided by the BG.

It can also be noted that after a change in the reward mapping, the inputs from the habit pathway hinder the selection, and therefore decrease the performance of the model, as the action promoted by the habit pathway was the appropriate one for the previous block, which the BG now try to prevent through strong activation of its D2 subpopulation.

Pathological habitual, or addictive, behaviours could result from the loss of the impact of the BG on the selection compared to the habit pathway, or from a decrease of the plasticity in the latter, or even from a too quick activation on the motor population, via the habit pathway, for the inhibition sent from the BG to affect the selection. In our current implementation, habitual behaviours could furthermore increase the chance of the system to end up in the situation where the RP system has adapted to the absence of a reward before discovering the correct action. With no RPE, the system would get stuck in selecting the same unrewarded action. This seems to call for some implementation of a model-based representation, or of a drive, to avoid such dead ends (Baldassarre et al., 2013).

Concerning the selection, a coherent or concurring contribution from the BG and the habit pathway could lead to shorter or longer, respectively, reaction time, as a putative threshold would be reached faster (Lo & Wang, 2006).

5.4 Delayed reward

In this task with delayed reward, the reward phase is postponed and occurs 1000 ms after the end of the activation of the state and of the efference copy (fig. 5.7). During this interval, the network is in a similar state as in the resting phase, i.e. there is only background activity. The time constant for the $E$ and $P$ traces were set to 2000 and 10000 ms, respectively, in order to allow the overlap of the traces and to prevent too much decay of the weights in-between trials. We also doubled the number of trials within a block, from 40 to 80. It has to be mentioned that the results presented for delayed reward come from a version of the model where the D1 and D2 weights were allowed to take negative and positive values.

The system is able to adapt and to learn to select the correct action for all the reward mappings (nine blocks). Compared to the results of the version
Figure 5.7: Raster plot of the activity of a network consisting of three states and three actions, during three trials (7.5 seconds) of a delayed reward learning task. The states (blue), D1 (green), D2 (red), GPi/SNr (purple), efference copy (cyan) and dopaminergic (black) populations, grouped by representation coding, are shown. The indices of the states and actions begin from the top. For example, neurons with an ID between 550-610 represent the D1 and D2 populations coding for action 1. A trial lasts for 2500 ms and starts with the onset of a new state, indicated by a vertical dashed grey line. The simultaneous higher phasic firing rates in the D1 and D2 populations correspond to the subpopulation coding for the selected action receiving inputs form the efference copy. Note that the x-axis, representing time, does not start from zero.
without delayed rewards, it requires more trials to stop selecting the previously correct action (fig. 5.8). The evolution of the weights in the D1 and D2 pathways is similar to the one of the version without the RP pathway. Indeed, the weights of the RP pathway never get positive and exhibit less variations compared to the implementation without delayed reward (results not shown). However, the RP pathway never learns to inhibit the reward signal. This failure of the RP pathway results from the reduced overlap between the traces of the pre-, striosomal, and postsynaptic, dopaminergic neurons. Indeed, striosomal MSNs fire at the same time as the efference copy is active, that is 1000 ms before the reinforcement signal, whereas state and action coding subpopulations fire simultaneously in the D1 and D2 pathways. Even with $\tau_e$ set to twice the interval between the RP activation and the reinforcement signal, the $P_{ij}$ traces do not grow enough compared to the $P_i$ and $P_j$ traces, resulting in negative weights. The obvious solution could be to increase $\tau_e$ to a value large enough to enable the RP weights to grow positive. However, with longer time constants come the problem of inter-trial overlaps. This illustrates the difficult problem of delayed reward and temporal credit assignment, even in a clean and controlled experimental set up. Another way to improve the situation would be to make the relevant striosomal MSNs fire at the time of the reinforcement signal. In the current state of our model, this would prevent the RP from being included in the action selection process. This further indicates the need for separate pathway in the RP circuit, one involved in the coding of the RPE, as it is implemented in the spiking model, and one in the selection, as tested in the abstract model by the condition Actor+RP (paper I).

Anyways, there still resides a problem in any of these cases. We have avoided the underlying computation of timing by conveniently setting the conduction delay from the striosomes to the dopaminergic neurons to the duration of the interval. Reflecting on the requirements of delayed reward learning and considering the functional architecture of the RP we have mentioned, it seems to call for the recruitment of additional loops than the striosomo-dopaminergic nuclei described. In order for the RP to be delivered at the time of the delivery of the expected reward, and because of the mono- or disynaptic connections from striosomal MSNs to dopaminergic neurons in the SNC and the VTA, a timing mechanism must take place outside of this pathway, if one excludes the conduction delay from the possible ways to adapt to the reward delay. Furthermore, studies have demonstrated the necessity of the striatum, and of its afferent projections from the SNC, for temporal perception and temporal production [Matell & Meck, 2004]. The timing information thus seem to be provided to the striatum, possibly specifically to the striosomes, which would project directly to the dopaminergic nuclei at the appropriate time. This agrees with studies linking the BG, and especially the striatum, along with the PFC.
Figure 5.8: Average success ratio in a delayed reward learning task. Vertical dashed grey lines denote the start of a new block, where the reward mapping was changed. Note that here the number of trials per block is doubled to 80.

and the cerebellum, with temporal processing, through lesions, pharmacological and imaging approaches (Harrington & Haaland, 1999; Matell & Meck, 2004).

