Modélisation computationnelle du rôle de la dopamine dans les boucles cortico-striatales dans l'apprentissage et la régulation de la sélection de l'action

Jean Bellot

Une Thèse de doctorat par

BELLOT JEAN

présentée à

Université Pierre et Marie Curie

Ecole doctorale Cerveau-Cognition-Comportement

Institut des Systèmes Intelligents et de Robotique / Equipe AMAC

pour obtenir le grade de

Docteur en Neurosciences Computationnelles

Modélisation computationnelle du rôle de la dopamine dans les boucles cortico-striatales dans l'apprentissage et la régulation de la sélection de l'action

2015

Jury

Thomas BORAUD	– Université de Bordeaux	Rapporteur
Nicolas ROUGIER	– Université Bordeaux - INRIA	Rapporteur
Arthur LEBLOIS	– Université Paris Descartes	Examinateur
Christelle BAUNEZ	– Aix-Marseille Université	Examinateur
Philippe FAURE	– Université Pierre et Marie Curie	Examinateur
Benoît GIRARD	– Université Pierre et Marie Curie	Encadrant
Mehdi KHAMASSI	– Université Pierre et Marie Curie	Encadrant

Mots clefs

Neuroscience Computationnelle; Dopamine; Ganglions de la Base; Apprentissage par Renforcement; Conditionnement Instrumental; Maladie de Parkinson

Résumé

Dans ce travail de thèse, nous nous proposons de modéliser le rôle de la dopamine dans l'apprentissage, ainsi que dans les processus de sélection de l'action en lien avec les ganglions de la base. La dopamine est un neurotransmetteur lié à de nombreuses pathologies telles que la maladie de Parkinson, impliquant une détérioration des capacités motrices, l'addiction ou encore la schizophrénie. Les travaux de Schultz et collègues durant les années 90 ont permis de faire le parallèle entre l'activité des neurones dopaminergique, enregistrée au cours de conditionnements pavloviens, et le signal d'apprentissage utilisé par les algorithmes d'apprentissage par renforcement, l'erreur de prédiction de la récompense (Schultz et al. 1998, Sutton et Barto 1998). Ainsi, l'activité phasique de ces neurones, projetant vers le striatum, est donc supposée guider le processus de sélection de l'action fait par les ganglions de la base (Mink, 1996), ce qui peut expliquer certains processus d'addiction.

Nous avons dans un premier temps analysé l'information encodée par les neurones dopaminergiques en la comparant à différentes informations utilisées par les modèles d'apprentissage par renforcement. Nous avons, à ces fins, modélisé la tâche utilisée dans Roesch et al. 2007 et comparé quantitativement la capacité des différents algorithmes d'apprentissage à reproduire l'activité dopaminergique enregistrée chez des rats. Ces analyses suggèrent que l'information encodée par les neurones dopaminergiques dans la tâche de Roesch et collègues n'est que partiellement compatible avec une erreur de prédiction et semble en partie dissociée du comportement.

Dans une deuxième partie, nous proposons de modéliser l'effet de la dopamine sur un modèle des ganglions de la base prenant en compte des connexions existant chez le primate, souvent négligées dans la littérature (Liénard et Girard 2014). En effet, la plupart des modèles actuels font l'hypothèse d'une séparation stricte de deux chemins dans les ganglions de la base : le chemin direct lié à la récompense et le chemin indirect lié à la punition (Frank et al. 2004). Cependant si cette dissociation semble légitime chez la souris, des études anatomiques montrent que ces chemins ne sont pas aussi dissociés chez le primate. Nous proposons ainsi d'étudier le comportement de ce modèle afin d'analyser sa capacité à reproduire les dysfonctionnements observés chez des patients parkinsoniens tels que l'apparition d'oscillation bêta dans le globus pallidus ou de différence dans l'apprentissage lié à la récompense et à la punition.

De par l'étude du signal phasique des neurones dopaminergiques ainsi que de l'étude de l'effet de la dopamine sur le processus de sélection de l'action des ganglions de la base, nous cherchons à mieux comprendre le rôle de ce neurotransmetteur dans l'apprentissage ainsi que son effet dans maladies liées à un dysfonctionnement dopaminergique.

Abstract

In this thesis work, we modelled the role of dopamine in learning and in the processes of action selection through its interaction with the basal ganglia. Dopamine is a neurotransmitter bound to numerous pathologies such as Parkinson's disease, involving a motor control abnormalities, drug addiction or even schizophrenia.

During the 90s, the work of Schultz and colleagues (Schultz et al. 1998) has led to major progress in understanding the neural mechanisms underlying the influence of feedback on learning. In these studies, the activity of dopaminergic neurons exhibited properties of the reward prediction error signal used in so-called Temporal Difference (TD) machine learning algorithms (Sutton and Barto 1998). Considering the strong connectivity between the DA system and the basal ganglia known for its action selection properties (Mink, 1996), DA has thus been thought to be the neural signal that help us to adapt our behavior.

In the first part of my PhD, we analyze the information encoded by dopaminergic neurons recorded during a multi-choice task (Roesch et al. 2007). In this purpose, we modeled the task and simulated different TD learning algorithms to quantitatively compare their ability to reproduce dopamine neurons activity. Our results show that the information carried out by dopamine neurons is only partly consistent with a reward prediction error and seems to be dissociated from behavioral adaptation.

In the second part of my PhD, we study the effect of different levels of dopamine in a biologically plausible model of primates basal ganglia that considers existing connections often neglected in the literature (Liénard and Girard, 2013). Indeed, most of current models of basal ganglia assume the existence of two segregated pathway : the direct pathway associated with reward and the indirect pathway associated with punishment (Frank et al., 2004). However, if this dissociation seems to exist in mice, anatomical studies in primates revealed that these two pathways are not dissociated (Parent et al., 1995). We study the ability of such a model to reproduce beta oscillations observed in Parkinsonian and the differences in reward and punishment sensitivity, with high or low-level of dopamine.

By the study of phasic dopaminergic neurons signal as well as the study of the effect of dopamine on the process of action selection performed by the basal ganglia, we seek to better understand the role of this neuromodulator in learning and decision making.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Contexte scientifique général

L'étude de l'intelligence bénéficie aujourd'hui de méthodes modernes, permettant l'enregistrement de l'activité neurale de sujets animaux (dont humains) réalisant une tâche d'apprentissage, et ainsi d'observer les processus neuronaux à loeuvre. De plus, la puissance croissante de calcul des ordinateurs permet la simulation de ces processus d'apprentissage, donnant lieu à de nombreuses interactions entre deux domaines scientifiques : les neurosciences et la modélisation informatique (et plus spécifiquement l'intelligence artificielle).

L'intelligence et l'apprentissage sont deux principes suggérant une interaction : la mise en relation entre plusieurs informations dans le premier cas et une adaptation entre un sujet et son environnement dans le deuxième cas. Afin de mieux comprendre comment un agent interagit avec et apprend de son environnement, il est courant en intelligence artificielle de schématiser une interconnexion entre un agent et son environnement (Sutton et Barto 1998). L'agent est capable d'agir sur son environnement et est capable de percevoir son environnement afin de s'y adapter. L'agent est ici utilisé au sens large et peut être un animal ou encore un robot. Pour évoluer dans son environnement, il doit être capable de percevoir, via divers senseurs, son environnement. À l'aide de cette observation, l'agent pourra agir sur son environnement pour assurer sa survie. Cette capacité à agir sur son environnement découle de capacités de sélection de l'action : étant donnée une observation, l'agent sélectionne un comportement. Parfois le comportement adopté par l'agent n'est pas ou plus adapté à son environnement. Le fait de changer progressivement de comportement afin de s'adapter à un environnement changeant démontre des capacités d'apprentissage. Même sans changement de l'environnement, la réponse de

l'agent à un contexte donné pourra changer selon son expérience passée. Par exemple, un enfant ayant eu l'expérience d'une brûlure ne se rapprochera probablement plus d'un feu de façon imprudente. L'expérience passée forme le comportement. Ainsi, en étudiant les capacités d'adaptation d'un sujet à son environnement, on peut observer et analyser ses capacités d'apprentissage et les processus mis en oeuvre pour l'apprentissage.

À la fin du XIXème siècle, le psychologue Edward Thorndike commença à étudier les capacités d'apprentissage de chats placés dans une cage dont ils devaient apprendre à sortir (Thorndike 1927). De par ces études, il conclut que les chats étaient capables de reproduire de plus en plus rapidement les actions qui permettaient leur libération, mettant en évidence un apprentissage lié à l'expérience de la cage. Cependant, il observa que les animaux ne comprenaient très certainement pas les conséquences directes de chaque action menant au final à l'ouverture de la cage. Il conclut notamment que les mécanismes sous-jacents échappent aux chats testés. Le fait d'avoir trouvé la séquence d'actions leur permettait de sortir, sans compréhension des mécanismes d'ouverture. Thorndike propose donc de dissocier le fait de réaliser une séquence d'actions pour un but donné au fait de comprendre la finalité des différentes actions.

Ces études lui ont fait supposer que les animaux apprennent par une suite d'essais et d'erreurs et que chaque comportement ayant été récompensé se trouvera renforcé par la suite et donc plus probablement adopté par l'animal (Thorndike 1927, hypothèse connue sous le nom de "*law of effect*"). Cette notion d'apprentissage par essais et erreurs et de renforcement lié à la récompense ont depuis été repris notamment par les algorithmes d'apprentissage par renforcement, aujourd'hui encore utilisés pour reproduire le comportement animal.

Conjointement, le chercheur Ivan Pavlov a mené des expériences de conditionnement aujourd'hui connues sous le nom de conditionnement pavlovien. Ces recherches ont permis de mettre en évidence les capacités de chiens à associer un stimulus à une récompense lorsque le stimulus précède un nombre suffisant de fois la récompense.

Ces différents travaux ont permis de montrer que les animaux comme le chien ou le chat, sont capables de renforcer des associations entre des actions ou des stimuli et des récompenses, mettant en évidence des capacités d'apprentissage. Depuis, le conditionnement pavlovien est utilisé dans de nombreuses études afin d'analyser les processus liés à l'association entre un stimulus et une récompense. Un autre type de conditionnement est utilisé pour étudier l'association entre une récompense et un comportement : le conditionnement instrumental.

Conditionnements pavlovien et instrumental

Dans un conditionnement pavlovien, l'animal perçoit un stimulus neutre (qui n'a pas pour l'animal de valeur intrinsèque), qui peut être auditif dans le cas de la cloche de Pavlov, suivit d'une récompense, tel que de la nourriture. Dans le jargon du conditionnement, on parle de stimulus non conditionné (US) pour désigner de façon générique la récompense. Ce stimulus doit donc provoquer une réponse spontanée de l'animal (la sali-

vation est la réponse non conditionnée du chien dans l'expérience de Pavlov – illustrée par les oreilles qui bougent dans la figure 1.1). Ce n'est pas une réponse réfléchie mais un réflexe inné de l'animal à la vue de la nourriture.

Lorsque le stimulus neutre est associé avec la récompense de façon répétitive un nombre suffisant de fois, l'animal associe au stimulus une valeur liée à sa capacité de prédiction de la récompense (du US). Lorsque le stimulus est associé à l'US, on parle alors de stimulus conditionné (CS). Ainsi on peut voir la réponse non conditionnée apparaître dès le moment de la présentation du CS indiquant que l'animal a appris la contingence CS→US. Cela se traduit par la salivation du chien de Pavlov au moment de la perception du son de la cloche prédisant la récompense. Ce type de conditionnement permet de mesurer l'évolution des réponses conditionnées de l'animal au cours de l'apprentissage et de la comparer avec l'évolution des enregistrements électrophysiologiques. Plusieurs études observent en effet les mouvements occulaires de l'animal vers le stimulus puis la récompense, mesurent la salivation de l'animal ou encore en mesurent combien de fois l'animal essaie de lécher la récompense par anticipation de celle-ci (Matsumoto et Hikosaka 2009 ; Waelti et al. 2001).

Le conditionnement instrumental repose sur le même principe. Cependant, au lieu de n'avoir qu'une contingence stimulus→ récompense (CS→US), l'animal doit effectuer une action qui lui permettra d'accéder à la récompense (voir Figure 1.1). Ainsi, l'animal apprend ici une association action→ récompense. Ce type de conditionnement permet d'étudier les capacités de l'animal à effectuer l'action le menant à la récompense, et son temps d'adaptation pour comprendre cette contingence action/récompense. De plus dans le cas où plusieurs actions sont proposées à l'animal, on peut observer son comportement en terme de sélection de l'action. Quelle action l'animal va-t-il choisir afin de maximiser sa récompense future ? Ce type de conditionnement est donc plus riche en ce sens que l'on peut directement observer les modifications du comportement dans les choix de l'animal. Cela permet de faire le lien entre les enregistrements électrophysiologiques et les capacités d'adaptation de l'animal (Morris et al. 2006 ; Roesch et al. 2007). Ces travaux sont donc particulièrement pertinents dans l'étude des capacités d'apprentissage et de sélection. Ainsi un très grand nombre d'études comportementales ou électrophysiologiques s'inspirent de ces types de conditionnement.

Dans cette thèse nous étudierons plus particulièrement les capacité d'adaptations de rats dans un contexte instrumental (Chapitre 3), afin de faire le lien entre adaptation comportementale et activité électrophysiologique enregistrée pendant la tâche. Nous chercherons à lier computationnellement l'information encodée par les neurones dopaminergiques avec la convergence du comportement (voir Chapitre 4). Puis nous élargirons ce travail à la modélisation du rôle de la dopamine dans les ganglions de la base dans l'apprentissage et la régulation de la sélection de l'action chez les primates (Chapitres 5 et 6).

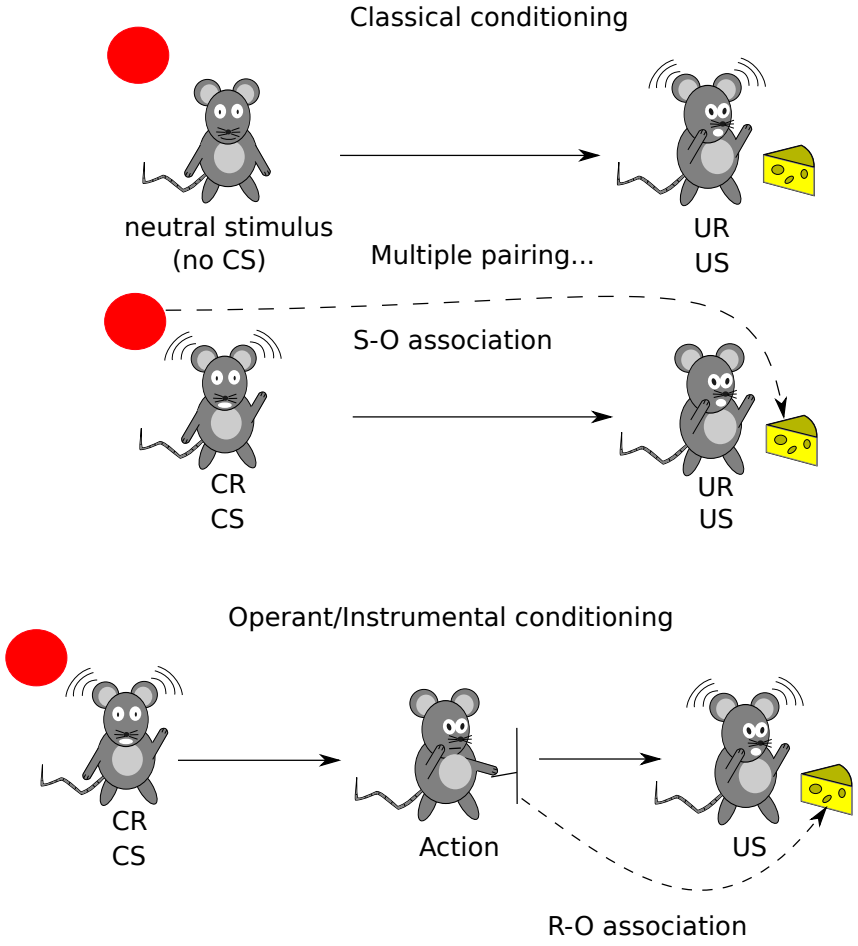

Figure 1.1 – *Présentation des principes liés aux conditionnements classique et instrumental. A Conditionnement classique. NS : stimulus neutre ; US : stimulus non conditionné, ici une récompense donnant lieu à une réponse non conditionnée de la part de l'animal ; UR : réponse non conditionnée (ici salivation) ; CS : stimulus conditionné, élicitant une réponse conditionnée de la part de l'animal comparable à une UR ; CR : réponse conditionnée comparable à une UR. Voir texte pour explications.*

Les neurosciences computationnelles

Les études comportementales basées sur les différents types de conditionnement ont permis de faire le constat d'un apprentissage et proposent une analyse, souvent qualitative, des processus de cette apprentissage. Plus tard, les recherches se sont portées sur les bases neurales permettant l'apprentissage. Notamment Hebb, 1949 (Hebb 1949), élabora une théorie proposant que les neurones déchargeant en même temps ont tendance à renforcer leur interconnection ("*fire together wire together*"). Ces recherches influenceront d'ailleurs les informaticiens déterminés à créer des formes d'intelligence artificielle, s'inspirant de ces travaux pour les appliquer à des problèmes de classification. Cela a mené à la création des premières formes des réseaux de neurones, tel que le perceptron (Rosenblatt 1958). Depuis, l'étude de l'intelligence et de l'apprentissage en neuroscience nourrit l'intelligence artificielle.

Ce type d'interactions entre neurosciences et intelligence artificielle ont mené à la création d'un nouveau champ de recherche : les neurosciences computationnelles. De façon générale, le but des neurosciences computationnelles est de comprendre les principes guidant la dynamique de la gestion de l'information dans le cerveau afin de les reproduire en simulation, et d'en déduire des prédictions expérimentales précises permettant de tester davantage l'hypothèse computationnelle proposée. Notons que la reproduction de ces principes de gestion de l'information n'est pas une fin en soi mais permet de tester ces hypothèses afin de vérifier si elles reproduisent les dynamiques observées *in vivo*. De plus comme proposé par l'approche ANIMAT, la validation de ces principes neurobiologiques permet de trouver de nouvelles stratégies pour la conception de robots inspirés de la biologie (Guillot et Meyer 2003 ; Wilson 1991).

Tout comme l'étude expérimentale des neurosciences, les neurosciences computationnelles peuvent se placer à différents niveaux de modélisation. Certains modèles permettent d'expliquer au niveau algorithmique le comportement animal dans une tâche donnée, en utilisant typiquement des modèles de décision bayésiens ou encore des algorithmes d'apprentissage par renforcement (Sutton et Barto 1998). Ces modèles sont d'assez haut niveau et ne rendent pas forcément compte d'une représentation fine de la biologie. Ils permettent toutefois de tester au niveau macro des hypothèses sur les processus d'apprentissage sous-jacents. D'autres modèles s'intéressent à des principes électrophysiologiques de plus bas niveau, en modélisant l'activité de plusieurs neurones, le fonctionnement détaillé d'un neurone, ou encore les échanges moléculaires intra et inter neurones. Chaque modèle ayant pour objectif d'expliquer à un niveau donné, l'évolution de processus observés chez des sujets, avec le plus de parcimonie dans les hypothèses utilisées pour la réalisation du modèle. Les modèles de haut niveau tel que les modèles bayesiens seront adaptés pour reproduire le comportement des animaux mais ne permettent pas d'expliquer les principes neuronaux sous-jacents. Un modèle bas niveau prenant en compte les interactions moléculaires permettant la génération de potentiels d'action auront à l'opposé une plausibilité biologique importante. Cependant, il sera difficile d'obtenir un système capable d'apprentissage avec un tel niveau de détail. Ainsi, selon le sujet de l'étude, on se placera à différents niveaux de modélisation.

On pourra citer George E. P. Fox, qui est connu entre autre pour avoir dit : *all models are wrong, but some are useful*. En effet, par essence, tout modèle est faux car ne peut représenter qu'une forme artificielle des mécanismes qu'il veut expliquer. On pourra apposer une citation du mathématicien Norbert Wiener qui complète bien la précédente disant que : *The best material model of a cat is another, or preferably the same, cat.* Lorsque l'on parle de modèle computationnel en neuroscience, nous ne pouvons que représenter des morceaux choisis de processus neuraux, compte tenu de l'impossibilité admise de reproduire actuellement l'intégralité d'un cerveau. Ainsi, chaque région modélisée est considérée dans un espace où l'influence des autres régions n'est pas prise en compte.

Notre approche scientifique dans cette thèse repose sur les neurosciences computationnelles. Nous étudierons les processus neurologiques de l'apprentissage et de la prise de décision en les modélisant et en les comparant aux résultats expérimentaux. La finalité de ce travail n'est pas la création d'un modèle robotique mais de tester par la modélisation plusieurs hypothèses sur les substrats neuronaux de l'apprentissage et la prise de décision, et de les comparer statistiquement.

1.2 Motivations

De nombreuses études ont permis aujourd'hui d'avoir une idée relativement détaillée sur les bases neurales de l'apprentissage et de la prise de décision. Au coeur de ces processus d'apprentissage se trouvent les neuromodulateurs comme la dopamine, la serotonine, la noradrénaline et l'acétylcholine. Doya 2002 propose ainsi que ces différents neurotransmetteurs jouent le rôle de paramètres permettant la mise à jour des connaissances de l'animal et ainsi de guider les processus de sélection de l'action et donc de changer le comportement de l'animal. On trouve ainsi dans la littérature un parallèle entre les algorithmes d'apprentissage par renforcement et l'apprentissage par essais et erreurs effectué dans les ganglions de la base via le signal dopaminergique (voir Figure 1.2).

Deux hypothèses principales ont permis l'établissement de ce parallèle :
- La dopamine encode un signal de type erreur de prédiction de la récompense permettant le renforcement d'actions récompensées (Schultz et al. 1997).
- Les ganglions de la base sont le substrat neural de la sélection de l'action (Mink 1996 ; Redgrave et al. 1999b).

Ainsi, la modulation du signal cortical envoyé aux ganglions de la base par la dopamine peut permettre le renforcement des actions ayant été suivies d'une récompense.

Dans cette thèse nous nous sommes intéressés plus particulièrement au rôle de la dopamine dans l'apprentissage, son lien avec l'adaptation comportementale et son rôle sur la sélection de l'action effectuée par les ganglions de la base. Nous verrons que ce neuromodulateur montre une activité compatible avec un signal d'apprentissage permettant de mettre à jour les prédictions du sujet sur la base de ses erreurs et se trouverait donc

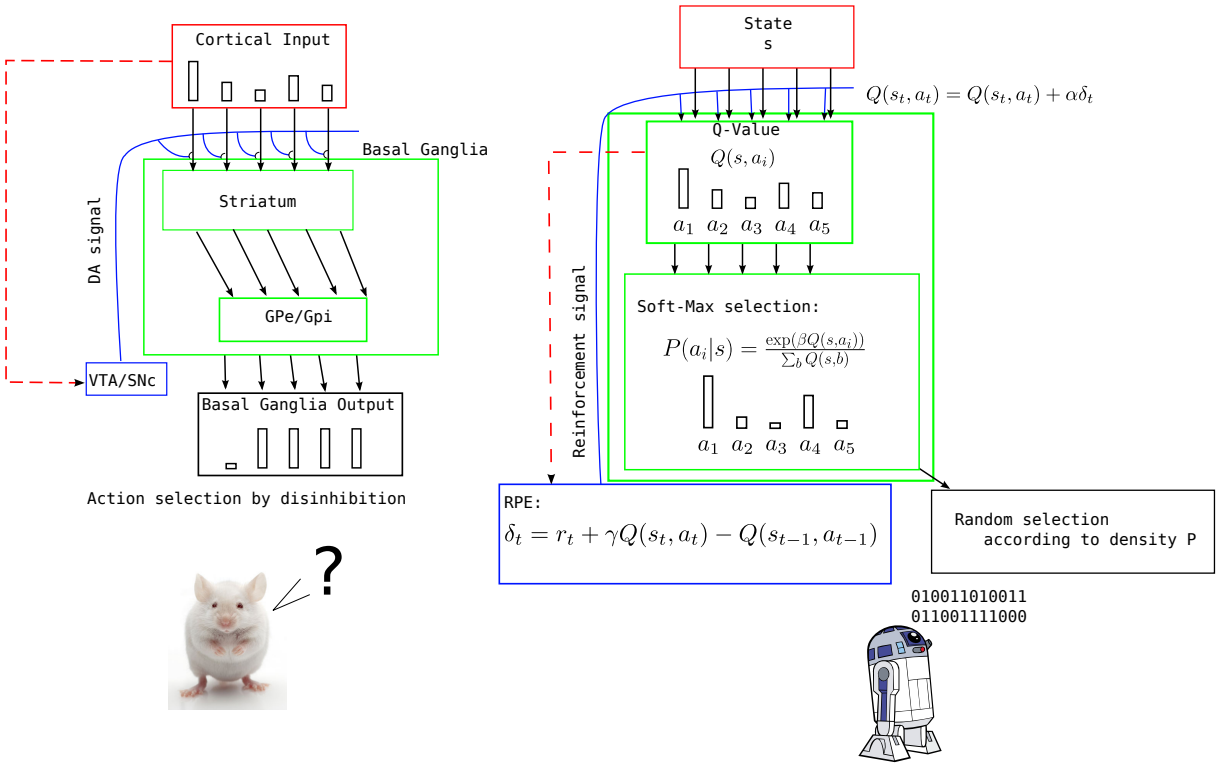

Figure 1.2 – *Comparaison d'un modèle anatomique des ganglions de la base et de la structure de l'algorithme d'apprentissage par renforcement issus de la littérature de l'intelligence artificielle.*

au coeur même des processus d'apprentissage décrits voilà plus d'un siècle par Thorndike. Nous allons utiliser différents niveaux de modélisation pour étudier la dopamine. Mais nous verrons également que nos travaux computationnels incitent à nuancer la théorie dominante à l'heure actuelle, selon laquelle le signal dopaminergique constitue purement et simplement un signal d'apprentissage, sans autre effet sur la sélection de l'action. Nous utiliserons, dans un premier temps, des algorithmes issus de la littérature de l'intelligence artificielle développés dans les années 90 et ne reposant pas sur une représentation biologiquement plausible. Ces algorithmes nous permettront de comparer directement l'information contenue dans des enregistrements électrophysiologiques de neurones effectués sur des rats en comportement par l'équipe de Geoffrey Schoenbaum et Matthew Roesch aux Etats-Unis, avec différentes informations utilisées par les algorithmes d'apprentissage (voir Chapitre 4). Dans un second temps nous étudierons son effet sur un modèle des ganglions de la base biologiquement plausible (Liénard et Girard 2014) afin d'étudier non plus la dopamine en tant qu'information mais en tant que neuromodulateur agissant sur un système de sélection de l'action (voir Chapitre 5 et 6).

Ces méthodes sont complémentaires, car pour que la dopamine soit effectivement considérée comme un signal d'apprentissage, il faut que l'information encodée par son activité soit compatible avec un signal de ce type, ce que les algorithmes issus de l'intelligence artificielle nous permettent de faire. De plus, l'effet de la dopamine sur les processus de sélection de l'action doit également être compatible avec l'hypothèse de guide de l'apprentissage.

Nous verrons que l'effet de la dopamine sur les ganglions de la base dépend des types de récepteurs présents sur les neurones recevant linflux dopaminergique. Les récepteurs peuvent se décomposer en deux types : D1 et D2 [1]. La dopamine a pour effet de renforcer les forces de connexions synaptiques des neurones ayant des récepteurs D1 et de diminuer celles des neurones ayant des récepteurs D2 (Shen et al. 2008). Mais en plus d'un rôle sur le renforcement, la dopamine a aussi un effet instantané qui n'est pas encore bien compris sur la transmission d'informations depuis le striatum jusqu'aux structures de sortie des ganglions de la base, affectant ainsi la sélection de l'action (Gurney et al. 2001a). Ainsi la place des récepteurs à la dopamine dans les ganglions de la base affecte fortement l'effet de la dopamine sur le système. Nous verrons que la vision classique des ganglions de la base, faisant l'hypothèse d'une ségrégation des neurones en fonction de leurs types de récepteurs dopaminergiques, est probablement une (trop forte) simplification (voir Chapitre 2). Ainsi nous testerons plusieurs niveaux de ségrégation pour observer comment cela influence l'effet de la dopamine sur le système dans sa capacité à apprendre et à sélectionner une action.

Enfin, nous proposerons d'étudier comment différents niveaux de dopamine, modélisant par exemple un état parkinsonien, affectent les ganglions de la base et la sélection qui en découle. Nous opposerons les prédictions des différents modèles computationnels, dont certains issus de la littérature (Girard et al. 2008 ; Humphries et al. 2012), avec des

1. En tout, 5 types de récepteurs dopaminergiques ont été catégorisés et chaque type de récepteur appartient à l'une des familles D1 ou D2.

résultats issus de la littérature.

Les problématiques posées dans cette thèse sont ainsi multiples :
- Quel type de signal est encodé par l'activité dopaminergique au cours d'une tâche impliquant un choix parmi plusieurs actions possibles ?
- Quelle est l'influence du niveau tonique de la dopamine sur la sélection de l'action ? Comment change-t-elle l'activité des noyaux des ganglions de la base ?
- Comment différents niveaux de ségrégation des neurones D1 et D2 affectent l'effet de la dopamine sur un modèle des ganglions de la base ?
- Comment modéliser les processus d'apprentissage liés à la dopamine sur un modèle reproduisant un état parkinsonien sous différents niveaux de traitement médical (tel que la LEVODOPA) ?

Globalement, cette thèse a pour objectif d'étudier la relation entre la dopamine et le comportement afin de mieux appréhender son rôle dans l'adaptation comportementale.

1.3 Organisation de la thèse

Cette thèse se compose de 5 chapitres principaux. Deux chapitres bibliographiques introduisant les hypothèses développées dans la littérature qui ont servi de base à nos travaux et trois chapitres présentant nos différentes contributions au domaine.

Le chapitre 2 introduit la place de la dopamine dans l'apprentissage en mettant en évidence son lien avec l'erreur de prédiction de la récompense utilisée par les algorithmes d'apprentissage par renforcement. Nous montrerons également la multiplicité des signaux dopaminergiques découverts notamment en réponse à la punition et discutons d'hypothèses alternatives présentes dans la littérature.

Le chapitre 3 introduit les ganglions de la base à la fois d'un point de vue anatomique et fonctionnel. De plus, nous présenterons plusieurs modèles computationnels de cette structure qui ont influencé de façon significative nos travaux.

Le chapitre 4 présente nos travaux sur le lien entre l'activité dopaminergique enregistrée par Roesch et al. 2007, chez des rats effectuant une tâche à choix multiple, et les fonctions de valeurs utilisées par différents algorithmes d'apprentissage par renforcement simulés sur cette même tâche. Ces travaux ont montré que l'information encodée par l'activité des neurones dopaminergiques ne présente qu'une partie des caractéristiques d'une fonction d'erreur de prédiction de la récompense et que cette information semble au moins partiellement décorrélée de l'adaptation comportementale. Ces résultats suggèrent la présence de systèmes d'apprentissage indépendants du système dopaminergique. Nous trouvons également que le signal dopaminergique peut être mieux modélisé dans ces conditions comme un signal qui encode une mixture entre une erreur de prédiction et un signal de valeur.

Le chapitre 5 présente nos résultats sur la simulation de différents niveaux de dopamine tonique dans notre modèle des ganglions de la base modifié de Liénard et Girard 2014. Nous avons observé comment la dopamine tonique affecte la sélection de l'action

et le compromis exploration/exploitation dans trois modèles différents des ganglions de la base. Nous montrerons que la prise en compte ou non d'une franche ségrégation des chemins direct et indirect dans les ganglions de la base change l'effet de la dopamine sur la sélection. Ceci montre que l'hypothèse de ségrégation forte, sur laquelle reposent des modèles computationnels des ganglions de la base de la littérature (Frank et al. 2004 ; Humphries et al. 2012), doit être au moins partiellement révisée.

Le chapitre 6 présente nos travaux réalisés avec notre modèle des ganglions de la base sur les différences d'apprentissage chez des patients atteint de la maladie de Parkinson sur une tâche introduite par Frank et al. 2004. La tâche implique un système de récompenses et punitions auquel plusieurs types de sujets sont soumis : des sujets atteints de la maladie de Parkinson avec ou sans traitement de remplacement dopaminergique (LEVODOPA) et des sujets sains. Plusieurs études ont observé des différences de sensibilité à la punition et à la récompense chez ces différents sujets (Frank et al. 2004 ; Shiner et al. 2012 ; Smittenaar et al. 2012). Nous montrerons que notre modèle ne prédit qu'une partie des résultats obtenus par l'équipe de Frank et al. 2004, et ne reproduit pas certains comportements qui n'ont pas toujours été reproduits par d'autres études expérimentales (Shiner et al. 2012 ; Smittenaar et al. 2012).

Chapitre 2

Dopamine, erreur de prédiction de la récompense et apprentissage

Sommaire

2.1 Introduction

Dans ce chapitre, nous nous intéressons au rôle d'un neuromodulateur dans l'apprentissage : la dopamine. La dopamine est un neuromodulateur sécrété par les neurones dopaminergiques situés majoritairement dans l'aire A9 et A10 du mésencéphale correspondant respectivement à la *substantia nigra par compacta* (*SNc*) et à l'aire tegmentale ventrale (*VTA*). Ces neurones projettent vers de nombreuses régions corticales et sous-corticales.

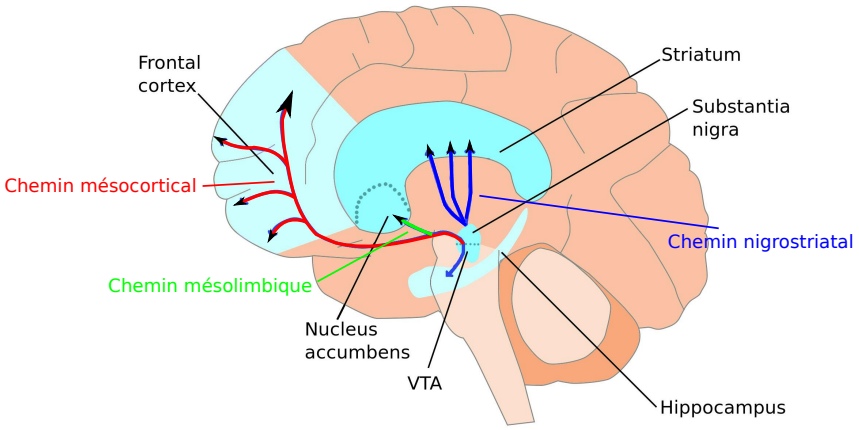

Figure 2.1 – *Illustration des principaux chemins dopaminergiques dans le cerveau.*

Les projections dopaminergiques peuvent être décomposées en trois chemins principaux : mésocorticale, mésolimbique et nigrostriatale (voir Figure 2.1).

Le chemin mésocortical connecte la *VTA* et le cortex frontal. Ce chemin est supposé avoir une implication dans le contrôle cognitif. Il est étudié dans le cadre de la dépression et du stress chronique (Furuyashiki 2012) et la schizophrénie (Masana et al. 2011). Le chemin mésolimbique connecte la VTA au striatum ventral (nucleus accumbens), ayant un rôle important dans l'encodage de la valeur (comme nous le verrons par la suite, la valeur désigne une prédiction de récompense future attendue ; Cromwell et Schultz 2003 ; Khamassi et al. 2008). Il est impliqué dans les mécanismes de l'addiction (Nestler 2001). Le chemin nigrostriatal connecte la SNc avec le striatum dorsal (correspondant au putamen et noyau caudé chez le primate). Ce chemin est associé à la formation des habitudes et aux associations sensori-motrices (Faure et al. 2005 ; Yin et al. 2005).

Les neurones qui sécrètent la dopamine sont particulièrement sensibles aux récompenses ou à des stimuli prédisant une récompense (Ljungberg et al. 1992). Le rôle de la dopamine dans l'évaluation de la récompense est intimement lié à de nombreuses addictions telles que l'addiction à la drogue ou aux jeux d'argent (Maia et Frank 2011). Certains patients ayant un traitement à la Lévodopa – traitement utilisé notamment chez des patients atteints de la maladie de Parkinson qui augmente le niveau de dopamine dans des régions liées à l'évaluation de la récompense –, développent une addiction pathologique aux jeux d'argent. De plus, de nombreuses drogues, telles que la cocaïne, visent les récepteurs dopaminergiques et entraînent une addiction forte. Ainsi, la dopamine est fortement liée à différents phénomènes d'addiction ainsi qu'à l'évaluation de récompense.

La dopamine est donc présente dans de nombreuses régions corticales et sous-corticales et joue un rôle dans la régulation du contrôle cognitif jusqu'au contrôle moteur.

Dans les années 80, Wise émet l'hypothèse que la dopamine est le neurotransmetteur

permettant d'évoquer le plaisir (Wise et Rompré 1989), liant ainsi la dopamine avec l'hédonisme. Avec les travaux de Schultz et collègues dans les années 90 (Hollerman et Schultz 1998 ; Schultz 1998), cette hypothèse de *liking* s'est transformé en *learning*. En effet, ces derniers montrent que l'information encodée dans une tâche de conditionnement Pavlovien, peut être comparée à un signal de renforcement utilisé dans les algorithmes apprentissage par renforcement et plus particulièrement les algorithmes d'apprentissage par différence temporelle (*TD*), développés par Sutton et Barto (Sutton et Barto 1998) – que nous allons décrire ci-après –, généralisant le principe d'apprentissage de Rescola-Wagner (Rescorla et Wagner 1972). Aujourd'hui, malgré un grand nombre d'études validant les résultats obtenus par Schultz et collègues, une troisième hypothèse, défendue notamment par Berridge et collègues, semble vouloir la remplacer ou du moins la compléter : l'hypothèse du *wanting* (Berridge 2007). Cette hypothèse place le signal dopaminergique non pas comme un signal d'apprentissage mais un signal motivationnel. Si la composante hédonique de la dopamine semble remise en question et peu présente dans la littérature actuelle, les hypothèses *learning* et *wanting* coexistent sans que l'on puisse réellement trancher entre les deux hypothèses.

Dans ce chapitre, nous introduirons les algorithmes *TD* permettant d'expliquer l'hypothèse d'encodage de RPE par les neurones dopaminergiques. Nous verrons ensuite les résultats de plusieurs études permettant de valider cette hypothèse. Puis, nous verrons que d'autres études ont trouvé des résultats qui semblent la remettre en question, notamment lors de l'évaluation de la punition. Nous discuterons également de l'hypothèse de Berridge sur le rôle de la dopamine dans la motivation et l'encodage de la salience (*wanting*) et tâcherons d'évaluer la relation entre dopamine et comportement.

2.2 Apprentissage par renforcement

L'apprentissage par renforcement est une classe de problèmes d'apprentissage automatique. Les algorithmes d'apprentissage par renforcement résolvent ce problème par une suite d'essais et erreurs afin de maximiser une fonction de récompense (qui est souvent définie comme la somme de récompenses accumulées sur le long-terme).

2.2.1 Le problème des Processus de décision Markovien (MDP)

Les problèmes que résolvent les algorithmes d'apprentissage par renforcement sont définis par les processus de décision markoviens. Ils permettent de modéliser un contexte où un agent doit apprendre à choisir l'action qui lui rapportera le plus, au sens d'une fonction de récompense, en fonction de son état courant. Son action le fera alors changer d'état et il devra donc de manière séquentielle résoudre le problème afin d'obtenir la meilleure récompense possible. Un état peut par exemple être une position dans un labyrinthe et la récompense de l'agent se situe à un endroit donné de ce labyrinthe. Pour trouver les déplacements – les actions – à effectuer afin de s'approcher de la récompense, l'agent apprendra la relation entre sa position et le déplacement à faire. Ce faisant, l'agent aura appris à associer pour chaque état une action qu'il devra effectuer pour espérer obtenir

une récompense.

Les MDP sont définis par cinq éléments (Sigaud et Buffet 2008) :

- l'espace d'état $\mathcal{S}$, qui représente dans notre exemple l'ensemble des positions de notre labyrinthe
- l'espace des actions $\mathcal{A}$, l'ensemble des directions que l'on peut prendre
- l'axe temporel $\mathcal{T}$.
- une mesure de probabilité $p()$ sur l'ensemble de nos transitions entre états nous donnant pour tout triplet état s_t, état s_{t+1} et actions a_t : $p(s_{t+1}|s_t, a_t)$. C'est-à-dire la probabilité d'arriver à l'état s_{t+1} au temps $t+1$ sachant que l'on est en s_t et que l'on fait l'action a_t. $p()$ donne en quelque sorte un modèle topologique de l'environnement : quels états sont connectés entre eux et par quelle action. Notons que selon le type d'algorithme étudié, l'agent apprendra ou négligera ce modèle de l'environnement.
- une fonction de récompense $r()$, qui associe à chaque transition, (s_t, s_{t+1}), une valeur représentant la récompense.

Les MDP reposent sur l'hypothèse de Markov qui considère que ce qui s'est passé avant l'état courant n'influence pas les futures décisions ou transitions. Dans le cas du labyrinthe, on comprend facilement que cette hypothèse se vérifie : le fait que je sois à une position donnée à un instant t détermine à lui seul le choix, l'action à faire pour aller à la sortie. Le fait que je provienne d'un endroit ou d'un autre ne change pas la solution de mon problème.

On peut également considérer une fonction de transition non stochastique $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, qui pour tout couple (état, action) associe l'état d'arrivée. De même la fonction de récompense peut être réduite à une fonction de $\mathcal{S}$ dans $\mathbb{R}$. La récompense survient lorsque l'on atteint un état quelque soit l'état précédent.

Le but des algorithmes d'apprentissage par renforcement est de résoudre les MDP. C'est-à-dire de trouver quelle action faire dans un état donné pour maximiser la fonction de récompense. Cette association entre état et action est appelée politique ou stratégie. La politique est donc une fonction qui associe à tout état une action : $\pi : s \in \mathcal{S} \to \pi(f) \in \mathcal{A}$.

On cherche, à partir d'une mesure de performance une politique π^* qui maximisera cette mesure. Plusieurs mesures de performance existent mais celle que nous utiliserons est le critère γ-pondéré qui définit la performance comme l'espérance de la somme des récompenses futures. Le facteur γ sert à donner plus ou moins d'importance aux récompenses éloignées dans le temps, il est compris entre 0 et 1. On définit ainsi le critère γ-pondéré comme : $E[\sum_{t=0}^{\infty} \gamma^t r_t]$.

Plus le facteur γ sera petit, moins les récompenses éloignées dans le temps seront importantes dans la décision. On aura ainsi un agent impulsif qui choisira systématiquement les récompenses immédiates. Avec un γ grand, proche de 1, on aura un agent capable de choisir une action lui permettant d'obtenir plus tard une récompense plus grande, plutôt

qu'une petite récompense immédiate. Ce paramètre influence donc grandement le comportement de l'agent apprenant.

Cette définition de la performance permet de définir la notion de fonction de valeur V, qui à chaque état, s, associe à partir d'une politique la valeur espérée de récompense cumulée sur le long-terme en prenant s comme état initial. On a donc pour le critère γ-pondéré :

$$\forall s \in S, V_{\gamma}^{\pi}(s) = E^{\pi}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s]$$

Le but des algorithmes d'apprentissage par renforcement est alors de trouver la politique optimale π^* telles que cette valeur - ou récompense cumulée - soit la plus haute possible. On parle alors de politique optimale $\pi^* = argmax_{\pi} V^{\pi}$.