An hypothesis could be that the coding of time in the BG relies on the integration of the time required for an action to be performed. These durations would be learned during development, and might involve the efference copy, as it could provide feedback about the execution. Indeed, one key role of the brain is to control movements (Wolpert et al., 2011). It could therefore make sense that the coding of time might be associated with the realisation of motor command and sequences. In an fMRI study where human participants were asked to modulate their attention to colour or to time, Coull et al. (2004) observed activity in visual areas of the occipital cortex when subjects had their attention on colour whereas when their attention was directed towards time, activity in the cortico-striatal circuits was reported. Furthermore, Jin et al. (2009) suggested that time information could emerge as a byproduct of event-coding in the BG. In a task where monkeys had to make saccades towards stimuli which appeared at various intervals, the authors argued that prefrontal and striatal neurons could encode time-stamp representations of time, that is a peak activity with task-dependent latencies.

We can here also expand the idea that the implementation of a second, indirect, pathway from the striosomes to the dopaminergic nuclei, that we
have mentioned, notably in section 5.2.2, could allow for n-ary conditioning. Thanks to the dopaminergic neurons burst triggered by a reward predicting event, a CS, at the time of this event, which is earlier than the reward delivery, the US, events associated with this prediction could in turns elicit a RP, a CR, which could be picked up by even earlier events etc. Ultimately, this would lead to the inhibition of that anticipatory phasic dopamine increase, but these earlier events, CS, would then trigger their own CR. The response transfer should follow TD learning models, where the transition of the prediction shifts directly from the time of the later associated event to the earlier one.

In relation to delay and time, Mobini et al. (2000) reported impulsive behaviours of rats favoring small immediate rewards instead of larger, but delayed ones, after depletion of the central serotonergic system. It has thus been suggested that the serotonergic signal could be involved in the computation of the discount factor of future rewards (Doya, 2002). Indeed, optogenetic activation of serotonergic neurons in dorsal raphe promoted mice’s patience for delayed reward (Miyazaki et al., 2014). Such results do not necessary imply the involvement of the RP in the selection in a model derived from TD learning, as the discount factor acts on action values. However, in our model, this would vouch for a recruitment of the RP in the selection.

The coding of the predicted reward is not only about the probability of getting the reward, which our model is able to learn, but it has to also match the timing and the duration of the change in the firing rate of dopaminergic neurons produced by the reinforcement feedback. We have set the stimulation time of a striosome equal to the one of the reinforcement (section 4.4.2). It seems, however, likely that in biology there is some temporal compression of the inhibitory prediction into the actual time window of the reinforcement signal (Yagishita et al., 2014).

5.5 Parkinson’s Disease

The simulation of PD was done in the spiking model by reducing the number of dopaminergic neurons during the simulation. The abstract version of our model does not feature an explicit tonic dopamine level that we could have biased in an attempt to simulate a similar effect. The reduction of the number of dopaminergic neurons during a simulation leads to a decrease in the performance of the model, worsening with the increase of the proportion of deleted neurons (PD16 = 16%, PD33 = 33% and PD66 = 66%, fig. 5.5 and fig. 5.9). The results suggest that the main cause of the bradykinesia observed in PD patients could be the incorrect contribution of the D1 pathway to the selection, caused by the lack of plasticity in the pathway, itself resulting from the reduced baseline dopamine level, translated to a bias towards a negative
RPE. Indeed, it appears that the D2 pathway, although exhibiting an overactive synaptic plasticity compared to the normal condition, is still able to fulfill its function, which is to prevent the selection of actions leading to a negative outcome.

The two conditions with the largest depletion of dopaminergic neurons, PD33 and PD66, have significantly worse results than the version of the model without D1 contributions (fig. 5.5). This could indicate that it would be beneficial to completely silence the D1 pathway in advanced Parkinsonian stages. In our simulations of the loss of dopaminergic neurons, most of the time the D1 pathway promotes an action which does not lead to the reward and thereby hinders the selection of the appropriate one. The constant negative RPE resulting from these two most severe conditions does not allow plasticity in the D1 pathway. It is thus frozen in the situation it was at the time of the, sudden (it occurs gradually in reality), decrease of the dopamine level. However, the action it promotes turns out to be the correct one in every third block, due to the particular reward mapping we used, and the model indeed exhibits better results during these specific blocks.