2.2.2 Algorithme d'apprentissage par différence temporelle

Certains algorithmes résolvent ce problème par une suite d'essais/erreurs. Ils mettent en continu à jour leur fonction de valeur à l'aide d'erreurs de prédictions de la récompense au cours des essais (voir Figure 2.2). À partir de ces fonctions de valeurs, ces algorithmes construisent une politique permettant de maximiser la récompense à long terme (dépendant du facteur γ comme indiqué précédemment). Ces algorithmes se basent donc sur une erreur de prédiction de la récompense ici définie comme l'erreur de différence temporelle (*TD error*), dont le calcul varie en fonction du type d'algorithme utilisé.

Trois algorithmes principaux sont étudiés dans cette thèse : Q-learning, Sarsa et Actor-Critic. Les algorithmes d'apprentissage par renforcement Q-learning et Sarsa reposent tous deux sur le même principe. Ils tiennent à jour une table de valeurs Q qui pour chaque couple (état, action) (noté (s, a)), associe une valeur représentant l'intérêt de choisir l'action a en étant dans l'état s, c'est-à-dire l'espérance de récompense. L'architecture Actor-Critic tient à jour en parallèle un critique, qui apprend à estimer la fonction de valeur V en associant à chaque état une valeur, et un acteur qui apprend une politique, P associant à tout couple (état,action) (noté (s, a)), la probabilité de choisir l'action a sachant que l'on est dans l'état s (i.e. $P(a|s)$).

Ces fonctions de valeur sont mises à jour à partir de l'erreur de différence temporelle δ de sorte que $\forall f \epsilon \{Q, V, P\} : f_{t+1} = f_t + \alpha \delta_t$. Cependant la *TD error*, δ, ne se calcule pas de la même façon selon l'algorithme utilisé :

– Q-learning :

$$\delta_t = r_{t+1} + \gamma \max_a (Q(s_{t+1}, a)) - Q(s_t, a_t) \tag{2.1}$$

– Sarsa :

$$\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{2.2}$$

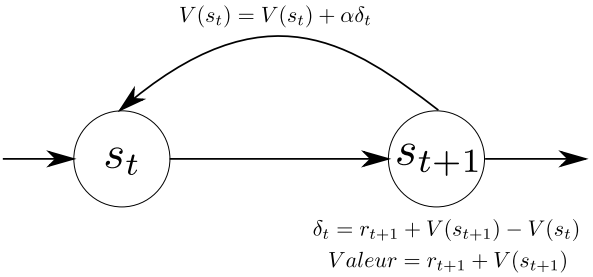

Figure 2.2 – *Illustration du principe d'apprentissage par erreur de prédiction de la récompense. Après chaque transition, l'algorithme évalue l'erreur de prédiction de la récompense et met à jour sa fonction de valeur de l'état précédent. Voir texte pour plus de détails.*

– Actor-Critic :

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{2.3}$$

Ainsi le calcul d'erreur de prédiction peut se décomposer en trois parties : récompense immédiate, r_{t+1}, récompense future attendue, qui varie en fonction de l'algorithme utilisé, et prédiction de récompense calculée dans l'état précédent, $Q(s_t, a_t)$ ou $V(s_t)$ [1]. La différence principale entre ces trois algorithmes repose sur la prédiction de récompense future attendue dans l'état courant. Q-learning fait une prédiction optimiste basée sur la meilleur action accessible, indépendamment du choix futur. Sarsa fait une prédiction réaliste basée sur le choix futur (menant à s_{t+2}), et Actor-Critic fait une prédiction moyenne basée sur l'état courant et indépendant des actions.

La politique est ensuite déduite de la table de valeur, $Q(s, a)$ pour Q-learning et Sarsa, ou par l'acteur P pour Actor-Critic. Il existe deux méthodes principales pour construire la politique : utilisation d'un *softMax* ou $\varepsilon - greedy$. La méthode *softMax* consiste à construire une densité de probabilité basée sur un *softMax* de la fonction de valeur. Ainsi on définit :

$$p(a|s) = \frac{e^{\beta Q(s,a)}}{\sum_b \beta Q(s,b)}$$

où β est la température inverse déterminant le compromis entre exploration et exploitation des connaissances.

La méthode $\varepsilon - greedy$ consiste à choisir la meilleure option prédite par la fonction de valeur $\varepsilon\%$ du temps et une action aléatoire le reste du temps.

Les différences algorithmiques de Sarsa et Q-learning entraînent des différences comportementales parfois importantes. C'est notamment le cas dans une tâche où un

1. Notons que dans cette notation, l'état courant est en t+1 et l'état précédent est en t. Ainsi δ_t est calculé en t+1 pour mettre à jour la prédiction de valeur de l'état t.

Algorithm 1 Apprentissage

Require: état initial : s_0, bloc : mdp

$s_t \leftarrow s_0$
$a_t \leftarrow choixAction(s_t)$
for $i = 0$ à $maxiter$ **do**
  while $s_t nonterminal$ **do**
    $s_{t+1} \leftarrow Transition(s_t, a_t)$
    $a_{t+1} \leftarrow choixAction(s_t)$
    Calcul de δ_t (voir équations 2.1, 2.2, 2.3)
    mise à jour de $Q(s_t, a_t)$ à partir de δ_t
    $s_t \leftarrow s_{t+1}$
    $a_t \leftarrow a_{t+1}$
  end while
end for

agent doit aller vers une récompense positionnée près d'un ravin. La chute dans le ravin étant extrêmement punitive. Dans ces conditions, SARSA choisira de s'éloigner le plus possible du ravin pour aller vers la récompense. Au contraire, Q-LEARNING aura tendance à longer le ravin, empruntant ainsi le chemin le plus court, qui est également le plus risqué lorsque les paramètres de la simulation conserve une part d'exploration. Ainsi, de temps en temps, la politique de Q-LEARNING fera tomber l'agent dans le ravin, alors que SARSA a une politique moins risquée et plus performante dans cette tâche (CHRISTIAN 2003 ; SUTTON et BARTO 1998).

Ces algorithmes sont dit *model free* car ils ne cherchent pas à apprendre un modèle des transitions T. En ce sens, ces modèles sont réactifs car ils n'apprennent pas la conséquence des actions et ne planifient pas à l'avance les actions futures. Par opposition, certains modèles, dit *model based*, apprennent le modèle des transitions et sont capables de planifier des séquences d'actions à l'avance.

2.3 Dopamine et erreur de prédiction de la récompense

La dopamine est aujourd'hui largement considérée comme un signal d'apprentissage encodant une erreur de prédiction de la récompense (RPE), telle que celle utilisée par les algorithmes d'apprentissage par renforcement temporel. Cette hypothèse repose sur le schéma de réponse de l'activité phasique de la dopamine à la récompense et au stimuli prédisant cette récompense. Les neurones dopaminergiques répondent aux récompenses inattendues en début d'apprentissage, ne répondent plus aux récompenses attendues en fin d'apprentissage, et répondent aux stimuli prédisant la récompense après apprentissage.

L'hypothèse de RPE repose principalement sur le transfert de l'activité dopaminergique du moment de la récompense en début de conditionnement au moment du stimulus prédis-

ant la récompense. Nous allons dans cette partie décrire les résultats soutenant cette hypothèse.

2.3.1 Schultz et collègues

Les travaux de Schultz et collègues durant les années 90 (Hollerman et Schultz 1998 ; Ljungberg et al. 1992 ; Mirenowicz et Schultz 1994 ; Schultz 1998 ; Schultz et al. 1993 ; Schultz et al. 1997 ; Waelti et al. 2001), ont permis de mettre en évidence le lien entre l'information portée par l'activité des neurones dopaminergiques et le signal d'erreur calculé par les algorithmes d'apprentissage par renforcement et plus particulièrement par les algorithmes de différence temporelle (TD) présentés précédemment (voir Figure 2.3). Ce signal joue un rôle central dans l'apprentissage et le système dopaminergique est supposé guider la sélection de l'action faite dans les ganglions de la base (Mink 1996 ; Redgrave et al. 1999b ; voir Chapitre 3), en reportant un signal de retour basé sur la différence entre la valeur attendue et la valeur perçue. Cette information de retour peut permettre de mettre à jour la connectivité du striatum afin de permettre l'encodage de la valeur des actions en compétition (Samejima et al. 2005).

Les travaux de Schultz et collègues reposent en majorité sur un même protocole expérimental : un singe est assis devant deux leviers, l'un est associé à la récompense et l'autre à aucune récompense. Après qu'un stimulus visuel a été présenté, le singe doit appuyer sur le levier gauche afin d'obtenir la récompense sous forme de jus de fruits. Si le singe appuie sur le levier de droite, aucune récompense ne lui est délivrée. Les chercheurs ont enregistré les cellules dopaminergiques à différents moments du conditionnement dans ce protocole expérimental. Ils ont observé qu'en début de conditionnement, lorsque le singe a un comportement encore exploratoire sur les deux leviers, les neurones dopaminergiques présentaient une excitation phasique [2] au moment de la récompense (voir Figure 2.3 haut). Cette réponse phasique des neurones dopaminergiques à la récompense a initialement mené à l'hypothèse que la dopamine encode le plaisir associé à la récompense (Wise 1985). Cependant, après que le singe a appris le comportement adapté à la tâche, se concentrant sur l'unique levier associé à la récompense, cette excitation phasique disparaît au moment de la récompense pour apparaître au moment où l'animal perçoit le stimulus prédicteur de la récompense, alors que le plaisir et la motivation de l'animal liés à la récompense sont établis comme toujours présents (voir Figure 2.3 milieu). L'hypothèse proposée est donc que les neurones dopaminergiques répondent aux récompenses non conditionnées (US) inattendues, ne répondent pas aux récompenses prédites mais transfèrent leur réponse au moment où la présentation d'un stimulus conditionné CS) saillant (lui-même inattendu) permet d'anticiper l'arrivée de la récompense. Ceci conduit à un transfert de la réponse phasique dopaminergique de l'US au CS au cours de l'apprentissage de la tâche. Ce transfert d'activité est comparable à l'apparition de salivation chez le chien de Pavlov. Sans conditionnement, le stimulus n'est pas prédictif et l'animal

2. Une activité phasique est une excitation ou inhibition des neurones sur une durée réduite. En opposition au niveau tonique qui est une activité soutenue sur de plus longues durées, par exemple plusieurs minutes voir plusieurs jours.

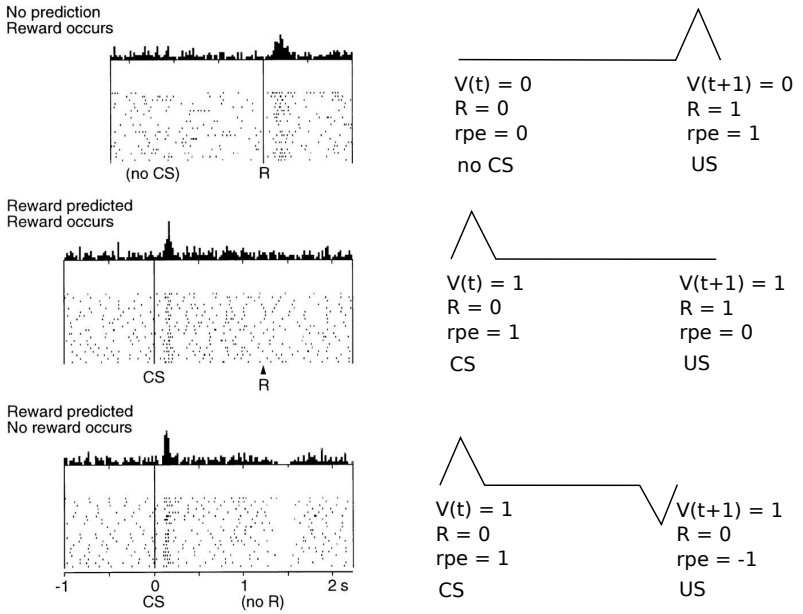

Figure 2.3 – *Comparaison entre l'activité dopaminergique enregistrée chez des singes par Schultz et collègues à différentes étapes du conditionnement de l'animal sur une tâche pavlovienne et la RPE théorique calculée sur les différents états. En haut, activité enregistrée avant conditionnement, milieu et bas après conditionnement dans des essais où l'animal a respectivement reçu la récompense prédite et la récompense a été omise. Enregistrements dopaminergiques adaptés de* Schultz *1998.*

ne salive pas. Lorsque l'animal a appris la contingence entre le stimulus (CS) et la récompense (US), au moment où l'animal perçoit le CS il se met à saliver par anticipation de la récompense.

Le transfert de l'activité dopaminergique au moment de la présentation du CS, au cours du conditionnement semble avoir un lien avec la surprise de l'animal, ce qui a amené à des interprétations du signal dopaminergique comme un signal de surprise et de nouveauté (Bunzeck et Düzel 2006 ; Smith et al. 2006). Cependant on peut expliquer ce type d'activité plus précisément dans le cadre de l'apprentissage et de l'encodage d'une information de type RPE. En effet, avant tout conditionnement, l'obtention d'une récompense chez l'animal est inattendue et imprévisible. Aussi, au moment de la récompense (US), au temps t+1, l'erreur de prédiction vaut : $\delta_t = r_{t+1} + \gamma V(t+1) - V(t) = r_{t+1} > 0$ car $V(t) = V(t+1) = 0$, indiquant une prédiction nulle de récompense au moment de la perception du stimulus, ainsi qu'une prédiction de récompense future nulle au moment de l'US puisque c'est la fin de l'essai (voir Figure 2.3 haut). Ce qui permet d'expliquer le pic d'activité au moment de la récompense.

Une fois le conditionnement effectué, au moment du CS la RPE vaut : $\delta_{t-1} = r_t +$

$\gamma V(t) - V(t-1) = R$ avec $R = 0$, $V_t = R$ car le stimulus est maintenant prédicteur de l'obtention de récompense et $V_{t-1} = 0$ au début de l'essai. Ainsi, l'hypothèse de RPE permet d'expliquer pourquoi l'activité du signal dopaminergique est transférée au CS. Lorsque l'animal reçoit la récompense, Schultz et collègues n'ont enregistré aucune activation phasique des neurones dopaminergiques (voir Figure 2.3), traduisant une RPE nulle. Et en effet, on a bien : $\delta_{t+1} = r_{t+1} + \gamma V(t+1) - V(t) = 0$, car $V(t+1)$ est toujours nul et $V(t) = r_{t+1} = R$ – i.e. la récompense est parfaitement prédite – (voir Figure 2.3 milieu).

Si au contraire l'animal ne reçoit pas de récompense alors qu'elle était prédite par la présentation d'un CS après apprentissage, les neurones dopaminergiques présentent une inhibition phasique de leur activité, au moment où la récompense était attendue par l'animal, ce qui correspond à une erreur de prédiction négative. Encore une fois ce résultat est compatible avec les prédictions de l'apprentissage par renforcement. En considérant que la valeur prédite soit égale à la récompense et qu'aucune récompense n'est donnée on a bien $\delta_t = -V(t) = -r_{t+1} = -R < 0$.

Cette activité reportée est donc bien compatible avec les différents cas de figures possible dans l'évaluation de l'erreur de prédiction de la récompense et présente donc une affinité forte avec cette information. De plus d'autres études ont montré plus tard que cette activité est dépendante de la valeur de la récompense (Tobler et al. 2005), ainsi plus la récompense est importante plus la réponse phasique des neurones dopaminergiques sera importante. Depuis, plusieurs études ont proposé des raffinements de cette hypothèse sur la base de ces résultats. Notamment, certains proposent que la dopamine encode une fonction RPE pondérée par un paramètre d'apprentissage (Daw 2013), ce qui correspond, dans le formalisme des algorithmes *TD*, à : $\alpha\delta$. Un des arguments en faveur de ce produit entre paramètre d'apprentissage et RPE est que si la dopamine a pour finalité le renforcement de la valeur encodée par le striatum, tel que décrit par les algorithmes *TD*, alors le signal doit être au final cette combinaison du paramètre d'apprentissage et de la RPE, puisque la fonction de valeur est mise à jour non pas avec juste δ, mais avec $\alpha\delta$. Il n'y a cependant pour l'instant à notre connaissance pas d'étude expérimentale qui ait validé cette hypothèse théorique. D'autant qu'il est relativement complexe de distinguer une information de RPE simple à une information de RPE pondérée par une valeur, qui ici refléterait le paramètre d'apprentissage.

D'autres ont émis l'hypothèse que ce signal reflète le produit de la pertinence du signal et de la surprise associée (Smith et al. 2006). Cependant, ces études ne font qu'apporter un raffinement quant à l'interprétation du signal mais ne remettent pas en question son rôle dans l'apprentissage ni dans l'encodage d'une information de type RPE. La principale critique de cette hypothèse repose sur le fait que la réponse phasique de la dopamine intervient après moins de 100ms après la réception de la récompense. Certains ont critiqué l'hypothèse d'encodage de RPE en argumentant que ce délais de réponse est trop court pour permettre l'encodage de cette information. Ils proposent que ce signal soit un signal attentionnel permettant un changement rapide de comportement (Redgrave et al. 1999a).

Une autre critique possible sur ces résultats est l'absence d'incertitude et de prise de décision influençant la récompense de la part de l'animal[3]. Or, plusieurs algorithmes d'apprentissage de *TD* proposent différents types de signaux de RPE (voir équations 2.1, 2.2 et 2.3) qui ne peuvent être discriminés sur la base des travaux de Schultz et collègues, car dans ces expériences les animaux n'ont pas à prendre de réelle décision parmi plusieurs possibilités d'actions. Nous présentons dans la suite de cette partie des études plus récentes qui ont enregistré l'activité dopaminergique d'une part au cours de tâches variant les niveaux d'incertitude sur la récompense (pour voir la réponse dopaminergique varier graduellement avec le niveau d'incertitude), et d'autre part au cours de tâches impliquant un choix de la part de l'animal, permettant ainsi de prédire si la RPE encodée par les neurones dopaminergiques contient ou non de l'information sur l'action effectuée par l'animal.

2.3.2 Dopamine et incertitude

Le travail de Fiorillo et al. 2003 a porté sur l'étude de la réponse dopaminergique à l'incertitude. Ils ont en effet enregistré l'activité dopaminergique de deux singes sur une tâche pavlovienne dans laquelle cinq stimuli distincts prédisaient la récompense avec une probabilité spécifique (P = 0 ; P = 0,25 ; P = 0,5 ; P = 0,75 et P = 1). Lorsque P = 0.5, l'incertitude de l'obtention de la récompense est maximale. L'incertitude est nulle dans les conditions P = 0 et P = 1. Les léchages anticipés des animaux montrent que les singes étaient capables de distinguer la récompense associée aux cinq stimuli. De plus, la réponse phasique des neurones dopaminergiques montre une sensibilité importante à la probabilité d'obtenir une récompense prédite par le stimulus (voir Figure 2.4).

En effet, au moment où les animaux perçoivent le CS, l'excitation phasique des neurones dopaminergiques reflète la probabilité d'obtention de récompense. La réponse dopaminergique est ainsi plus importante lorsque le stimulus prédit une récompense importante et une réponse faible lorsque le stimulus ne prédit pas de récompense. On observe l'effet inverse lors de la réception de la récompense. Plus celle-ci était prévisible moins la réponse des neurones dopaminergiques à la récompense était importante.

Ces résultats montrent que l'information encodée par les neurones dopaminergiques est cohérente avec la RPE des algorithmes *TD* qui encode au moment du stimulus $\delta_{CS} = V(CS) = p(r_{t+1}|CS)R$ et au moment de la réception de la récompense $\delta_{US} = r - p(r_{t+1}|CS)R = R(1 - p(r_{t+1}|CS))$. Ainsi au moment de la présentation du CS, les modèles d'apprentissages prédisent bien une RPE proportionnelle à la probabilité d'obtenir la récompense à l'instant d'après. Par opposition, au moment de la présentation de la récompense (US), la RPE théorique est proportionnelle à la probabilité de ne pas obtenir la récompense connaissant le CS. Les résultats sont donc compatibles avec l'hypothèse de RPE, encodant la probabilité d'obtention de la récompense.

La particularité de l'activité enregistrée dans cette étude est l'activité des neurones entre la présentation du CS et l'obtention de cette dernière. En effet, ils ont observé une

3. Dans le protocole expérimental de Schultz et collègues, on peut relever la présence de deux leviers. Or, un des deux leviers n'est pas associé à une récompense et l'animal apprend à ne plus l'utiliser. Il ne présente donc pas un choix valide pour l'animal.

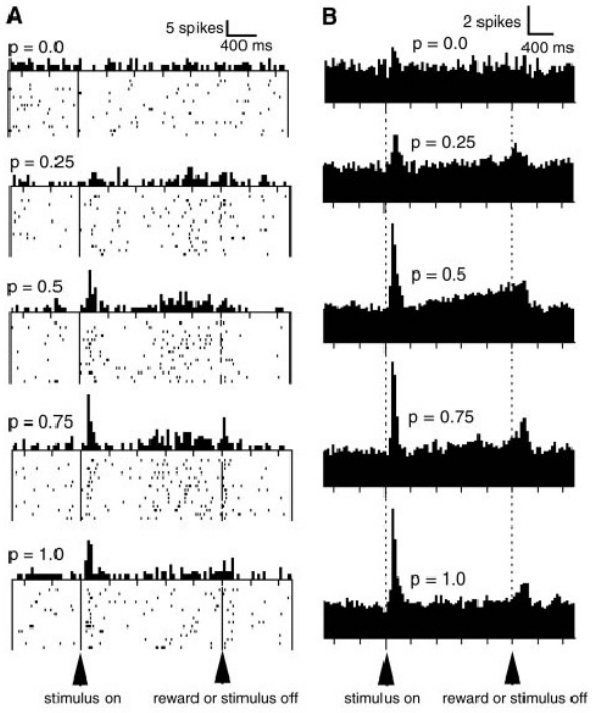

Figure 2.4 – *Activité dopaminergique enregistrée sur une tâche impliquant l'incertitude de la récompense. Plus le CS prédit la récompense avec une forte certitude, plus la réponse des neurones dopaminergiques au moment du CS est forte. D'autre part, une activité dite "rampante" entre le CS et l'US a une amplitude qui est d'autant plus forte que le niveau d'incertitude est fort (p=0.5). A. Histogramme illustrant l'activité d'un neurones au cours des différents essais. B. Histogramme montrant l'activité moyenne de tous les neurones enregistré. Figure reprise de* Fiorillo *et al. 2003*

rampe d'activité croissante durant cette période. Ils ont également pu déterminer que l'amplitude de cette rampe est proportionnelle à l'incertitude de la récompense. Lorsque la récompense (ou l'absence de récompense) est prédite avec certitude, cette rampe n'est pas présente (P = 0 ou P = 1). Cependant, lorsque l'incertitude est maximale (P = 0.5) cette rampe est plus importante. Pour les cas d'incertitude intermédiaire (P=0.25 ou P=0.75) l'amplitude de la rampe est de niveau intermédiaire. Plus il y a d'incertitude sur la récompense, plus cette activité est maintenue (voir Fig. 2.4 droite).

Cet effet de rampe a été examiné par l'étude de Niv et al. 2005, dans laquelle ils défendent l'idée que cette rampe peut être expliquée comme un artefact dû à l'encodage asymétrique de la RPE par les neurones dopaminergique en raison de leur faible activité tonique. En effet, la faible activité des neurones dopaminergiques fait qu'une excitation phasique des neurones dopaminergiques est jusqu'à 6 fois plus importante en amplitude que l'inhibition en cas d'omission.

Ainsi, les résultats de Fiorillo et al. 2003 sont en accord avec l'hypothèse d'encodage de RPE et, au travers de l'explication computationnelle de Niv et al. 2005, mettent

en évidence un encodage asymétrique de la RPE par les neurones dopaminergiques. Ce travail fait également le parallèle entre l'encodage de l'incertitude et le niveau d'activité des neurones dopaminergiques.

La rampe d'activité enregistrée par Fiorillo et collègues rappelle une étude plus récente réalisée par Howe et al. 2013 dans laquelle ils ont également trouvé une rampe d'activité durant l'attente de la récompense. Cependant la rampe d'activité enregistrée dans Howe et al. 2013 ne corrèle pas avec l'incertitude, mais avec la distance et la valeur de la récompense. Ces derniers lient ainsi la dopamine non pas avec l'encodage de RPE, mais à un signal motivationnel (Howe et al. 2013).

Ces résultats semblent donc privilégier un encodage de la valeur et non une erreur de prédiction. Cependant, Gershman (Gershman 2013) argumente que cette rampe d'activité peut également être expliquée dans le cadre de l'encodage de la RPE en ajoutant la prise en compte de l'encodage de la proximité de l'animal à la récompense.

2.3.3 Dopamine : RPE positive et négative ?

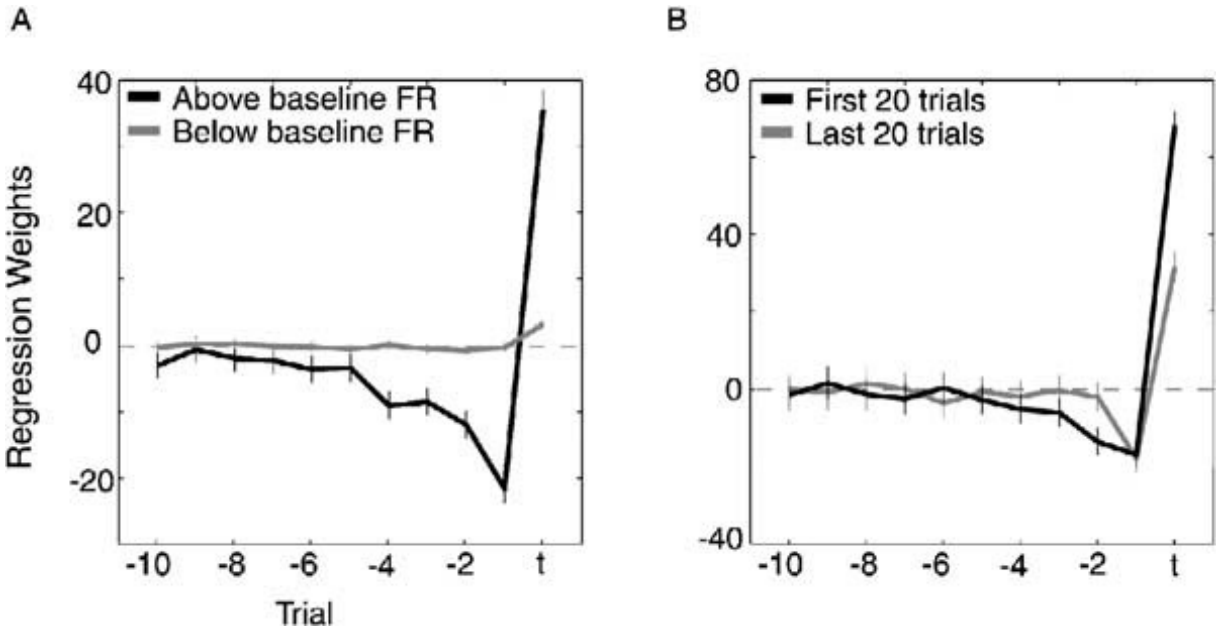

Figure 2.5 – *Poids de régressions obtenus pour reproduire l'activité dopaminergique enregistrée dans la tâche de* Bayer *et* Glimcher *2005. L'axe des abscisses représente les essais précédents et l'ordonnée le poids, w, associé aux récompenses obtenues dans ces essais par régression afin de reproduire l'activité dopaminergique de l'essai courant par la somme pondérée des récompenses précédentes. A. Poids de régressions obtenus sur les essais avec une activité : supérieur à l'activité de base en noir et inférieur à l'activité de base en gris. B. Poids de régressions obtenus sur les 20 premiers essais en noir et les 20 derniers essais en gris. Figure extraite de* Bayer *et* Glimcher *2005.*

Le travail réalisé dans Bayer et Glimcher 2005 a permis de montrer la relation entre les récompenses précédemment obtenues et l'activité dopaminergique enregistrée. À l'aide de régressions, ils illustrent comment les récompenses perçues au essais précédents permettent d'expliquer l'activité dopaminergique courante. La tâche utilisée dans l'étude repose sur un conditionnement instrumental. Les singes doivent faire un mouvement des yeux à un moment précis afin de maximiser la récompense.

Les auteurs ont montré que l'activité dopaminergique observée au cours de cette tâche peut être modélisée par une différence entre la récompense immédiate reçue et une somme pondérée des récompenses obtenues dans les essais précédents, sous la forme : $DA\ r_t - \sum_i w_i r_{t-i}$. Les poids, w, étant le résultat d'une régression (voir Figure 2.5 A et B) et i l'ième récompense obtenue précédemment. Ainsi, on peut identifier la forme du calcul d'une RPE qui fait la différence entre la récompense actuelle moins la prédiction de récompense, qui est calculée en fonction des récompenses obtenues dans le passé. Cependant, cette formulation ne se retrouve pas lorsque la récompense réelle est pire que prévue (voir Figure 2.5 A). Dans ce cas, les poids de régression sont presque tous nuls et ne représentent pas un calcul de RPE. Cela semble indiquer une limitation de la capacité de l'activité des neurones dopaminergiques à encoder une RPE dans le cas où celle-ci est négative. Cette limitation peut s'expliquer par la faible activité tonique de ces neurones, suivant le même argument de Niv et al. 2005. En effet, l'activité tonique des neurones dopaminergiques est d'environ 5Hz, ce qui empêche ces neurones d'encoder une erreur de prédiction négative avec la même variation d'amplitude qu'une erreur de prédiction positive. Ce constat a conduit à différentes hypothèses sur l'encodage possible de la RPE négative qui pourrait être codé par la sérotonine (Daw et al. 2002) ou par la durée de l'inhibition et non son amplitude (Bayer et al. 2007).

Cette étude permet ainsi d'étendre les résultats de Schultz et collègues renforçant l'hypothèse d'encodage de RPE positive. Cependant, elle pose le problème de l'encodage de l'erreur de prédiction négative de la récompense qui, aujourd'hui encore, reste controversé.

Cependant, malgré cette controverse, des résultats d'enregistrements récents suggèrent que la concentration dopaminergique dans le noyau accumbens du rat encode à la fois une RPE positive et négative (Hart et al. 2014). Ainsi, bien que la corrélation entre RPE et activité des neurones dopaminergiques puisse ne pas être évidente lorsque la première est négative, il est possible que l'inhibition de ces neurones soit toutefois suffisante pour modifier significativement la concentration de la dopamine au niveau striatal. Une faible inhibition phasique du signal dopaminergique pourrait en effet avoir des conséquences importantes sur le niveau de concentration de ce neurotransmetteur au niveau du striatum.

2.3.4 Dopamine, valeur et RPE

Comme nous allons le voir, la fonction de valeur et la RPE sont des fonctions relativement proches (ne serait-ce que parce que théoriquement la RPE est calculée à partir de la valeur) et peuvent sous certaines conditions être sensiblement identiques. Enomoto et al. 2011 ont enregistré l'activité dopaminergique au cours d'une tâche dans laquelle il y a deux types d'essais : les essais d'exploration et les essais d'exploitation (voir Fig-

ure 2.6). Au cours de chaque essai, l'animal est confronté à trois stimuli et un seul est associé à la récompense. Le stimulus récompensant reste le même jusqu'à ce que l'animal trouve ce stimulus et pendant deux essais d'exploitation supplémentaires. Ainsi, on peut décomposer l'expérience en groupes d'essais dans lesquelles la récompense est associée au même stimulus. Chaque groupe d'essais comprend une première phase dans laquelle le singe cherche le stimulus associé à la récompense (jusqu'à trois essais si le singe choisit en premier lieu les stimuli non récompensant), et une deuxième phase de deux essais durant laquelle la contingence stimulus-récompense reste inchangée.

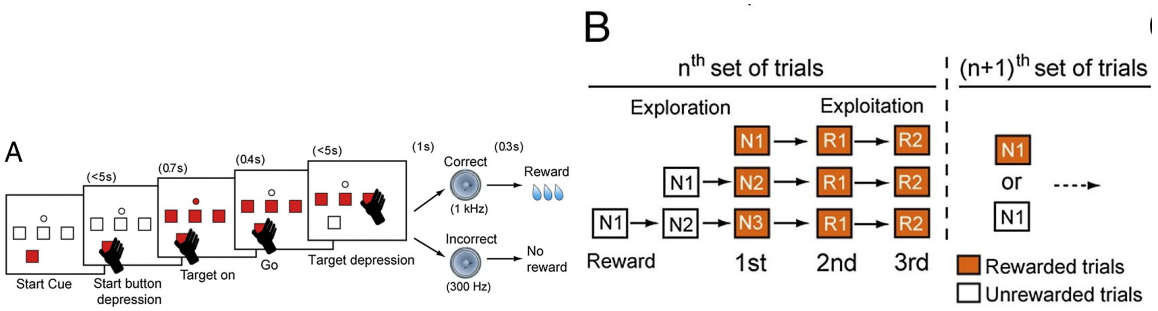

Figure 2.6 – *Tâche à étapes multiples utilisée par* Enomoto *et al. 2011 permettant de montrer que le signal dopaminergique encode les valeurs futures de récompenses attendues sur plusieurs essais. A. Illustration du déroulement d'un essai de la tâche expérimentale. Après un stimulus indiquant le début d'un essai, l'animal se voit présenter trois stimuli représentant les choix de l'animal. Après avoir indiqué son choix, l'animal perçoit un signal auditif indiquant la validité de son choix et reçoit ou non la récompense. B. Déroulement des groupes d'essais comprenant un nombre variable d'essais d'exploration (N1-3) et deux essais d'exploitation (R1-2). La tâche est composée d'une suite de groupes d'essais.*

Les auteurs ont enregistré l'activité dopaminergique au moment du début de chaque essai. Ils ont remarqué que cette activité reflète la valeur des récompenses futures attendues sur les prochains essais. En effet, lors du troisième essai exploratoire, l'activité dopaminergique est maximale prédisant les trois récompenses certaines à venir. Notamment, cette activité est plus forte que durant les deux derniers essais d'exploitation, suggérant que cette forte activité indique une anticipation des récompenses futures. Les auteurs ont montré que ce type d'activité est bien reproduite par une fonction de valeur et non pas par une RPE.

Cette reproduction de l'activité dopaminergique avec une fonction de valeur et non une RPE contraste avec les travaux de Schultz et collègues ainsi qu'avec les résultats de Bayer et Glimcher 2005. Cependant, il faut noter que seule l'activité au début de chaque essai a été reproduite. Or, il est important de noter qu'en début d'essai, la prédiction de récompense précédente, $V(t-1)$, est nulle. On a donc : $\delta_t = r_t + V(t) - V(t-1) = V(t)$. Ainsi, il n'est pas surprenant que dans ces conditions la fonction de valeur et la RPE soient identiques.

Ces résultats montrent pour la première fois les capacités d'anticipation de la récompense sur plusieurs essais. Ainsi l'information encodée par la dopamine prend en compte

la valeur des récompenses obtenues sur plusieurs essais et non pas uniquement sur l'essai en cours. Quoiqu'il en soit, cette étude valide une fois de plus la théorie de l'encodage de RPE par l'activité des neurones dopaminergiques.

2.3.5 Dopamine et encodage de la valeur future attendue

La plupart des études présentées précédemment utilisent une tâche pavlovienne ou n'impliquent pas de la part de l'animal un choix actif entre deux récompenses différentes. Depuis, plusieurs études ont analysé cette activité lors d'une prise de décision afin de comparer l'activité en fonction du choix de l'animal (Day et al. 2010 ; Gan et al. 2010 ; Morris et al. 2006 ; Roesch et al. 2007). Le but de ces études est entre autres de déterminer si l'information de RPE encodée par les neurones dopaminergiques dépend ou non du choix de l'animal. Au moment de la présentation du choix – apparition de stimuli associés aux récompenses –, soit le signal reflète la valeur du choix future, soit l'activité est la même indépendamment du choix de l'animal. Ces deux prédictions correspondent respectivement à l'encodage d'une RPE telle que calculée par Sarsa d'une part ou bien par Actor-Critic ou Q-learning d'autre part.

Ainsi dans l'ensemble des études citées ci-dessus, les animaux (rats chez Day et al. 2010 ; Gan et al. 2010 ; Roesch et al. 2007 et singes chez Morris et al. 2006) sont confrontés à deux types d'essais. Des essais dits choix forcés, dans lesquels seule une option est récompensante, et des essais dits choix libre, dans lesquels les deux options sont associées à une récompense différente, dans lesquels l'animal apprendra à faire le meilleur choix parmi ces deux options pour maximiser sa récompense. Les essais choix forcés servent de contrôle afin de comparer l'activité durant ces choix, dans laquelle l'activité dopaminergique de l'animal est censée refléter une RPE en fonction de la valeur de la récompense associée au stimulus actif, avec l'activité dans les essais choix libre en fonction du choix de l'animal.

Chez Morris et al. 2006, chaque stimulus est associé à une probabilité de récompense (voir Figure 2.7 gauche). Ainsi, si la dopamine encode le choix futur de l'animal, au moment du choix, l'activité dopaminergique reflétera cette probabilité de récompense (à l'image des résultats obtenus par Fiorillo et al. 2003).

C'est en effet ce qu'ont observé les auteurs. L'activité au moment de la présentation des stimuli est dépendante du choix de l'animal. Ainsi, ils ont observé une plus forte activité phasique lorsque l'action que l'animal sapprête à effectuer conduit à la récompense avec une forte probabilité, et une réponse phasique plus faible quand le choix porte sur la récompense associée avec une probabilité plus faible. Ces résultats semblent favoriser une RPE calculée par l'algorithme Sarsa (voir section 2.2), prenant en compte l'information du choix que l'animal sapprête à effectuer. On notera cependant que malgré l'utilisation de probabilités de récompense impliquant une incertitude, les auteurs ne reportent pas de rampe d'activité entre le stimulus et la récompense.

Roesch et al. 2007 testent différentes valeurs de récompense et différents délais dacquisitions de récompenses (voir Figure 2.7 droite). Les animaux apprennent rapidement à choisir les récompenses les plus grandes ou étant accessibles le plus rapidement. Ils mon-

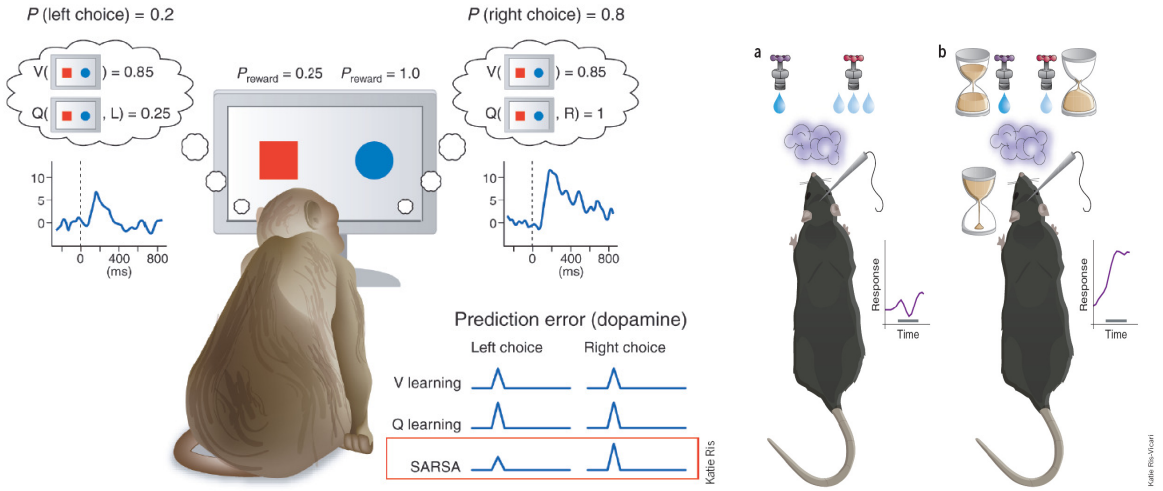

Figure 2.7 – *Illustration des tâches à choix multiples utilisées dans* Morris *et al. 2006 à gauche et* Roesch *et al. 2007 à droite. Gauche. Dans la tâche de* Morris *et al. 2006, le singe se voit présenter deux options associées à une probabilité d'obtention de la récompense. Les auteurs ont observé que l'activité dopaminergique reflète la probabilité d'obtention de la récompense au moment du choix, compatible avec les prédictions de* Sarsa. *Droite. Dans* Roesch *et al. 2007, des rats sont confrontés à un choix entre des récompenses plus ou moins grandes ou accessibles après un délai plus ou moins long. L'activité enregistrée est compatible avec les prédictions de* Q-learning. *Illustration issue de* Daw *2007. Illustration reproduite de* Niv *et al. 2006.*

trent que l'activité au moment de la présentation du stimulus dans les essais libres est identique quelque soit le choix futur de l'animal (donc incompatible avec Sarsa). De plus les auteurs ont montré que cette activité est comparable à l'activité enregistrée durant les essais forcés associés à la récompense la plus attractive, reflétant ainsi une information optimiste. Dans les essais forcés associés à la récompense la moins attractive, l'activité dopaminergique est plus faible, montrant que même dans cette tâche la dopamine encode la valeur de la récompense attendue.

Ces résultats ont permis de conclure que l'information encodée par l'activité des neurones dopaminergique est reproduit une RPE calculée par l'algorithme Q-learning. Ce dernier calcule la RPE à partir de la prédiction la plus optimiste sur les choix futurs (voir section 2.2), et donc propose une RPE indépendante du choix effectif de l'animal, contrairement à Sarsa.

Une partie des travaux réalisés au cours de cette thèse, a consisté à étudier les résultats de cette étude de façon plus fine à l'aide d'une simulation systématique des différents modèles computationnels en compétition pour l'interprétation de ces résultats. Nous avons ainsi montré que l'activité dopaminergique n'est pas tout à fait en accord avec certains critères d'encodage de RPE (voir Chapitre 4).

Dans Gan et al. 2010 et Day et al. 2010, les auteurs ont testé si l'activité dopaminergique encode l'effort à fournir pour obtenir une récompense. Ces deux études parta-

gent un protocole expérimental assez proche, dans lequel les rats doivent fournir un certain nombre d'appuis sur le levier afin d'obtenir la récompense. Contrairement aux deux études décrites précédemment, qui présentaient des enregistrements unitaires de neurones dopaminergiques, ces études enregistrent la concentration de la dopamine dans le noyau accumbens de rats par *fast scan cyclic voltametry*. Ils regardent ainsi directement le niveau de concentration en dopamine dans le striatum ventral et non uniquement l'activité des neurones. Le nombre de pressages de levier nécessaires à l'obtention de la récompense varie de 1 à 32. Les animaux apprennent à choisir le levier associé avec le moins de pressage afin de minimiser l'effort.

Day et al. 2010 montrent que lors des choix forcés, la concentration dopaminergique dans le noyau accumbens est sensible à la fois au nombre de pressages de levier requis et au délais d'obtention de la récompense. Cependant, dans les choix libres, les auteurs ne relèvent aucune différence de concentration dopaminergique en fonction du choix de l'animal et suggèrent que la concentration dopaminergique reflète la meilleure option, validant l'hypothèse de Roesch et al. 2007 selon laquelle la RPE encodée par les neurones dopaminergiques est compatible avec l'algorithme Q-learning.

Gan et al. 2010 ont également montré que bien qu'il y ait une différence significative de concentration dopaminergique pour les signaux indiquant 16 et 2 pressages de levier dans les choix forcés, il n'y a pas de différence significative entre 16 et 32 pressages de levier. Dans cette dernière condition, même s'il y a une préférence comportementale claire pour l'option 16 pressages de levier, la dopamine ne reflète pas cette préférence. Ce résultat est particulièrement intéressant puisqu'il montre que le comportement peut être en partie dissocié de la concentration dopaminergique enregistrée dans le noyau accumbens.

La concentration dopaminergique enregistrée dans le noyau accumbens semble ainsi ne refléter l'effort associé à l'obtention de la récompense que pour un effort modéré. Il semble qu'au-delà d'un certain seuil à déterminer (plus de 16 pressages ?), l'information contenue dans le signal dopaminergique semble ne pas être capable de refléter les différences d'effort, bien que l'animal soit conscient de la différence d'accessibilité des deux récompenses (comme le suggère sa préférence comportementale pour la récompense associée au moindre effort). Il n'est également pas clair si l'information dopaminergique encode l'effort en tant que tel ou bien si la prise en compte de l'effort est intrinsèque à l'encodage d'un signal d'utilité de la récompense.

Dans cette étude, l'activité dopaminergique semble en partie refléter le choix de l'animal, cependant les auteurs se gardent de poser des conclusions définitives sur ce point étant donné l'absence de différence significative dans certains cas lors des choix libres.

Bilan

Ces différentes études n'arrivent donc pas à des conclusions similaires quant à l'implication de l'action future de l'animal dans le codage de la RPE par l'activité des neurones dopaminergiques. Morris et al. 2006 et Gan et al. 2010 observent une activité reflétant au moins partiellement le choix de l'animal, tandis que Roesch et al. 2007 et Day et al. 2010 observent une activité indépendante de ce choix et encodant la valeur de la meilleure

option accessible [4].

Autant on pourrait expliquer plausiblement la différence entre les résultats de Morris et al. 2006 et de Day et al. 2010 ; Roesch et al. 2007 par une différence inter-espèce singes/rats ou des récompenses différentes, autant Day et al. 2010 ont un protocole et une méthode d'enregistrement très similaires à Gan et al. 2010. Une hypothèse pouvant expliquer ces différences est la présence explicite des différentes options chez Morris et al. 2006 et Day et al. 2010 dans l'environnement, avec deux stimuli visuels différents représentant les deux options. Par opposition, Roesch et al. 2007 et Day et al. 2010 n'ont au moment du choix qu'un unique stimulus.

Ainsi l'encodage ou non de la valeur future par la dopamine peut être influencée par la présence explicite ou non dans l'environnement des différentes options au moment du choix. De prochains travaux seront nécessaires pour trancher avec certitude jusqu'à quel point la valeur future est prise en compte par l'information encodée par la dopamine.

Nous avons dans le Chapitre 3 étudié plus précisément les données de Roesch et al. 2007 et nous avons comparé leurs données avec les informations de RPE calculées par Sarsa, Actor-Critic et Q-learning, trouvant que l'interprétation du signal en terme de Q-learning n'était que partiellement défendable, et trouvant une dissociation entre la vitesse d'évolution du comportement des animaux et le degré de convergence (i.e. d'apprentissage) révélé par le signal dopaminergique à l'aide des modèles computationnels.

2.4 Les multiples signaux dopaminergiques en réponse à la punition

Nous avons vu dans les parties précédentes que de nombreuses études ont montré la pertinence de l'hypothèse d'encodage de RPE des neurones dopaminergiques. Cependant, la plupart de ces études ont des protocoles expérimentaux n'impliquant que des récompenses. Or, nous avons vu que l'encodage de RPE négative est possiblement limité à cause de la faible activité tonique des neurones dopaminergiques. Afin de tester plus avant les capacités d'encodage de RPE négative de la dopamine, plusieurs études ont enregistré l'activité dopaminergique dans des tâches impliquant des punitions.

Si l'activité dopaminergique reflète une RPE, alors elle doit également répondre à des punitions par des inhibitions phasiques reflétant d'une manière ou d'une autre l'amplitude de l'erreur de prédiction. Ces études montrent une multiplicité de réponses des neurones dopaminergiques à des punitions.

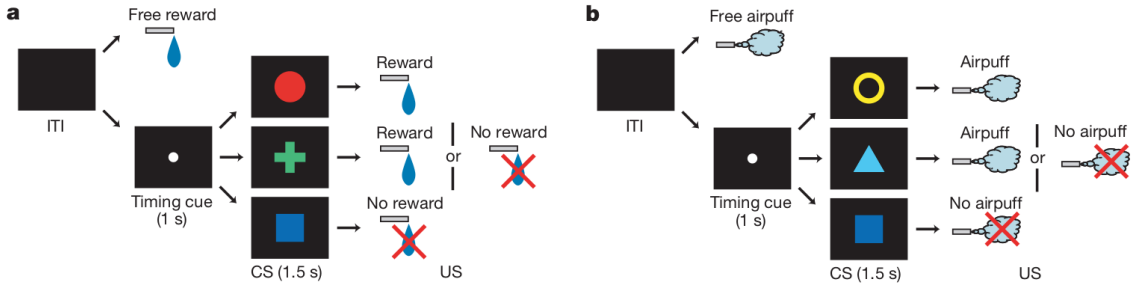

Figure 2.8 – *Illustration de la tâche de conditionnement utilisée dans* Matsumoto *et* Hikosaka *2009. Gauche : bloc appétitif avec trois stimuli indiquant trois probabilités d'obtenir une récompense. Droite : bloc aversif avec trois stimuli associés à une probabilité d'obtenir une punition. Reproduit de* Matsumoto *et* Hikosaka *2009.*

2.4.1 Plusieurs populations de neurones dopaminergique : RPE et |RPE| ?

En utilisant le paradigme de Pavlov, Matsumoto et Hikosaka 2009 ont comparé la réponse des neurones dopaminergiques en réponse à une punition et à une récompense afin d'analyser les différences d'encodage de RPE positive et négative.

Ils ont utilisé deux types de blocs : un bloc où est délivrée une récompense et un bloc aversif où des jets d'air (*air puff*) sont délivrés. Dans chaque bloc trois stimuli différents indiquent la probabilité d'obtenir la récompense – dans les blocs appétitifs – et la punition – dans les blocs aversifs – (voir Figure 2.8). Deux signaux déterministes sont utilisés indiquant 100% et 0% de chance d'obtenir la récompense/punition et un signal stochastique indiquant une récompense/punition avec une probabilité de 50%.

Sur la base de leur activité en réponse à la punition, les auteurs ont identifié deux types de neurones dopaminergiques :

- *airpuff-inhibited* : excité par la récompense et inhibée par les jets d'air.
- *airpuff-excited* : excité à la fois par la récompense et les jets d'air.

Les neurones inhibés par les jets d'air montrent une activité compatible avec l'hypothèse RPE, ayant une excitation phasique forte (resp. inhibition) aux stimuli qui prédisent une récompense (resp. punition) avec 100% de probabilité, une réponse modérée aux stimuli indiquant 50% de chance d'obtenir une récompense ou une punition et aucune réponse aux stimuli qui ne prédisent pas de récompense. Au moment de la récompense, si celle-ci est prédite, il n'y a pas d'activité phasique des neurones, mais une activité phasique importante si la récompense ou la punition sont imprévues. L'activité de ces neurones est donc en accord avec les résultats présentés précédemment et montrent l'encodage de la punition en tant que récompense négative dans un signal de RPE.

Par opposition, les neurones excités par les jets d'air ont une réponse similaire aux

4. Reprenant ici la terminologie employée dans ces articles, par valeur on entend plutôt RPE calculée à partir de la valeur de la meilleure option. Cependant comme vu précédemment, RPE et valeur peuvent dans certaines conditions être proches.

différents signaux indiquant une récompense ou une punition. La réponse dépend surtout de la probabilité du résultat – qu'il soit positif ou négatif. Ces neurones sont en effet excités à la fois par les jets d'air et par la récompense. Cependant, ils semblent moins sensibles à la récompense qu'à la punition. En effet, même lorsque la récompense n'est prédite qu'à 50%, ces neurones n'ont pas de réponse phasique au moment de la réception de la récompense. Au contraire, même si le jet d'air est entièrement prédit, les neurones montrent une importante excitation phasique suivie d'une courte inhibition phasique. Par opposition, moins la punition est prédite, plus la réponse phasique est importante.

Ces neurones montrent une activité incompatible avec l'hypothèse d'encodage de RPE : la punition n'est pas forcément encodée sous la forme d'une récompense négative, s'éloignant ainsi des prédictions de la RPE. Ils semblent encoder une valeur absolue du signal de RPE indiquant une erreur de prédiction certes, mais pas de la récompense. Notons néanmoins qu'un signal de RPE en valeur absolue devrait prédire un pic positif au moment d'une omission (que celle-ci soit d'une récompense ou d'une punition), ce que les auteurs ne trouvent pas dans leurs analyses présentées dans les *supplementary material*. On peut donc penser qu'il s'agit d'un signal de saillance de la RPE mais non strictement de valeur absolue de RPE.

Les auteurs pointent des différences de localisation anatomique de ces différents types de neurones :

- les neurones inhibés par la punition sont localisés dans la région ventromédiale de la *SNc* et la *VTA* et projettent principalement vers la partie ventrale du striatum associée à l'encodage de la valeur (Samejima et al. 2005).
- les neurones excités par une punition sont localisés dans la région dorso-latérale de la *SNc* et projettent principalement vers le striatum dorsal associé aux fonctions motrices.

Cette différence anatomique peut nous donner un indice sur le rôle de ces neurones. Les neurones encodant une RPE (e.g. inhibés par la punition), peuvent permettre d'apprendre à encoder la valeur. La partie dorsale du striatum est, elle, plus souvent associée à la construction de la politique du système guidant le comportement et l'exécution de la commande motrice (Yin et al. 2004, 2005). L'information encodée par les neurones excités par la punition, et encodant une sorte de valeur absolue de RPE, pourrait ainsi servir de base au signal d'erreur de prédiction sur les transitions entre les états observés par *irm* fonctionnelle chez l'homme (Gläscher et al. 2010). Elle est notamment utile pour se représenter un modèle interne de l'environnement et permettant de planifier une suite d'action pour atteindre la récompense le plus efficacement. Dans la théorie de l'apprentissage, cette information est utilisée par les algorithmes basé sur un modèle de l'environnement (couramment nommé *model based*). Il est également possible que les neurones excités par la punition encodent un signal d'erreur global mais pas uniquement lié à la récompense.

Une critique faite à cette étude part du constat que les jets d'air dans le visage ne sont pas des événements réellement aversifs (Fiorillo et al. 2013b). Si cette interprétation est vraie, cela signifie que les deux groupes de neurones dopaminergiques de Matsumoto et

Hikosaka 2009 encodent tous les deux des RPE positives (donc de manière compatible avec la théorie classique) mais un groupe généralisant à deux types d'événements (air-puffs et jus) tandis que l'autre serait spécifique d'un seul type (jus). Toutefois, d'autres études ont confirmé des réponses hétérogènes de la part de différents groupes de neurones dopaminergiques en lien avec des punitions de différentes natures. Wang et Tsien 2011 ont utilisé le conditionnement pavlovien avec deux événements punitifs différents : tremblement de la cage de l'animal et la chute libre ; et une récompense. Ils ont observé l'activité des neurones dopaminergiques de la *VTA* utilisant un enregistrement multi-tétrodes [5]. Sur la base de leurs réponses à la punition et à la récompense, les auteurs ont identifié trois types différents de neurones dopaminergiques dans la *VTA* (voir Fig 2.9).

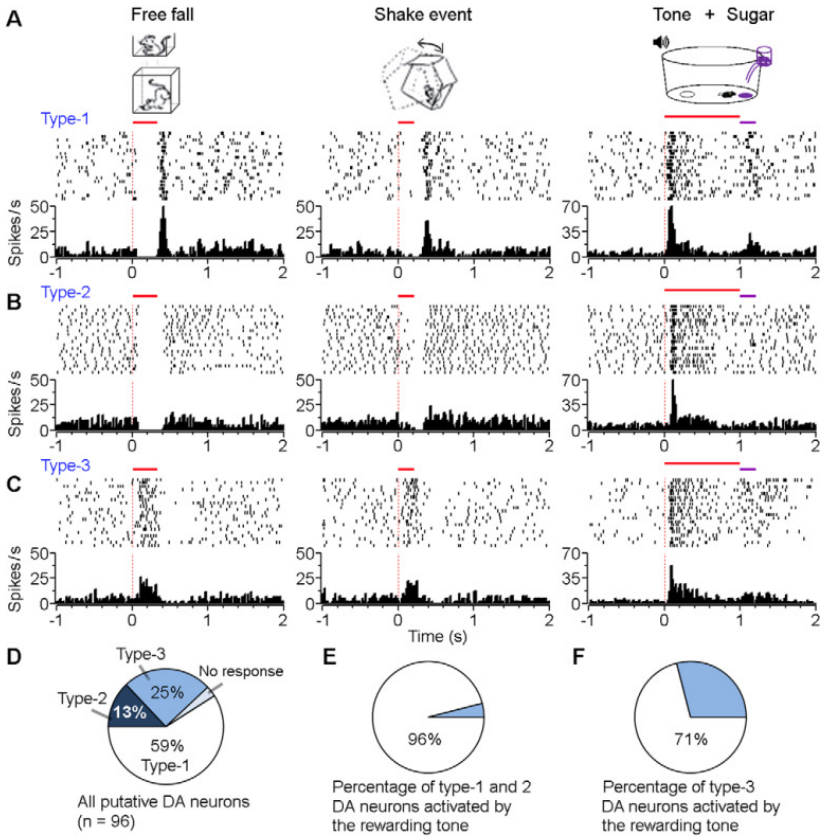

Figure 2.9 – *Activité des différents types de neurones identifiés dans* Wang *et* Tsien *2011. A. Activité des neurones de type 1. B. Activité des neurones de type 2. C. Activité des neurones de type 3. D. Proportion de chaque type de neurone. Figure reprise de* Wang *et* Tsien *2011.*

– Type-1 (59% des neurones enregistrés, 57/96) : montrent une inhibition phasique à la chute libre et aux tremblements, avec une forte excitation de rebond après ces

5. L'enregistrement multi-tétrodes permet d'enregistrer simultanément l'activité de plusieurs neurones d'une même région.

événements. Ces neurones ont également une excitation phasique pour la récompense.
- Type-2 (13%, 12/96) : ont une activité identique aux neurones de Type-1 si ce n'est qu'ils n'ont pas d'excitation phasique de "rebond" à la fin des événements négatifs.
- Type-3 (25%, 24/96) : augmentent leur activité durant les événements négatifs et sont excités par les signaux de récompense, à l'image des neurones *airpuff-excited* observé dans Matsumoto et Hikosaka 2009.

Les neurones de *type-1* semblent considérer la fin des événements négatifs de la même manière qu'ils jugent la récompense. Ainsi la fin de l'événement négatif pourrait être interprété par ces neurones comme un événement positif. Dans cette étude ils montrent également que la réponse des neurones dopaminergiques à un stimulus spécifique varie en fonction du contexte. Lorsque l'animal est dans la cage associée au tremblement, le stimulus produit en effet une réponse phasique différente que lorsqu'il est présenté dans un environnement neutre. Cela montre la capacité des neurones dopaminergiques à encoder une information contextuelle et non uniquement associée au stimulus.

Comme dans Matsumoto et Hikosaka 2009, Wang et Tsien 2011 ont trouvé plusieurs types de neurones dopaminergiques en fonction de leur réponse à des événements aversifs. Cependant, et par opposition aux résultats de Matsumoto et Hikosaka 2009, les neurones enregistré dans Wang et Tsien 2011, se trouvent tous dans la *VTA* ; on ne peut donc pas faire de déduction sur les spécificités anatomiques des neurones dopaminergiques. À noter que les auteurs indiquent que ces neurones pourrait ne pas être dopaminergiques et d'autres expériences sont nécessaires pour confirmer si c'est ou non le cas. Ils pourraient ainsi être des neurones GABAergiques de la *VTA* et servir au calcul de l'erreur de prédiction. Cohen et al. 2012 ont par exemple montré que les neurones GABAergiques de la *VTA* semblent encoder une information de valeur qui pourrait servir au calcul de la RPE.

D'autres études ont également mis en évidence que certains neurones dopaminergiques montrent une excitation phasique lors de la réception d'une punition qui peut prendre différentes formes (Brischoux et al. 2009 ; Cohen et al. 2012 ; Fan et al. 2012). Ces résultats remettent donc au moins en partie en question l'hypothèse d'erreur de prédiction pure avancée par Schultz et collègues et montrent que certains neurones semblent encoder une forme de valeur absolue de l'erreur de prédiction ou de saillance.

De plus les résultats de ces études semblent indiquer que le signal dopaminergique n'est pas unique et qu'il faut bien dissocier quels types de neurones dopaminergiques sont enregistrés pour analyser l'information qu'ils encodent. En particulier certaines études montrent que différents sous-groupes de neurones dopaminergiques sont impliqués dans différents circuits comportementaux liés à l'approche ou à l'évitement (Lammel et al. 2011, 2012).

2.4.2 Fiorillo 2013 : un signal multi-phasique

Les travaux de Fiorillo et collègues (Fiorillo 2013 ; Fiorillo et al. 2013a,b) portent également sur l'étude des réponses à un événement aversif afin de mieux comprendre comment l'information de la punition est prise en compte par les neurones dopaminergiques. Comme mentionné précédemment, les auteurs remettent en question l'aspect aversif des punitions utilisées dans la plupart des études précédentes (notamment chez Matsumoto et Hikosaka 2009). Ils ont ainsi évalué l'aspect aversif des jets d'air dans le visage utilisé dans Matsumoto et Hikosaka 2009 et ont analysé plus précisément le signal dopaminergique lors de la réception de la punition. Ils ont ainsi montré que certains singes ne considèrent pas les jets d'air comme étant aversifs. Pour ce faire, ils ont mesuré la quantité de jus que l'animal était prêt à sacrifier pour éviter les jets d'air. Ils ont observé que certains ne sacrifient pas de jus pour éviter les jets d'air dans le visage. Ainsi il semble que l'aspect aversif de la punition employé par Matsumoto et Hikosaka 2009 soit discutable. Fiorillo et collègues ont donc préféré délivrer les jets d'air directement dans le nez de l'animal ; qui alors était prêt à sacrifier de la récompense pour éviter les jets d'air.

Dans Fiorillo et al. 2013b, les auteurs argumentent que la temporalité fine de l'activité des neurones doit être prise en compte afin de décoder l'information portée par ceux-ci. Leur thèse défend que l'activité des neurones dopaminergiques peut être décomposée en trois phases distinctes. Lors de la première phase (<150ms), l'activité refléterait la salience du stimulus sans prendre en compte la composante motivationnelle du signal. Lors de la deuxième phase d'activité (150-400ms) est encodée la valence (positive ou négative) du stimulus suivie d'une dernière phase de rebond (>400ms) de l'activité avant de revenir à l'activité de base du neurone.

Ce constat remet en question l'analyse de nombreuses études ne se basant que sur une activité moyennée sur plusieurs centaines de millisecondes. Notamment, dans Fiorillo et al. 2013a, l'analyse de l'activité des neurones dopaminergiques sur la phase encodant la composante motivationnelle du signal, les auteurs n'ont pu déceler de différences significatives entre les neurones enregistrés, suggérant que les différences d'activités obtenues chez Matsumoto et Hikosaka 2009 reposent en grande partie sur la première phase d'activité encodant la salience du stimulus. Leur argument porte notamment sur le fait qu'ils ont observé les mêmes différences de réponse pour un stimulus aversif et neutre. Ainsi, cette nouvelle analyse du signal semble montrer que les différences observées dans les études présentées dans la section précédente ne refléteraient non pas un encodage de la punition différent chez les différents neurones mais une prise en compte différente du signal.

Leurs résultats les ont ainsi mené à l'hypothèse que la récompense et la punition ne sont pas encodées sur une dimension unique mais sont deux dimensions séparées, avec la composante punitive possiblement encodée par d'autres groupes de neurones que les neurones dopaminergiques (Fiorillo 2013).

Ces études sont d'une importance cruciale dans l'analyse de l'activité dopaminergique car aucune n'a auparavant analysé le signal dopaminergique de façon multi-phasique, ce qui doit pourtant aujourd'hui être pris en compte.

Cependant, si les travaux de Fiorillo et collègues remettent en partie en question les travaux montrant un encodage de la punition différent par différents groupes de neurones dopaminergiques sur la phase motivationnelle du signal, ils n'adressent pas le constat qu'il existe tout de même une multiplicité de réponses à un même signal.

2.5 Dopamine, salience et comportement

Si l'hypothèse principale pour expliquer l'information portée par les neurones dopaminergiques est aujourd'hui encore l'hypothèse d'erreur de prédiction de la récompense, une autre théorie portée notamment par Berridge semble avoir également une capacité explicative de cette information. Cette théorie est que la dopamine encode la salience des événements et est liée à un renforcement motivationnel plus qu'à un signal d'apprentissage.

Dans les études présentées jusqu'à présent, le lien entre dopamine et comportement n'a pas été directement adressé. Nous avons vu que la dopamine et le comportement évoluent au cours de l'apprentissage mais pas le lien de causalité entre les deux. Si la dopamine encode une erreur de prédiction de la récompense et est, comme dans la théorie de l'apprentissage par renforcement, le signal d'apprentissage, alors une stimulation du système dopaminergique doit entraîner un apprentissage. Une critique principale que fait Berridge (Berridge 2007) est que, le fait que la dopamine encode un signal similaire à une RPE utilisée par les algorithmes *TD*, n'implique pas qu'elle ait le même rôle de signal de renforcement. Il faut vérifier si le comportement suit ou pas le lien de causalité dopamine→comportement, supposé par la théorie. La question n'est plus de savoir quel type d'information est encodé, mais plutôt quel est le rôle de cette information ? Le signal de RPE encodé par l'activité des neurones dopaminergiques est-il nécessaire pour l'apprentissage ou bien l'apprentissage forge-t-il le signal ? Il s'agit ici de savoir qui du signal dopaminergique ou de l'apprentissage est créé par le second.

Dans la littérature actuelle, il est encore difficile de déterminer la place exacte de la dopamine par rapport au comportement et si le signal dopaminergique est réellement un signal d'apprentissage ou non. Berridge 2007 passe en revue les différentes hypothèses de la dopamine et développe un argumentaire selon lequel la dopamine n'est pas un signal d'apprentissage mais un signal motivationnel, passant de l'hypothèse *learning* à l'hypothèse *wanting*.

Notons toutefois qu'après que Berridge a défendu son hypothèse, plusieurs études ont montré un lien de causalité entre dopamine et comportement (Kravitz et Huber 2003 ; Steinberg et al. 2013). La dopamine est en effet suffisante pour créer une association stimulus-récompense et l'absence de dopamine affecte l'apprentissage. Cependant, il n'y a pas de lien quantitatif et graduel rigoureux entre dopamine et comportement. Pour être plus précis, comme nous le verrons dans les résultats du Chapitre 4, dans certaines études, l'évolution du signal dopaminergique ne semble pas indiquer la même vitesse d'évolution que celle du comportement ; il semble au contraire que la vitesse d'adaptation du comportement observée expérimentalement s'explique mieux comme le résultat d'une influence de plusieurs systèmes d'apprentissage et de décision dont un seul des systèmes

serait celui impliquant les signaux RPE dopaminergiques.

L'hypothèse de Berridge est que la dopamine encode la salience et augmente l'envie, la motivation pour l'obtention de la récompense (Berridge 2007). Il faut ainsi dissocier la notion de plaisir de l'envie. Dans un grand nombre d'études, la dopamine a été interprétée comme encodant une composante hédonique de la récompense ce qui a mené à l'hypothèse portée par Wise (Wise 1985). Dans le contexte de l'hypothèse d'encodage de la salience, la dopamine code pour un signal motivationnel sans composante hédonique.

On peut considérer le *wanting* comme l'envie qu'a une personne souffrant d'addiction de consommer de la drogue. La consommation ne procure pas forcément du plaisir, mais le besoin de la consommer est grand. Ce besoin se traduit ainsi en *wanting* et non en *liking*. La cocaïne entraînant l'arrêt de la recapture de la dopamine, entraîne une augmentation progressive de la concentration dopaminergique dans les synapses et une motivation surdimensionnée à la consommation de la drogue.

Cependant, ce phénomène d'addiction peut également s'expliquer avec l'hypothèse de l'erreur de prédiction (Keramati et Gutkin 2013 ; Redish 2004). L'augmentation disproportionnée du niveau dopaminergique entraîne un renforcement grandissant de la valeur associée à la consommation de la drogue conduisant à un comportement addictif.

Pour déterminer le rôle de la dopamine entre apprentissage et motivation, il faut prendre en compte l'évolution du comportement liée à la dopamine. Berridge et collègues se basent sur la considération que si la dopamine joue un rôle dans l'apprentissage et non dans la motivation, alors un changement lié à l'activation de la dopamine sera plus long puisqu'il nécessitera un temps d'apprentissage. Si au contraire la dopamine est liée à un signal motivationnel, alors le changement sera plus immédiat (voir Figure 2.10).

Ils montrent par exemple que lorsqu'un animal apprend une contingence entre un *CS* et un résultat négatif, le fait de rendre le résultat soudainement attractif entraîne un changement comportemental immédiat chez l'animal. Sa motivation nouvelle lui fera montrer une réponse inconditionné d'envie face au CS de façon quasi-instantanée. Cela met en évidence le fait que le changement de comportement ne se fait pas après un réapprentissage de la valeur du CS, mais est immédiate et due à un changement motivationnel face à la récompense.

Cependant, ce constat est possiblement critiquable dans le sens où l'apprentissage de la contingence CS-US peut se faire de manière liée à l'homéostasie. Ainsi, la valeur apprise peut dépendre de l'état physiologique de l'animal. L'apprentissage porte alors sur le type de récompense et l'effet qu'elle aura sur l'état de l'animal. Le fait que celle-ci devienne désirable créera un changement immédiat. De plus ce type de changement brutal de comportement peut être dû à une décision corticale n'empêchant pas un processus sous cortical de désapprentissage plus lent (cela peut être le cas avec l'hypothèse de changements contextuels – *task set* – développée notamment par Collins et Koechlin ; Collins 2010).

Un autre constat de Berridge est que la dopamine n'est pas nécessaire à l'apprentissage. Certaines études ont montré que des souris incapables de synthétiser de la dopamine ont des capacités d'apprentissage (Cannon et Palmiter 2003). Ces résultats mettent

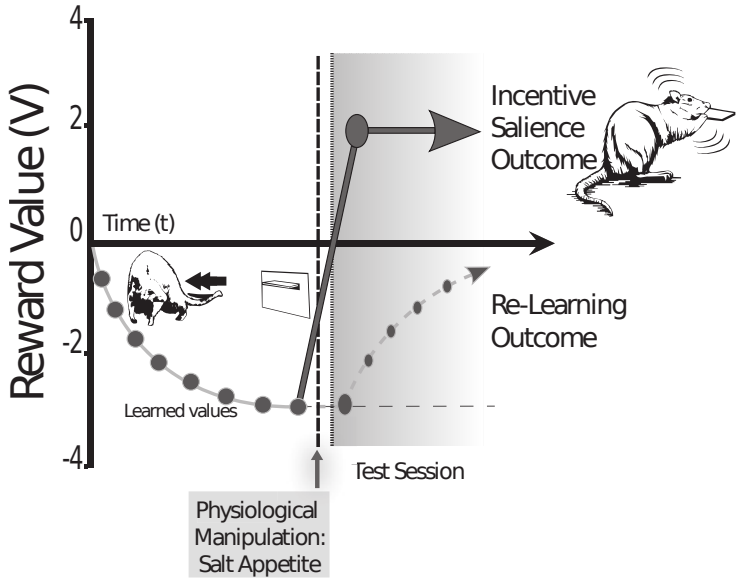

Figure 2.10 – *Illustration du changement comportemental pour l'hypothèse motivationnelle ou lié à l'apprentissage après avoir rendu un résultat attractif. Figure reproduite de* Robinson *et* Berridge *2013.*

en évidence le fait que l'apprentissage n'est pas forcément porté par la dopamine. De plus, des études d'apprentissage chez des souris ayant un niveau dopaminergique élevé montrent que l'augmentation de la concentration en dopamine crée une augmentation de la motivation pour l'obtention de la récompense mais n'augmente pas les capacités d'apprentissage liées à la récompense (Yin et al. 2006). Ainsi, il est probable qu'il y ait plusieurs systèmes d'apprentissage et de décision ; possiblement un cortical moins sensible au signal dopaminergique et un sous cortical (Lesaint et al. 2014).

La relation entre dopamine et comportement est donc possiblement plus complexe qu'initialement prédite par l'hypothèse de RPE de Schultz et collègues. L'optogénétique est une technique permettant chez des souris ou rats mutant(e)s d'activer de façon sélective certaines populations de neurones à l'aide d'une activation lumineuse sans changer l'activité des neurones environnants. Cette technique a permis de montrer que l'activation de neurones dopaminergiques seule permet de renforcer une association entre un stimulus et une récompense. Notamment, Steinberg et al. 2013, montrent que l'activation du système dopaminergique permet de supprimer le blocage de l'association stimulus-récompense lorsque celle-ci est déjà prédite par un premier stimulus [6]. Ce résultat montre que la dopamine seule est suffisante pour créer une association CS-US (sans forcément dire qu'elle est nécessaire, puisqu'un autre mécanisme d'apprentissage pourrait faire cette association dans une autre partie du cerveau ; Daw et al. 2005 ; Khamassi et Humphries 2012). Cela permet de montrer que la dopamine seule peut être la conséquence d'un

6. Le *blocking* est un phénomène qui fait qu'un stimulus ne peut être associé à une récompense si il est couplé avec un autre stimulus prédisant déjà cette récompense.

changement comportemental.

Ainsi, on est en droit de penser que la dopamine permet de renforcer des comportements, bien que tous les comportements ne soient pas nécessairement la conséquence d'un apprentissage lié à la dopamine.

On notera que McClure et al. 2003 ont proposé que la dopamine fasse la passerelle entre *liking* et *wanting*. Ils émettent l'hypothèse que le signal dopaminergique est à la fois utilisée comme signal d'apprentissage au niveau phasique et est utilisée comme un biais pour la sélection de l'action au niveau tonique. Le niveau tonique de la dopamine, biaisant la sélection, étant la composante motivationnelle de la dopamine.

2.6 Conclusion

De nombreuses études se sont penchées sur le problème de la fonction des neurones dopaminergiques. Aujourd'hui il est couramment admis dans la littérature que la dopamine encode un signal de RPE (voir Colombo 2014 ; Glimcher 2011 pour des synthèses sur cette hypothèse). Cependant l'effet de ce neurotransmetteur sur le comportement reste encore sujet à discussion et l'hypothèse motivationnelle de la dopamine semble avoir une réelle force d'explication. Des travaux permettant la comparaison simultanée de l'évolution du signal dopaminergique et du comportement seront très certainement nécessaires pour permettre de mieux comprendre le rôle de la dopamine. C'est en partie ce que nous avons tenté de faire par des analyses fondées sur les modèles computationnels sur des enregistrements dopaminergiques précédement effectués par Roesch et al. 2007 et qui seront présentés dans le chapitre 3.

Chapitre 3

Les ganglions de la base

Sommaire

3.1 Introduction

Les ganglions de la base, autrement appelés noyau gris centraux, sont un ensemble de noyaux sous-corticaux interconnectés situés dans les régions télencéphalique et diencéphalique. La structure des ganglions de la base est présente dans de nombreuses

espèces animales telles que le primate, le rongeur ou la lamproie. Elle est plus généralement commune à toutes les espèces vertébrées (Reiner 2009 ; Stephenson-Jones et al. 2011 ; voir Figure 3.1). Cependant, selon les espèces, les ganglions de la base peuvent présenter une anatomie différente, mais il semble que leur fonction reste inchangée.

L'étude des ganglions de la base remonte à plus d'un siècle pour son implication dans le contrôle moteur. Dès 1928, Wilson (Wilson 1928), suggère que les ganglions de la base sont un système moteur extrapyramidal, donc parallèle et indépendant du système moteur pyramidal. Ce n'est que bien plus tard que des études permettront de montrer que les ganglions de la base sont plutôt un système lié au contrôle moteur, de façon non pas parallèle, mais en relation directe, via le thalamus, avec le cortex moteur. À la fin des années 80, deux études proposent que les ganglions de la base contrôlent l'excitation et l'inhibition de régions corticales (Albin et al. 1989 ; DeLong 1990). Ces études mettent en évidence l'implication des ganglions de la base dans les maladies de Parkinson et de Huntington, toutes deux liées à des troubles moteurs.

L'étude des ganglions de la base n'est aujourd'hui plus limitée au cadre du contrôle moteur. En effet, cette structure est étudiée pour son implication dans l'apprentissage, la sélection de l'action et pour son rôle dans de nombreuses pathologies, liées notamment à des dérèglements du système dopaminergique.

Les travaux de Mink 1996 et Redgrave et al. 1999b ont popularisé dans la littérature l'idée que les ganglions de la base forment un système de sélection de l'action. Nous verrons que la position anatomique des ganglions de la base, ainsi que leur relation avec les différentes régions corticales en font, en effet, un candidat idéal pour l'intégration d'un grand nombre d'informations nécessaires à la sélection de l'action. De plus, son rôle dans le contrôle moteur place cette structure dans une position privilégiée pour la génération des commandes motrices associées à la sélection. Nous verrons également que les ganglions de la base sont une cible majeure du système dopaminergique, qui est au centre de nombreuses pathologies pouvant impliquer le contrôle moteur ainsi que des capacités cognitives de plus haut niveau, tels que des troubles dans l'apprentissage et la prise de décision (voir Chapitre 2).

Dans ce chapitre, nous nous intéresserons plus particulièrement aux fonctions et à l'anatomie des ganglions de la base en lien avec la sélection de l'action et au travers de son interaction avec le système dopaminergique. Nous verrons comment la maladie de Parkinson affecte le fonctionnement des ganglions de la base. Nous introduirons plusieurs modèles computationnels des ganglions de la base et discuterons de leur anatomie par le prisme des récepteurs dopaminergiques présents dans les neurones épineux moyens (*MSN*) du striatum.

3.2 Les noyaux des ganglions de la base

Les ganglions de la base sont composés de quatre noyaux principaux : le striatum (*Str*), le noyau subthalamique (*STN*), le globus pallidus (*GP*) et la substance noire (*SN*). Ces noyaux sont fortement interconnectés et forment une structure complexe encore largement

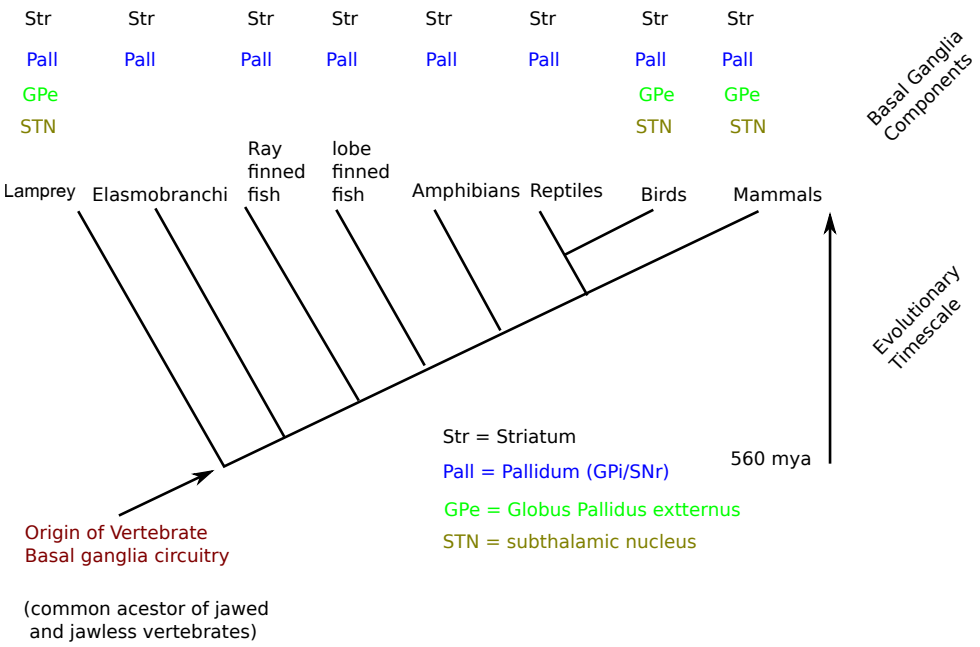

Figure 3.1 – *Arbre phylogénétique des ganglions de la base chez les vertébrés (illustration modifiée de* Stephenson-Jones *et al. 2011).*

débattue dans la littérature. Dans cette partie, nous allons détailler plus précisément les connexions intrinsèques et extrinsèques des ganglions de la base, qui sont illustrées dans la Figure 3.4. Nous verrons ainsi les connexions anatomiques des ganglions de la base avec les régions corticales ainsi que la connectivité interne des ganglions de la base. Nous décrirons comment les ganglions de la base sont organisés de façon topographique en conservant plusieurs boucles parallèles correspondant à différentes entrées corticales.

À noter que nous ne décrirons pas les relations entre les ganglions de la base et les autres régions sous corticales. Pourtant, les ganglions de la base sont également fortement connectés avec l'hippocampe, le cervelet, l'amygdale et le colliculus supérieur, qui interviennent à différents niveau comportementaux. Le colliculus supérieur par exemple, via son interaction avec les ganglions de la base, permet la génération et la sélection de saccades oculaires (N'Guyen et al. 2014 ; Thurat et al. 2015). Cependant la description de ces interactions va au-delà du champ de cette thèse.

3.2.1 Le striatum

Le striatum est le noyau le plus massif des ganglions de la base. Il est estimé être composé de 30.4 millions de neurones chez le macaque et 110 millions chez l'homme (Yelnik et al. 1991). Par comparaison, les autres noyaux des ganglions de la base sont composés de moins d'un million de neurones chacun (Hardman et al. 2002).

Chez le primate, le striatum est composé du putamen, du noyau caudé et du noyau

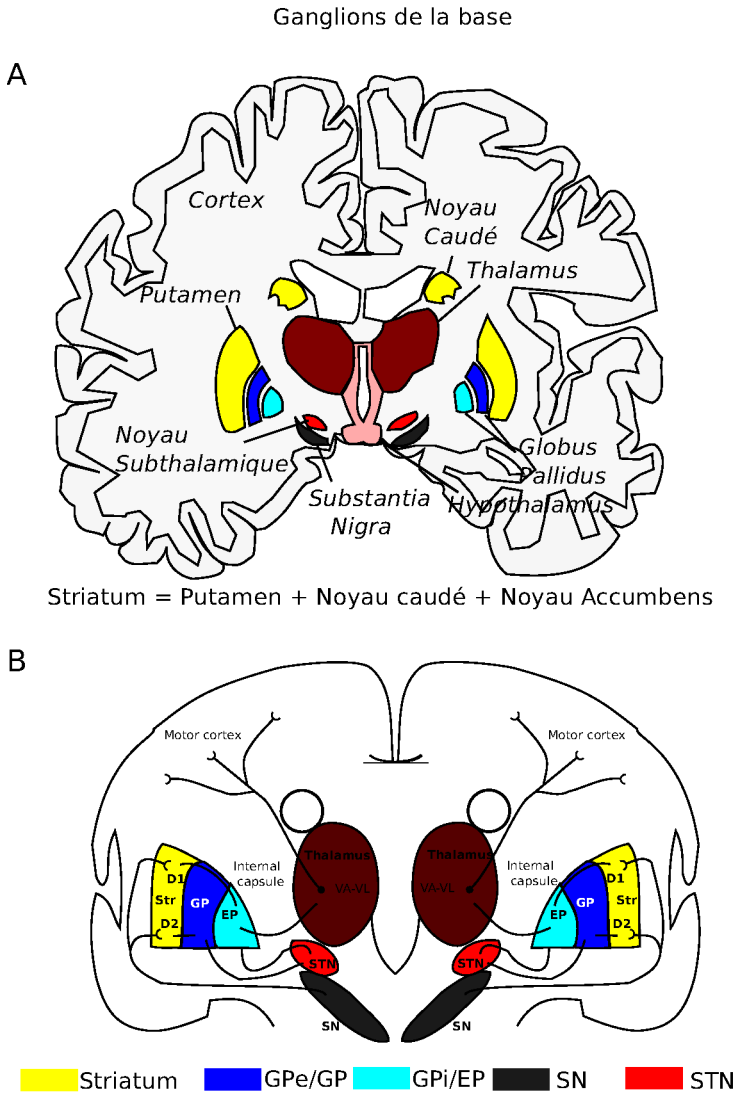

Figure 3.2 – *Illustration de l'anatomie des ganglions de la base. A. Ganglions de la base chez le primates. On notera que le striatum est composé de plusieurs noyaux distinct chez le primates : le putamen, le noyau caudé et le noyau accumbens. B. Ganglions de la base chez le rongeur. Str : striatum ; GPe : globus pallidus externe ; GPi : globus pallidus interne ; EP : noyau entopédonculaire ; STN : noyau subthalamique ; SN : substance noire.*

accumbens. On peut plus généralement diviser le striatum d'un point de vue fonctionnel en striatum dorsal (ou néostriatum), qui inclut la majeure partie du putamen et du noyau caudé, et striatum ventral composé du noyau accumbens et les parties ventro-médiales du putamen et du noyau caudé. Chez le rat, le striatum est formé d'un seul noyau et est également étudié selon le gradient ventromédian-dorslatéral (Voorn et al. 2004). On dissociera particulièrement la partie ventrale du striatum, le noyau accumbens, des parties dorsales, en distinguant la partie dorso-médiale de la partie dorso-latérale.

Les neurones du striatum

Le striatum contient différents types de neurones. Les neurones les plus nombreux sont les neurones épineux moyens (*MSN*) qui représentent environ 95% des neurones du striatum (Gerfen et Wilson 1996 ; Tepper et al. 2008).Ces neurones sont globalement silencieux et ont une activité assez faible (1Hz). Comme de nombreux neurones des ganglions de la base ils sont gabaergiques. Ces neurones projettent vers l'extérieur du striatum par opposition aux autres neurones qui ne projettent que de façon interne au striatum. Les 5% de neurones restant sont constitués de différentes sous population d'interneurones : les interneurones à taux de décharge rapide (*FSI*), les interneurones cholinergique (*TAN*), les interneurones immunoréactifs à la tyrosine hydroxylase (TH+), les interneurones réactif à la calretinin (*CR+*) et les interneurones à taux de décharge à seuil bas (*PLTS*) (Gittis et Kreitzer 2012 ; Kawaguchi et al. 1995). À l'exception des *TAN*, les interneurones sont tous gabaergiques. La connectivité interne au striatum entre les différents types de neurones est encore mal connue. Cependant tous les interneurones projettent vers les *MSN* (Nakano et al. 2000).

Afférences corticales

Le striatum est le noyau principal d'entrée des ganglions de la base et reçoit une excitation glutamatergique de l'ensemble des régions corticales. Ainsi, il reçoit des projections des parties limbiques du cortex, telles que le cortex orbito-frontal ou le cortex préfrontal ventro-médial, jusqu'à des parties motrices, telles que l'aire motrice supplémentaire et le cortex moteur. L'organisation des projections cortico-striatales se fait de façon ordonnée et similaire chez le primate et le rat. En effet, les différentes régions corticales projettent vers des parties bien définies du striatum, le subdivisant en parties motrice, associative et limbique (Alexander et al. 1986 ; Haber 2003 ; Haber et Calzavara 2009 ; Haber et al. 2000 ; Joel et Weiner 2000 ; Parent 1994 ; Parent et Hazrati 1995 ; Redgrave et al. 2011). Cette organisation topographique est préservée dans tous les noyaux des ganglions de la base (voir Figure 3.3), formant des circuits cortico-basaux parallèles associés au contrôle de différents niveaux comportementaux.