Additionally, the PD16 and the noSF conditions show similar results in learning (fig. 5.5). We have already seen that it is the D1 pathway that is affected in the latter. In PD16, the dynamics of the evolution of the weights in the D1 pathway are affected, exhibiting slower learning, resulting from the bias towards a negative RPE caused by the dopaminergic neurons loss.

Parkinsonian patients off medication have an easier time learning to avoid negative outcomes than learning to approach positive ones, with opposite effects once on medication [Frank et al., 2004]. Results from our simulations are consistent with this, as it is only the D2 pathway that contribute beneficially to the selection.

During PD and with the decrease of the tonic dopamine level, D2 MSNs show an increased sensitivity to glutamatergic inputs although the number of receptors decreases [Bamford et al., 2004] [Gerfen, 2003]. Additionally, the increased sensitivity of the D1 receptors has been suggested to correspond to a compensatory mechanism to the reduction of the dopamine level during PD [Iravani et al., 2012]. Such mechanisms would delay the onset of the motor symptoms in PD, usually starting when the dopaminergic neuron loss reaches 50 to 70% [Obeso et al., 2004], which is later than in our model. However, it should be noted that our results refer to learning of new tasks, which exacerbates the troubles compared to performance in well established day to day activities. This could be supported by the condition where the simulated dopaminergic neurons loss occurs not concomitantly with a new block, but at the middle of a block. There, the drop in performance is less severe and can only be noticed in the PD66 condition, which is more in line with the be-
Figure 5.9: Evolution of the weights in the D1, D2 and RP pathways in simulations of PD degeneration of dopaminergic neurons. After 8 blocks (black arrow), 16% (left column) or 66% (right column) of the dopaminergic neurons were rendered useless until the end of the simulation. In A and D, the weights represented are the average weight to each action coding population in D1 and D2 from state 1. B and E correspond to the average weights for each state-action pairing striosomal sub-population to the dopaminergic neuron population. C and F display the moving average success ratio of the model over the simulation. Vertical dashed grey lines denote the start of a new block. In A, B, D and E, the black line is the total average of the plotted weights. Shaded areas represent standard deviations. From paper III.
havioral observations. It should here be stressed that, in reality, this loss of dopaminergic neurons occurs gradually (section 3.4.1). Also, and it can be noted also for the noD1 condition, the results might be improved by the fact that, in our implementation, an action is always selected, thereby exhibiting the inappropriate selection. We have seen that optogenetic activation of the D2 MSNs does not elicit a motor response but rather suppresses or decreases motor activity (Kravitz et al. 2010; Sippy et al. 2015; Tai et al. 2012), thereby supporting the idea that the resulting global or possibly focal inhibition of the GPi/SNr is not enough to trigger a motor response.

In the simulation of the loss of dopaminergic neurons in SNc observed in PD patients, we have only modified the number of active neurons, while the remaining ones kept the same parameters and dynamic as before. As we have mentioned, the intrinsic excitability of MSNs, and possibly of interneurons, is affected during the disease. Implementing this in our model, for example as Frank (2005) did, where the dopamine level modulates the gain and the activation threshold of MSNs, would further improve the match between our model and biology.

In our model, the independence of the RP weights to the sign of the RPE might have prevented the development of further disruptions in the dynamics of the weights, or in the performance, as it could have had in, for example, a sign-dependent direct and indirect RP pathways implementation.

Finally, the implementation of the remaining nuclei of the BG, the STN and the GPe, could allow us to simulate DBS, which prominently targets these structures. This would be inspired from the work made for our second paper, applied to the spiking model of paper III.
6. Summary and Conclusion

Our computational model of the BG, based on the BCPNN learning rule, which computes the probabilities of activation of a neuron based on the presence of input features, is able to learn to select the rewarded action in various reinforcement learning situations. We have investigated the functional contributions that the main pathways and properties of the BG have on the selection and on the learning performance through simulated lesions and reorganisation of the model. During the development of our model to a spiking neuron implementation, we faced situations where the lack of available data challenged us to consider the possible alternatives and their theoretical and biological implications. Reviewing the literature in order to find evidence or clues that would guide or implementation was a seizable part of this work, to filter the relevant information among the vast ocean of data available, and to adapt and build the model in accordance. This integration led to the discussion of various hypotheses concerning biological substrates of functions, which themselves where commented from a theoretical point of view. We have notably suggested that the phasic increases of dopamine, observed in conditioning tasks, might require an additional pathway than the one from striosomal MSNs to dopaminergic nuclei. We have further described possible loops and pathways emerging from striosomes that could be involved in both the RP and the selection, and could also account for conditioning phenomena. Our results suggest that the D2 pathway plays a major role in the suppression of a previously correct action, suppression which eventually enables the system to select a different action. The D1 pathways seems to be critical in enhancing the saliency of the correct action, which ultimately leads to habit formation, and possibly to shorter reaction times. Our simulations of the degeneration of dopaminergic neurons observed in PD have also led to the assessment of the role of these pathways during the disease. A prominent result was the dysfunctional synaptic plasticity in the D1 pathway while the functional contributions of the D2 one was mostly preserved in the PD simulation.
6.1 Outlook