Chez le primate, la partie motrice du striatum comprend les parties dorso-latérales du putamen et du noyau caudé. Chez le rat, la partie motrice du striatum correspond au striatum dorso-latéral. Cette partie motrice reçoit des projections des neurones du cortex moteur primaire, cortex prémoteur et cortex moteur supplémentaire. On peut noter que de nombreuses études ont montré que cette partie du striatum est nécessaire pour faire

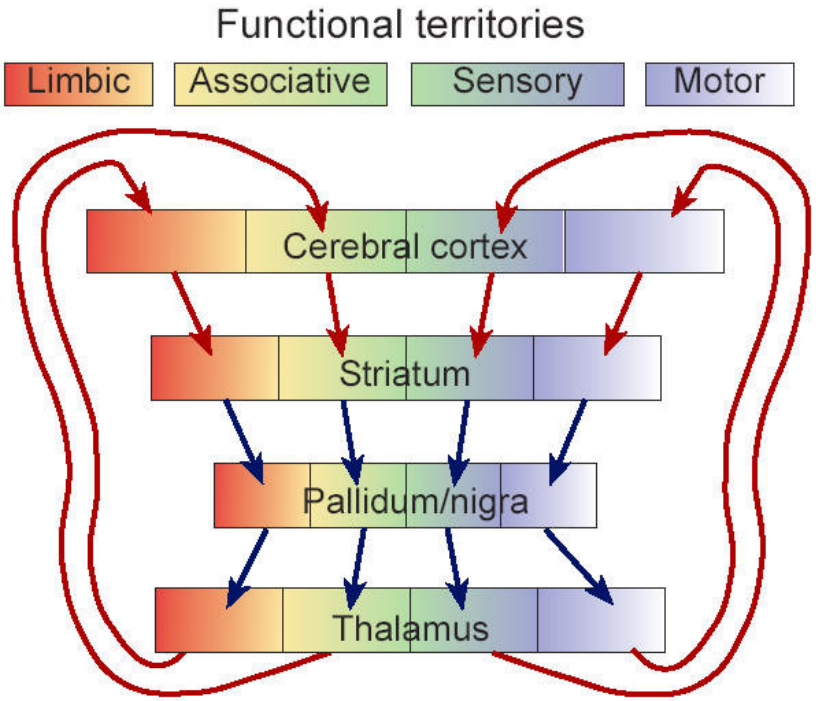

Figure 3.3 – *Organisation topographique des ganglions de la base. Figure reprise de* Redgrave *et al. 2011, décrivant les boucles parallèles limbique, associative et motrice présentent dans les ganglions de la base. On pourra noter que bien que le noyau subthalamique ne soit pas représenté dans la figure, cette organisation topographique est conservée dans ce noyau (*Haynes *et* Haber *2013).*

la sélection de l'action et l'apprentissage de comportements habituels – non sensibles à la dévaluation de la récompense (Balleine et O'Doherty 2010 et voir section 3.3.2).

La partie associative du striatum chez le primate comprend une large partie du putamen ainsi que la tête et la queue du noyau caudé. Elle correspond chez le rat au striatum dorso-latéral. Ces régions du striatum reçoivent chez le primate les entrées corticales du cortex préfrontal dorso-médian (aires 8,9,10 et 46 ; Joel et Weiner 2000). Chez le rat, le striatum dorso-latéral reçoit l'entrée de l'aire cingulaire antérieure qui est l'équivalent du cortex préfrontal dorso-latéral chez le primate. Chez le rat, et dans une moindre mesure chez le primate, ces régions sont associées au contrôle des comportements dirigés vers un but par opposition aux comportements habituels (voir section 3.3.2).

La partie limbique du striatum est la partie ventrale chez le rat, chez le primate il s'agit des parties ventrales du putamen et du noyau caudé, ainsi que du noyau accumbens. Ces régions reçoivent principalement les entrées de régions limbiques du cortex tels que le cortex orbito-frontal et le cortex préfrontal ventro-médian. Cette partie de la boucle est associée à l'évaluation de la valeur et des préférences de l'individu.

Afférences thalamiques et pallidales

En plus de ces entrées corticales, le striatum reçoit également une excitation glutamatergique de la part du thalamus, qui représente la deuxième entrée des ganglions de la base après le cortex. À l'image des projections cortico-striatales, ces projections sont également organisées de façon topographique en conservant l'indépendance des boucles limbiques, associatives et motrices. Ainsi, les différentes sous régions du thalamus se connectent avec les différentes régions du striatum associées à la même boucle cortico-basale. Le thalamus projette également vers le cortex et reçoit en entrée le signal de sortie des ganglions de la base (du *GPi* et *SNr*). Les projections vers le cortex et le striatum gardent cette organisation parallèle entre régions limbiques, associatives et motrices (Haber et Calzavara 2009).

De plus le striatum reçoit une afférence du *GPe*. Les projections pallido-striatales sont plus diffuses que les projections striato-pallidales décrites par la suite. En effet, dans Haber et Calzavara 2009, les connexions pallido-striatales sont décrites comme s'étendant sur une plus large partie du striatum que les projections striatopallidales sur le pallidum. Cela suggère que la séparation en boucles parallèles présentes au niveau cortico-striatales n'est pas forcément maintenue dans les connections pallido-striatales.

Dans les circuits ventraux des ganglions de la base, le striatum a également des afférences de l'hippocampe et de l'amygdale. De plus, le striatum reçoit des projections du *STN*.

Efférences

Les seuls neurones du striatum projetant vers les autres noyaux des ganglions de la base sont les *MSN*. Ils projettent massivement vers les deux parties du *GP*. Les projections

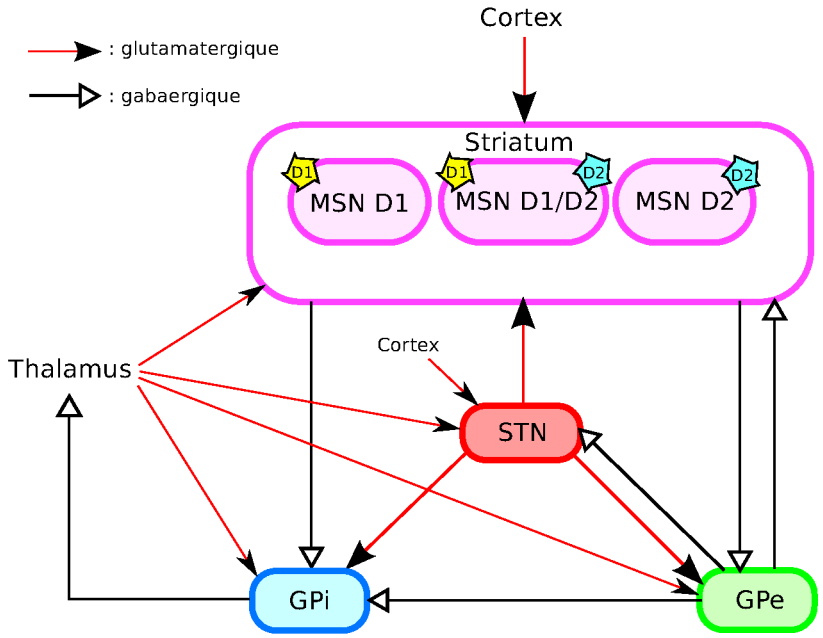

Figure 3.4 – *Connectivité interne des ganglions de la base et afférences corticales et thalamiques. Les projections gabaergiques sont représentées en noir et les projections glutamatergique sont représentées en rouge. MSN : neurones épineux moyens ; STN : noyau subthalamique ; GPe : globus pallidus externe ; GPi : globus pallidus interne.*

striatopallidales sont également organisées de façon topographique de sorte que de l'entrée à la sortie des ganglions de la base les connexions se font de façon à garder la topographie du signal cortical d'entrée (Haber et Calzavara 2009 ; Parent et Hazrati 1995 ; Yelnik et al. 1996). Seules les projections rétrogrades du *GP* vers le striatum semblent ne pas garder cette topographie.

Les *MSN* projettent essentiellement vers les parties interne et externe du *GP*. La vision classique des projections striato-pallidales considère deux populations distinctes de *MSN* : ceux projetant vers le *GPe* et ceux projetant vers le *GPi* (Albin et al. 1989 ; Alexander et Crutcher 1990). Ces deux populations sont considérées comme étant essentiellement ségrégées et diffèrent de part leurs récepteurs dopaminergiques. En effet, les neurones projetant vers le *GPi* ont des récepteurs dopaminergiques de type *D1* (*MSN D1*) alors que les *MSN* projetant vers le *GPe* ont des récepteurs dopaminergiques de type *D2* (*MSN D2*). Cette vision est aujourd'hui encore largement acceptée dans la littérature, mais certaines études remettent en question cette vision et montrent une superposition plus ou moins prononcée de ces chemins, en particulier chez le primate (voir sections 3.4.1 et 3.4.4).

3.2.2 Le noyau subthalamique (*STN*)

Le *STN*, contrairement aux autres noyaux des ganglions de la base, est composé de neurones glutamatergiques, et donc excitateurs. Il est le deuxième noyau d'entrée des ganglions de la base et reçoit, à l'image du striatum, des signaux excitateurs des différentes régions corticales (Nambu et al. 2002). Une étude récente a montré que l'organisation topographique limbique, associative et motrice présente dans les autres noyaux des ganglions de la base est également présente dans le *STN* (Haynes et Haber 2013). Il reçoit de plus des projections thalamiques et forme une boucle fermée avec le *GPe*.

La boucle fermée entre le *STN* et *GP* est possiblement à la base de troubles moteurs et de l'apparition d'oscillations β dans l'activité des noyaux des ganglions de la base dans la maladies de Parkinson (voir Chapitre 5). Comme nous le verrons par la suite, la place centrale du *STN* dans les ganglions de la base l'a mis au coeur de différents traitements de la maladie de Parkinson et de l'addiction (Lardeux et al. 2013 ; Witjas et al. 2005).

Les neurones du *STN* projettent vers les noyaux de sortie des ganglions de la base (*GPi/SNr*), ainsi que vers les neurones du striatum.

Le *STN* est donc à une place centrale dans l'organisation des ganglions de la base en recevant l'entrée corticale et étant connecté à tous les noyaux internes. Il forme le troisième chemin des ganglions de la base : le chemin hyperdirect (voir section 3.4.1).

3.2.3 Le Globus Pallidus (*GP*) et la substance noire

Les neurones du globus pallidus (externe et interne) sont des neurones gabaergiques, inhibiteurs. Le *GP* est composé de deux parties distinctes : la partie interne (*GPi*) et la partie externe (*GPe*). Chez le rongeur, le *GP* n'est pas séparé en partie interne et externe et est l'équivalent du *GPe*. Le noyau entopédonculaire est, chez le rongeur, l'équivalent du *GPi* du primate.

La substance noire est également subdivisée en *pars compacta* (*SNc*), dopaminergique, et *pars reticulata* (*SNr*), gabaergique. La *SNc* ne fait pas partie, à proprement parler, des ganglions de la base mais, via son signal dopaminergique, est fortement liée à cette structure.

La *SNr* et le *GPi* sont souvent considérés comme une population homogène compte tenu de leur proximité anatomique et physiologique. Ils forment la sortie des ganglions de la base et ne projettent vers aucun autre noyau des ganglions de la base. Ils reçoivent les projections glutamatergiques du *STN* et du thalamus, et gabaergique du *GPe* et des *MSNs*.

La balance entre les entrées glutamatergiques, excitatrices et les entrées gabaergiques, inhibitrices détermine la sélection faite par les ganglions de la base. La sélection se fait par déshinibition : plus une action est inhibée en sortie du *GPi/SNr*, plus cette action a de chance d'être sélectionnée (Chevalier et Deniau 1990 ; Mink 1996 ; voir section 3.3).

Le *GPe* est un noyau essentiellement intrinsèque aux ganglions de la base puisqu'il reçoit majoritairement des projections d'autres noyaux des ganglions de la base et projette également presque exclusivement de façon interne aux ganglions de la base. En effet,

il reçoit des projections inhibitrices des *MSNs* (supposées porteurs de récepteurs dopaminergiques de type D2), et excitatrices du *STN* et du thalamus. Il projette vers le *STN* et les neurones du striatum. Il est donc, à l'image du *STN*, connecté avec tous les autres noyaux de ganglions de la base.

3.3 Ganglions de la base : sélection et contrôle comportemental

Dans cette partie nous allons discuter du rôle des ganglions de la base dans la sélection de l'action, l'aquisition et la génération de comportement. Nous discuterons ainsi de la place centrale de ce système dans la prise de décision et la génération de la commande motrice choisie.

3.3.1 Sélection de l'action par désinhibition

Dans leur article fondateur faisant le lien entre sélection de l'action et ganglions de la base, Redgrave et al. 1999b définissent la sélection de l'action comme un problème d'accès aux ressources motrices. En effet, compte tenu de la limitation des ressources motrices, l'activation simultanée de plusieurs commandes motrices peut être impossible. La limitation de ces ressources fait que l'on doit privilégier une action plutôt qu'une autre, ce qui revient à sélectionner une action et en mettre d'autres en attente.

Cette capacité à sélectionner une action est essentielle pour la survie chez les êtres vivants. En effet, chaque individu doit conserver son homéostasie en contrôlant différents facteurs tels que la faim, la soif, la température corporelle et de nombreux autres paramètres vitaux. Il doit de plus veiller à conserver son intégrité physique et donc se protéger d'un éventuel prédateur. Lorsque la sensation de soif augmente, l'animal ira s'abreuver à un point d'eau. Cependant, si un son soudain lui parvient alors qu'il est en train de s'hydrater, une décision vitale se présentera : partir et ne pas étancher sa soif ou considérer que le son n'indique pas un éventuel prédateur et continuer de s'abreuver. Il est, dans ces conditions, nécessaire d'avoir un système de sélection permettant de trancher rapidement entre ces deux options. L'individu doit donc avoir des outils permettant, au travers de son expérience passée, de choisir l'option la plus profitable (Cohen et al. 2007).

Si la capacité de prise de décision peut paraître naturelle, elle implique la présence d'un système neural permettant la sélection de l'action. Ce système doit être capable d'évaluer les options possibles en compétition et de sélectionner la plus profitable afin d'assurer la survie de l'individu.

D'un point de vue théorique, pour qu'une sélection ait lieu, il est nécessaire que les différentes options entrent en compétition. Le gagnant de cette compétition sera l'option sélectionnée. Redgrave et al. 1999b décrivent la théorie de la sélection de l'action et passent en revue les données des ganglions de la base conduisant à l'hypothèse que les ganglions de la base sont un système de sélection de l'action. Ils proposent principalement trois visions théoriques d'architecture de cette possible compétition (voir Figure 3.5).

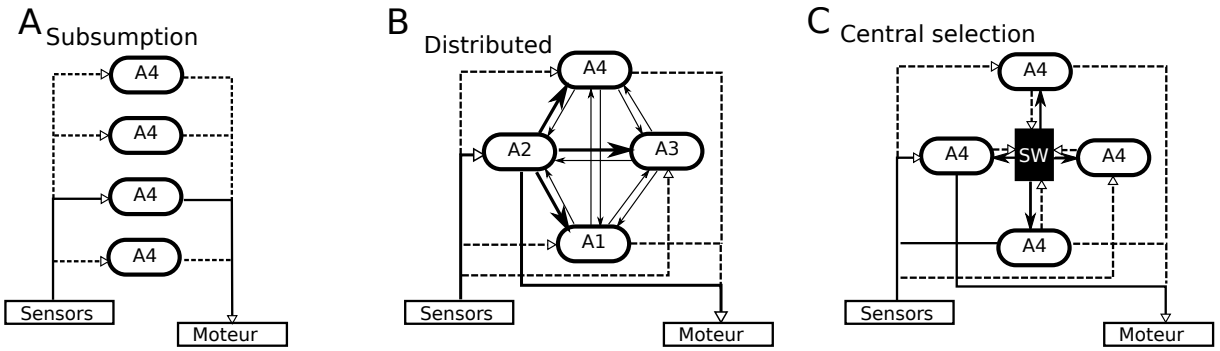

Figure 3.5 – *Description de différentes architectures de sélection de l'action. A. Architecture de subsomption où les différentes actions sont en compétition directe. Elle a une organisation hiérarchique figée des actions. B. Architecture de sélection distribuée dans laquelle chaque action exerce une inhibition latérale sur les autres actions. C. Sélection centralisée, où chaque action est connectée à un système central permettant une inhibition latérale tout en minimisant le nombre de connexions. Figure reproduite de* Redgrave *et al. 1999b.*

En premier lieu, on peut considérer une compétition simple dans laquelle chaque option est indépendante et est directement liée à l'exécution motrice (voir Figure 3.5 A). Cette architecture est une subsomption. Elle suppose un ordre hiérarchique figée des actions qui permet de déterminer quelle action sera sélectionnée lorsque plusieurs modules sont actifs simultanément. L'ordre hiérarchique des actions fixé a priori biaise donc la sélection de façon importante. Cette architecture est peu complexe, mais intègre des connaissances a priori sur l'ordre de priorité à associer aux actions.

Une deuxième architecture théorique de sélection revient à considérer une inhibition latérale des différentes options de façon distribuée (voir Figure 3.5 B). L'inhibition latérale fait que chaque option inhibe les autres options ; la plus active empêche donc la sélection de ses concurrentes. Il en résulte que seule une option est sélectionnée. Cependant cette architecture implique que toutes les options soient interconnectées. Le nombre de connexions nécessaires à cette sélection augmente donc fortement avec le nombre d'options en compétition ($n(n-1)$ connexions nécessaires avec n le nombre d'options en compétition).

Une troisième architecture de sélection, dite centralisée, permet à la fois d'avoir une inhibition latérale et d'éviter une explosion du nombre de connexions en fonction du nombre d'options en compétition (voir Figure 3.5 C). Pour ce faire, il faut ajouter un noyau central auquel chaque option est connectée. Ce système central permet dexercer une forme d'inhibition latérale en centralisant l'intensité des différentes options. Ici, le nombre de connexions est linéaire en fonction du nombre d'options en compétition ($2n$ connexions).

L'architecture centralisée a donc de nombreux avantages en terme de sélection à la fois pour l'efficacité de la sélection et la parcimonie du nombre de connexions nécessaires. Les ganglions de la base reflètent en partie ce type d'architecture. En effet, la place des ganglions de la base, qui leur permet d'intégrer de nombreuses informations corticales, met ce système à la croisée des informations limbiques, sensorielles et motrices et en fait

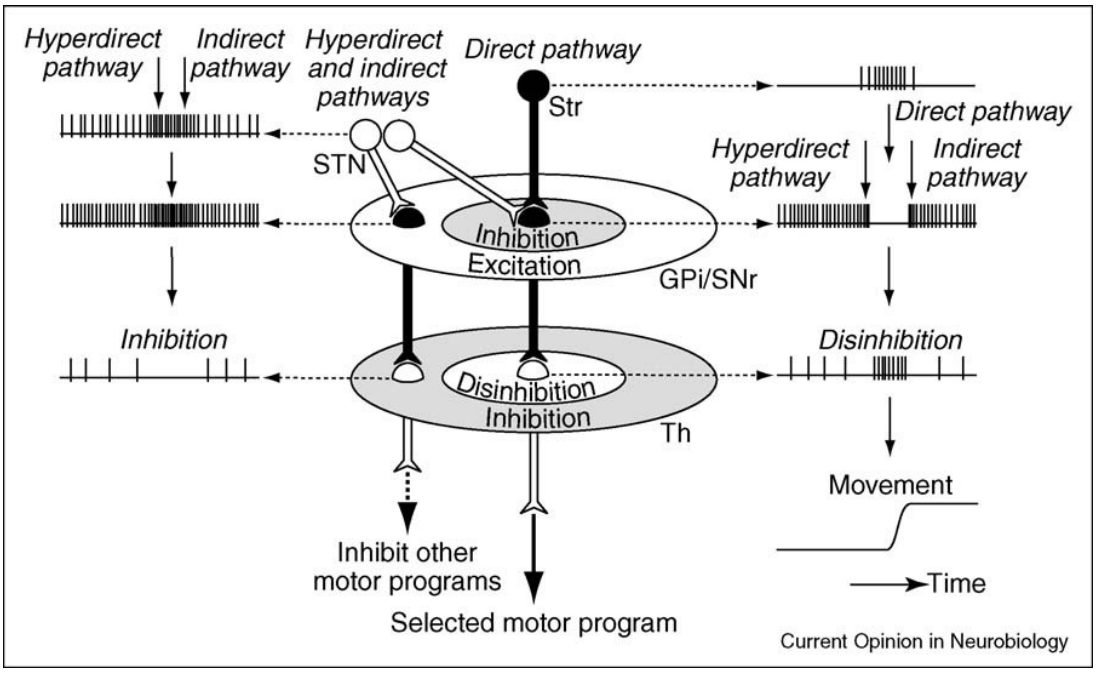

Figure 3.6 – *Illustration des mécanismes de sélection par désinhibition depuis l'entrée corticale des ganglions de la base jusqu'à la sortie de la commande motrice. L'inhibition du noyau de sortie des ganglions de la base entraîne une désinhibition du thalamus ce qui permet l'excitation du cortex moteur et de la commande motrice désirée. Figure reprise de* Nambu *2008.*

donc un système central adapté à la prise de décision.

Les mécanismes physiologiques de ce système de sélection par les ganglions de la base ont été notamment décrits par Mink 1996 dans un autre article fondateur de cette hypothèse. Les ganglions de la base exercent une inhibition continue des commandes motrices corticales via le thalamus (Chevalier et Deniau 1990 ; voir Figure 3.6). En effet, les noyaux de sortie des ganglions de la base (*GPi/SNr*) sont actifs de façon tonique et exercent en continu une inhibition du thalamus et par transitivité du cortex moteur, qui reçoit le signal thalamique. L'inhibition de la sortie des ganglions de la base permet la suppression de cette inhibition tonique du thalamus et la génération d'une action motrice par une déshinibition du cortex moteur (voir figure 3.6).

L'envoi d'un signal excitateur du cortex au striatum permet l'inhibition du complexe *GPi/SNr*, entraînant la levée de l'inhibition tonique du schéma moteur et la génération de la commande motrice désirée.

De plus, lorsqu'un mouvement est généré, les régions motrices du cortex envoient un signal au noyau subthalamique qui cause une excitation globale des noyaux de sortie des ganglions de la base, inhibant ainsi les actions motrices concurrentes, via ce que l'on appellera la voie hyperdirecte (voir section 3.4.1). Dans cette vision de la sélection proposée par Mink 1996, les ganglions de la base permettent donc à la fois de prévenir la génération de commandes motrices non désirées tout en permettant la génération de l'action désirée. On parle donc de sélection de l'action par désinhibition.

Depuis ces travaux, cette proposition a été développée et reprise dans de nombreuses études (Berns et Sejnowski 1998 ; Daw et Doya 2006 ; Frank 2005 ; Girard et al. 2003 ; Girard et al. 2008 ; Gurney et al. 2001b ; Humphries et Prescott 2010 ; Humphries et al. 2012 ; Khamassi et al. 2005 ; Leblois et al. 2006). Ces nombreux modèles computationnels ont également testé et validé cette hypothèse de système de sélection par désinhibition (voir section 3.5). Cependant, certains aspects de l'anatomie des ganglions de la base sont encore débattus et le processus de sélection réalisé par le système donne aujourd'hui lieu à de nombreuses études et collaborations entre anatomistes, expérimentalistes et modélisateurs.

3.3.2 Striatum et génération de comportements

Depuis les travaux de Mink 1996 et Redgrave et al. 1999b, de nombreuses études ont montré la place prépondérante des ganglions de la base dans la génération de différents types de comportements. Un grand nombre d'entre elles se sont particulièrement intéressées à la place du striatum.

Le striatum reçoit massivement le signal dopaminergique de la *VTA* et la *SNc*. L'analogie entre le signal dopaminergique et un signal d'apprentissage (Schultz 1998 ; voir Chapitre 2), ainsi que les capacités du striatum à encoder une information de valeur (Samejima et al. 2005) a permis de développer un modèle des ganglions de la base issus de la théorie de l'apprentissage par renforcement (Barto 1995 ; Doya 1999 ; Gillies et Arbuthnott 2000 ; Houk et Wise 1995 ; Joel et al. 2002). Cette hypothèse suppose que les différentes régions du striatum sont impliquées dans la génération de différents comportements (Atallah et al. 2007 ; Balleine et al. 2007 ; O'Doherty et al. 2004 ; Yin et al. 2004, 2005).

On peut ainsi distinguer la partie ventrale des parties dorsales du striatum. En effet, des études de lésions ont montré que le striatum ventral est nécessaire à l'apprentissage d'une tâche mais pas à la mise en oeuvre de l'action. Le striatum dorsal est nécessaire pour la performance de l'action mais pas pour l'apprentissage (Atallah et al. 2007). Ce résultat a permis de renforcer la vision Actor-Critic des ganglions de la base (Houk et Wise 1995 ; Joel et al. 2002 ; Khamassi et al. 2005) dans laquelle le striatum ventral joue le rôle du critique, encodant les valeurs d'action (Samejima et al. 2005), et le striatum dorsal joue le rôle de l'acteur.

De plus, et comme mentionné précédemment, le striatum dorsal peut être subdivisé en striatum dorso-médian (*DMS*) et striatum dorso-latéral (*DLS*). Ces deux sous régions du striatum dorsal interviennent en effet pour l'expression de différents types de comportement. Les études de conditionnements distinguent deux types de comportement : les comportements flexibles orientés vers un but et les comportements rigides et habituels (Khamassi et Humphries 2012). Pour différencier ces deux types de comportements, de nombreuses études utilisent une technique de dévaluation de la récompense. Ce procédé consiste à rendre la récompense donnée à l'animal moins ou plus du tout attractive. Par exemple, en faisant en sorte que l'animal n'ait plus faim, toute récompense sous forme de nourriture perdra son attractivité. Dans certaines procédures, la récompense peut également être empoisonnée pour que l'animal l'associe à la maladie. Cette dévaluation de la

récompense fait que l'animal ne devrait plus montrer d'intérêt à réaliser les actions lui permettant d'y accéder.

Le test permettant de voir si un animal développe un comportement habituel ou dirigé vers un but se fait en trois phases distinctes. Lors de la première phase, le sujet – i.e. l'animal – apprend à réaliser l'action menant à la récompense (presser un levier lorsqu'un stimulus apparaît, appuyer sur le bon interrupteur, etc...). Après cette première phase d'apprentissage de la contingence action-récompense, la récompense est dévaluée. Lors de la troisième phase, on replace le sujet dans l'environnement de la première phase et on observe s'il continue à réaliser l'action menant à la récompense.

Si le sujet ne continue pas à réaliser l'action cela indique que son comportement est dirigé vers un but. En effet, dans ce cas l'action n'a pour l'animal aucune valeur intrinsèque et c'est pour l'obtention d'une récompense attractive qu'il réalise l'action. En l'absence de récompense l'animal ne voit pas dintérêt à réaliser l'action. Cependant si le sujet continue à faire l'action de façon répétitive, aussi fréquemment qu'à la phase 1, malgré la perte d'attractivité de la récompense, cela indique qu'il n'est pas sensible à la dévaluation de la récompense et son comportement est dit habituel. Plus l'animal aura eu l'expérience de la contingence action-récompense présente durant la première phase d'apprentissage plus la probabilité pour qu'il développe un comportement habituel est grande. Ce type de comportement agit comme si l'animal avait associé une valeur intrinsèque à l'action indépendamment de la récompense.

Des études de lésion (Yin et al. 2004, 2005) ont montré que l'inactivation du *DLS* empêche le comportement de l'animal de devenir habituel et que l'inactivation du *DMS* déclenche une apparition prématurée d'un comportement habituel chez l'animal. Ainsi, le *DLS* est associé à l'expression des comportements habituels et le *DMS* l'expression des comportements dirigés vers un but.

Ces résultats permettent de montrer de façon directe l'implication et le rôle majeur de différentes parties du striatum dans la génération du comportement et donc la sélection de l'action ; renforçant ainsi l'idée que les ganglions de la base ont une place centrale dans la prise de décision (Balleine et al. 2007). De plus cela montre la place des ganglions de la base dans l'apprentissage des comportements et dans l'exécution de ces comportements avec un gradient ventro-dorsal et latéro-médial dans la partie dorsale.

Cependant, plusieurs visions des ganglions de la base se superposent actuellement, avec notamment des propositions que les différentes régions du striatum, sont associées à différents niveaux de représentation de la prise de décision (Ito et Doya 2011 ; Keramati et Gutkin 2014), avec ce même gradient ventro-dorsal. La partie ventrale du striatum étant associée au contexte, la partie dorso-médiale à des signaux sensoriels et la partie dorso-latérale à l'aspect sensori-moteur (Redgrave et al. 2011). Cette hypothèse repose notamment sur les boucles parallèles limbique, associative et motrice des ganglions de la base et donc du striatum (voir Figure 3.3).

Bout à bout, si ces études montrent également que le rôle précis des différentes régions du striatum n'est pas encore admis unanimement, elles ont pu mettre en évidence la place centrale du striatum dans lacquisition et l'expression de comportements nouveaux.

3.3.3 Représentation multi-canaux des actions en compétition

Si les ganglions de la base forment un système central de prise de décision, il est primordial que les options à évaluer soit représentées dans leurs différents noyaux. La plupart des modèles de sélection des ganglions de la base font l'hypothèse de la présence de différents canaux parallèles, chacun associé à une option comportementale. Chaque canal est représenté dans les différents noyaux par un sous groupe de neurones encodant tous une information (de valeur, de salience,...) sur l'option associée.

Ces canaux sont principalement parallèles les uns des autres mais entrent en compétition via des activations latérales des différents canaux. En effet, les connexions entre deux noyaux peuvent être de deux types : focalisée ou diffuse. Une connexion focalisée entre deux noyaux indique que les neurones du noyau envoyant le signal projettent vers une sous partie des neurones du noyau recevant le signal, correspondant donc au même canal. Au contraire si une connexion est diffuse, les neurones du noyau envoyant le signal auront des connexions avec virtuellement l'ensemble des neurones du noyau receveur. Ainsi, si l'on considère la présence de canaux, cela implique qu'un canal influencera l'activité de tous les autres canaux dans le cas diffus, ou uniquement de son propre canal dans le cas focalisé. Les modèles actuels font l'hypothèse de la présence de ces canaux parallèles pour représenter la compétition entre les différentes actions (Girard et al. 2008 ; Gurney et al. 2001b ; Humphries et al. 2006 ; Khamassi et al. 2005 ; Liénard et Girard 2014 ; Prescott et al. 2006).

Cette vision doit également prendre en compte la réduction de dimension qui a lieu dans les ganglions de la base avec des noyaux de sortie ayant un nombre de neurones bien moins important que les noyaux d'entrée des ganglions de la base (Bar-Gad et Bergman 2001 ; Bolam et al. 2000). Il est ainsi probable qu'un canal soit représenté par un nombre important de neurones dans le cortex, plus faible dans le striatum et encore plus faible dans *GP* et *STN*.

3.4 Architecture interne des ganglions de la base et dopamine

Dans la section précédente, nous avons vu les différents noyaux des ganglions de la base et avons décrit leurs interactions avec les autres noyaux. Nous avons également évoqué la connectivité entre les neurones de projection du striatum – les *MSN* – et les parties interne et externe du *GP*. Cette connectivité entre striatum et *GP* est aujourd'hui encore soumise à débat dans la littérature et définit en grande partie l'architecture interne des ganglions de la base. Nous allons dans un premier temps décrire la vision classique des ganglions de la base en explicitant ce que sont les chemins direct, indirect et hyperdirect. Nous montrerons comment cette théorie a permis de donner une interprétation du rôle des différents noyaux des ganglions de la base. Nous verrons également comment la perte de neurones dopaminergiques dans la maladie de Parkinson change la dynamique de ce système.

Dans un second temps nous montrerons que de plus en plus d'études questionnent cette

interprétation de l'architecture interne des ganglions de la base. Elles remettent ainsi en question une partie de notre compréhension de l'influence de la dopamine sur les ganglions de la base ; et donc de notre compréhension des mécanismes menant à la perte du contrôle moteur dans Parkinson ou des différences d'apprentissage par la récompense et punition sur des sujets ayant différents niveaux dopaminergiques.

3.4.1 Les chemins direct, indirect et hyperdirect

De nombreux signaux corticaux parcourent les ganglions de la base jusqu'aux noyaux de sortie *GPi/SNr*. Trois voies différentes allant de l'entrée des ganglions de la base à la sortie ont été identifiées dans la littérature : la voie directe, indirecte et hyperdirecte (voir Figure 3.7). Selon la voie empruntée par le signal cortical, il facilitera ou non le mouvement en sortie.

Comme introduit dans la section 3.2.1, les neurones de projection du striatum (les *MSN*) peuvent être regroupés en deux catégories de neurones présentant des caractéristiques physiologiques différentes (Albin et al. 1989 ; Alexander et Crutcher 1990 ; Gerfen et Wilson 1996 ; Obeso et al. 2000 ; Utter et Basso 2008). Un premier groupe de *MSN* projette directement vers *GPi/SNr*. Ils forment le chemin ou voie directe puisque le signal traverse les ganglions de la base en ne passant que par le striatum et est envoyé directement à la sortie. Ce chemin exerce ainsi une inhibition gabaergique directe sur la sortie des ganglions de la base et promeut la génération du mouvement (voir section 3.6). Frank et al. 2004, propose l'appellation 'Go' pour désigner le fait que ce chemin direct favorise l'action. Ces *MSN* expriment des récepteurs dopaminergiques D1 et co-expriment la substance P et le dynorphin (voir Figure 3.7).

Chemin direct :
Cortex^{+} $\rightarrow$ *MSN D1*$^{-}$ $\rightarrow$ *GPi/SNr*$^{-}$ $\rightarrow$ Thalamus^{+} $\rightarrow$ Cortex

Retracer le parcours du signal cortical au travers des ganglions de la base nous permet de mieux comprendre comment l'information corticale est traitée par ce système. Ainsi, l'excitation corticale du chemin direct, crée une désinhibition du noyau de sortie des ganglions de la base, levant l'inhibition du thalamus, permettant une excitation du cortex. Le renforcement du chemin direct a donc un effet excitateur sur le cortex et facilite la génération de commande motrice.

Par opposition, les *MSN* qui projettent vers le *GPe* forment le chemin indirect. Ils portent des récepteurs dopaminergiques de type D2 et expriment la peptide enképhaline (ENK).

Chemin indirect :
Cortex^{+} $\rightarrow$ *MSN D1*$^{-}$ $\rightarrow$ *GPe*$^{-}$ $\rightarrow$ *STN*$^{+}$ $\rightarrow$ *GPi/SNr*$^{-}$ $\rightarrow$ Thalamus^{+} $\rightarrow$ Cortex

L'information passe donc des *MSN* au *GPe* qui à son tour projette sur le *STN* qui envoit le signal au *GPi*. Aussi, un renforcement de l'activité dans les *MSN* du chemin direct a globalement un effet excitateur sur le *GPi* prévenant l'activation du mouvement

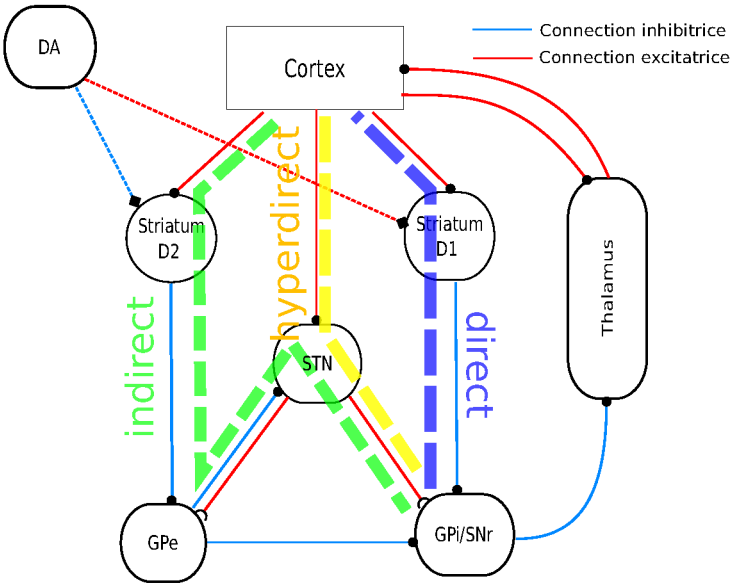

Figure 3.7 – *Représentation des chemins direct, indirect et hyperdirect dans les ganglions de la base.*

(voir Figure 3.7).

Des études ont permis de valider que l'activation du chemin indirect bloque le mouvement (Sano et al. 2013) alors que l'activation du chemin direct favorise le mouvement (Nakanishi et al. 2014).

Frank et al. 2004 ont suggéré que l'équilibre entre les chemins direct et indirect [1] permet d'apprendre à la fois à approcher la récompense et à éviter les punitions. Dans son modèle Go-NoGo, le chemin direct, 'Go', est associé à l'approche de la récompense et le chemin indirect, 'NoGo', est associé à l'évitement de la punition.

Le chemin hyperdirect est la troisième voie des ganglions de la base. Dans cette voie, le signal cortical passe par le *STN* qui, à son tour, projette directement vers le *GPi/SNr*. Elle est donc indépendante du striatum.

Chemin hyperdirect :
Cortex^{+} $\rightarrow$ *STN*$^{+}$ $\rightarrow$ *GPi/SNr*$^{-}$ $\rightarrow$ Thalamus^{+} $\rightarrow$ Cortex

Le chemin hyperdirect a un effet excitateur sur le noyau de sortie des ganglions de la base agissant comme un frein à la génération de commande motrice (Frank 2006). Son rôle semble être d'éviter la génération prématurée d'une action en freinant la sélection et

1. Dans Frank et al. 2004, le chemin indirect n'est pas pris en compte dans son entièreté puisque le *STN* n'est pas modélisé.

en régulant l'activité des deux parties du *GP*.

Cette description des ganglions de la base se basant sur une forte ségrégation des projections striato-pallidales, a permis de mieux appréhender le rôle de la dopamine, à la fois dans l'apprentissage via la modulation des poids synaptiques cortico-striataux, mais également de l'effet d'un changement du niveau dopaminergique dans la dynamique du système (voir Figures 3.8 A-C). Cependant cette description est aujourd'hui remise en cause par plusieurs études (voir section 3.4.4).

3.4.2 Plasticité synaptique

Les différents types de récepteurs dopaminergiques que portent les *MSN* entraînent un effet opposé de la dopamine sur la connectivité cortico-striatale (Gerfen et Surmeier 2012 ; Shen et al. 2008). Il y a en effet 5 types de récepteurs dopaminergiques que l'on peut répartir en 2 familles. La première catégorie, communément appelée D1, comprend les récepteurs dopaminergiques de types D1 et D5. Ces récepteurs sont positivement renforcés par la dopamine. En effet, la dopamine entraîne une potentialisation à long terme (LTP) chez les neurones ayant des récepteurs de type D1, ce qui a pour conséquence de renforcer les connexions synaptiques. Dans le cas des *MSN*, cela renforce notamment la connectivité cortico-striatale, rendant les *MSN* plus sensibles au signal cortical.

La deuxième famille, regroupée sous le nom de récepteurs D2, est composé de l'ensemble des récepteurs de type D2, D3 et D4. Ces récepteurs ont une affinité opposée à la dopamine. Elle entraîne chez eux une dépression à long terme (LTD) qui amoindrit la connection synaptique. Cela rend donc les *MSN* moins sensibles au signal cortical.

De récentes études (Kravitz et al. 2012 ; Paton et Louie 2012 ; Sano et al. 2013) utilisant des techniques d'optogénétique chez la souris ont permis de valider l'effet de renforcement et d'évitement de la dopamine sur les différents groupes de *MSN*. L'optogénétique permet de stimuler de façon exclusive certaines catégories de neurones et ainsi de tester l'influence de différentes populations de neurones sur le comportement. Notamment, ces techniques permettent de stimuler uniquement les neurones ayant des récepteurs D1 ou D2 à la dopamine pour ainsi mieux observer leur rôle dans le comportement.

La stimulation des *MSN D1* favorise des comportements d'approche et semble interprétée de la part de l'animal de façon similaire à l'obtention de récompenses (Kravitz et al. 2012 ; Paton et Louie 2012). La stimulation du chemin direct est également associée à une augmentation de la vélocité de l'animal et à l'inhibition de neurones de la *SNr*[2](Freeze et al. 2013). Par opposition, la stimulation des *MSN D2* entraîne des comportements d'évitement et une diminution de la vélocité de l'animal (Freeze et al. 2013 ; Kravitz et al. 2012 ; Paton et Louie 2012).

Ces études ont ainsi mis en évidence que : (1) l'activation du chemin direct via les *MSN D1* a bien un effet de renforcement et de récompense, menant au mouvement, et que (2) l'activation du chemin indirect via les *MSN D2* bloque la locomotion et entraîne des comportements d'évitement. Toutefois, il faut bien noter que ces études ont été faites chez la souris. Elles ne valident donc la vision fortement discriminante D1/chemin-direct/Go

2. La *SNr* est chez les rongeur l'équivalent du *GPi*.

versus D2/chemin-indirect/No-Go que chez cette espèce. Le niveau de discrimination entre ces deux chemins reste donc à valider chez d'autres espèces en particulier chez le primate.

De plus, d'autres études suggèrent que l'activation du chemin direct et indirect est nécessaire pour la sélection de l'action (Friend et Kravitz 2014 ; Keeler et al. 2014), se plaçant en opposition au modèle classique des ganglions de la base.

3.4.3 Parkinson et ganglions de la base

La maladie de Parkinson se traduit par la mort de neurones dopaminergiques dans *SNc* (Obeso et al. 2000). Son étude permet donc de mettre à l'épreuve nos connaissance sur le rôle de la dopamine dans les ganglions de la base. De plus cette maladie entraîne un déséquilibre entre les chemins direct et indirect et permet de tester leurs rôles dans cette pathologie.

La maladie de Parkinson est notamment associée à des mouvements incontrôlés entraînant des tremblements. Cependant les conséquences de la maladie de Parkinson ne se limitent pas à des troubles moteurs mais entraînent également des troubles cognitifs – troubles de l'attention, démence – et psychiatriques – dépression, axiété – (Bartels et Leenders 2009).

Avec la vision classique des ganglions de la base supposant la séparation des chemins direct, indirect et hyperdirect, il est possible d'évaluer les conséquences de la mort des neurones dopaminergiques dans la *SNc*, chez les patients parkinsoniens, sur l'activité des ganglions de la base (Albin et al. 1989 ; Bartels et Leenders 2009 ; Frank et al. 2004 ; Parent et al. 2000). La mort neuronale dans *SNc* entraîne une diminution du niveau dopaminergique au niveau des ganglions de la base, créant un déséquilibre entre le chemin direct et indirect. Comme on peut le voir sur la figure 3.8 A-C, en condition normale les deux chemins sont en position d'équilibre ce qui permet la sélection ou l'évitement de comportement.

Si on considère une diminution du niveau dopaminergique, cela aura pour conséquence de renforcer le chemin indirect via les *MSN D2* et de diminuer la force du chemin direct, via les *MSN D1* (voir Figure 3.8). La sortie des ganglions de la base sera alors globalement plus excitée, ce qui engendre une difficulté accrue à sélectionner des actions.

Si on considère au contraire une augmentation du niveau dopaminergique – ce qui est supposé être le cas chez les sujets atteints de la maladie de Parkinson sous traitement Levo-Dopa –, le niveau tonique élevé doit entraîner un renforcement du chemin direct et une diminution de la force du chemin indirect. Cela a pour effet d'inhiber plus fortement le *GPi* et ainsi d'augmenter les chances de sélectionner une action et d'enclencher un mouvement.

L'hypothèse des chemins direct, indirect et hyperdirect a donc permis de mieux appréhender le fonctionnement des ganglions de la base lors d'un changement du niveau dopaminergique et ainsi de mieux comprendre les conséquences de la maladie de Parkinson. Cependant, cette hypothèse repose essentiellement sur la séparation nette entre les *MSN D1* projetant vers le *GPi* et les *MSN D2* vers le *GPe*. Nous verrons que cette séparation n'est pas unanimement admise dans la communauté scientifique.

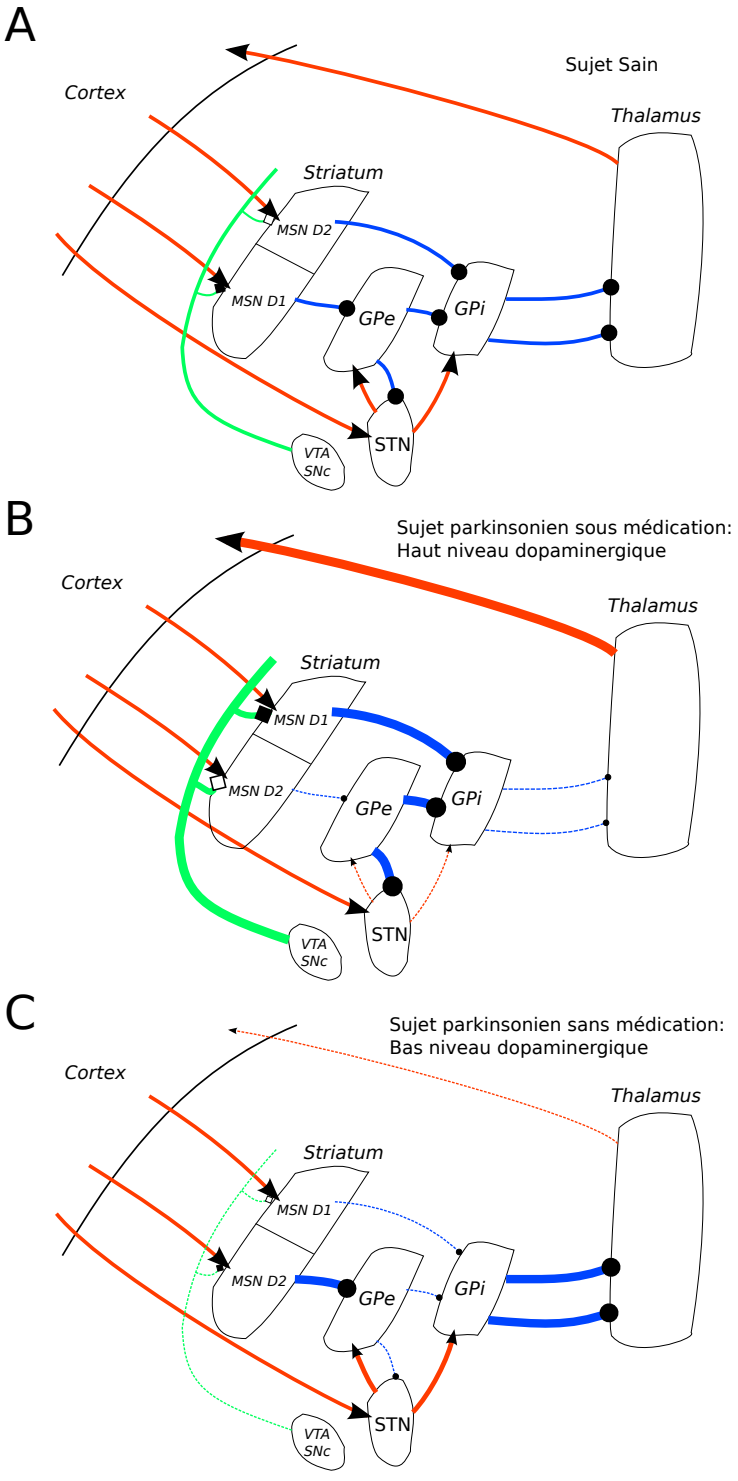

Figure 3.8 – *A.B.C. Illustration de l'hypothèse du chemin direct, porté par les MSNs ayant des récepteurs D1, et du chemin indirect, porté par les MSNs ayant des récepteurs D2, utilisée dans le modèle de Frank et collègues sous différents niveaux de dopamine tonique. A. fonctionnement classique des ganglions de la base chez un patient sain . B. Sujet atteint de la maladie de Parkinson sous traitement Levo-Dopa qui augmente le niveau tonique de dopamine C. Sujet atteint de la maladie de Parkinson sans médication et présentant un niveau dopaminergique plus faible qu'un patient sain.*

3.4.4 Une ségrégation D1/D2 partielle

L'hypothèse de la séparation des neurones du striatum (MSN) en chemin direct et indirect (Albin et al. 1989 ; Alexander et Crutcher 1990) est aujourd'hui encore questionnée par des études chez le rat (Calabresi et al. 2014) comme chez le primate (Lévesque et Parent 2005 ; Nadjar et al. 2006 ; Nambu 2008). Elles remettent en question le fait que les projections striato-pallidales sont parfaitement ségrégées en *MSN D1→GPi* et *MSN D2→GPe*.

Lévesque et Parent 2005 ont montré par traçage de cellule unique dans le striatum que la plupart des *MSN* ont des axones terminant à la fois dans *GPe* et *GPi* chez le primate. Ils ont observé que 82% des *MSN* projettent vers *GPi* et que tous les *MSN* projettent vers *GPe*. Ces résultats mettent donc en évidence un recouvrement important des chemins direct et indirect et invalide l'hypothèse d'une franche ségrégation D1/D2 dans cette espèce. Cependant cette étude montre également que certains *MSN* projettent uniquement vers le *GPi* ce qui peut dans une certaine mesure laisser place à la présence partielle du chemin direct. Plusieurs autres études ont également montré un recouvrement au moins partiel chez le primate (Lévesque et Parent 2005 ; Nadjar et al. 2006 ; Nambu 2008) appelant à la reconsidération de l'architecture interne des ganglions de la base (notamment celle décrite en section 3.4.1).

Des résultats similaires ont été trouvés chez le rats suggérant que 70% des neurones du chemin direct projettent également sur le chemin indirect Fino 2007 ; Fujiyama et al. 2011 ; Gerfen et Surmeier 2012 ; Kawaguchi et al. 1990 ; Wu et al. 2000. D'autres études évoquent que seuls 3% des neurones du striatum chez le rat projettent uniquement vers la *SNr* (Calabresi et al. 2014).

De plus, plusieurs études montrent une proportion non négligeable (jusqu'à 60%) de *MSN* présentant à la fois des récepteurs D1 et D2 (Nadjar et al. 2006), ce qui rend l'effet de la dopamine sur ces neurones incertain.

Ces résultats mettent à l'épreuve la vision classique des chemins direct et indirect dans les ganglions de la base et, ainsi, remettent en question notre compréhension du fonctionnement des processus de sélection, d'apprentissage des ganglions de la base. On notera que la plupart de nombreux modèles computationnels actuels font pourtant l'hypothèse de la présence, au moins partielle, de ces deux chemins (Albada et Robinson 2009 ; Albada et al. 2009 ; Collins et Frank 2014 ; Frank et al. 2004).

3.5 Modèles computationnels

Depuis une vingtaine d'années, de nombreuses études ont développé des modèles computationnels des ganglions de la base afin d'étudier ses présupposées capacités de sélection de l'action. Ils permettent d'étudier plus généralement la dynamique du système, que ce soit chez un sujet sain, ou en modélisant certaines pathologies telles que Parkinson ; qui modifie fortement la connectivité et affecte le contrôle moteur. Ces modèles ont cha-

cun leur propre architecture, se basant sur différentes hypothèses anatomiques que nous discuterons. Le but de cette thèse étant notamment de proposer un nouveau modèle des ganglions de la base prenant en compte les données anatomiques récentes sur la séparation partielle D1/D2.

Chaque modèle tente de répondre à un certain nombre de questions sur les ganglions de la base, d'un point de vue soit fonctionnel, soit anatomique ou électrophysiologique. Nous présentons dans la suite de ce chapitre plusieurs modèles computationnels des ganglions de la base, qui ont une pertinence particulière quant aux recherches présentées dans la suite de cette thèse.

3.5.1 Le modèle Gurney, Prescott, Redgrave (*GPR*)

Le modèle *GPR* (Gurney,Prescott,Redgrave) a été introduit par Gurney et al. 2001a et implémenté dans Gurney et al. 2001b en utilisant des neurones intégrateurs à fuite. Ce type de neurones est assez simple à simuler car ils ont une dynamique du premier ordre et sont notamment utilisés pour modéliser le comportement d'un ensemble homogène de neurones. Dans ce modèle, comme dans la plupart des autres modèles que nous présenterons, chaque option est représenté par un canal indépendant des autres. Tous les noyaux sont ainsi formés de plusieurs neurones, chacun codant pour un canal. Le cortex envoi en entrée du système la salience de chaque option.

D'un point de vue anatomique, ce modèle considère une ségrégation importante des *MSN D1* et *D2* (voire Figure 3.9). Toutefois, les auteurs proposent une nouvelle interprétation fonctionnelle de l'anatomie des ganglions de la base. Ils proposent de séparer deux modules : le premier sert à la sélection et le deuxième a un rôle de contrôle.

Le module de sélection comprend le chemin direct, Cortex→*MSN D1*→*GPi*, et le chemin hyperdirect, *Cortex*→*STN*→*GPi*. La connexion directe entre les *MSN D1* et le *GPi* est supposée focalisé, ayant un effet *off center*, car elle donne lieu à l'inhibition directe des différents canaux en fonction de leur salience. Le chemin hyperdirect est supposé avoir un effet *on surround* car il a pour effet d'exciter les canaux en compétition globalement de part la connexion diffuse entre *STN* et *GPi*. Ce module seul est ainsi capable de sélection.

Cependant, selon l'intensité du signal cortical, le module de sélection seul peut être incapable de sélection si l'effet *on surround* du module est trop fort ou trop faible. Ainsi, les auteurs supposent que le module de contrôle, composé du chemin indirect, permet de réguler l'activité du système. Cette effet de régulation proviendrait de la boucle *GPe*/*STN* qui converge vers une valeur moyenne indépendante du nombre de canaux en compétition. Sans ce processus de régulation, l'augmentation du nombre de canaux en compétition augmente également l'excitation qu'exerce le *STN* sur le *GPi*, rendant l'effet de sélection de la connection *MSN D1*→*GPi* trop faible pour la contrebalancer.

En ce sens, le modèle *GPR* se démarque de la vision classique des ganglions de la base. Il existe aujourd'hui plusieurs versions du modèle *GPR* (Humphries et al. 2012 ; Humphries et Gurney 2002). Ce modèle a notamment été prolongé pour ajouter la boucle thalamo-corticale (Humphries et Gurney 2002). Cela a permit de montrer que l'intégration de la boucle thalamo-corticale améliore les capacités de sélection du modèle.

Ce modèle a également été testé dans un cadre robotique (Prescott et al. 2006),

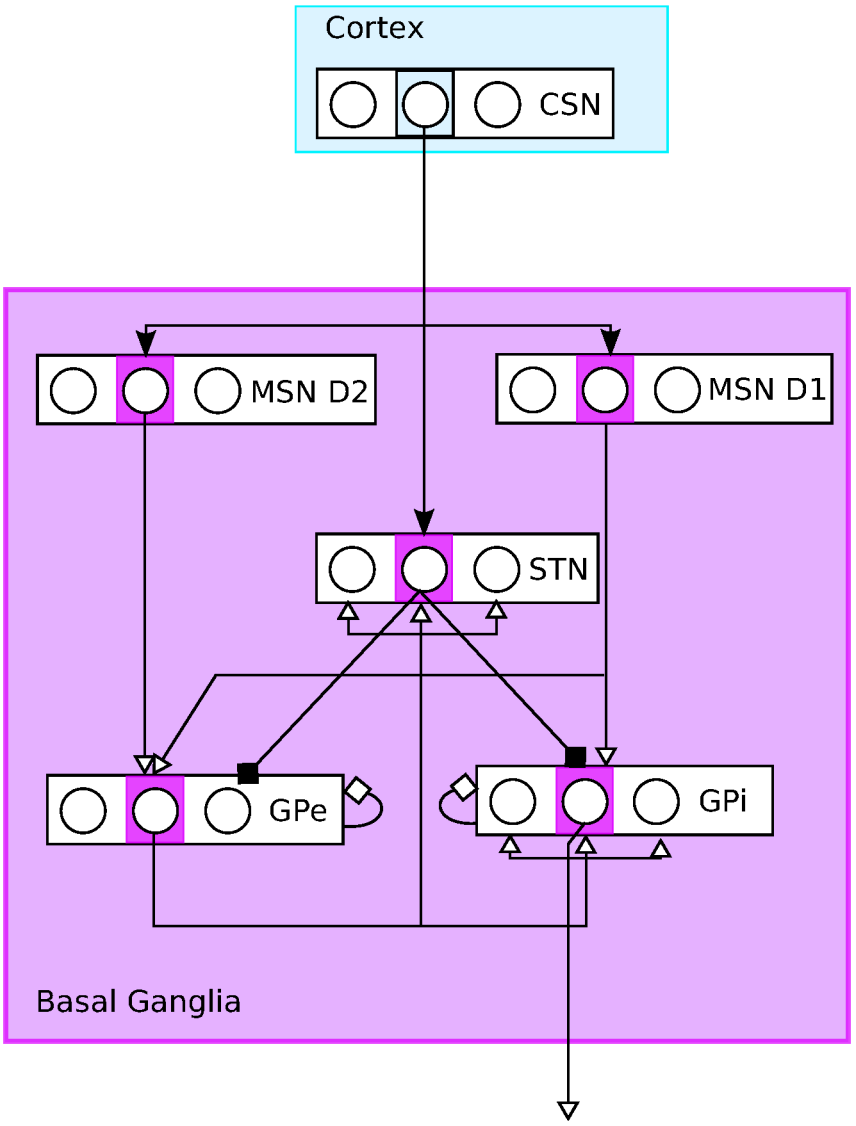

Figure 3.9 – *Illustration des connexions du modèle* GPR *utilisé dans* Humphries *et al. 2012. Les flèches noires indiquent une connexion gabaergique, inhibitrice et les flèches blanches une connexion glutamatergique, excitatrice. Pour plus de clarté, certaines connexions diffuses ont été symbolisées par des diamants. La couleur des diamants respecte le code utilisé pour les flèches. Note : la connection entre les* MSN D1 *et le* GPe *n'est pas présente dans le modèle original (*Gurney *et al. 2001a), mais a été ajoutée par la suite dans* Humphries2012.

mettant en évidence des capacités intéressantes de sélection de l'action. Ce modèle a donc l'avantage d'être à la fois simple d'un point de vue computationnel et d'avoir des capacités de sélection importantes.

Ce modèle souffre toutefois de plusieurs défauts. Tout d'abord nous avons vu que l'hypothèse de séparation franche des chemins direct et indirect semble être une simplification de la connectivité des ganglions de la base. De plus, les poids de connexion de ce modèle n'ont aucune réalité physique et ont été optimisés à la main afin que le modèle exhibe des capacités de sélection. On peut donc se poser la question si l'optimisation des poids n'est pas, à elle seule, responsable des capacités de sélection du modèle. D'autant que les poids ne sont contraints sur aucune base biologique.

3.5.2 Frank et collègues

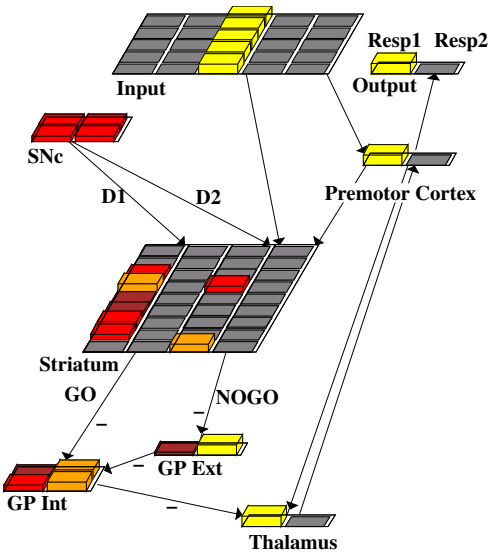

Figure 3.10 – *Modèle de ganglions de la base développé par Frank et collègues (*Frank *et al. 2004 ;* Frank *2005.*

Frank et collègues ont proposé un modèle des ganglions de la base, reposant sur le modèle des chemins direct et indirect (Frank et al. 2004 ; Frank 2005 ; voir Figure 3.10). Le modèle suppose la compétition entre le chemin direct, *Go*, facilitant la sélection de l'action et le chemin indirect, *NoGo*, qui prévient la sélection. Ainsi, lorsqu'une action mène a une récompense, l'excitation phasique des neurones dopaminergique vient renforcer les *MSN D1* et donc la composante *Go* de l'action. Au contraire, si l'action mène à une punition, l'inhibition phasique des neurones tend à renforcer les *MSN D2* et donc le chemin *NoGo* de l'action.

Ce modèle prédit que des patients parkinsonien seront plus sensibles à la punition et moins sensibles à la récompense. En effet, sous un faible niveau dopaminergique, le

modèle prédit un renforcement du chemin *NoGo* et une diminution de la force du chemin *Go*. Le modèle prédit l'effet opposé chez des sujets ayant un haut niveau dopaminergique (tels que les patients parkinsonien sous médication). Frank et al. 2004, montrent que les prédictions du modèle sont vérifiés de façon expérimentale chez des patients atteints de Parkinson. Ainsi, ce modèle a permis d'obtenir une explication simple et élégante du rôle de l'architecture interne des ganglions de la base, et notamment des changements de la dynamique du système observé chez des patients parkinsoniens. Cependant, certaines études expérimentales semblent remettre en questions cette interprétation (Shiner et al. 2012 ; Smittenaar et al. 2012 ; voir Chapitre 6).

La particularité de ce modèle est d'intégrer une part importante d'apprentissage, qui ce fait sous deux formes. D'une part, les poids synaptiques sont initialisés de façon aléatoire et sont optimisés par une variante de l'apprentissage hebbien implémenté dans la plateforme Leabra (O'Reilly 1998). D'autre part, l'apprentissage a lieu via le signal dopaminergique qui est positif lorsque la récompense est reçue à la fin de l'essai et négative sinon. Ce modèle a été utilisé dans de nombreuses études et existe en plusieurs versions dont une comprenant le noyau subthalamique (Frank 2006), ce qui a permis de reproduire les oscillations β que l'on observe lors d'une diminution du niveau dopaminergique, conséquence de la maladie de Parkinson. Ils ont également proposé que le rôle du noyau subthalamique, et de la voie hyperdirecte, est de fonctionner comme un frein lors de la prise de décision et aurait donc un rôle de régulation de la sélection, à l'image de la proposition du modèle *GPR*.

Ce modèle a une grande force d'explication des biais cognitifs dans l'apprentissage par la récompense et la punition, chez les sujets parkinsoniens. Cependant, les poids de connexion synaptique n'ont aucune base biologique. De plus, leur interprétation repose sur la ségrégation D1/D2 des neurones du striatum ce qui peut être aujoud'hui considéré comme une simplification de l'anatomie des ganglions de la base.

3.5.3 Apprentissage parallèle : *model based* et *model free*

Le modèle d'apprentissage développé dans Daw et al. 2005 repose sur la théorie de l'apprentissage par renforcement (Sutton et Barto 1998). Il permet de modéliser comment les comportement habituels et dirigés vers un but (voir section 3.3.2) sont appris en parallèle. Ce modèle n'est pas à proprement parlé un modèle des ganglions de la base mais permet d'évaluer les processus intégrés par les différentes boucles cortico-basales.

Le modèle utilise deux systèmes d'apprentissage : un premier réactif, *model free* et un second capable de planifier les actions, *model based* (voir Figures 3.11 A-B). Le système *model free* modélise les comportements habituels, tandis que le système *model based* modélise les comportements dirigés vers un but. Chaque module est capable d'effectuer la prise de décision à lui seul.

Les modèles sont mis à jour en continu et en parallèle en fonction de l'arrivée ou non de la récompense. Le système *model free* est réactif et n'apprend que des associations état/action. Le système *model based* apprend également un modèle des transitions sous forme d'arbre. Ce dernier peut donc planifier les conséquences de ces actions.

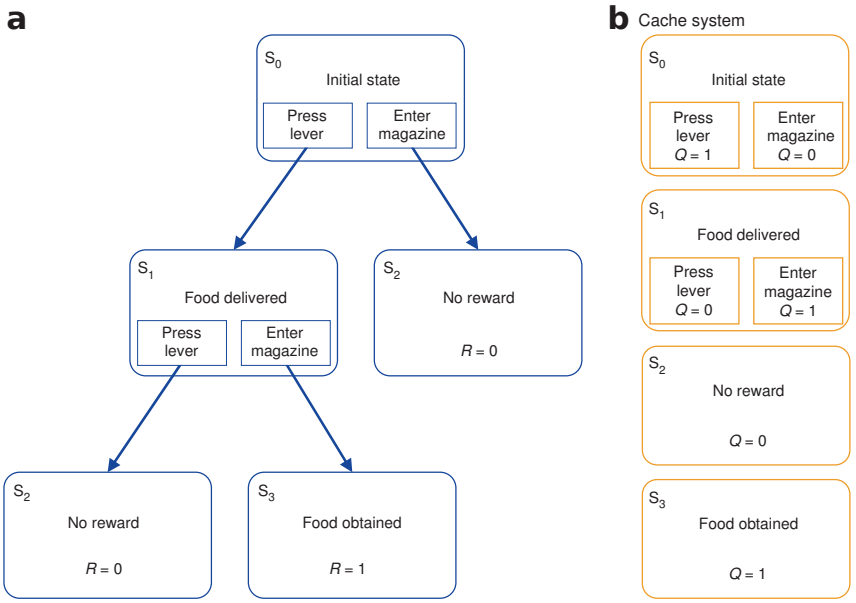

Figure 3.11 – *Illustration du modèle proposé par* Daw *et al. 2005. A. Module* model based, *apprenant un modèle des transitions. B. Module* model free. *Illustration reprise de* Daw *et al. 2005.*

Pour sélectionner le comportement à effectuer, leur modèle tiens à jour l'incertitude lié à chaque fonction de valeur des systèmes. Le système ayant l'incertitude la plus faible décide de l'action à effectuer.

L'hypothèse principale de cette architecture est que la partie *model free* est peu flexible et aura tendance a persister dans son comportement même après un changement de contingence. Au contraire, la partie *model based* étant capable d'évaluer les conséquences des actions aura une capacité plus grande à changer son comportement.

Ce modèle permet ainsi d'interpréter comment les comportements habituels et dirigés vers un but sont appris en parallèle dans les différentes boucles cortico-striatales. Il suggère également que le comportement résulte d'une interaction entre plusieurs système d'apprentissage.

3.5.4 Le modèle *CBG*

Le modèle *Contracting Basal Ganglia* (*CBG*) a été développé dans Girard et al. 2008.À l'instar du modèle *GPR*, il utilise un modèle neurones ayant une dynamique simple, de type lPDS. Ces neurones sont proches des neurones intégrateurs à fuite avec une borne interne, de sorte que l'activité du neurone ne peut diverger. Ce modèle vérifie des propriétés de contraction qui permettent d'assurer une dynamique stable du modèle et prévenant des phénomène d'hystérésis.

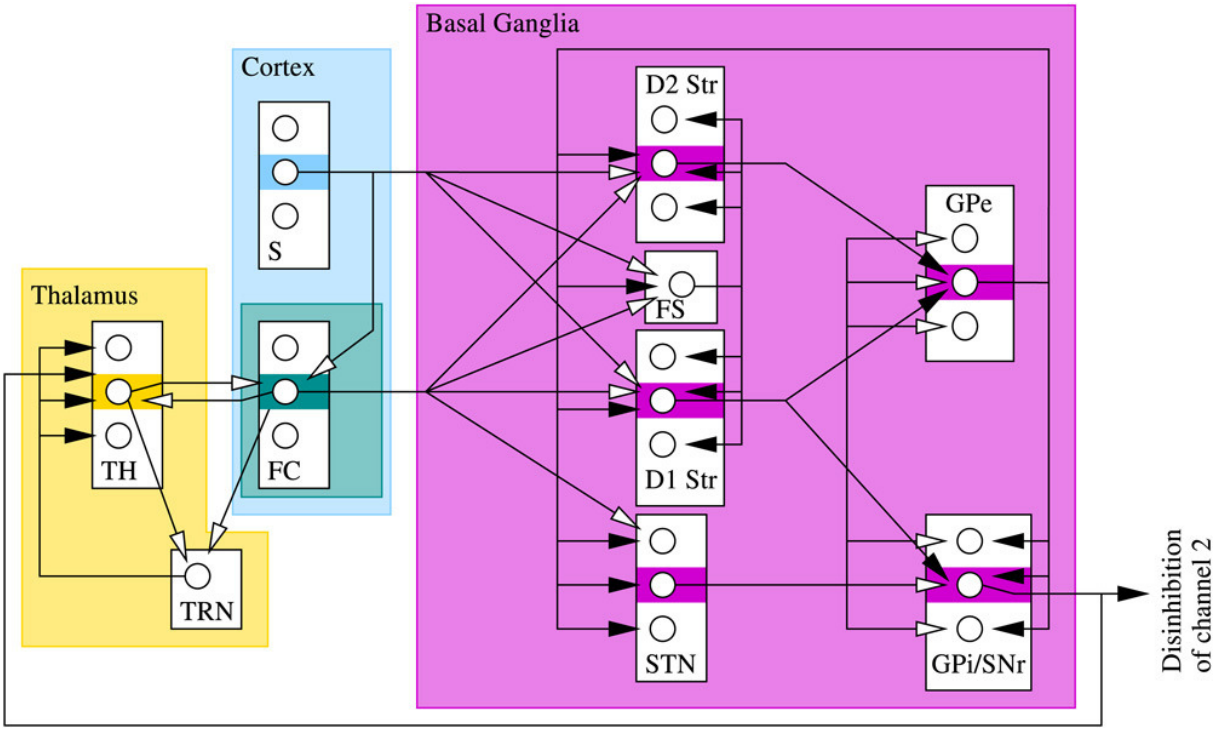

Figure 3.12 – *Illustration des connexions du modèle* CBG *(figure présente dans* Girard *et al. 2008). Les codes concernant les couleurs et formes des connexions sont les même que ceux utilisés pour le modèle* GPR.

D'une manière générale, le modèle *CBG*, comme la majorité des modèles de la littérature, suppose une séparation des neurones du striatum en *MSN D1* et *D2*. Cependant la séparation en chemin direct et indirect n'est que partielle, puisque si les *MSN D2* ne projettent que vers le *GPe*, les *MSN D1* projettent à la fois vers *GPe* et *GPi* (voir Figure 3.12). La connexion entre le *GPe* et *GPi* est considérée comme diffuse, ce qui lui prévient de détériorer la sélection et lui donne un rôle de modulation du *GPi*.

On notera la présence d'interneurones dans le striatum, les neurones à taux de décharge rapide du striatum qui ont un effet *off center* sur les *MSN*. De plus la boucle thalamo-corticale est présente, permettant de représenter dans sa totalité l'interaction cortico-basale supposée améliorer la sélection (Humphries et Gurney 2002).

Ce modèle a été étudié dans un tâche de survie introduite dans Girard et al. 2003. Dans la tâche, un robot doit se déplacer dans un environnement tout en contrôlant son homéostasie afin d'assurer sa survie. Il a deux variable à gérer : énergie potentielle et énergie. Si l'énergie arrive à zéro alors l'agent est considéré comme mort. Il peut la remplir en consommant son énergie potentielle lorsqu'il est dans une position spécifique de l'environnement. De plus, il peut regagner de l'énergie potentielle en se rechargeant à une autre position de l'environnement. Notons que, pour que la recharge soit effective, il faut que l'agent 1) soit au bon endroit et 2) choisisse l'action "se recharger" de façon explicite (la recharge ne se fait pas automatiquement quand l'agent se trouve dans la zone).

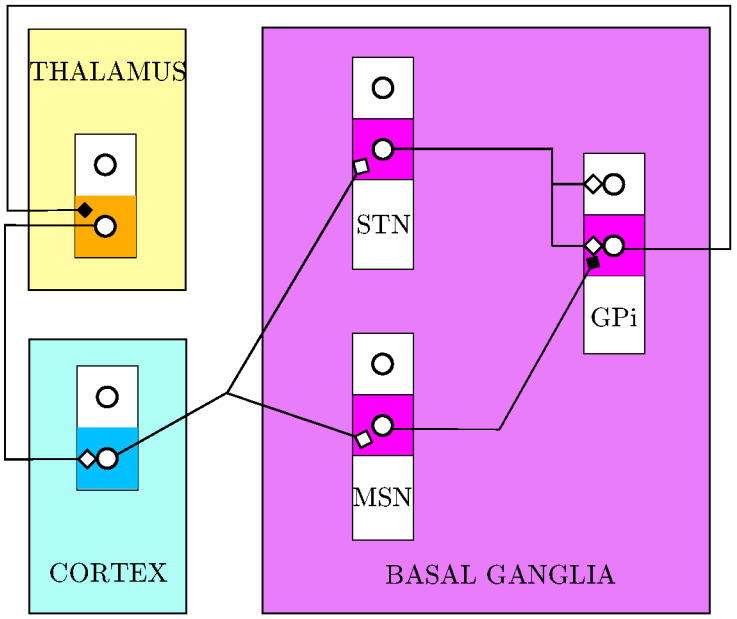

Figure 3.13 – *Architecture de la boucle cortico-basale du modèle développé dans* Leblois *et al. 2006.*

Cette tâche de survie a permis de mettre en évidence les meilleures capacités du modèle à contrôler l'homéostasie du robot, en comparaison avec une architecture simple de contrôle, basée sur un ensemble de condition *if...then...else* (*ITE*). Le *CBG* a permis une gestion de l'homéostasie du robot moins consommatrice en énergie que la règle de décision *ITE*.

Ce modèle présente donc des capacités de sélection de l'action suffisantes pour le contrôle robotique. Ces propriétés de contraction permettent de plus d'assurer une stabilité du modèle quelque soit l'entrée soumise au système.

3.5.5 Leblois et al. 2006

Le modèle de Leblois et al. 2006 propose d'étudier conjointement les capacités de sélection de l'action des ganglions de la base et de l'apparition d'oscillations dans un état parkinonien.

Le modèle a été simulé de deux façons différentes : dans la première chaque population est modélisée par 1000 neurones de type intégrateur à fuite et dans la seconde chaque population est représentée par un unique neurone intégrateur à fuite.

Le modèle propose une architecture épurée des ganglions de la base car il ne comprend pas le noyau *GPe* (voir Figure 3.13). De plus, il considère une unique population de *MSN*. Le modèle ferme la boucle thalamo-corticale avec le *GPi* projetant vers le thalamus connecté au noyau cortical.

Tel que dans le *GPR*, la sélection se fait par un effet *on surround* grâce à la connexion diffuse $STN \rightarrow GPi$ et par un effet *on center*, dû à la projection focalisée des *MSN* sur *GPi*.

Ce modèle a permis de montrer qu'un manque dopaminergique a un effet négatif sur la sélection, que la capacité de sélection est réduite avant l'apparition d'oscillations et que ces oscillations sont dues à la boucle Cortex→*STN*→*GPi*→Thalamus→Cortex.

Ce modèle prédit donc que les oscillations sont la conséquence de la boucle cortico-basale. Notre modèle présenté dans le Chapitre 5 propose plutôt que ces oscillations sont dues à l'interaction entre le *STN* et le *GPe*.

3.5.6 van Albada et collègues

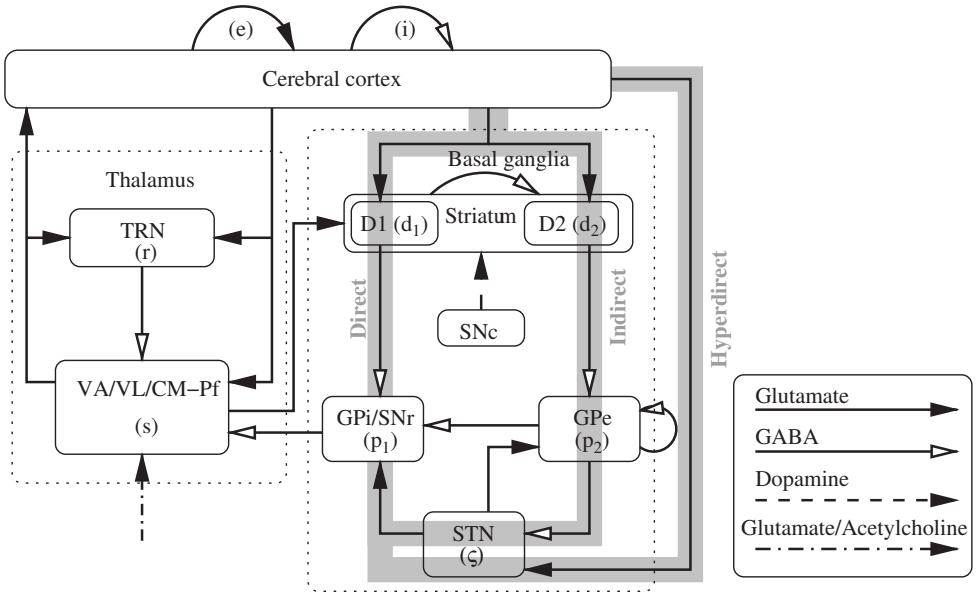

Figure 3.14 – *Connectivité du modèle des ganglions de la base développé par van Albada et collègues. Figure reprise de* Albada *et al. 2009.*

L'architecture de ce modèle à champs moyens reprend l'hypothèse des chemins direct, indirect et hyperdirect (voir Figure 3.14). Le modèle intègre également la boucle thalamo-corticale. Le thalamus est composé de deux modules. Le premier est le noyau thalamic réticulé (*TRN*). Le deuxième module représente la partie ventrolatérale (*VL*) du noyau thalamique, le noyau ventral antérieur *VA* et le thalamus centromédian-parafasciculaire (*CM-Pf*). Il reçoit l'inhibition du *TRN* et du *GPi* et une excitation corticale. Le *TRN* est excité par le signal cortical et par le complexe *VA/VL/CM-Pf*.

Ce modèle a été construit afin d'étudier la dynamique du système dans le contexte de Parkinson. À l'opposé des modèles présentés précédemment, celui-ci n'a pas été étudié dans un cadre de la sélection de l'action. Le but des deux études où il a été utilisé (Albada et Robinson 2009 ; Albada et al. 2009), a été d'observer l'activité des différents noyaux du système lors de la modélisation d'une déplétion de la dopamine dans les ganglions de la base, ainsi que de l'apparition d'oscillations dans la maladie de Parkinson.

L'approche de ce modèle est donc radicalement différente de la plupart des modèles de la littérature car il s'intéresse moins à la fonction du système qu'à sa dynamique, en prenant la maladie de Parkinson comme exemple de dysfonctionnement. Le but est ici d'observer comment la modification de linflux dopaminergique crée un déséquilibre dans l'activité du système.

Chaque paramètre du modèle a une unité correspondant à une réalité physique (force de connexions en mV.s, délais axonaux en ms, etc...). Cette description physiologique du modèle permet d'observer l'activité des noyau en Hz. Il a ainsi une activité directement comparable aux enregistrements électrophysiologiques de la littérature. Après un réglage des paramètres, les auteurs montrent que l'activité des noyaux est compatible avec les données de la littérature et que la modélisation d'un état parkinsonien entraîne des modifications de l'activité plausible. Notamment ils observent l'apparition d'oscillation β dans ces conditions.

Les autres modèles présentés précédemment ont des taux de décharge abstraits ne permettant pas la comparaison quantitative avec des valeurs biologiques d'activité neurale. Cependant cette modélisation plus fine des taux de décharge a un coût computationnel, puisqu'elle utilise des équations plus complexes – comparés aux modèles à intégrateur à fuite du *GPR* ou aux *lPDS* du *CBG* – et prend en compte des constantes d'intégration temporelle du signal. De plus, ce type de modèle requière de régler de nombreux paramètres et constantes de temps, contraints par la biologie, les rendant complexes à mettre en oeuvre.

Comme nous allons le voir avec le modèle *BCBG*, cette approche est complémentaire avec les approches de sélection de l'action et permet également d'avoir une vision peut-être moins simpliste de la dynamique des ganglions de la base.

3.5.7 Le modèle *BCBG*

Le modèle *BCBG* a été développé par Liénard et Girard (Liénard et Girard 2014) et repose sur des neurones à champs moyens. Ce modèle a, à l'image du modèle de van Albada et collègues, l'ambition d'intégrer un maximum de contraintes anatomiques et physiologiques de la littérature des ganglions de la base du macaque. Ils ont utilisé des algorithmes génétiques afin d'optimiser les paramètres du modèle tout en gardant un maximum de contraintes biologiques et d'obtenir ainsi une plausibilité maximale. Cette approche se différencie beaucoup des autres modèles qui ont souvent des poids abstraits (typiquement entre 0 et 1 sans réalité physique ; tels que les modèles de Frank et collègues, le *GPR*, le *CBG*) et des paramètres optimisés à la main pour faire émerger une capacité de sélection. Il se différencie également de l'approche utilisée dans van Albada et collègues par l'utilisation d'outils d'optimisation afin de fixer les paramètres.

Ainsi, comme dans le modèle de van Albada et collègues, chaque paramètre du modèle *BCBG* est exprimé en unité physique et non arbitraire. Cela permet d'établir des critères de plausibilité biologique et de discuter des forces de connexions du modèle au regard de données anatomiques de la littérature.

Le calcul de l'activité de chaque noyau dépend de nombreux paramètres biologiques tels

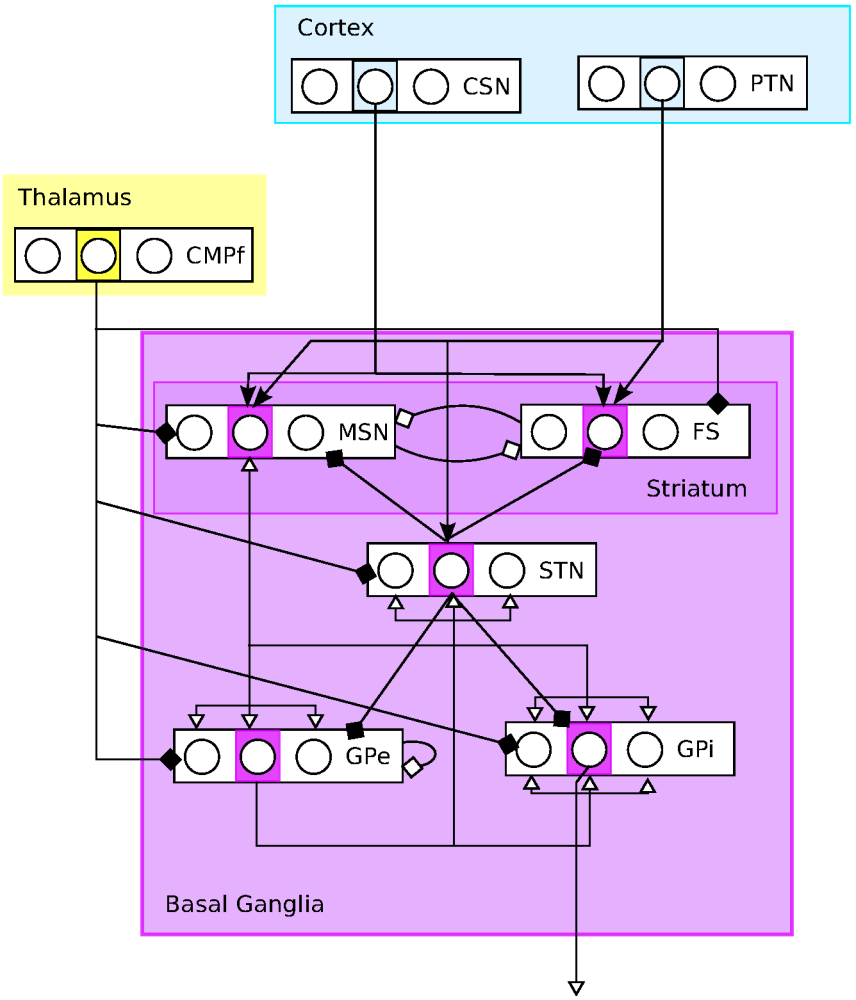

Figure 3.15 – *Illustration de la connectivité dans les ganglions de la base du primate reproduite dans le modèle* BCBG. *Les flèches noires indiquent une connexion gabaergique, inhibitrice et les flèches blanches une connexion glutamatergique, excitatrice. Pour plus de clarté, certaines connexions diffuses ont été symbolisées par des diamants. La couleur des diamants respecte le code utilisé pour les flèches.* CSN *: neurones cortico-striataux;* PTN *: neurones du faisceau pyramidal.*

que le nombre de neurones du noyau, la probabilité de projection d'un noyau à l'autre, la distance moyenne au soma des projections,... (voir Chapitre 5 pour le détail computationnel du modèle). Certains paramètres du modèle ont été fixé d'après les données de la littérature. D'autres, moins connus, ont été optimisé sous des contraintes de plausibilités anatomiques.

L'optimisation par algorithme génétique utilisée pour l'obtention des paramètres non fixés du modèle *BCBG* a permis d'obtenir de nombreuses (plus d'un millier) paramétrisations permettant à la fois de respecter des critères anatomiques et électrophysiologiques. Cette optimisation est multi-critères et permet la création d'un front de Pareto des solu-

tions optimales[3]. Cette approche a donc un double avantage par rapport à une optimisation à la main puisqu'elle est moins fastidieuse et permet d'obtenir plusieurs solutions lorsque l'on a plusieurs critères biologiques à respecter. Cela permet ainsi de tester et discuter plusieurs paramétrisations et d'évaluer des valeurs de paramètres étant encore assez peu connus dans la littérature biologique.

Par exemple, un point soulevé par Liénard et Girard 2014 est la connectivité entre *GPe* et *GPi* qui ne peut être définie comme diffuse ou focalisée grâce aux seules données de la littérature, et qui a été supposée diffuse dans le *BCBG* pour améliorer les capacités de sélection du modèle.

En plus de cette méthode d'optimisation du modèle pour la production de taux de décharge plausible, ce modèle se distingue également par l'anatomie qu'il considère, car néglige la ségrégation des *MSN D1* et *D2* en se basant sur les études questionnant cette aspect de l'anatomie des ganglions de la base, discuté dans la séction 3.4.4.

L'approche utilisée pour la construction du modèle ainsi que son architecture font de ce modèle l'un des plus complet par rapport à la littérature actuelle des ganglions de la base. Cependant, le formalisme computationnel utilisé entraîne une complexité certaine. Nous proposerons dans le Chapitre 5 une version simplifiée de ce modèle permettant de conserver les propriétés biologiques du système tout en simplifiant le formalisme et en le rendant comparable aux autres modèles fonctionnels des ganglions de la base. Nous ajouterons également dans le Chapitre 6 des capacités d'apprentissage au modèle se basant sur le modèle de Frank et collègues présenté précédemment.