Thrilling developments to this work can be considered. A relatively basic implementation would be that of the negative RP, and the inclusion of a pathway involving the LH. Additionally, as we discussed, it seems the RP does in fact integrate multiple circuits, each of them possibly involved in the coding of, possibly specific, reward-related dimensions. Computational theories and models of these pathways are needed to encompass the existing data and to produce testable hypotheses, which, thanks to advances in technology, will soon be confronted with novel biological findings.

Thanks to the level of biological abstraction of this work, we can consider multiple options for the future development of the model. Increasing the biological relevance could be obtained through the implementation of the missing structures of the BG, such as the STN and the GPe, but also of the different neurons types, e.g. interneurons, in the current model. These nuclei are also those commonly targeted by DBS in PD (section 3.4.1). A first step toward this implementation could be to reproduce the simulation of optogenetic stimulation in the spiking model. An additional option is to investigate the meta-effects of dopamine on the activity of the MSNs and on the performance of the model, in order to clarify its role. Another possibility is to investigate system level dynamics and the relation between populations. Part of this analysis could consider the variations in entropy and information occurring during learning, and what could impact these changes.

We suggest that our model can be enhanced to account for and to test the various hypotheses we mentioned concerning the RP pathway. It has already shown encouraging results in delayed learning task and in a habit formation condition.

A major evolution in the approach would be to shift to a model-based implementation. The model could learn in which state it ends up when performing a specific action from a given state. Enabling the system to carry this out recursively, and to use the associated reward prediction, would allow the model to make selections based on a rich and organized representation of the world. Furthermore, it would allow associations between states and could thus account for some classical conditioning phenomena.

Related to that last part is the implementation of the RPE and the comments we made in section 5.2.2. The increased activity observed at the occurrence of a CS seems to call for an additional pathway, in relation to the RP. Moreover, the mechanisms involved in the coding of the precise temporal materialisation of the reward, inhibiting the dopaminergic neurons by the estimated value of that reward, are still to be explained and taken into account in the model.

My background in clinical and cognitive psychology vouch for my interest
in mental disorders, and as we have seen in section 3.4, computational psychology and computational psychiatry could help bridge the explanatory gap, and who knows, true AI might end up facing true artificial mental troubles?

6.2 Conclusion

With the spiking neurons implementations we have increased the biological relevance of the model, of the results and of the predictions. Furthermore, we faced implementation dilemmas during the development of the spiking version of the model. As some functional and structural biological substrates were not known, we could formulate hypotheses constrained by the framework and the results of our models and by relevant biological data. We have tried to constrain our top-down approach with biological data. I hope this work gave a good overview of the general work process and of the train of thoughts that spanned over the last five years, as they constitute the substrate for the development of the models. Finally, the previous section has underlined the challenges that lie ahead.

6.3 Original Contributions

I have contributed to the implementation of the multiple factors learning rule and the definition of its interdependent effects.

I have implemented a complex computational model of the BG with spiking neurons and online learning. It is able to learn to select the correct action for a given state and to predict the reward, and shows good similarities with biological data.

I have detailed and tested the functional relevance of pathways and features of the BG in learning, and assessed the results with statistical analyses.

I have suggested and discussed different hypotheses related to the structural and functional connectivity in relation to the BG. Based on the results of the model, I have challenged the dominant view in PD, which attributes most of the motor troubles to dysfunctions in the D2 pathway of the BG.

Through the selection of open access journals for the publication of our works, and the openly accessible source code of the simulations (abstract model: see supplementary information of Paper II; spiking model: https://github.com/pierreberthet/bg_dopa_nest), my scientific contributions are available to the general population.
References


DOYA, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? 21 38


DOYA, K. (2000b). Reinforcement learning in continuous time and space. Neural computation, 12, 219–45. 82


GERFEN, C.R. (1992). The neostriatal mosaic: multiple levels of compartmental organization. *Trends in Neurosciences, 15*, 133–139. 44, 47


IZHIKEVICH, E.M. (2007). Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling. *Cerebral Cortex, 17*, 2443–2452. 61, 64, 80, 95


WATSON, J.B. (1913). Psychology as the behaviorist views it.