3. En optimisation multi-critères, une solution est optimale au sens de pareto si elle est optimale sur au moins l'un des critères.

Chapitre 4

Les neurones dopaminergiques n'encodent pas un pur signal de RPE.

Sommaire

Comme nous l'avons vu au chapitre 2, dans la littérature actuelle, l'activité phasique des neurones dopaminergiques (DA) est supposée encoder une erreur de prédiction de la récompense telle que définie dans les algorithmes d'apprentissage par différence tem-

porelle. Cette hypothèse est basée sur de nombreuses études électrophysiologiques, investiguant l'activité phasique des neurones dopaminergiques lors de tâches de conditionnement pavlovien. Toutefois, la nature exacte de ce signal reste encore à définir, notamment lorsque la tâche nécessite de faire un choix entre plusieurs options. Dans ce dernier cas, deux études ont conduit à des conclusions contradictoires, l'une suggérant que la dopamine encode des informations sur l'action future de l'animal, l'autre suggérant que cette information n'est pas prise en compte. Afin de démêler ces deux conclusions, nous avons simulé différents algorithmes d'apprentissage de différence temporelle (TD) issus de la littérature, dans une tâche multi-choix, et nous avons étudié leur capacité à reproduire le signal phasique de la dopamine enregistré dans une précédente étude (Roesch et al. 2007). Nos résultats indiquent que le signal dopaminergique ne peut pas être reproduit avec précision par une erreur de prédiction de récompense pure, et est mieux reproduit avec un mélange de valeur et d'erreur de prédiction de la récompense. En outre, nous montrons que l'information portée par les neurones dopaminergiques est, au moins en partie, dissociée du comportement et est mieux reproduit par l'algorithme d'apprentissage Q-learning et l'algorithme Actor-Critic, suggérant que l'activité dopaminergique ne prend pas en compte l'information de l'action future.

4.1 Introduction

Le travail de Schultz et collègues (Hollerman et Schultz 1998 ; Ljungberg et al. 1992 ; Mirenowicz et Schultz 1994 ; Schultz 1998) a conduit à d'importants progrès dans la compréhension des mécanismes neuronaux portant l'influence de l'information de retour sur l'apprentissage. Dans ces études, l'activité des neurones dopaminergiques (DA) présentait quatre propriétés clés du signal d'erreur de prédiction de récompense (RPE) utilisé dans les algorithmes de Différence Temporelle (TD) issus de l'intelligence artificielle (Doya 2007 ; Schultz et al. 1997 ; Sutton et Barto 1998 ; voir Chapitre 2) : (1) ils répondent aux récompenses inattendues ; (2) ils répondent à des signaux prédisant une récompense ; (3) ils ne répondent pas à des récompenses attendues ; (4) ils montrent une diminution de l'activité en réponse à l'omission d'une récompense attendue. Dans la littérature de l'intelligence artificielle, ce signal de RPE agit comme un signal d'apprentissage, permettant aux algorithmes TD d'apprendre à prédire les récompenses futures basées sur l'état et l'action en cours. Compte tenu de la forte connectivité entre le système DA et les ganglions de la base, connus pour leurs propriétés de sélection d'action (Mink 1996) ; voir Chapitre 4), une telle analogie a conduit à l'hypothèse que le signal DA joue un rôle crucial dans l'adaptation du comportement basé sur un apprentissage par essais et erreurs.

Cette hypothèse a été confirmée et étendue par de nombreuses études montrant la pertinence des algorithmes d'apprentissage *TD* à expliquer les mécanismes de sélection de l'action et de l'adaptation comportementale impliquant les neurones dopaminergiques et les ganglions de la base (Bayer et Glimcher 2005 ; Fiorillo et al. 2003 ; Niv et al. 2005 ; Tanaka et al. 2004). Cependant, les informations précises codées par les signaux DA restent incertaines. Une raison à cela est que l'activité DA a été principalement en-

registrée au cours de tâches où l'animal est passif, ou bien doit effectuer une action simple – par exemple toucher un levier au début de chaque essai (Schultz 2001) – sans avoir à choisir entre différentes possibilités. Ainsi, les résultats ne peuvent pas révéler le lien entre ce signal et le choix d'une action. Cette distinction est pourtant d'importance puisque les différents algorithmes d'apprentissage *TD* proposés dans la littérature de l'apprentissage automatique traitent l'importance du comportement ou des actions différemment : dans l'algorithme Sarsa, le signal de récompense d'erreur de prédiction dépend du choix de l'action future tandis que dans d'autres algorithmes populaires comme Actor-Critic et Q-learning, le signal d'erreur de prédiction de récompense est indépendant de l'action future (Sutton et Barto 1998). De récentes études électrophysiologiques ont abordé cette question, en mesurant l'activité de DA lors de tâches à choix multiples (Daw 2007 ; Day et al. 2010 ; Morris et al. 2006 ; Niv et al. 2006 ; Roesch et al. 2007). Cependant, ces études sont arrivées à des conclusions divergentes concernant l'algorithme expliquant le mieux l'influence de l'action sur l'activité de DA : Morris et al. 2006 ont trouvé que le signal phasique des neurones DA contient une information encodant l'action future, donc compatible avec Sarsa, tandis que Roesch et al. 2007 ont trouvé une information indépendante de l'action future, conforme à l'algorithme Q-learning (voir Chapitre 1).

Dans cette étude, nous avons cherché à répondre à cette question en proposant une analyse quantitative de l'information codée par les neurones dopaminergiques. Nous effectuons une comparaison du signal de RPE calculé par des modèles d'apprentissages *TD*, avec un ensemble de données électrophysiologiques enregistrées dans Roesch et al. 2007. Dans cette étude, les auteurs ont enregistré des neurones dopaminergiques de l'aire tegmentale ventrale[1] (*VTA*), chez des rats ayant à choisir entre deux actions conduisant à différents types de récompense. Elles sont associées à différents délais d'acquisition ou différentes tailles (voir Figure 4.1).

Nous avons effectué des simulations avec les principaux algorithmes d'apprentissage *TD* et extrait à la fois le signal de RPE et l'information de valeur calculés au cours de l'apprentissage afin d'analyser quel mélange précis de ces deux informations simulées permettaient la meilleure reproduction du schéma d'activité DA. Il est intéressant de noter qu'un modèle d'activité DA contenant un unique signal de RPE ne permet pas de reproduire les schémas d'activation phasique des neurones DA observés, contredisant ainsi le lien direct théorique supposé entre l'activité DA et le signal de RPE calculés par les algorithmes d'apprentissage TD.

Néanmoins, nous avons constaté que les schémas d'activité DA observés peuvent être modélisés comme un mélange de RPE et de valeur. En outre, libérer les contraintes comportementales sur les algorithmes a permis d'obtenir un signal mixte valeur et RPE, calculé par le modèle Q-learning, optimal pour la reproduction de l'activité DA. Nous avons de plus montré que la version Actor-Critic peut également reproduire ce genre d'activité de façon satisfaisante. Dans l'ensemble, ces résultats suggèrent une interaction plus complexe entre la capacité d'apprendre à prédire la récompense par le signal DA et le comportement, et sont compatible avec l'hypothèse que l'adaptation comportementale

1. Techniquement sur 19 neurones enregistré, seuls 2 appartiennent à la substance noire, le reste se trouve dans l'aire tegmentale ventrale

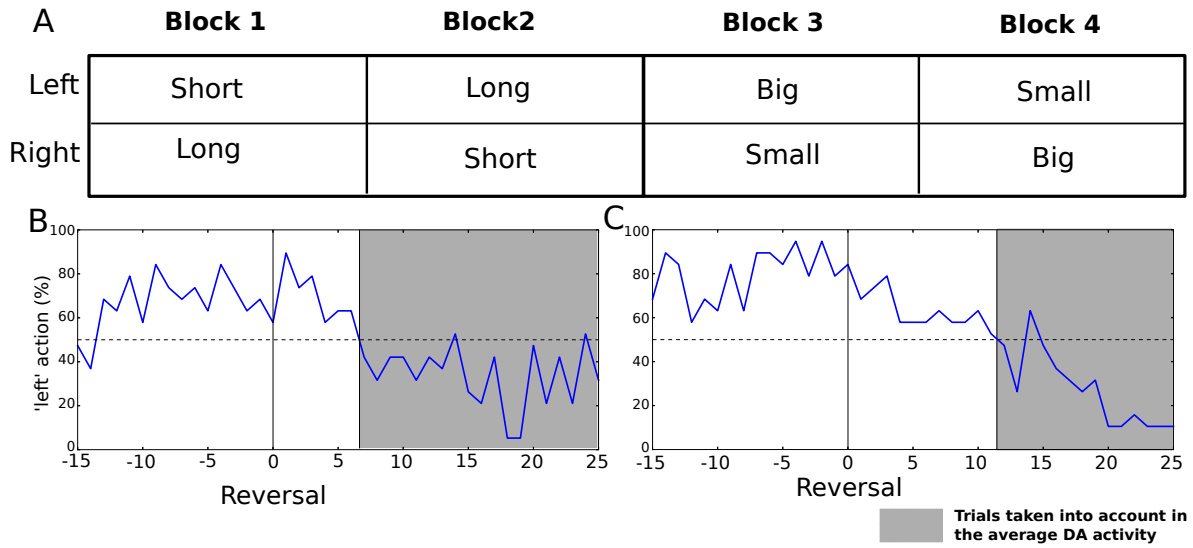

Figure 4.1 – Description de la tâche utilisée dans de Roesch et al. 2007. *A. Chaque session est composée de quatre blocs différents. Chaque bloc a une contingence différente et les changements de blocs sont non signalés. Les deux premiers blocs sont les blocs* delay*. Dans le premier bloc, la récompense* short *(e.g. délivrée après un court délai) est délivrée dans le puits de gauche et la récompense* long *(e.g. délivrée après un délai de plusieurs secondes, voir texte) est délivrée dans le puits de droite. Le deuxième bloc a la contingence inverse. Les blocs 3 et 4 sont les blocs* size*. Dans le bloc 3, la grande récompense est livrée dans le puits de gauche et la petite dans le puits de droite. Le bloc 4 a la contingence inverse. B-C. Évolution du comportement de l'animal enregistrée au cours des changements de blocs* delay *et* size *(respectivement* $1 \rightarrow 2$ *et* $3 \rightarrow 4$*). En gris sont représentés les essais à partir desquels l'activité DA a été enregistrée.*

n'est pas le résultat d'un système d'apprentissage unique, mais semble refléter une interaction plus complexe entre multiples systèmes d'apprentissages parallèles (Balleine et O'Doherty 2010 ; Corbit et Balleine 2011 ; Daw et al. 2005 ; Lesaint et al. 2014 ; White et McDonald 2002).

4.2 Méthode

4.2.1 Procédure expérimentale

Au cours de la tâche, les rats effectuent plusieurs blocs d'essais dans lesquels ils doivent apprendre à choisir la meilleure option entre deux puits fournissant différentes récompenses (voir la figure 4.1). Dans les blocs 1 et 2, appelés blocs *delay*, un puits est associé à une récompense immédiate (option *short*), l'autre avec une récompense retardée de plusieurs secondes (option *long*). Afin d'empêcher l'animal d'abandonner s'il est soudainement exposé à un délai élevé, l'attente associée à l'option *long* est augmentée progressivement : 1s lors du premier essai où l'animal sélectionne l'option *long*, 2s au second essai,

jusqu'à un maximum de 7s. En revanche, s'il choisit plus de 8 fois au cours des 10 derniers essais d'aller vers l'option *short*, alors le retard pour la récompense *long* est raccourci. Dans les blocs 3 et 4, appelés blocs *size*, un puits est associé à une grande récompense (option *big*), l'autre avec une petite récompense (option *small* ; voir la Figure 4.1). On notera que les récompenses *short* et *small* sont identiques. Les blocs sont organisés de telle sorte que la meilleure option est alternativement à gauche puis à droite : par exemple, gauche = *short* durant le bloc 1, gauche = *long* pendant bloc 2, gauche = *big* pendant le bloc 3, gauche = *short* pendant le bloc 4. Les changements de bloc ne sont pas signalés, forçant les rats à apprendre à changer leur préférence en apprenant de leurs propres erreurs au cours des différents blocs. Ainsi, dans chaque bloc, les rats doivent choisir entre le puits gauche ou droite, et apprendre par essais et erreurs quel puits a le meilleur rapport coût/bénéfice (e.g. grande récompense dans le cas *size* et la récompense à court terme dans le cas *delay*).

Une odeur parmi trois est présentée à chaque essai pour aider le rat à faire son choix. Cette odeur est le stimulus conditionné (CS) avec lequel le puits récompensant est associé. L'odeur 1 indique toujours que le puits de gauche contient une récompense (*short, long, big, small* selon le bloc courant) tandis que la droite est vide. L'odeur 2 indique toujours que le puits droit contient une récompense (*short, long, big, small* selon le bloc courant), tandis que le puits de droite est laissé vide. Ainsi les essais dans lesquels l'odeur 1 ou l'odeur 2 sont présentées sont appelés essais *forced choice* parce que l'animal ne peut être récompensé qu'avec une seule option réduisant ainsi son choix. L'odeur 3 indique que chaque puits est récompensant ; cependant, la qualité de la récompense varie en fonction du bloc courant. Ainsi, les essais où l'odeur 3 est présentée sont appelés essais *free choice* puisque l'animal est libre de choisir entre deux options récompensantes.

Tandis que les rats font l'expérience de ces différents blocs, les expérimentateurs ont enregistré le comportement et l'activité de 17 neurones dopaminergiques dans la *VTA*. Deux neurones DA supplémentaires ont été enregistrés dans la *SNC*. Les neurones dopaminergiques ont été identifiés en utilisant à la fois les critères de forme d'onde et sur l'effet d'une injection de l'agoniste à la DA, l'apomorphine, sur leur activité (plus de détails peuvent être trouvés dans l'expérience originale de Roesch et al. 2007).

4.2.2 Modélisation de la tâche expérimentale

Nous avons modélisé les expériences de Roesch et al. 2007 par un processus de décision de Markov (MDP) (voir la figure 4.2 A). Chaque état représente un événement marquant dans la tâche d'origine, qui déclenche une réponse phasique de la DA : le début de l'essai, le *nosepoke*[2], la perception de l'odeur et de la livraison ou de l'omission de la récompense (voir la figure 4.2 B). Ainsi il existe une correspondance entre les états du MDP et les événements vécus par les rats au cours d'un essai. La transition entre l'état *nosepoke* et l'état *odeur* ne dépend pas d'une action active de l'animal et est générée par la simulation afin de présenter chaque odeur le même nombre de fois, comme dans l'expérience originale.

Les états récompense nommés "R {L pour gauche ou R pour droite}{numéro de l'odeur}-

2. *nosepoke* : moment ou l'animal pose son museau dans le port délivrant l'odeur.

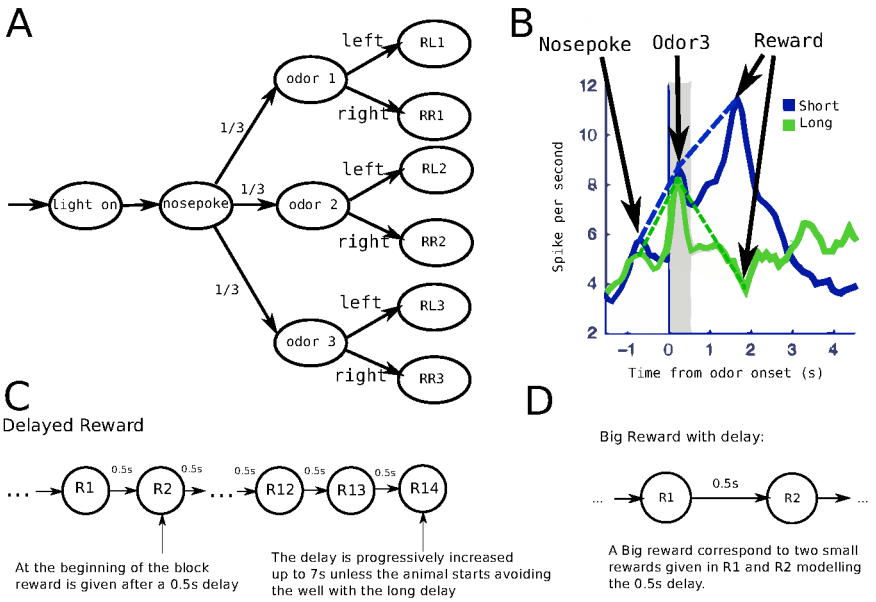

Figure 4.2 – *Modélisation des états de la tâche utilisée dans Roesch et al. (2007). A. Processus de décision Markovien utilisé pour modéliser la tâche ; RL3, récompense à gauche suivant l'odeur 3 ; RR3, récompense droite suivant l'odeur 3. Les autres états représentent le délai. B. Décomposition des états illustrée sur l'activité de DA rapporté par Roesch et collègues. Nous avons extrait l'activité DA de l'enregistrement original à trois moments saillants d'un essai : le moment du nosepoke, la perception de l'odeur et l'état de récompense RL ou Rnuméro de l'odeur. C. Modélisation du délai pour les récompenses du cas delay. D. Modélisation de la grande récompense, représentée comme deux petites récompenses successive.*

$\{i \in [1:14]$ représente le délai d'obtention de la récompense$\}$" – par exemple, RL31 sur la figure 4.2 A – représentent une succession d'états modélisant les différents schémas de récompenses (voir la figure 4.2 C et D). Le modèle tient compte de tous les schémas de récompenses, à savoir les différents délais de récompenses et les différentes tailles de récompenses. Pour passer d'un bloc à l'autre, la simulation manipule le délai de la récompense en ajoutant des états consécutifs sans récompense (cas *delay* voir la figure 4.2 C) ou ajoute un nouvel état récompense pour modifier la taille de la récompense (cas *size* ; voir la figure 4.2 D). Le délai est donc modélisé comme une succession d'états sans récompense, et une transition entre deux états correspond à environ 0.5s dans la tâche réelle.

De plus, la grande récompense est modélisée comme deux petites récompenses délivrés dans deux états consécutifs (voir 4.2 D). Il simule le délai de 0.5 secondes entre les deux petites récompenses utilisé dans l'expérience originale.

Comme l'odeur 1 et l'odeur 2 sont des choix forcés indiquant une récompense à droite ou à gauche respectivement, aucune récompense n'est donnée dans les états RL1 et RR2. En outre, comme dans l'étude originale, le schéma de récompense est le même dans RR1

et RR3, et dans RL2 et RL3.

La valeur de la récompense est réglée à 5 dans notre simulation pour modéliser les 0,05 ml de solution de saccharose à 10% donnée aux rats en guise de récompense.

4.2.3 Algorithmes étudiés

L'algorithme RL général utilisé dans le modèle est représenté dans l'algorithme 2. Nous avons comparé trois algorithmes : Q-learning, Sarsa et Actor-Critic. Q-learning et Sarsa sont basés sur les mêmes principes. En effet, tous deux mettent à jour pour chaque paire état-action, (s, a), une table de valeur (Q-table), qui stocke l'utilité attendue d'effectuer l'action a dans l'état s.

Cette Q-table est appelé *critique* : elle indique a quel point il est préférable de choisir une action particulière dans un état donné afin de maximiser la fonction récompense. Cette information est suffisante pour décider quoi faire dans n'importe quel état, si bien qu'un agent n'a pas besoin d'une autre structure pour déterminer sa politique.

En revanche, si l'architecture Actor-Critic contient aussi un *critique*, elle contient également une structure différente appelé *acteur*, qui représente la politique de l'agent. Dans cet algorithme, le *critique* contient moins d'informations que dans Q-learning et Sarsa. En effet, au lieu de stocker dans une Q-table l'utilité attendue pour effectuer toutes les actions dans l'état s, il stocke uniquement, dans un vecteur V, l'utilité attendue de chaque état s indépendamment de l'action.

L'acteur est représenté comme une table $\mathcal{P}$ qui associe à toute paire (s, a) une valeur à partir de laquelle la probabilité d'effectuer une action a dans l'état s est déduite, soit $\mathcal{P}(s, a) \rightarrow P(a|s)$. L'action effectivement choisie est déterminée par une fonction *SoftMax*, calculée à partir des valeurs de l'acteur.

Les structures de critique, Q et V, sont mises à jour à partir de l'*erreur TD* (information de RPE utilisée par les algorithmes *TD*), δ, en considérant $\forall f \in \{Q, V\} : f_{t+1} = f_t + \alpha\delta_t$. Mais le calcul de l'erreur de *TD* varie en fonction de l'algorithme :

- Q-learning : $\delta_t = r_{t+1} + \gamma \max_a(Q(s_{t+1}, a)) - Q(s_t, a_t)$
- Sarsa : $\delta_t = r_{t+1} + \gamma(Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t)$
- Actor-Critic : $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$

Dans Sarsa, le critique est mis à jour en tenant compte de la valeur de l'action qui sera exécutée dans le futur, alors que dans Q-learning il est mis à jour en supposant que l'agent effectuera la meilleure action, sans exiger que cette hypothèse soit vérifiée dans la pratique. Cette distinction est essentielle dans l'analyse du signal dopaminergique effectuée dans différentes études (Morris et al. 2006 ; Niv et al. 2006 ; Roesch et al. 2007). En effet, alors que Roesch et al. 2007 ont constaté que l'activité DA reflète la meilleure option, compatible avec la RPE calculé par Q-learning, Morris et al. 2006 ont constaté que la DA reflète le choix futur de l'animal, compatible avec Sarsa.

Dans l'architecture Actor-Critic, l'acteur $\mathcal{P}(S_T, a_t)$ doit également être mis à jour avec

$$\mathcal{P}(s_t, a_t) \;=\; \mathcal{P}(s_t, a_t) \;+\; \alpha'\,\delta_t,$$

où le taux d'apprentissage α' peut être différent de celui utilisé pour mettre à jour le

critique. Par exemple, fixer α' inférieur à α, peut induire une convergence plus lente du comportement à l'égard de l'erreur *TD*, qui est supposée correspondre à un signal dopaminergique dans le présent document. Plus généralement, en ayant une structure différente pour l'acteur et pour le critique, l'architecture Actor-Critic rend plus facile la représentation d'un comportement qui n'est pas sous le contrôle strict du critique.

Afin de choisir l'action a effectuer, la même fonction *SoftMax* est utilisée pour déduire la politique π quel que soit l'algorithme :

$$\pi(a|s_t) = \frac{\exp(\beta Q(s_t, a) \, or \, \mathcal{P}(s_t, a))}{\sum_b \exp(\beta Q(s_t, b) \, or \, \mathcal{P}(s_t, b))}.$$

Algorithm 2 Learning

Require: initial state : s_0, block : *mdp*

$s_t \leftarrow s_0$
$a_t \leftarrow softMax(s_t)$
for $i = 0$ to max_iter **do**
  while $s_t nonterminal$ **do**
    $s_{t+1} \leftarrow Transition(s_t, a_t)$
    $a_{t+1} \leftarrow softMax(s_{t+1})$
    update $Q(s_t, a_t)$ or $[V(s_t) \; and \; \mathcal{P}(s_t, a_t)]$ from $(s_t, a_t, s_{t+1}, a_{t+1})$
    $s_t \leftarrow s_{t+1}$
    $a_t \leftarrow a_{t+1}$
  end while
end for

4.2.4 Principe méthodologique général

Les modèles d'apprentissage utilisés dans cette étude partagent trois paramètres variables : le taux d'apprentissage α, la température inverse d'exploration β et le facteur de dépréciation γ. Ces paramètres ont une forte influence sur la dynamique du processus d'apprentissage, à la fois au niveau de la fonction de valeur (et donc sur le signal de RPE) et au niveau du comportement. Pour adapter la valeur de ces paramètres sur les données de Roesch et al. 2007, nous pouvons utiliser des données comportementales, les données dopaminergiques, ou les deux.

Si l'activité DA reflète un processus d'apprentissage qui contrôle directement le comportement de l'animal, comme supposé dans de nombreuses études basées sur des modèles (Pessiglione et al. 2006), il serait souhaitable de reproduire à la fois le comportement et l'activité DA. Ainsi, nous avons exploré l'espace des paramètres du modèle jusqu'à trouver un ensemble de paramètres qui décrit au mieux le comportement appris de l'animal. Nous avons ensuite extrait l'évolution essai par essai de différentes variables (δ, V) dans ce modèle optimisé et comparé ces données à l'activité DA afin d'observer si ils partagent un ensemble de propriétés qualitatives, et donnent quantitativement une bonne reproduction

des données. La qualité de la reproduction des données a été évaluée en effectuant une reproduction quantitative basée sur la régression des données de la figure 6 dans Roesch et al. 2007 et en testant statistiquement si le signal DA reproduit les propriétés décrites dans Roesch et al. 2007. Ces deux approches sont décrites dans les sections 4.2.7 et 4.2.8.

D'autres processus, qui ne seraient pas sous le contrôle des neurones dopaminergiques, peuvent également influencer le comportement de l'animal. Dans ce cas, la dynamique du comportement peut être partiellement déconnectée des variations de l'activité DA. Pour étudier cette possibilité, nous avons également optimisé les paramètres de chaque modèle uniquement sur la reproduction de l'activité neuronale, en utilisant une politique comportementale fixe extraite du comportement des animaux dans la tâche. Les animaux (ou agents) simulés apprennent juste la partie critique de leur modèle en utilisant cette politique fixe. La procédure utilisée dans ce cas est décrite ci-dessous.

4.2.5 Reproduction du signal DA et du comportement

Afin de reproduire les résultats comportementaux de Roesch et al. 2007, 50 agents simulés ont appris la contingence de chaque bloc d'une session au cours de 90 essais, en utilisant les algorithmes présentés précédemment. Chaque odeur a été présentée une fois tous les trois essais. Ainsi, dans un bloc, 30 essais de chaque odeur ont été présentés au modèle, ce qui correspond en moyenne au nombre d'essais par blocs que les rats ont effectué dans l'expérience. Le comportement de chaque modèle, décrit dans la section précédente, a été calculée comme le nombre de choix *left* au cours de chaque essai libre (odeur 3), et a été comparé à celui des rats. Plus précisément, le comportement de chaque modèle a été enregistré au cours des 15 derniers essais du premier bloc et au cours de 30 essais dans le bloc suivant (le comportement est enregistré après le changement de bloc entre les blocs 1 et 2 et entre les blocs 3 et 4), moyenné sur 50 agents faisant l'expérience d'une session. Ce comportement simulé est ainsi comparable au comportement rapporté dans l'expérience de Roesch et al. 2007.

Chaque algorithme a trois paramètres variables, α, β et γ (voir la section précédente), qui influencent le comportement. Pour chaque algorithme, nous avons effectué une recherche de grille et testé toutes les combinaisons de valeurs suivantes pour ces paramètres :

- α : de 0.1 à 0.9 avec un pas de 0.05 (des valeurs plus faibles entre 1E-4 et 0.1 ont également été testées),
- β : de 0.1 à 1 avec un pas de 0.05 ; nous avons également testé 1.5 et 2 pour des valeurs plus élevées,
- γ : de 0.1 à 0.9 avec un pas de 0.1 ; 0.99 a aussi été testé.

Les résultats obtenus ont été comparés, pour chaque jeu de paramètres, à ceux de Roesch et al. 2007 en minimisant la distance entre le comportement simulé et expérimental. Les points dans les courbes de Roesch et al. 2007 dans le cas *size* et le cas *delay* sont le pourcentage de choix de gauche lors de chaque essai des sessions au cours desquelles les mesures électrophysiologiques ont été effectuées (voir la figure 4.1B-C). Nous avons

cherché les paramètres qui optimisent l'adéquation entre nos modèles et ces données dans les deux cas (*size* et *delay*).

Nous avons également effectué une optimisation essai par essai du comportement (Daw 2011) en évaluant la probabilité que le comportement observé dans Roesch et al. 2007 soit généré par les différents algorithmes.

Nous avons dans ce cas utilisé le critère de vraisemblance permettant d'évaluer la probabilité que le comportement expérimental ait été généré par ces jeux de paramètres. La vraisemblance, $\mathcal{L}$ peut alors s'écrire comme :

$$\mathcal{L}(X, \theta) = \prod_{i=1}^{n} p(c_i = X_i|\theta),$$

où X représente le comportement observé des rats lors de la tâche, X_i le choix de l'animal au ième essais, c_i le choix des algorithmes d'apprentissage et θ les paramètres des algorithmes (ie α, β et γ).

Comme précédemment, nous avons effectué une recherche de grille pour trouver les paramètres qui maximisent cette probabilité. Nous avons également utilisé plusieurs descentes de gradient pour valider les paramètres trouvés dans la recherche de la grille. Cette optimisation a été effectuée au cours des 20 sessions où les enregistrements électrophysiologiques ont été réalisés. Cette méthode utilise une politique comportementale fixe décrite dans la section suivante.

4.2.6 Reproduction du signal DA avec une politique comportementale fixe

Dans la deuxième partie de ce travail, nous avons étudié la capacité des règles d'apprentissage précédemment décrites à reproduire l'activité DA sans exiger que les algorithmes reproduisent également le comportement de l'animal. L'action à effectuer n'est donc pas choisie avec un *SoftMax* basé sur les valeurs apprises par les algorithmes, comme décrit dans l'algorithme 1, mais avec une fonction dédiée *chooseAction* construite pour reproduire le comportement des rats de Roesch et al. 2007.

Cette fonction imite les choix des rats lors de l'inversion des 20 sessions qui ont été utilisées pour faire les enregistrements électro-physiologiques. Le modèle est ensuite mis à jour en fonction de ces choix.

En dehors des essais où les données comportementales sont entièrement accessibles (avant les 15 derniers essais dans le premier bloc des cas *delay* et *size*), nous avons supposé que l'animal a choisi l'action la plus attractive dans, en moyenne, 70% des essais au cours des choix libres (Roesch et al. 2007). Au cours des essais à choix forcé (odeur 1 et 2 odeurs), les animaux ont appris à effectuer la meilleure action sur presque tous les essais, donc l'agent choisit la meilleure action 99% du temps dans ces essais. Toutes les autres procédures d'optimisation (quantitatifs et qualitatifs) sont les mêmes que pour le cas avec les contraintes comportementales. Cette méthode nous permet de séparer totalement le processus de décision du processus d'apprentissage de la valeur.

4.2.7 Critère quantitatif de reproduction du signal DA

Comme l'activité DA enregistrée dans cette expérience montre une forte réponse phasique à la récompense, malgré la stabilisation du comportement appris (e.g. convergence comportementale), nous avons supposé que ce signal pourrait être mieux reproduit par une fonction de valeur (e.g. la somme de la récompense immédiate, r_t, plus les récompenses futures prédites $V(s_t)$ pour ACTOR-CRITIC, $Q(s_t, a_t)$ pour SARSA et $\max_a(Q(s_t, a))$ pour Q-LEARNING), au lieu d'une RPE classique. Notons que nous n'avons pas modifié le fonctionnement interne des algorithmes, nous avons uniquement cherché la combinaison des variables internes de ces algorithmes qui pourraient au mieux expliquer l'activité des neurones DA. Pour tester cette hypothèse, pour chaque simulation (e.g. ensemble de paramètres), nous avons testé la capacité de 10 *mixtures* différentes, M_w, avec $w \in [0, 1]$ avec un pas de 0.1, de fonction de valeur et de RPE, à reproduire l'activité DA précédemment enregistrée. Nous avons défini :

$$\begin{aligned} M_w(t) &= wValeur(t) + (1-w)RPE(t-1) \\ &= w[r_t + \gamma V(s_t)] + (1-w)[r_t + \gamma V(s_t) - V(s_{t-1})] \\ &= r_t + \gamma V(s_t) + (1-w)V(s_{t-1}). \end{aligned}$$

Cette équation illustre la *mixture* dans le cas ACTOR-CRITIC. La récompense future attendue $V(s_t)$ est remplacée par $\max_a(Q(s_t, a))$ pour Q-LEARNING et par $Q(s_t, a_t)$ pour SARSA. Bien entendu, le signal RPE intègre également le signal de valeur puisque la RPE est la différence entre la valeur actuelle et la prédiction de la valeur précédente. Ainsi, la *mixture* que nous avons utilisée comme variable explicative pour l'activité DA peut être interprétée comme une RPE déformée ou optimiste, c'est-à-dire où la partie négative –la prévision précédente de la valeur – est sous-pondérée.

Dans le but de comparer l'activité DA avec une *mixture* calculée par les différents algorithmes, nous avons fait correspondre trois états du MDP avec les données expérimentales enregistrées pendant les trois événements saillants présentés précédemment (voir Figure 4.2 B) : le *nosepoke*, la perception de l'odeur et la présentation ou omission de la récompense.

Comme l'activité DA et les *mixtures* simulées ne partagent pas une échelle commune, ni le même niveau de base, nous avons autorisé une transformation linéaire du signal simulé pour s'adapter à l'activité DA enregistrée *in vivo*. Nous avons minimisé la différence entre les deux valeurs par la méthode des moindres carrés (LS) en minimisant l'erreur $e = ||(aM_s + b) - DA_s||^2$ où DA_s est l'activité DA expérimentale moyennée dans l'état s, sur les essais après que les performances du rat passent au dessus de 50% et M_s la mixture moyenne calculée en s au cours des essais d'un bloc.

Ainsi, nous avons : $M_w(s) = \frac{1}{n}\Sigma_{e=0}^{n} M_w^e(s)$, où n est le nombre d'essais considérés et $M_w^s(e)$ est la *mixture* calculée à l'essai e en s. La paire (a, b) est déterminée par la méthode des moindres carrés. L'erreur reportée dans cette étude, notée erreur *LS*, est l'erreur obtenue avec une *mixture* moyennée sur tous les agents simulés. Cette erreur *LS* nous donne une évaluation quantitative de la capacité de notre modèle à reproduire l'activité DA.

4.2.8 Critère qualitatif de reproduction du signal DA

Dans l'étude originale, les auteurs ont comparé l'activité DA lorsque la récompense est délivrée ou omise et entre les essais en début et fin de chaque bloc. Cette analyse a été réalisée pour montrer que, comme dans les travaux antérieurs (Bayer et Glimcher 2005 ; Fiorillo et al. 2003 ; Hollerman et Schultz 1998 ; Schultz 1998 ; Schultz et al. 1997 ; Tanaka et al. 2004), l'activité DA était significativement plus faible pendant les premiers essais d'omission que durant les derniers essais d'omission d'un bloc ; et était significativement plus élevée pendant les premiers essais de livraison que pendant les derniers essais de livraison. En outre, ils ont effectué des tests statistiques sur l'activité DA au moment de l'odeur, en comparant les essais où l'animal choisit la meilleure option avec les essais dans lesquels l'animal choisit la moins attrayante, et ce dans toutes les conditions. La question initiale était de voir si le signal DA est influencé par l'action future de l'animal ou non. Les auteurs ont constaté qu'il n'y avait pas de différence statistique sur le niveau d'activité DA en fonction de l'action choisie dans les essais libres. En outre, l'activité DA enregistrée au cours des choix libres n'est pas statistiquement différente de l'activité DA enregistrée au cours des choix forcés qui ont conduit à la meilleure option dans chaque bloc. Cependant l'activité au moment du choix dans les essais libres est statistiquement différente de l'activité DA enregistrée au cours des choix forcés menant à la pire option.

Ainsi, pour évaluer de façon plus qualitative la capacité des trois modèles à reproduire l'activité DA, nous avons également inclus une analyse statistique de l'activité prévue par les modèles au moment de l'odeur dans les différentes conditions (choix libre *long/short* et *small/big* et choix forcé *long/short* et *small/big*). Cette analyse a porté sur 12 tests statistiques différents (voir le tableau 4.1), afin de reproduire le schéma d'activité observé au moment de la perception de l'odeur. Deux tests supplémentaires ont été ajoutés afin d'évaluer l'évolution de l'activité simulée au cours de l'omission et la livraison de la récompense après une inversion de contingence telle que mentionnée précédemment. Comme dans l'étude initiale, nous avons utilisé des t-tests pour déterminer si l'activité dans ces différents cas était la même ou non. Si la p-valeur était supérieure à 0.05, alors nous n'avons pas rejeté l'hypothèse nulle d'activité moyenne identique, sinon nous avons considéré l'activité des deux cas comme statistiquement différente.

Pour prendre ces résultats en compte dans notre étude, nous avons attribué à chaque modèle un score tests statistiques (ST), qui a été défini comme le nombre de tests statistiques que satisfait un modèle donné sur les tests effectués.

En utilisant cette méthode, nous avons pu analyser à la fois le motif précis de l'activité prévu par nos modèles au moment du choix et celui de l'évolution de cette activité au cours de l'omission et de la livraison de la récompense au cours des essais.

4.3 Résultats

4.3.1 Reproduction de l'adaptation comportementale

Dans le travail de Roesch et collègues, les rats ont appris à choisir plus souvent le puits associé à la meilleure option disponible (la récompense *big* dans le cas *size* et la récompense

	free Big	free Small	forced Big	forced Small
free Big	X	=	=	>
free Small	=	X	=	>
forced Big	=	=	X	>
forced Small	<	<	<	X

	free Short	free Long	forced Short	forced Long
free Short	X	=	=	>
free Long	=	X	=	>
forced Short	=	=	X	>
forced Long	<	<	<	X

Table 4.1 – *Description et résultats des différents t-tests utilisés dans l'étude originale. = : indique que $p > 0.05$ et les données ne sont pas statistiquement différentes. > ou < : indique que $p < 0.05$ et que les données en ligne sont respectivement supérieures ou inférieures aux données colonne.*

short dans le cas *delay*). Dans les essais libres, après avoir appris la contingence d'un bloc, les rats ont choisi la meilleure option 75-80% du temps. Quinze à vingt essais ont été nécessaires aux animaux pour s'adapter à la nouvelle contingence après une inversion de celle-ci. Dans les essais forcés, cependant, les rats ont rapidement appris à choisir le puits conduisant à la meilleure récompense de façon systématique. (Roesch et al. 2007).

Nous avons examiné la capacité des algorithmes d'apprentissage *TD*, présentés antérieurement, à reproduire l'adaptation comportementale des rats au cours des choix libres, qui exigeait de la part des animaux un changement de leur comportement après chaque inversion de contingence pour maximiser leur récompense. Nous avons testé différentes combinaisons de taux d'apprentissage α, de températures inverses β pour un compromis entre exploration/exploitation adéquate, et de facteurs de dépréciation γ requis pour générer avec différents algorithmes (Sarsa, Q-learning et Actor-Critic) un comportement proche de celui produit par les rats dans l'expérience de Roesch et al. 2007.

Les figures 4.3 A, D et G rapportent l'erreur *LS* entre le comportement expérimental et simulé pour Q-learning, Sarsa et Actor-Critic en fonction des différents paramètres. Ces résultats montrent que Q-learning et Sarsa minimisent l'erreur *LS* dans une région spécifique de l'espace des paramètres. En effet, avec un taux d'apprentissage α et une température inverse β proche de 0.3, le comportement est reproduit avec une erreur faible (voir les figures 4.3A et D). Un paramètre γ élevé semble également aider les algorithmes à mieux reproduire le comportement. Q-learning et Sarsa minimisent la distance entre leur comportement et le comportement de rats avec les mêmes paramètres fixes : $\alpha = 0.35$, $\beta = 0.25$ et $\gamma = 0.9$. Avec ces paramètres ils reproduisent l'adaptation comportementale de façon satisfaisante. En effet, ils choisissent l'action *left* avant un changement de bloc 70-80% du temps et à la fin du bloc post inversion, environ 20-30% ; ce qui correspond au comportement des rats dans cette tâche (voir les figures 4.3 BC et EF pour le meilleur ajustement respectivement Q-learning et Sarsa). Nous

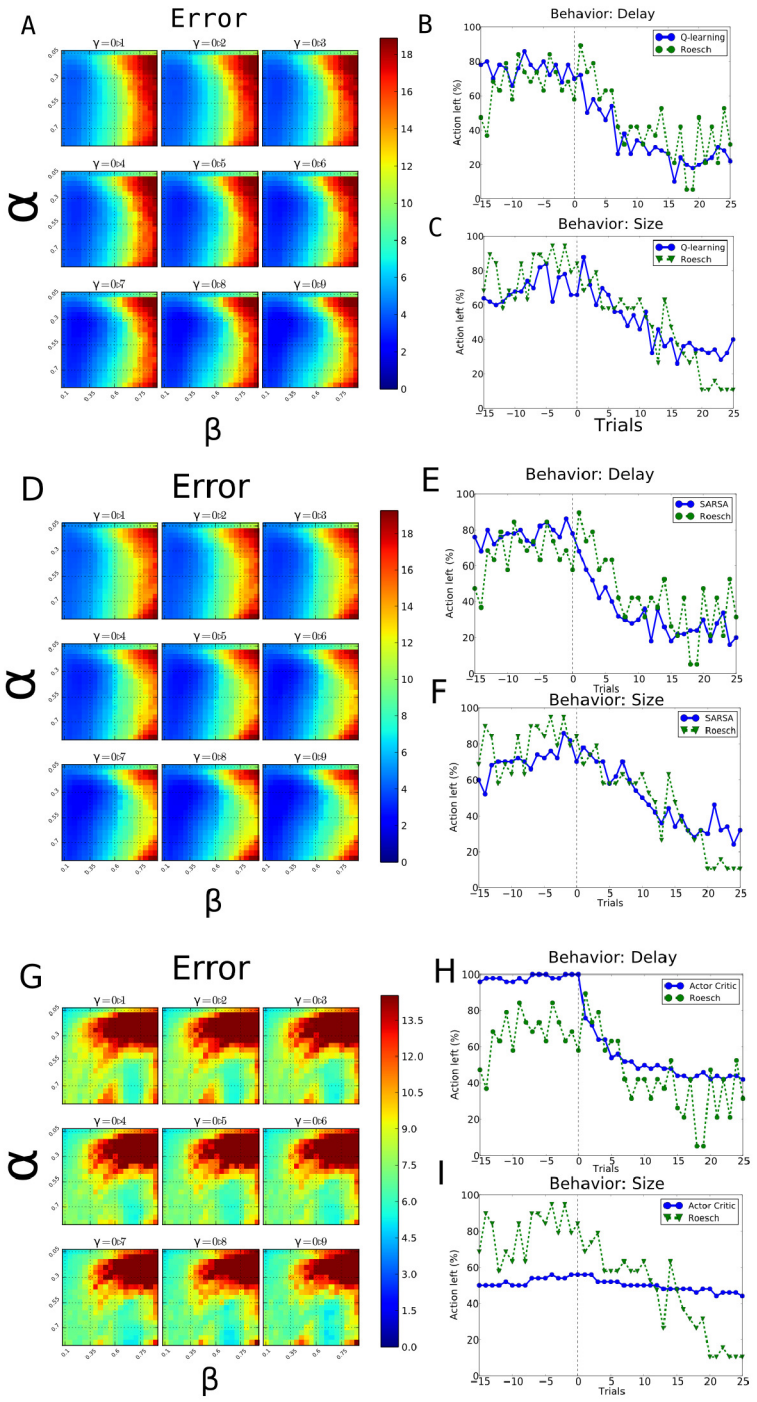

Figure 4.3 – *Reproduction du comportement expérimental avec* Q-LEARNING, SARSA *et* ACTOR-CRITIC. *A, D et G. Erreur LS en fonction de paramètres : le paramètre d'apprentissage α ; la température inverse β et le facteur de dépréciation γ pour respectivement* Q-LEARNING,SARSA *et* ACTOR-CRITIC. *B, E et H. Meilleure reproduction du comportement dans les blocs* delay *pour respectivement* Q-LEARNING, SARSA *et* ACTOR-CRITIC. *C, F et I. Meilleure reproduction du comportement dans les blocs* size *pour respectivement* Q-LEARNING, SARSA *et* ACTOR-CRITIC.

avons également optimisé le comportement sur la base d'une analyse essai par essai (voir méthode) et nous avons trouvé des jeux de paramètres similaires sans modifier les conclusions de ce document ($\alpha = 0.45$, $\beta = 0.25$ et $\gamma = 0.9$ pour Q-LEARNING et $\alpha = 0.35$, $\beta = 0.3$ et $\gamma = 0.9$ pour SARSA).

Ces résultats mettent en évidence la similarité importante de ces deux algorithmes. Bien que le calcul de la RPE soit légèrement différent, les deux algorithmes construisent leurs politiques et calculent la RPE à partir de Q-valeur (voir Méthode pour plus de détails), et ne montrent pas de différences comportementales significatives dans cette tâche.

Par ailleurs, le modèle ACTOR-CRITIC montre une sensibilité très différente aux paramètres par rapport à Q-LEARNING et SARSA. En effet pour ce modèle, l'erreur est minimisée avec un taux d'apprentissage, α, beaucoup plus grand (environ 0,8) et une plus grande température β (voir Figures 4.3 A,D et G) par rapport à ceux permettant de reproduire le comportement des rats pour Q-LEARNING et SARSA. Cependant, même en considérant les paramètres qui produisent le meilleur ajustement ($\alpha = 0.85$, $\beta = 0.7$ et $\gamma = 0.9$), le comportement généré n'est pas satisfaisant. En effet, alors que dans le premier bloc *delay*, le comportement converge à 100% de choix *left*, dans les blocs suivants le comportement moyen est bloqué à 50% d'actions *left*, qui ne correspond pas à l'adaptation comportementale des rats .

D'autres études (BERTSEKAS et TSITSIKLIS 1995 ; LLOYD et al. 2012) montrent également que l'algorithme ACTOR-CRITIC a une capacité limitée à reproduire le comportement de l'animal lors d'inversions de la contingence, comme celle décrite dans ROESCH et al. 2007. BERTSEKAS et TSITSIKLIS 1995 met notamment en évidence, que d'un point de vue théorique, ACTOR-CRITIC est moins apte à prendre des décisions dans un environnement incertain. Contrairement à SARSA et Q-LEARNING, l'algorithme ACTOR-CRITIC doit apprendre à la fois une fonction de valeur qui code la future récompense attendue sachant l'état actuel, $V(s_t)$ et une valeur $P(s_t, a_t)$ dont la politique est inférée. Il semble que cette architecture a besoin de plus de temps pour s'adapter à un changement de bloc et est moins adaptée à effectuer de multiples inversions de la contingence. Cette architecture tend à créer un comportement optimal (voir la figure 4.3 H) avant l'inversion en créant une plus grande différence dans les P-valeurs que dans les Q-valeurs calculées par SARSA et Q-LEARNING. Comme la politique est déduite de P-valeurs, s'il existe une grande différence entre elles, la probabilité de choisir la meilleure action est augmentée.

De plus, étant donné un politique π, nous avons :

$$V(o3) = \pi(o3,'left')[r_{left} + V(r_{left})] + \pi(o3,'right')[r_{right} + V(r_{right})]$$

Ici, o3 signifie odeur 3 et indique l'état dans lequel les modèles reçoivent le signal indiquant un choix libre. $V(o3)$ est donc inférieure à la valeur de la meilleure option et supérieure à la valeur de la pire option si la politique π est stochastique (si la politique choisit toujours l'une ou l'autre action alors V converge vers la valeur de cette action, a, qui est $Q(s, a)$). Par conséquent, lorsque la pire action est choisie, la RPE est négative : $\delta_{worst} = r_{worst} + V(r_{worst}) - V(s) < 0$ donc $P(s, worst)$ est diminuée en fonction de la règle des mises à jour d'ACTOR-CRITIC (voir méthode pour plus de détails). Nous avons l'effet inverse pour la meilleure action. Ainsi, la P-valeur pour la meilleure et la pire option ne

converge pas vers la valeur réelle de la récompense future, mais, lorsque la politique est stochastique, ces valeurs divergent formant un écart important entre elles. Cette différence importante rend une inversion de la contingence plus difficile à apprendre parce que les modèles utilisés ici ont besoin de désapprendre les valeurs précédemment apprises avant d'être en mesure d'en apprendre de nouvelles.

Cependant, au cours de ces simulations nous avons forcé le taux d'apprentissage du critique à être le même que le taux d'apprentissage de l'acteur. Certaines études théoriques suggèrent que le taux d'apprentissage du critique devrait être inférieur au taux d'apprentissage de l'acteur (Bhatnagar et al. 2007 ; Konda et Tsitsiklis 1999). Pour tester davantage la capacité de l'architecture Actor-Critic à reproduire le comportement des rats dans cette tâche, nous avons effectué des simulations additionnelles pour tester si un autre réglage des paramètres, et surtout si deux taux d'apprentissage différents pour l'acteur et le critique, pourraient générer une meilleure reproduction des données. Cependant, la meilleure solution trouvée a un faible taux d'apprentissage (pour l'acteur et/ou critique) et/ou une basse température inverse ce qui donne lieu à une politique très exploratoire, incompatible avec le comportement des rats (voir Annexe 8). Cela montre ainsi une limitation de cette architecture en tant que telle à reproduire des comportements expérimentaux dans un environnement changeant.

Nos résultats montrent que Q-learning et Sarsa sont capables de reproduire fidèlement le comportement des rats au cours des changements de blocs *delay* et *size* avec le même ensemble de paramètres, alors qu'Actor-Critic est incapable de reproduire ce comportement en raison de son architecture différente.

4.3.2 Reproduction de l'activité DA avec les contraintes comportementales

Sur la base des paramètres obtenus à partir de la reproduction du comportement, nous avons examiné si les simulations utilisant la RPE ou une *mixture* de valeur et RPE pourraient correspondre à l'activité DA observée chez les rats. Si l'activité DA reflète le signal RPE de l'algorithme par lequel les rats apprennent la tâche, un algorithme d'apprentissage paramétré pour reproduire le comportement devrait montrer le même type de signal de RPE que celui enregistré chez les neurones dopaminergiques. Pour évaluer la capacité d'un signal à reproduire l'activité DA, nous avons deux critères : 1) le score ST basé sur la capacité du signal à reproduire le modèle d'activité DA enregistrée au moment de l'odeur et pendant l'omission ou la livraison de la récompense durant l'apprentissage (voir Méthodes pour plus de détails) et 2) l'erreur de reproduction LS obtenue en minimisant la distance entre le signal DA et une transformation linéaire du signal simulé.

Nos résultats montrent que le signal de RPE simulé avec Q-learning ou Sarsa a trop convergé pour expliquer la réponse DA élevée à la récompense observée dans les données expérimentales (voir les figures 4.4 A et C). La meilleure reproduction obtenue avec un signal de RPE pur calculé par Sarsa ou Q-learning est un signal presque

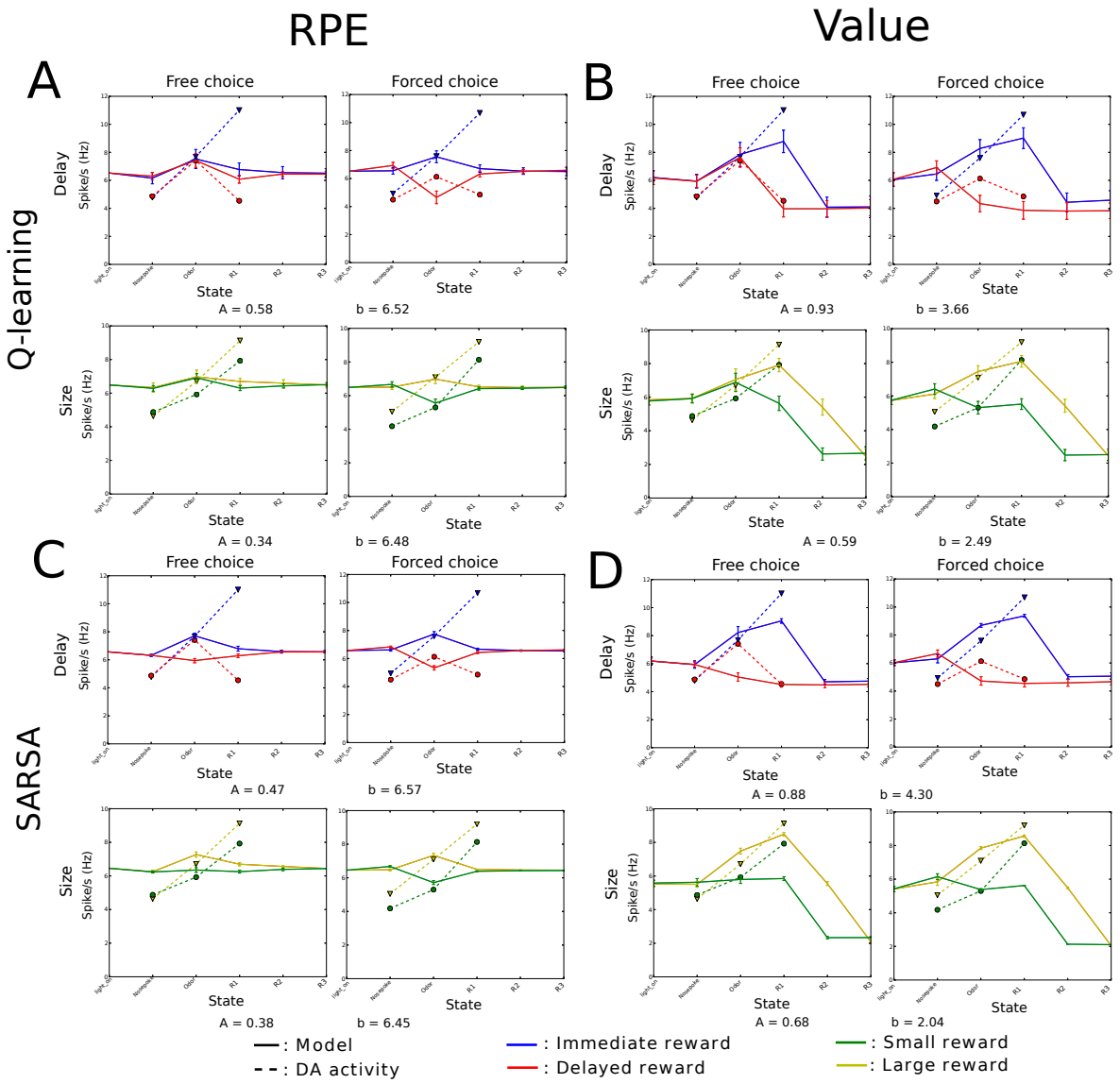

Figure 4.4 – *Reproduction de l'activité DA avec des paramètres qui reproduisent au mieux le comportement expérimental. Chaque sous-figure illustre la meilleur reproduction de l'activité DA pour les cas delay et size, pour les choix libres et forcés. Deux optimisations ont été réalisées : une pour le cas size et une autre pour le cas delay résultant en deux paires (A, b) qui minimisent la distance entre le signal DA enregistré et A*(signal simulé) + b. Cette transformation linéaire est nécessaire afin de comparer le signal DA avec des valeurs simulées, car elles ne partagent pas la même échelle ni la même "activité" de base. Le signal simulé a été moyenné sur 50 sessions. A-B. Reproduction de l'activité DA avec respectivement, la RPE et la valeur calculée par* Q-LEARNING. *C-D. Reproduction de l'activité DA avec, respectivement, la RPE et la valeur calculée par* SARSA.

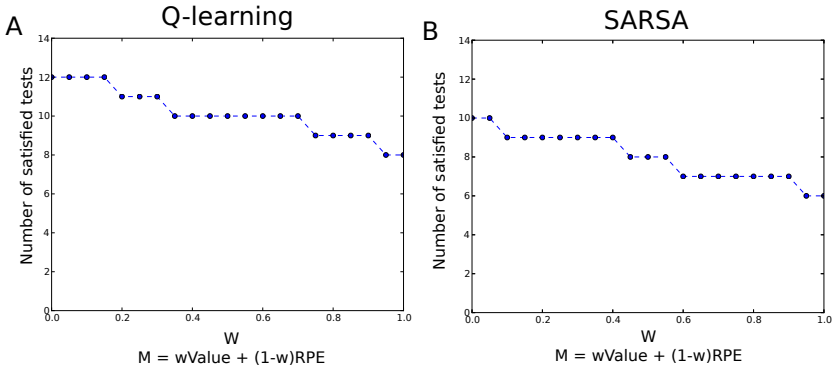

Figure 4.5 – *Nombre de tests valides (score* ST*) en fonction du poids, w, de la* mixture *pour* Q-LEARNING *(A) et* SARSA *(B) lors de la reproduction de l'activité DA sous contrainte comportementale. Si w = 1, le mélange représente la somme de la récompense immédiate et future attendue (c'est-à-dire une fonction de valeur pure) et si w = 0, alors le mélange représente un signal RPE pur.*

plat indiquant que les algorithmes ont appris à prédire pleinement la récompense et n'ont donc pas fait d'erreur de prédiction de récompense. Cette réponse très faible au moment de la récompense n'est pas compatible avec l'activité DA enregistrée dans les mêmes conditions. Cependant, à la fois pour Q-LEARNING et SARSA, un signal de RPE pur obtient un meilleur score *ST* que toutes autres *mixtures* (voir la figure 4.5 A et B). Cela indique que, conformément à l'interprétation de ROESCH et al. 2007, pour ces algorithmes et dans le cas de paramètres contraints par le comportement, un signal de RPE pur peut mieux reproduire le schéma de l'activité au moment de choix et que les deux signaux RPE peuvent reproduire l'évolution de l'activité DA au cours de l'omission et la livraison de la récompense. Un mélange avec un poids trop important sur la fonction de valeur ne permet pas reproduire une telle évolution (voir la Figure 4.5).

D'autre part, la valeur simulée permet de mieux reproduire le schéma global d'activité DA (voir la figure 4.4 B et D) pour SARSA et Q-LEARNING en reproduisant l'activité croissante au cours de l'essai, lorsque l'agent simulé se rapproche de la récompense. L'erreur *LS* est donc plus faible pour la fonction de valeur que pour un signal de RPE ou toutes autres mixtures. Plus globalement, le score *ST* de la fonction de valeur est inférieur à celui du signal de RPE. Cependant la fonction de valeur ne peut pas reproduire l'évolution de l'activité DA observée expérimentalement pendant la livraison ou omission de la récompense.

La plus faible erreur de *LS* a été observée lorsque l'on considère un signal de valeur pur, alors que le score le plus élevé *ST* a été observé avec un signal de RPE pur (voir la figure 4.5). En outre, Q-LEARNING semble être mieux adapté pour reproduire les données observées, puisqu'il prédit une activité qui ne dépend pas de l'action dans les essais libres (voir la figure 4.4 A et B). En revanche, SARSA prédit des signaux différents dans les essais libres (et forcés), en fonction de l'action choisie (voir la figure 4.4 C et D).

Nos résultats montrent que le signal DA n'est ni un signal de RPE pur, ni un signal de

valeur pur. Bien que Q-learning obtienne de meilleurs résultats que Sarsa, conformément à l'interprétation de Roesch et al. 2007, lorsqu'il est ajusté sur le comportement du rat, le signal de RPE des deux algorithmes convergent trop pour reproduire le schéma d'activité DA observé au moment de la récompense. Ainsi aucune *mixture* ne reproduit intégralement le modèle de l'activité au moment de l'odeur, la forte réponse à la récompense, et l'évolution du signal pendant l'omission et la livraison de la récompense au cours de l'apprentissage.

Une explication possible est que le signal DA enregistré ici ne reflète que partiellement le processus d'apprentissage qui sous-tend le comportement observé. Cette hypothèse est basée sur de nombreuses études qui suggèrent la présence de plusieurs systèmes d'apprentissage parallèles impliquant différentes parties des ganglions de la base et du cortex (Daw et al. 2005 ; Ito et Doya 2011 ; Samejima et Doya 2007 ; Yin et al. 2004, 2005). Si certains processus d'apprentissage ne sont pas sous le contrôle du signal DA – ou du moins pas que de VTA – alors il est possible de laisser le signal s'écarter de la prédiction des modèles contraints par des changements comportementaux. Par conséquent, dans la prochaine partie de notre travail, nous avons relâché la contrainte comportementale, ce qui nous permet d'explorer la capacité des modèles à reproduire la seule activité DA.

4.3.3 Reproduire l'activité DA avec une politique fixée

Pour approfondir l'étude de la nature de l'information codée par les neurones dopaminergiques, nous avons relâché la contrainte comportementale afin de se concentrer sur la comparaison entre le signal DA enregistré dans l'expérience de Roesch et al. 2007 et plusieurs signaux générés par différentes *mixtures* de valeur et de RPE des différents modèles simulés.

Nous avons simulé l'activité DA en utilisant différents modèles définis par : (1) les algorithmes utilisés : Q-learning, Sarsa ou Actor-Critic ; (2) les paramètres de l'algorithme choisi : taux d'apprentissage α et facteur de dépréciation γ (β n'étant plus nécessaire puisque la politique est fixe) ; (3) le paramètre de *mixture*, w, qui définit la *mixture* (w = 1 si le signal est un signal de valeur pure ; w = 0 si c'est un signal de RPE pur). Comme dans la section précédente, nous avons attribué un score à chaque modèle en fonction du nombre de tests statistiques qu'il peut satisfaire (score ST), et de l'erreur lors de la reproduction du signal DA (erreur LS). Les Figures 4.6 A, B et C rapportent les résultats des modèles qui ont réussi à reproduire l'évolution de l'activité au cours de l'omission et de la livraison de la récompense.

Comme prédit par Roesch et al. 2007, Q-learning a été capable de reproduire l'ensemble des 14 tests, obtenant un score ST parfait. De façon importante, et contrairement aux conclusions antérieures, le modèle Actor-Critic a pu obtenir un score similaire avec notre analyse qualitative. Le fait qu'Actor-Critic puisse satisfaire tous ces tests est en partie inattendu car, en moyenne, nous nous attendions à ce que la réponse au moment de l'odeur pendant les essais libres soit inférieure à la réponse à l'odeur pendant les essais forcés conduisant à la meilleure option. Néanmoins, c'est cohérent avec le travail théorique sur l'apprentissage par renforcement montrant que la fonction de valeur V apprise par le critique converge progressivement vers la valeur maximale des Q-valeurs

apprises par Q-learning (Jaakkola et al. 1994). Ce résultat montre que la partie critique de l'architecture Actor-Critic peut aussi être un bon candidat pour expliquer ces données. Toutefois, Q-learning semble la meilleure option pour expliquer cette activité. Il obtient en effet une erreur *LS* inférieure à Actor-Critic (voir les figures 4.6 A, B).

Sarsa d'autre part ne peut reproduire que 10 tests au maximum. Le fait que Sarsa soit moins adapté à la reproduction de l'activité DA au moment du choix, confirme que, dans ce cadre de comportemental particulier, le signal DA enregistré ne code aucune information sur l'action choisie, ce qui est encore une fois conforme aux conclusions initiales de Roesch et al. 2007.

Ainsi, le modèle qui peut à la fois satisfaire tous les tests statistiques et minimiser l'erreur *LS*, résulte d'un mélange de valeur et de RPE calculée par Q-learning. Les paramètres utilisés par le meilleur de ces modèles sont $\alpha = 0.8$, $\gamma = 0.4$ et $w = 0.8$ (voir la figure 4.7 A). Le mélange utilisé par le meilleur modèle contient plus de valeur que d'information de RPE. En outre, le fait que le taux d'apprentissage soit assez élevé indique une convergence rapide de la RPE et de la valeur. Par conséquent, le signal de RPE disparaît rapidement tandis que le signal de valeur devient plus fort rapidement. Cela renforce l'idée que le schéma d'activité observé dans Roesch et al. 2007 semble plus compatible avec un codage de signal de valeur.

Le meilleur modèle Actor-Critic satisfait également tous les tests statistiques, mais contrairement au meilleur modèle Q-learning, il s'appuie davantage sur le signal de RPE. En effet, les paramètres de ce modèle sont : $\alpha = 0.05$, $\gamma = 0.1$ et $w = 0.6$ (voir la figure 4.7 B pour la reproduction du signal DA). Le taux d'apprentissage est très faible par rapport à celui utilisé par le modèle de Q-learning, ce qui empêche le signal de RPE de converger rapidement. Toutefois, de manière cohérente avec les résultats précédents montrant qu'un signal de valeur obtient une erreur *LS* inférieure, le meilleur modèle Actor-Critic obtient une erreur *LS* plus élevée que le meilleur modèle Q-learning basé sur un signal analogue à une fonction valeur.

Une de nos hypothèses était que ce modèle d'activité DA pourrait être reproduit avec une RPE convergeant plus lentement que le comportement observé. Cela permettrait d'éviter que le signal de RPE ne disparaisse trop vite pour reproduire la forte activité DA phasique observée au moment de la récompense dans Roesch et al. 2007. Mais de façon cohérente avec les résultats précédents, aucun signal de RPE pur n'a pu obtenir une faible erreur *LS*, même sans aucune contrainte comportementale. En effet, nous pouvons voir que les modèles avec un très haut w, indiquant une valeur presque pure, obtiennent une erreur *LS* faible par rapport à des mélanges avec un plus faible paramètre w (voir les figures 4.6 E, F, G). De plus, nos résultats montrent qu'un fort w n'est pas incompatible avec un score *ST* élevé (points rouges dans les figures 4.6 AC). Bien que, si nous nous concentrons uniquement sur les modèles de RPE pures ($w = 0$), nous observons que Actor-Critic semble être mieux adapté pour reproduire l'activité DA que Q-learning puisqu'il obtient une erreur inférieure, plus à même de reproduire l'activité DA pour le meilleur score (voir Figure 4.6 D).

Une caractéristique intéressante du modèle Actor-Critic qui pourrait aider à reproduire le signal avec une RPE pure est qu'il calcule la RPE basée sur les valeurs des

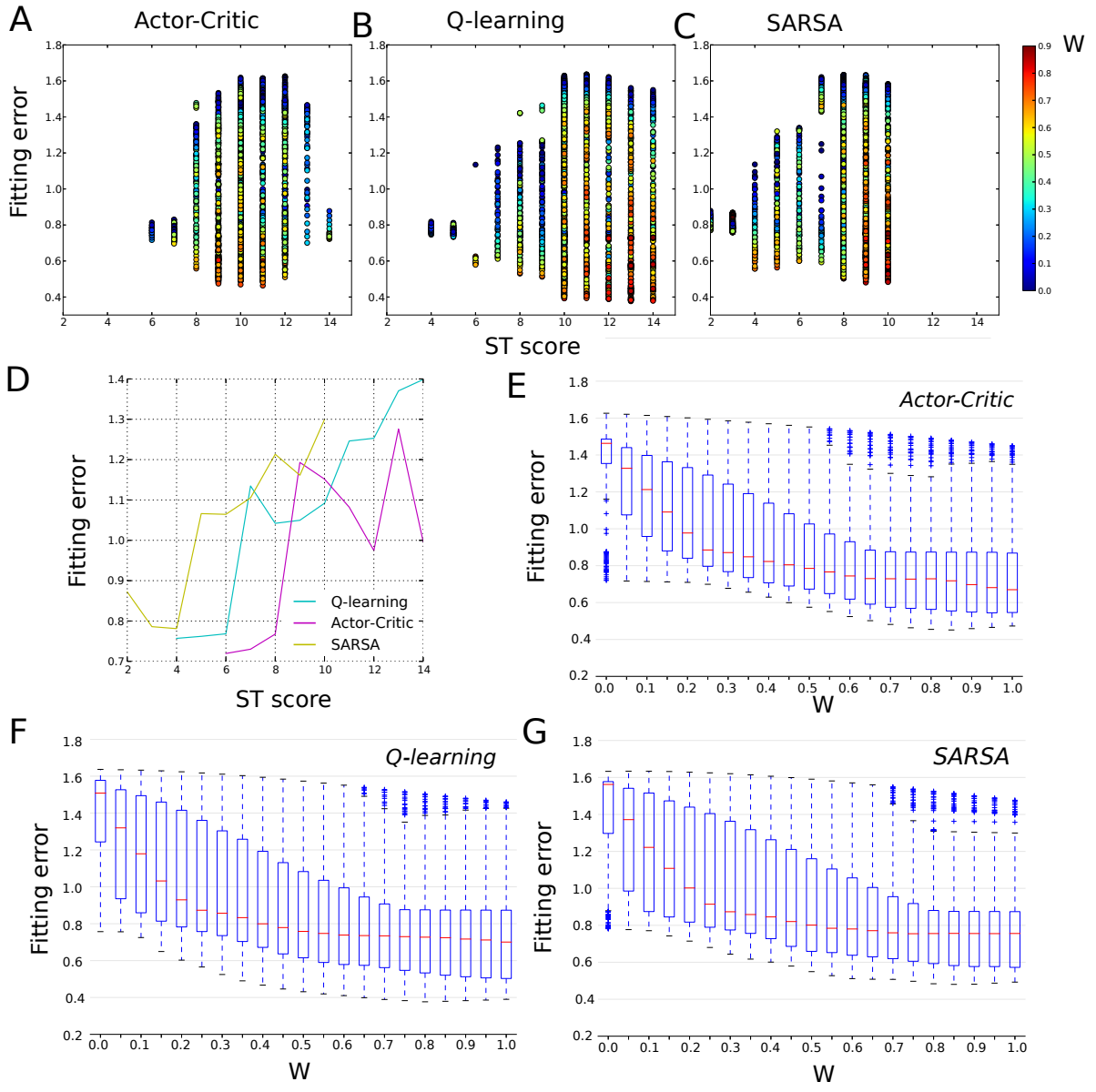

Figure 4.6 – Performances des différents modèles dans la reproduction des tests statistiques et erreurs *LS*. *Chaque figure affiche le score* ST *et l'erreur* LS *de chaque modèle généré par l'un des trois algorithmes étudiés : A.* Actor-Critic*, B.* Q-learning *et C.* Sarsa*. Chaque point de la figure représente un seul modèle. La couleur du point représente le poids, w, de la mixture utilisé pour calculer son score* ST *et son erreur* LS*. Les points bleus représentent des modèles avec un signal de RPE pur (w = 0) et les points rouges représentent les modèles avec 90% valeur (w = 0,9). Les modèles qui ne pouvaient pas reproduire l'évolution de l'activité DA au cours de la livraison et omissions de la récompense ont été exclus. Ainsi aucun modèle de pure valeur n'est affiché dans ces figures. D. Évolution de l'erreur* LS *du modèle RPE pur (w = 0) en fonction du score* ST *pour les trois algorithmes. Seules les erreurs* LS *inférieures ont été rapportées pour chaque score* ST *de chaque algorithme. E-F-G. Évolution de l'erreur* LS *médiane en fonction du paramètre de* mixture*, w, pour respectivement* Actor-Critic*,* Q-learning *et* Sarsa*. L'erreur* LS *a été moyennée pour chaque poids w sur toute les combinaisons testée de paramètres α et γ.*

différents états plutôt que sur des valeurs état-action. Par conséquent, même si le taux d'apprentissage est important, il existe toujours un signal de RPE restant au moment de la récompense. En effet, comme la politique est basée sur le comportement de l'animal, dans les essais libres, la récompense la moins attractive est choisie environ 30% du temps. Comme V représente la prédiction de récompense moyenne future, lorsque la pire (respectivement la meilleure) option est choisie, la RPE est négative (respectivement positive) au moment de la récompense. Comme Q-learning et Sarsa calculent la RPE basée sur des valeurs d'état-action, il n'y a pas de signal restant RPE après convergence de la Q-valeur.

Ensemble, ces résultats confirment que l'information codée par l'activité rapportée par Roesch et al. 2007 ne contient pas d'informations sur l'action à venir. Cette information est soit calculée avec une évaluation de la valeur de l'état actuel – ce qui correspond a une évaluation de la récompense moyenne à laquelle l'animal doit s'attendre, comme prédit par Actor-Critic– soit basée sur une prévision plus optimiste sur la base de la meilleure quantité de récompense que l'animal peut recevoir dans le futur – ce que prévoit Q-learning. Nos résultats conduisent également à une révision importante de l'information codée par les neurones dopaminergiques car ils montrent qu'aucun signal de RPE pur, avec ou sans contrainte comportementale, ne peut reproduire avec précision le motif d'activité de ces neurones DA de VTA. L'hypothèse la plus parcimonieuse explorée ici est que cette information contient également un signal de valeur.

4.4 Discussion

Dans cette étude, nous avons effectué une nouvelle analyse, basée sur des modèles d'apprentissage, de l'activité de neurones dopaminergiques précédemment enregistrés chez des rats dans une tâche multi-choix (Roesch et al. 2007). Notre volonté a été de vérifier les prédictions des auteurs de l'étude originale, selon lesquelles l'algorithme Q-learning permettrait de mieux expliquer cette activité neuronale. Pour ce faire, nous avons comparé la capacité des différents algorithmes d'apprentissage *TD* – Q-learning, Sarsa et Actor-Critic– à reproduire cette activité selon deux critères différents : (1) des tests statistiques comparant l'activité neuronale et l'activité des modèles au moment de la perception de l'odeur (score ST) ; (2) la distance euclidienne entre les deux courbes tracées de ces activités enregistrées à ces différents points dans le temps pendant tout l'essai (erreur LS).

L'hypothèse testée, basée sur l'enregistrements des neurones DA chez les singes passifs (Schultz et al. 1997), est que l'activité des neurones dopaminergiques pourrait refléter une information de type RPE (Bayer et Glimcher 2005 ; Fiorillo et al. 2003 ; Hollerman et Schultz 1998 ; Tanaka et al. 2004), compatible avec le signal de renforcement utilisé dans les algorithmes d'apprentissage *TD* (Doya 2007 ; Schultz et al. 1997 ; Sutton et Barto 1998). Nous avons testé l'hypothèse selon laquelle l'apprentissage dans cette tâche serait dépendant ou contraint par ce signal de renforcement ; ce qui implique que l'adaptation comportementale devrait être liée à des changements dans le signal dopaminergique enregistré. Nous avons également spécifiquement évalué le rôle

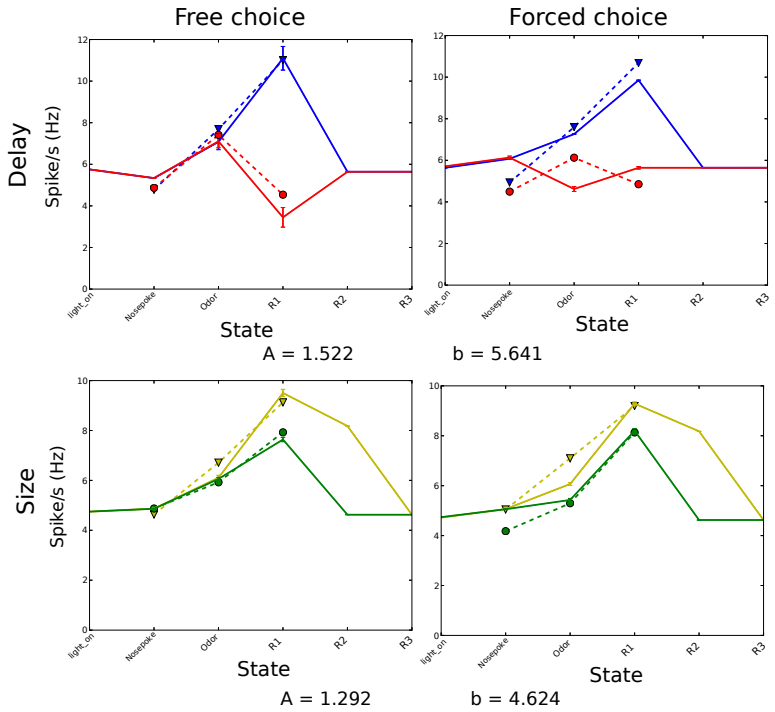

Figure 4.7 – *Reproduction de l'activité DA avec le meilleur modèle simulé obtenue avec* Actor-Critic. *Le score* ST *de ce modèle est 13 et obtient l'erreur* LS *minimale pour un modèle avec ce score. Les paramètres du modèle sont :* $\alpha = 1$, $\gamma = 0.3$ *et* $w = 0.55$.

de l'action future dans l'encodage du signal, ce qui correspondait à la question initiale soulevée dans différentes études (Morris et al. 2006 ; Roesch et al. 2007) : Est-ce que l'algorithme Sarsa – qui code explicitement des informations sur l'action future dans le signal de RPE – permet de mieux reproduire l'activité de la dopamine ? Ou au contraire les prédictions des algorithmes Q-learning ou Actor-Critic – dans lesquels le signal RPE ne porte pas d'informations sur l'action future – correspondent-elles mieux à cette activité ?

Nous avons constaté que (1) contraindre les modèles pour s'adapter au comportement des rats dans cette tâche a donné l'avantage à Q-learning dans la reproduction de l'activité DA, comme prédit par Roesch et al. 2007, (2) toutefois, le score *LS* était alors très faible, ce qui indique que l'activité DA ne peut encoder un signal de RPE pur et que le comportement observé peut ne pas être uniquement formé par l'apprentissage basé sur l'activité DA, (3) l'élimination de la contrainte comportementale produit un bien meilleur ajustement sur l'activité de DA lors de l'utilisation d'une combinaison de RPE et de valeur dans les algorithmes testés, ce qui indique que l'activité pourraient refléter un mélange de ces deux informations, et (4) dans ces conditions, l'algorithme Q-learning reste le meilleur modèle pour reproduire l'activité DA, bien qu'Actor-Critic s'avère également être un bon candidat.

4.4.1 Dissociation entre adaptation comportementale et activité des neurones DA

Dans la théorie de l'apprentissage par renforcement standard, l'apprentissage se fait exclusivement à partir du signal de RPE, ce qui suggère une évolution conjointe entre le comportement et la RPE. Dans la plupart des études antérieures, l'activité DA a été enregistrée avant et après un conditionnement intensif dans des tâches qui n'impliquent pas de choix actifs de la part des animaux (Bayer et Glimcher 2005 ; Fiorillo et al. 2003 ; Matsumoto et Hikosaka 2009 ; Schultz et al. 1997), rendant la comparaison entre la convergence du comportement et le signal DA hasardeuse. Le protocole expérimental utilisé dans Roesch et al. 2007 inclus de fréquentes inversions de la contingence, ce qui a forcé les rats à surveiller constamment leurs choix et à adapter leur comportement lors des changements de blocs. Dans ce contexte, si l'activité DA reflète une RPE et si seule la RPE guide l'apprentissage, alors la RPE calculée dans les simulations contraintes par le comportement devrait mieux reproduire l'activité DA observée.

Ainsi, dans la première partie de notre travail, nous avons paramétré les algorithmes étudiés de manière à reproduire le comportement observé, comme fait couramment dans la littérature (Ito et Doya 2009, 2011 ; Lesaint et al. 2014 ; Takahashi et al. 2011), afin de voir si l'activité DA serait, telle que prédite par les modèles, liée à l'adaptation comportementale.

De façon surprenante, et donc intéressante, le modèle Actor-Critic était incapable de reproduire le changement de comportement observé lors des changements de contingences (voir Figure 4.3 G, H et I), ce qui d'un point de vue théorique était inattendu. Toutefois, cette incapacité, en raison des multiples inversions de contingences utilisées dans cette

tâche, est compatible avec une étude qui a révélé que l'algorithme Actor-Critic est moins adapté à reproduire le comportement des animaux après un tel changement (Lloyd et al. 2012).

Ce résultat semble toutefois en contradiction avec l'interprétation Actor-Critic du fonctionnement des ganglions de la base. Une possible limitation de l'algorithme standard Actor-Critic est qu'il n'y a qu'une seule information de RPE pour mettre à jour à la fois l'acteur et le critique. La différence d'activité DA entre les neurones DA de la *VTA* et de la *SNc*, que nous avons vu dans le Chapitre 1, peut suggérer que différents signaux DA mettent à jour la partie acteur (striatum ventral) et critique (striatum dorsal) des ganglions de la base ; ce qui expliquerait pourquoi l'architecture Actor-Critic biologique pourrait changer son comportement lorsque l'algorithme Actor-Critic standard ne le peut pas. Il serait donc intéressant de regarder la différence entre l'activité des neurones DA de la *VTA* et de la *SNc* pour voir s'ils codent différentes informations et éventuellement différents types de RPE, en fonction de leur localisation anatomique (Matsumoto et Hikosaka 2009). Néanmoins, nous avons également testé la simulation d'un modèle Actor-Critic comportant des paramètres d'apprentissage différents pour l'acteur et le critique – permettant ainsi un degré de liberté supplémentaire par rapport aux autres algorithmes –, mais ce modèle ne pouvait toujours pas obtenir un bon ajustement sur le comportement des rats dans cette tâche (voir Annexe 8), ce qui indique que notre conclusion sur Actor-Critic n'est pas due à une trop forte contrainte sur le taux d'apprentissage utilisé.

Bien que Sarsa et Q-learning peuvent reproduire le comportement des rats, le signal de RPE simulé dans ces conditions n'est pas compatible avec l'activité DA enregistrée. Au lieu de cela, le signal de RPE montre une convergence plus rapide que la réponse phasique DA à la récompense (voir la figure 4 A et C). La fonction de valeur – définie comme la somme de la récompense attendue immédiate plus la récompense future attendue – a généré un signal qui a été quantitativement meilleur à s'adapter à la tendance d'activité enregistrée qu'un signal de RPE pur (voir Figure 4 B et D), mais ne peut pas expliquer l'évolution de l'activité DA au cours de l'omission et la livraison de la récompense.

Ces résultats soulignent de fortes limitations des algorithmes standard d'apprentissage *TD* pour reproduire à la fois l'activité DA et l'adaptation comportementale des rats. Ils suggèrent que l'information portée par l'activité de DA pourrait ne pas être entièrement responsable des changements de comportement, contrairement à ce que supposent les algorithmes d'apprentissage par renforcement.

Cela soulève une question sur le lien entre la convergence conjointe du comportement et de l'information codée par les neurones dopaminergiques. Il serait intéressant d'examiner l'évolution conjointe des deux depuis les premiers essais de l'apprentissage jusqu'à une convergence totale du comportement dans un environnement stable. Cela nous permettrait de voir si l'écart observé ici est dû à l'instabilité introduite par les fréquentes inversions de contingence stimulus-récompense.

Bien qu'en contradiction avec les modèles computationnels, la dissociation apparente entre l'activité DA et l'adaptation comportementale est cohérente avec l'idée que le comportement est le résultat de plusieurs systèmes d'apprentissage parallèles, dont certains

potentiellement indépendants du système DA. Plusieurs études neurobiologiques ont suggéré que le comportement est le résultat d'interactions complexes entre plusieurs systèmes neuronaux parallèles (White et McDonald 2002).

Un point de vue populaire dans la littérature est que les différents circuits corticaux et sous-corticaux contrôlent les comportement habituels vs dirigés vers un but (Balleine et O'Doherty 2010 ; Daw et al. 2005 ; Dollé et al. 2010 ; Keramati et al. 2011 ; Khamassi et Humphries 2012 ; Yin et Knowlton 2006). En accord avec cela, nos résultats suggèrent que le signal DA de la *VTA* peut être une partie d'un système d'apprentissage, et que le comportement résulte d'une interaction entre ce système et d'autres systèmes parallèles. Ces systèmes peuvent dépendre de signaux DA non évalués ici, par exemple les signaux DA de la *SNc*, ou ils peuvent également être indépendants de la DA, tel que suggéré par de récents travaux (Flagel et al. 2010 ; Lesaint et al. 2014).

4.4.2 L'activité DA encode une *mixture* de RPE et valeur

Compte tenu de la dissociation apparente entre le comportement et l'activité DA, nous avons relaché les contraintes comportementales sur les paramètres afin de se concentrer sur le décodage de l'information représentée par l'activité DA. Cela nous a permis de lancer une recherche dans l'espace des paramètres d'une manière plus exhaustive afin de reproduire le signal DA. Nous avons examiné si l'activité DA pouvait refléter une RPE avec une convergence lente, ce qui expliquerait le maintien de la forte réponse phasique DA au moment de la récompense, ce qui indiquerait, dans le cadre de l'hypothèse d'encodage de RPE, que la récompense n'est pas encore totalement prédite à ce stade de l'apprentissage. Nous voulions ainsi tester si seule une fonction de valeur ou un mélange des deux informations permettaient de mieux expliquer cette activité. Nos résultats montrent qu'un mélange de valeur et de RPE calculé par Q-learning est effectivement le meilleur modèle pour reproduire l'activité DA observée dans cette tâche (voir Figure 4.6 et 4.7). Bien qu'en contradiction avec la forme la plus forte de l'hypothèse largement acceptée que l'activité DA reflète un signal RPE pur (Hollerman et Schultz 1998 ; Schultz et al. 1997), ces résultats sont compatibles avec des études rapportant une de rampe d'activité DA lorsque des animaux se rapprochent de la récompense (Howe et al. 2013).

Toutefois, l'hypothèse de RPE n'est pas la seule à être discutée dans la littérature actuelle. Il existe certaines preuves que l'activité des neurones DA pourrait également refléter une *salience* (Berridge 2007 ; Berridge 2012 ; voir Chapitre 2) et dans une certaine mesure, l'information de valeur que nous injectons dans notre signal de mixture peut être comparée à une forme de *salience* ; la tâche utilisant uniquement des récompenses positives. De plus cette hypothèse ne considère pas la dopamine comme le signal de renforcement lié à l'apprentissage, considérant ce signal uniquement comme motivationnel et non comme un signal d'apprentissage. Ainsi, la dissociation entre convergence du signal DA et adaptation comportementalle est naturelle dans le cadre de cette hypothèse.

La tâche implique de nombreuses inversions de la contingence stimulus-récompense, ce

qui pourrait dans une certaine mesure expliquer pourquoi le signal de DA peut encore être fort au moment de la récompense. Pourtant, certaines récompenses sont données dans de nombreux blocs consécutifs et doivent ainsi être prévisibles. En effet, quand on regarde le déroulement d'une session (voir la Figure 4.1 A), dans le puits de droite, du bloc 2 à 4 – soit pendant 90 essais libres consécutifs – au moins une récompense est donnée (pour les récompenses *sort, small* et *big*) et sans incertitude temporelle. Le fait que la réponse phasique des neurones dopaminergiques reste élevée, fait qu'il est impossible de l'expliquer seulement avec l'hypothèse de RPE et suggère la présence de multiples informations telles que la valeur.

On peut également argumenter sur une possible présence de deux sous-populations distinctes de neurones dopaminergiques enregistrées avec des schémas distincts d'activité. Cela pourraient expliquer nos résultats selon lesquels l'activité DA ne peut être reproduite que par un mélange de valeur et de RPE. Or, les 17 neurones VTA DA utilisés ici ont été choisis pour leurs propriétés de codage de RPE : (1) ils sont sensibles à la récompense et au signaux prédisant la récompense ; (2) leur activité tend à diminuer à mesure que la récompense devient prévisible ; (3) leur réponse est d'abord inhibée par une omission de la récompense et tend à revenir à l'état initial lorsque l'omission devient prévisible. Bien que ces neurones puissent être séparés en deux groupes différents de neurones, dont l'un est plus sensible à des signaux de prédiction de la récompense et l'autre à la récompense elle-même, les deux catégories présentent en moyenne une activité similaire et sont toutes deux mieux expliquées par un mélange de RPE et la valeur.

Des études récentes de Fiorillo et ses collègues (Fiorillo et al. 2013b) ont révélé que le signal DA peut avoir une dynamique temporelle plus complexe que précédemment supposé. En effet, ils ont constaté que l'activité phasique des neurones dopaminergiques peut être décomposée en trois phases distinctes, chacune reflétant différents types d'informations. Au cours de la première phase – entre 40 et 120 ms après le stimulus – l'activité des neurones reflète l'intensité sensorielle. Pendant la deuxième phase – entre 150 à 250 ms – le signal rapporte une valeur motivationnelle et la troisième phase a été interprétée comme une phase de rebond due à l'activation ou l'inhibition précédente des neurones. Par conséquent, on peut se demander si la présence d'un signal de *mixture* observé dans notre étude pourrait être considéré comme une conséquence de la propriété multiphasique du signal DA observé dans ces études. Cependant, nous n'avons pas observé ces différentes phases dans le signal DA enregistré dans Roesch et al. 2007. Une explication possible de l'absence du signal phasique multiple dans nos données est que les stimuli utilisés dans la tâche ont une intensité sensorielle modérée. Or, Fiorillo et collègues (Fiorillo et al. 2013b) ont trouvé que les stimuli d'intensité sensorielle peu élevée ne déclenchent pas une réponse phasique pendant la première phase. Ainsi nous pouvons considérer que l'activation éventuelle de la première phase des neurones dopaminergiques est négligeable dans cette étude et que le signal de mélange n'est pas la conséquence de la séquence triphasique de l'activité DA.

Une question qui reste en suspend est de savoir si ces deux types d'informations seraient encore présents après la convergence totale du comportement des rats. En effet, la tâche que nous avons utilisé pour cette comparaison ne semble pas laisser suffisamment de temps pour que les rats apprennent pleinement les différentes contingences. Ainsi, nous

ne savons pas exactement comment le signal DA finirait par converger si l'environnement était stable. Nos résultats peuvent faire plusieurs types de prédictions quant à l'évolution de l'activité DA, et notamment après convergence complète, en fonction de la façon dont nous interprétons la fonction de *mixture* utilisée ici. À cet égard, il est important de noter que la *mixture* peut être interprétée soit comme un signal de RPE ayant une partie valeur plus forte ($r_{t+1} + V(s_{t+1})$ renforcée), soit une partie de prédiction plus faible ($V(s_t)$ diminuée). Si le signal de valeur est sur-représenté dans les neurones DA, alors même après un entraînement intensif sur la tâche, les neurones DA devraient encore répondre à la récompense attendue. Alternativement, si le signal de prédiction est tout simplement plus faible dans les premiers essais ou dans un environnement incertain, après un entraînement intensif sur la tâche, le signal de prédiction augmenterait jusqu'à ce que le signal global de DA se rapproche d'un signal de RPE pur. Ces neurones devraient alors cesser de répondre à la récompense prédite, ce qui suggère que la partie de valeur et de prédiction du signal de RPE ne sont pas apprises à la même vitesse.

De plus, comme discuté dans la partie précédente, le codage de RPE pur n'implique pas forcément une convergence totale du signal au moment de la récompense. En effet, l'algorithme Actor-Critic peut toujours obtenir une RPE non nulle au moment de la récompense, même après un apprentissage approfondi sur la tâche. Pour obtenir un signal nul de RPE au moment de la récompense, l'estimation de la valeur des algorithmes Actor-Critic doit être égale à la valeur de la meilleure option et l'acteur doit toujours choisir cette option, ce qui n'était pas le cas ici, puisque les rats n'ont jamais atteint une phase de pure exploitation et choisissent encore régulièrement d'explorer le choix associé à une récompense moins attractive.

4.4.3 Les neurones DA de VTA ne reflètent pas le choix futur de l'animal

Nous avons également étudié la dépendance de l'activité aux actions futures. Sarsa, qui calcule sa prédiction de récompense future basée sur l'action future, ne peut pas reproduire l'activité, indiquant que le choix futur de l'animal n'est pas pris en considération par l'information DA. Ainsi nos résultats confirment que l'information codée par les neurones dopaminergiques ne reflète pas l'action future de l'animal dans cette tâche.

Q-learning était le meilleur modèle pour reproduire cette activité, cependant, nos résultats montrent également qu'Actor-Critic est capable de la reproduire. Ces résultats suggèrent que l'information codée par les neurones DA de la *VTA* est basée uniquement sur l'état actuel, soit par l'évaluation des $V(s)$ – c'est-à-dire une moyenne de la récompense attendue, comme prédit par Actor-Critic–, soit par une évaluation optimiste de la meilleure récompense qui peut être obtenue, comme prédit par Q-learning.

Cependant, au cours du processus d'apprentissage, $V(s)$ converge vers la valeur de la meilleure option, car le comportement devient optimal, en enlevant la différence entre Actor-Critic et Q-learning après convergence. Le fait que Morris et al. 2006 aient trouvé une activité dépendante du choix de l'animal chez le singe peut s'expliquer par des différences dans la tâche, les espèces utilisées, ou même dans la localisation de

l'enregistrement (Daw 2007 ; voir Chapitre 1). Dans Roesch et al. 2007, les neurones dopaminergiques ont été enregistrés dans la *VTA* alors que dans Morris et al. 2006, les neurones dopaminergiques ont été enregistrés dans la *SNc*. Les neurones DA de la *VTA* projettent préférentiellement vers la partie ventrale du striatum (Haber 2003 ; Haber et Calzavara 2009 ; Haber et al. 2000 ; Joel et Weiner 2000), qui est considérée comme la partie critique de l'architecture Actor-Critic des ganglions de la base (Joel et al. 2002). Nos résultats montrent que l'information codée par les neurones DA de la *VTA* sont compatibles avec un signal construit pour mettre à jour la valeur sur la base de l'état courant, tel que dans la partie critique de l'architecture Actor-Critic. La politique, d'autre part, est supposée être calculée dans une partie plus dorsale du striatum (Joel et al. 2002 ; Yin et al. 2005) qui reçoit en majorité les signaux DA de la *SNc*. Ainsi, l'activité des neurones DA de la *SNc* pourrait encoder une RPE responsable de mettre à jour la partie acteur des ganglions de la base.

Des études récentes montrent également la pertinence de déterminer précisément la localisation des neurones dopaminergiques afin d'être en mesure d'interpréter les informations qu'ils véhiculent (Matsumoto et Hikosaka 2009). En effet, les neurones DA projettent vers de nombreuses régions corticales et sous-corticales telles que l'amygdale et le cortex préfrontal médian. Il est très probable qu'en fonction de la zone ciblée, les neurones dopaminergiques encodent des informations différentes et aient des rôles différents dans l'apprentissage par la punition et la récompense comme suggéré dans Lammel et al. 2012.

4.4.4 Conclusion

En résumé, notre travail montre la limitation des algorithmes standard d'apprentissage *TD* à reproduire à la fois l'activité DA et l'adaptation comportementale. Il interroge également l'hypothèse de RPE, aujourd'hui communément admise dans la littérature, en montrant que l'information codée par les neurones dopaminergiques dans une tâche multi-choix ne peut pas être reproduite par un signal de RPE pur, même sans aucune contrainte comportementale. En outre, nous avons montré que l'activité phasique des neurones DA est mieux reproduite par un mélange de RPE et de valeur. Notre travail montre également que si l'algorithme Q-learning est le meilleur modèle pour reproduire l'activité DA enregistrée par Roesch et al. 2007, Actor-Critic ne peut être exclu de façon certaine avec nos données et est toujours un bon candidat pour expliquer l'apprentissage de la valeur dans le circuit ventral des ganglions de la base (Joel et al. 2002 ; Khamassi et al. 2005).

Tout en questionnant la validité de l'hypothèse de RPE, notre étude ne la contredit pas fortement. Il y a en effet de plus en plus de preuves indiquant que l'activité des neurones dopaminergiques ne code pas un signal unique, mais semble montrer une hétérogénéité de réponses aux stimuli aversifs (Brischoux et al. 2009 ; Matsumoto et Hikosaka 2009 ; Schultz et Romo 1987 ; Wang et Tsien 2011), même s'il n'y a toujours pas de consensus sur la question (Fiorillo et al. 2013a). Ces signaux différents peuvent faire partie de circuits différents en fonction de leur localisation anatomique ; comme suggéré par les travaux de Lammel et al. 2011, 2012 montrant la présence de différentes

voies liées à la récompense ou à la punition, impliquant différentes sous-populations de neurones dopaminergiques.

Chapitre 5

rBCBG : Un modèle réduit du BCBG pour la sélection de l'action

Sommaire

5.1 Introduction

5.1.1 Résumé des chapitres précédents

Les ganglions de la base sont un ensemble de noyaux sous-sorticaux interconnectés étudiés depuis près d'un siècle pour leur relation avec le contrôle moteur (Mink et Thach 1991 ; Wilson 1928). Depuis une dizaine d'années avec notamment les travaux de Mink

1996 et de Redgrave et al. 1999b, cet ensemble de noyaux sont vus dans la littérature comme le substrat neural de la sélection de l'action (voir Chapitre 3).

Les ganglions de la base reçoivent, via le striatum et le noyau subthalamique, des projections de nombreuses régions corticales, des régions frontales du cortex (orbito frontal, cortex préfrontal ventro-médial), aux parties associatives (cortex préfrontal dorso-latéral/dorso-médiale), jusqu'aux parties motrices (Haber 2003 ; Haber et al. 2000 ; Joel et Weiner 2000). Les ganglions de la base intègrent donc un grand nombre d'informations et sont l'une des cibles majeures des neurones dopaminergiques de l'aire tegmentale ventrale (VTA) et de la substance noire pars compacta (SNc). Ainsi, ils ont accès aux informations nécessaires à la sélection de l'action via les différentes entrées corticales et sont guidés par un apprentissage principalement influencé par les signaux de renforcement dopaminergique (voir Chapitre 2).

De nombreux modèles des ganglions de la base ont été développés et étudiés pour tester leur capacité de sélection de l'action (Albada et al. 2009 ; Berns et Sejnowski 1998 ; Frank et al. 2004 ; Girard et al. 2008 ; Gurney et al. 2001b ; Humphries et al. 2012 ; Leblois et al. 2006 ; Liénard et Girard 2014 ; voir Chapitre 3). Ces études ont permis de valider l'hypothèse de sélection de l'action avancée par Mink 1996 et Redgrave et al. 1999b. Cependant, bien que fondés sur un sous-ensemble de données biologiques, la plupart des modèles négligent une grande partie des données anatomiques et physiologiques, et ont donc fixé arbitrairement un certain nombre de paramètres de connexion entre noyaux modélisés des ganglions de la base de façon à obtenir de bonnes capacités de sélection de l'action.

Par opposition, le modèle *Biologically Constrained Basal Ganglia*(*BCBG*) développé par Liénard et Girard 2014 a été construit de façon à satisfaire un maximum de propriétés biologiques connues des ganglions de la base. En effet, les poids de connexion entre les différents noyaux sont déduits de propriétés telles que le nombre de neurones présents dans les différents noyaux, les longueurs et diamètres moyens des dendrites des neurones, les proportions de neurones projetant d'un noyau vers un autre ou encore l'activité maximale de chaque noyau. Si certains paramètres ont été fixés par la littérature sur l'anatomie des ganglions de la base, d'autres, par un manque de données ont été optimisés afin de reproduire au mieux les taux de décharge des différents noyaux observés électrophysiologiquement. Par construction, le modèle *BCBG* comprend donc un grand nombre de connaissances sur l'anatomie et la physiologie des ganglions de la base de la littérature actuelle.

De façon marquante, Liénard et Girard 2014 ont pu montrer que les capacités de sélection de l'action de leur modèle émergent, non pas d'une paramétrisation *ad hoc* du modèle, mais d'une optimisation des paramètres de connexions sur des critères de plausibilité anatomique et électrophysiologique. Cependant cette fidélité biologique entraîne également une certaine complexité computationnelle.

5.1.2 Objectif scientifique

Dans ce chapitre nous nous intéressons à la construction et à l'étude d'un modèle réduit du modèle *BCBG*. Nous comparons ses capacités de sélection à deux autres modèles

de la littérature : le modèle Gurney, Prescott, Redgrave (GPR ; Gurney et al. 2001b ; Humphries et al. 2012) et le modèle *Contracting Basal Ganglia* (CBG ; Girard et al. 2008) pour vérifier que le modèle réduit ne perd pas ses propriétés fonctionnelles.

En plus de sa capacité à produire des taux de décharge plausibles au repos, le modèle *BCBG* se démarque de la plupart des autres modèles des ganglions de la base de la littérature de par sa connectivité. L'anatomie des ganglions de la base est encore largement discutée dans la littérature et notre connaissance de la connectivité entre les différents noyaux est encore soumise à débat. Une hypothèse est principalement questionnée dans le modèle *BCBG* : la séparation des neurones du striatum (*MSN*) en chemins direct et indirect (Alexander et Crutcher 1990), en particulier chez le macaque. Cette hypothèse suggère que les *MSN* du striatum ayant des récepteurs à la dopamine de type D1 (*MSN D1*) projettent vers le *GPi* alors que les *MSN* ayant des récepteurs à la dopamine de type D2 (*MSN D2*) projettent vers le *GPe*. La plupart des modèles actuels font l'hypothèse de la présence, au moins partielle[1], de cette ségrégation (Albada et Robinson 2009 ; Albada et al. 2009 ; Frank et al. 2004 ; Girard et al. 2008 ; Gurney et al. 2001b).

Le questionnement de cette hypothèse repose sur différentes études qui remettent en question cette vision des ganglions de la base à la fois chez le primate (Lévesque et Parent 2005 ; Nadjar et al. 2006) et chez le rat (Calabresi et al. 2014).

Ainsi, contrairement à la majorité des modèles actuels de la littérature, le modèle *BCBG* fait l'hypothèse de l'absence de ségrégation des neurones *MSN D1* et *MSN D2* en chemin direct et indirect en accord avec ces résultats et considère les *MSN* comme une population de neurones uniformes.

L'absence de ségrégation des *MSN* devrait avoir un impact sur l'effet de la dopamine sur le système. Aussi le premier objectif de notre travail est d'étudier cet impact et en particulier de comprendre l'influence de différents niveaux de dopamine tonique sur les capacités de sélection de l'action des différents modèles. Le travail de Humphries et al. 2012 a montré que le niveau tonique de dopamine dans un modèle *GPR* des ganglions de la base peut être interprété comme un facteur gérant le compromis entre exploration et exploitation à l'image du paramètre de température utilisé par le *softMax* des algorithmes d'apprentissage par renforcement (Sutton et Barto 1998 ; voir Chapitre 2). Un niveau élevé de dopamine tonique augmente la probabilité que le système choisisse la meilleure action et un faible niveau promeut l'exploration. Le modèle utilisé dans cette étude, basé sur le modèle *GPR*, fait l'hypothèse d'une ségrégation importante entre chemin direct et indirect. Nous testerons ces prédictions avec notre modèle ne faisant pas cette hypothèse et nous testerons également si la présence partielle du chemin direct peut permettre de retrouver des effets similaires de la dopamine sur la sélection.

Nous nous intéresserons également dans ce chapitre aux capacités du modèle réduit à garder les propriétés du modèle original, telles que la reproduction fidèle des taux de décharge, les capacités de sélection de l'action ou encore de l'apparition d'oscillations lors de la modélisation d'une diminution du niveau de dopamine tonique. Nous comparerons de plus ses capacités de sélection avec le *GPR* et le *CBG*.

Nous montrerons que ces différents modèles des ganglions de la base se comportent

1. Plusieurs modèles considèrent toutefois une connexion entre les *MSN D1* et le *GPe*

différemment les uns des autres dans des tâches de sélection de l'action. De plus, selon le modèle utilisé, l'impact de différents niveaux de dopamine tonique produit différents effets sur la sélection, ce qui génère donc des prédictions expérimentales susceptibles de départager ces modèles. Nous verrons également que la réduction de l'intégration temporelle du signal synaptique du modèle *BCBG* permet de garder la plupart des propriétés du modèle complet telle que l'apparition d'oscillations β lorsque le niveau dopaminergique est trop faible.

5.2 Méthode

Le modèle utilisé dans cette étude est basé sur le modèle des ganglions de la base du primates *BCBG* (*Biologically Constrained Basal Ganglia*) développé par Jean Liénard au cours de sa thèse (Liénard et Girard 2014 ; Liénard 2013) et introduit dans le Chapitre 3 (se référer à ce chapitre pour la connectivité détaillée du modèle ainsi que celle des modèles *GPR* et *CBG*). Nous verrons dans cette partie comment le modèle computationnel intègre de nombreuses propriétés biologiques des ganglions de la base. De plus, nous montrerons comment nous avons proposé de réduire la dynamique d'intégration du signal post-synaptique afin d'obtenir un nouveau modèle : le *rBCBG*. Dans ce modèle réduit nous introduisons également une dissociation des neurones épineux moyens du striatum selon le récepteur à la dopamine qu'ils portent, ce qui n'est pas le cas dans le modèle original.

Dans le but de comparer les capacités de sélection de l'action du modèle *rBCBG* aux modèles *GPR* et *CBG* introduits dans le Chapitre 3, nous avons implémenté et soumis ces modèles à plusieurs tests de sélection de l'action qui sont également présentés dans cette partie.

5.2.1 *BCBG* : le modèle computationel

Le modèle prend en compte un grand nombre de critères biologiques et a donc une complexité computationnelle assez importante. Nous allons dans cette section le décrire ainsi que la réduction que nous avons effectuée.

Le *BCBG* est un modèle de neurones à champ moyen (Deco et al. 2008). L'hypothèse de ces modèles est que tous les neurones d'une population reçoivent le même signal d'entrée. Ainsi une population est modélisée par un unique neurone artificiel dont l'activité est représentative de l'activité moyenne de la population. Le nombre moyen de potentiels d'actions pré-synaptiques émis par des neurones émettant un neurotransmetteur n de la population y projetant vers la population x, ψ_x^n est :

$$\psi_x^n(t) = \nu_{x \leftarrow y} \phi_y(t - \tau_{y \rightarrow x})$$

Avec $\nu_{x \rightarrow y}$ le nombre moyen de synapses d'un neurone de la population x sur la population y, $\tau_{y \rightarrow x}$ le délai axonal entre les deux populations et $\phi_k(t - \tau_{y \rightarrow x})$ l'activité de la population y au temps $t - \tau_{y \rightarrow x}$.

Le calcul du nombre moyen de synapses dans la population x recevant un signal de la population y est déduit du nombre de neurones dans chaque population N_x et N_y, de la proportion de neurones d'une population y avec un axone visant les neurones d'une population x (noté $\mathcal{P}_{y\to x}$; voir Figure 5.1) et le nombre de variscosités de ces neurones $\alpha_{y\to x}$. Ainsi :

$$\nu_{x\leftarrow y} = \frac{\mathcal{P}_{y\to x}\mathcal{N}_y}{\mathcal{N}_x}.\alpha_{y\to x}$$

L'activité moyenne d'une population de neurones est :

$$\phi_x(t) = \frac{S_x^{MAX}}{1 + exp(\frac{\theta_x - \Delta V_x(t)}{\sigma'})}$$

Où $\Delta V_x(t)$ représente le potentiel d'entrée moyen au soma au temps t, S_x^{Max} le taux de décharge maximal. θ_x est la différence moyenne entre le taux au repos et le seuil d'activité. $\sigma' = \sigma\frac{\sqrt{3}}{\pi}$, avec σ l'écart type du taux de décharge (Tsirogiannis et al. 2010).

La génération d'un potentiel d'action post-synaptique par un neurotransmetteur n dû à l'arrivée d'un potentiel d'action sur une synapse donnée se fait au cours du temps par :

$$V_0^n(t) = \frac{A_n}{D_n}e^{-\frac{D_n}{t}}$$

où A_n représente l'amplitude et D_n la demie vie.

Le modèle prend également en compte l'atténuation du signal en fonction de la distance au soma en modélisant les dendrites comme un câble. On peut ainsi exprimer pour une population x :

$$V_{soma}^n(t) = V_0^n(t)\frac{cosh(L_x - Z_x)}{cosh(L_x)}$$

avec, L_x la constante électronique du neurones et Z_x la distance moyenne des récepteurs synaptique le long de la dendrite qui est exprimée comme un pourcentage de la longueur dendritique maximale :

$$Z_{x\leftarrow y} = p_{y\to x}L_x$$

avec,

$$L_x = l_x\sqrt{\frac{4}{dx}\frac{R_i}{R_m}}$$

où $p_{y\to x}$ le pourcentage de contact entre la population x et y, R_i est la résistance intracellulaire, R_m la résistance de la membrane et d_x le diamètre moyen des axones.

Par commodité, nous poserons par la suite : $C(x,y) = \frac{cosh(L_x - Z_x)}{cosh(L_x)}$.

Ainsi, on peut finalement exprimer l'évolution de l'activité d'une population x par :

$$\Delta V_x(t) = \sum_{(y,n)} \psi_x^n(t)V_0^n(t)C(x,y)$$

Paramètres	Unités	Symboles	Valeurs
Activité au repos fixe du modèle	(Hz)		
CSN		ϕ_{CSN}	$2Hz$
PTN		ϕ_{PTN}	$15Hz$
CMPf		ϕ_{CMPf}	$4Hz$
Activité au repos désirée du modèle	(Hz)		
MSN		ϕ_{MSN}	$0.5Hz \pm 0.5Hz$
FSI		ϕ_{FSI}	$10Hz \pm 10Hz$
STN		ϕ_{STN}	$19Hz \pm 3.8Hz$
GPe		ϕ_{GPe}	$65Hz \pm 9.4Hz$
GPi/SNr		ϕ_{GPi}	$69Hz \pm 10.2Hz$
Propriété neurale	(mV)		
Threshold spread		σ	3.8
Amplitudes PSP	(mV)		
AMPA		A_{AMPA}	1
$GABA_A$		A_{GABA_A}	0.25
NMDA		A_{NMDA}	0.025
Demi-vie des PSP	(ms)		
AMPA		D_{AMPA}	5
$GABA_A$		D_{GABA_A}	5
NMDA		D_{NMDA}	100
Délais axonaux	(ms)		
$Ctx \rightarrow Str$		$\tau_{Ctx \rightarrow Str}$	7
$Ctx \rightarrow STN$		$\tau_{Ctx \rightarrow STN}$	3
$Str \rightarrow GPe$		$\tau_{Str \rightarrow GPe}$	7
$Str \rightarrow GPi$		$\tau_{Str \rightarrow GPi}$	11
$STN \rightarrow GPe$		$\tau_{STN \rightarrow GPe}$	3
$STN \rightarrow GPi$		$\tau_{STN \rightarrow GPi}$	3
$GPe \rightarrow STN$		$\tau_{GPe \rightarrow STN}$	10
$GPe \rightarrow GPi$		$\tau_{GPe \rightarrow GPi}$	3

Table 5.1 – *Paramètres utilisés par le modèle* BCBG *fixés a priori d'après les données de la littérature. L'activité des noyaux CSN, PTN et CMPf est constante dans le modèle. L'activité des CSN reflète la salience des différents canaux dans les tâches de sélections. L'activité des autres noyaux du modèle ne dépend que de la dynamique du système. CSN : neurones cortico-striataux ; PTN : neurones du faisceau pyramidal ; CMPf : noyau centro-médian parafascicular du thalamus ; FSI : interneurones à taux de décharge rapide.*

Paramètres	Unités	Min	Max	Solution utilisée
Poids de connexions	$(\mu V.s)$			
$f_{CSN \to MSN}$		1132.07	1465.67	1374.84
$f_{CSN \to FSI}$		777.39	1482.94	780.05
$f_{PTN \to STN}$		1050.69	1116.78	1050.69
$f_{MSN \to GPe}$		12814.49	15504.70	13914.55
$f_{MSN \to GPi}$		14075.90	26790.20	16668.73
$f_{STN \to GPe}$		331.26	591.66	591.66
$f_{STN \to GPi}$		180.80	298.79	232.01
$f_{STN \to MSN}$		0.00	0.40	0.00
$f_{STN \to FSI}$		0.00	11.36	8.54
$f_{GPe \to STN}$		39.60	50.73	39.96
$f_{GPe \to GPi}$		22.33	29.91	23.24
$f_{GPe \to MSN}$		0.00	0.15	0.00
$f_{GPe \to FSI}$		7.18	35.52	14.14
$f_{GPe \to GPe}$		46.26	46.88	46.26
$f_{FSI \to MSN}$		59.38	118.98	104.25
$f_{FSI \to FSI}$		28.22	118.05	115.19
$f_{MSN \to MSN}$		146.04	456.01	16.04
$f_{CMPf \to MSN}$		29.55	93.16	93.16
$f_{CMPf \to FSI}$		626.32	1449.21	1027.01
$f_{CMPf \to STN}$		316.55	424.82	424.82
$f_{CMPf \to GPe}$		66.52	300.00	203.57
$f_{CMPf \to GPi}$		177.51	299.95	239.88
$f_{PTN \to MSN}$		20.07	37.56	20.07
$f_{PTN \to FSI}$		14.97	163.48	16.66

Table 5.2 – *Poids de connexions extrêmes des différents modèles* BCBG *reportés dans* Liénard *et* Girard *2014 ainsi que poids de connexions du modèle utilisés dans les résultats présentés, sauf mention du contraire.*

On a donc :

$$\Delta V_x(t) = \sum_{(y,n)} \nu_{x \leftarrow y} \phi_y(t - \tau_{y \to x}) V_0^n(t) C(x,y)$$

$\phi_y(t - \tau_{y \to x})$ représentant l'activité de y projetant vers x. Ainsi le poids de connexion entre deux populations de neurones, $w_{x \to y}$ est défini par :

$$\begin{aligned} w_{x \to y} &= \nu_{x \leftarrow y} V_0^n(t) C(x,y) \\ &= (\frac{\mathcal{P}_{y \to x} \mathcal{N}_y}{\mathcal{N}_x} . \alpha_{y \to x})(A_n D_n e^{-D_n t}) C(x,y) \end{aligned}$$

La transmission d'un signal entre deux populations n'est donc pas instantanée dans le modèle et dépend du type de neurotransmetteur relâché dans la synapse. Ainsi le changement de potentiel de la population post-synaptique dépend non seulement de l'activité de la population pré-synaptique et des forces de connexions entre les deux populations, mais également du délai de transmission du signal. Ainsi, si la transmission via les récepteurs *AMPA* est assez rapide, la transmission est bien plus lente avec les récepteurs *NMDA*. Aussi, si cette propriété permet d'atteindre une plausibilité plus grande de la dynamique du système, elle complexifie l'étude du modèle et de l'évolution possible des poids synaptiques au cours de l'apprentissage.

Afin de passer outre cette complexité nous avons négligé ces délais de transmission synaptique en considérant la transmission comme instantanée et négligeant les différences de transmission dues au type de neurotransmetteur. Ainsi, les poids de connexions synaptiques du modèle simplifié du *BCBG* sont définis par :

$$w_{x \to y} = \int_0^\infty \nu_{y \leftarrow x} V_{soma}^n(t) dt = \nu_{y \leftarrow x} C(x,y) A_n \int_0^\infty \frac{t}{D_n} e^{-\frac{t}{D_n}} dt$$

Or,

$$\int_0^\infty \frac{t}{D_n} e^{-\frac{t}{D_n}} dt = -\int_0^\infty (-\frac{t}{D_n}) e^{-\frac{t}{D_n}} dt = -[[t e^{-\frac{t}{D_n}}]_0^\infty - \int_0^\infty e^{-\frac{t}{D_n}} dt]$$

De plus $[t e^{-\frac{t}{D_n}}]_0^\infty = 0$, d'où :

$$\int_0^\infty \frac{t}{D_n} e^{-D_n t} dt = \int_0^\infty e^{-\frac{t}{D_n}} dt = D_n$$

On obtient ainsi :

$$w_{x \to y} = \nu_{x \leftarrow y} C(x,y) A_n D_n = \frac{\mathcal{P}_{y \to x} \mathcal{N}_y}{\mathcal{N}_x} . \alpha_{y \to x} C(x,y) A_n D_n$$

Les poids de connexions du modèle simplifié sont donc définis par les poids $w_{x \to y}$, obtenus avec les différentes paramétrisations des modèles optimisés.

Version simplifiée :

$$\begin{cases} \Delta V_x(t) = \sum_{(y,n)} w_{x \to y} \phi_y^n(t) \\ a(t+dt) = a(t) + \frac{dt}{\tau}(\Delta V_y - a(t)) \\ \phi_x(t) = \frac{S_x^{MAX}}{1+exp(\frac{\theta_x - a(t)}{\sigma'})} \end{cases}$$

Le modèle réduit du *BCBG* permet donc une expression computationelle simple et compacte dans le formalisme standard des intégrateurs à fuite utilisé, entre autre, dans le modèle *GPR*. De plus, via la paramétrisation des poids de connexions directement dérivée du *BCBG*, il garde une plausibilité biologique forte. Cependant, la réduction fait perdre au modèle une finesse au niveau de la différence d'intégration temporelle du signal par les récepteurs AMPA, NMDA et GABA. Cela peut entraîner des différences dans la dynamique du système. Nous étudierons dans ce chapitre le possible impact de cette réduction et montrerons qu'elle affecte peu les propriétés du modèle en terme de dynamique de sélection de l'action, ou dans ses régimes oscillatoires permettant de reproduire des troubles liés à la maladie de Parkinson.

Modélisation de l'effet de la dopamine

Afin de modéliser l'effet de la dopamine tonique dans le système, nous avons ajouté le niveau de dopamine λ, qui est nul lorsque le modèle n'est sous aucun dérèglement dopaminergique. Chaque population de neurones est donc modulée par la dopamine en fonction de leurs types de récepteurs dopaminergique (voir Figure 5.1). Une population x ayant des récepteurs D1 (respectivement D2) est modulée positivement (respectivement négativement) par la dopamine, aussi nous multiplions $\Delta V_x(t)$ par $(1+\lambda)$ (respectivement $(1-\lambda)$).

Cette modélisation suppose que la dopamine affecte les différents noyau avec la même force.

5.2.2 Tests de sélection de l'action

Nous avons soumis les modèles à différents tests de sélection de l'action afin de quantifier leur capacité à sélectionner une option lorsque plusieurs sont en compétition. Pour cela nous avons implémenté différents tests de sélection de l'action impliquant 2 ou plusieurs options. Comme dans de nombreuses études, chaque noyau du modèle a été subdivisé en canaux parallèles (voir Figure 3.15). Chaque canal représente une option/action dans le test de sélection. La capacité des modèles à sélectionner une action est observée au niveau de l'activité du noyau de sortie *GPi/SNr*.

Efficacité de la sélection

Nous avons cherché à mesurer l'efficacité de la sélection de l'action des modèles et à en observer leur limite dans un contexte où deux actions sont en compétition, reproduisant en cela les tests utilisés dans Prescott et al. 2006 et Girard et al. 2008. Cinquante saliences pour chaque action sont testées : entre 0 et 1, par pas de 0.02 pour les modèles

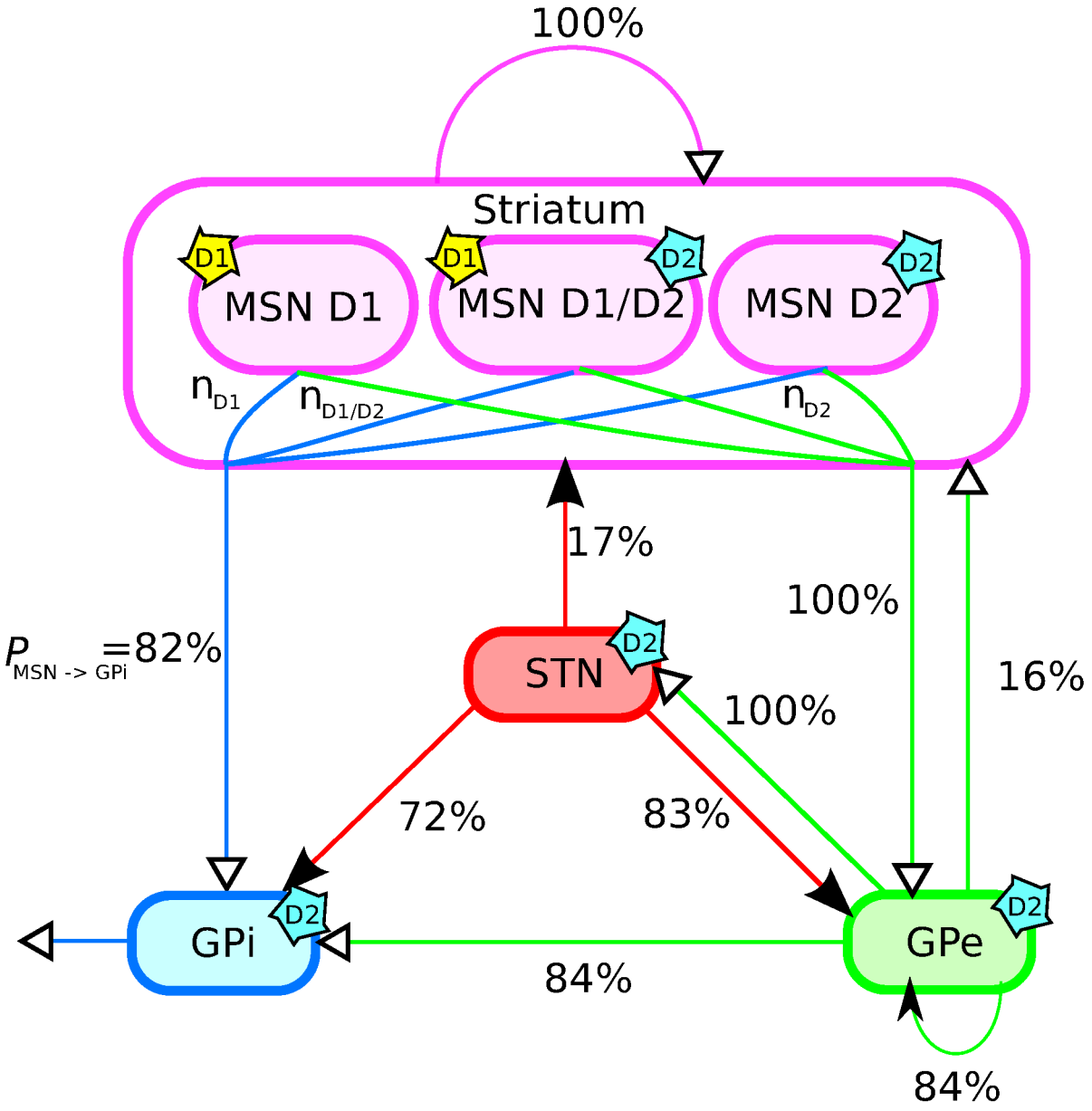

Figure 5.1 – *Illustration de la connectivité des ganglions de la base. A coté de chaque connexion est notée la proportion de neurones de la population projetant vers la population cible (noté $\mathcal{P}_{x\to y}$ dans le modèle computationel). Les types de récepteurs à la dopamine sont indiqués par une étoile. La proportion de neurones* MSN *projetant vers le* GPi, $p_{MSN\to GPi}$, *est de 82%. Les* MSN *projetant vers* GPi *expriment différents types de récepteurs à la dopamine : D1, D2 ou les deux. Ainsi $\mathcal{P}_{MSN\to GPi}$ peut se décomposer en nombre de* MSN D1 *projetant vers* GPi, *noté n_{D1} sur la figure, de* MSN D2, *noté n_{D2} et* MSN D1/D2, *noté n_{D2}. On a donc $\mathcal{P}_{MSN\to GPi} = p_{MSND1\to GPi} + p_{MSND2\to GPi} + p_{MSND1/D2\to GPi} = 82\%$; $p_{MSND1\to GPi}$, $p_{MSND2\to GPi}$ et $p_{MSND1/D2\to GPi}$ n'étant pas connus.*

GPR et *CBG* et entre 0 et 20Hz [2], par pas de 0.4Hz pour le modèle *BCBG* (la différence est due au fait que les modèles *GPR* et *CBG* ont des taux de décharges compris entre 0 et 1 alors que le modèle *BCBG* produit des taux de décharge qui peuvent aller jusqu'à plusieurs centaines de Hz).

La salience est intégrée dans ces modèles par l'entrée corticale (*CSN* pour le *GPR* et *rBCBG* et *S* pour le *CBG*). Pour chaque test, nous évaluons les capacités de sélection selon deux critères principaux : (1) l'activité de l'option choisie et (2) le contraste en sortie du *GPi* des différentes actions. Le contraste est défini par la différence entre la sortie la moins inhibée et les autres sorties (critère utilisé dans Girard et al. 2003 ; Gurney et al. 2001a). Comme nous comparons des modèles ayant des taux de décharge très différents, nous normalisons les différents critères par l'activité au repos du noyaux de sortie.

Ainsi, la capacité à inhiber une action en sortie du *GPi* est : $s = \frac{\phi^i_{GPi}}{\phi^{repos}_{GPi}}$.

Le contraste est défini par :

$$c = \frac{\sum_j^n |\phi^i_{GPi} - \phi^j_{GPi}|}{\phi^{repos}_{GPi}}$$

pour i l'action la plus inhibée.

Nous regardons également quelle action est choisie et si elle est choisie seule ou bien si les deux actions sont choisies simultanément. Afin de déterminer si deux actions sont choisies simultanément nous regardons si le contraste est supérieur à 1% de l'activité au repos (ceci afin d'éviter que des écarts trop faibles d'activités entraînent une sélection).

Test de Gurney et al. 2001b

Afin de tester les capacités de sélection de l'action du modèle lorsque plus de deux options sont en compétition, nous avons implémenté le test utilisé dans Gurney et al. 2001b. Ce test permet de montrer comment le modèle est capable de sélectionner le canal ayant la salience (e.g. l'entrée corticale) la plus importante lorsque six options sont en compétition. La salience des canaux 1 et 2 change au cours du temps alors que celle des 4 autres reste nulle tout au long du test. Dans ce test, on observe si les différents canaux sont sélectionnés seuls ou partiellement. On considère un canal comme sélectionné si son activité en sortie du *GPi* est inférieure à l'activité au repos et un canal est partiellement sélectionné si plusieurs canaux sont sélectionnés en même temps.

Le test comprend cinq phases de 2000 ms. Lors de la première phase, la salience de tous les canaux est nulle, permettant d'observer le système au repos. Dans la deuxième phase, seule la salience du premier canal est augmentée à 0.4. On considère cette phase comme valide si le modèle sélectionne uniquement le canal 1. Lors de la troisième phase, le deuxième canal est augmenté à 0.6, soit un niveau supérieur du canal 1 gardé à 0.4. Cette phase est complètement validée si le canal 2 est le seul à être sélectionné. Elle est partiellement validée si deux canaux sont partiellement sélectionnés mais que le canal 2

2. Cette borne supérieure de 20Hz a été fixée après une analyse de la réponse du modèle *rBCBG* avec des saliences élevées. Une salience supérieure à 20Hz sur plusieurs canaux entraîne une déshinibition totale du *GPi* de ces deux canaux, ne permettant pas de sélection.

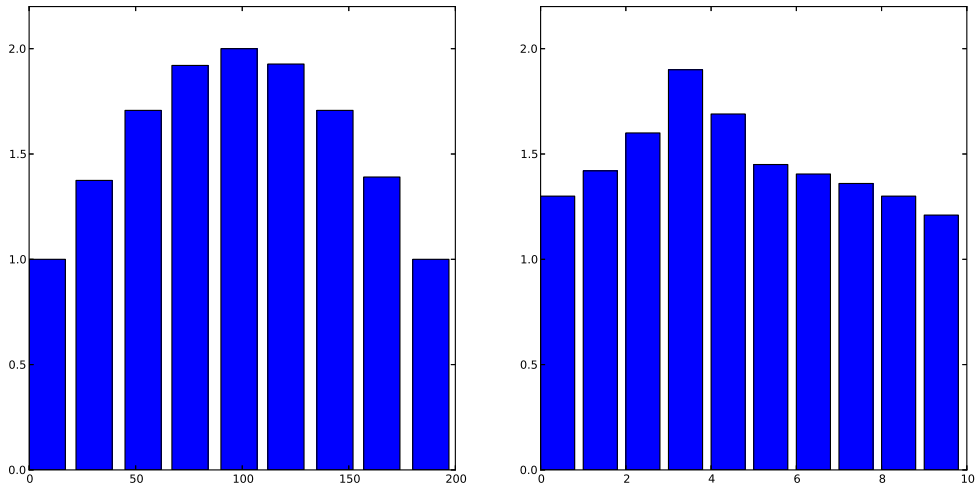

Figure 5.2 – *Saliences utilisées dans A la tâche de* Georgopoulos *et al. 1982 et B la tâche de* Humphries *et al. 2012.*

est plus inhibé que les autres canaux. Dans la quatrième phase, la salience du canal 1 est augmentée à 0.6 (soit au niveau du canal 2). Cette phase est valide si les canaux 1 et 2 sont similairement inhibés en sortie du *GPi* (i.e. similairement partiellement sélectionnés). Dans la dernière phase, la salience du canal 1 est remise à 0.4 et est donc identique à la troisième phase. Elle permet de voir comment le système dé-sélectionne le canal 1.

Différentes saliences

Nous avons testé deux types de saliences avec de multiples canaux. La fonction salience est injectée dans les modèles qui sont simulés jusqu'à convergence de l'activité des noyaux.

La première fonction de salience testée comprend 10 canaux et est basée sur une distribution $\Gamma(2, 0.1)$ (à l'image de Humphries et al. 2012 ; voir Figure 5.2 B). Cette salience permet un test théorique de la sélection et la comparaison avec les travaux de Humphries et al. 2012, mais ne s'appuie sur aucune donnée électrophysiologique. Il n'est ainsi pas évident de déterminer la plausibilité d'un tel signal cortical. Afin de tester une sélection basée sur un signal biologiquement plausible nous avons modélisé une tâche motrice étudiée dans la littérature et pour laquelle la forme de l'entrée corticale est connue (Georgopoulos et al. 1982 ; Kalaska et al. 1989). La tâche consiste à faire un mouvement du bras dans 8 directions différentes équidistantes de la position initiale (8 cm) et réparties avec un intervalle de 45°. Ces études ont permis d'établir que les neurones du cortex moteur d'un macaque ont une direction préférée pour laquelle ils montrent une activité maximale et ont une activité qui diminue lorsque l'on s'éloigne de cette direction selon une forme sinusoïdale.

En accord avec ces études et en reprenant la modélisation utilisée dans Liénard et

Paramètre	MSN D1→ GPi	MSN D2 → GPi	MSN D1+D2 → GPi
	0.2	0.02	0.6
	0.25	0.07	0.5
	0.3	0.12	0.4
	0.35	0.17	0.30
	0.4	0.22	0.20
	0.5	0.17	0.15
	0.6	0.07	0.15
	0.7	0.02	0.1
	0.82	0.	0.

Table 5.3 – *Proportion de neurones projetant vers GPi. Chaque ligne du tableau représente une configuration de ségrégation partielle permettant d'avoir une influence sur le chemin direct. Une configuration est définie par la proportion de MSN D1, MSN D2 et de neurones avec colocalisation projetant vers le GPi. Dans chaque configuration, 82% des neurones projettent vers le GPi. Nous modélisons ainsi un chemin direct partiel en supposant qu'une majorité des neurones projetant vers le GPi ont des récepteurs D1.*

Girard 2014, nous avons fixé l'entrée corticale à :

$$\phi^i_{CSN} = (1 + sin(\frac{i}{2} + 180))$$

avec i la direction entre $[-180°, +180°]$ discrétisée par pas de 45° (voir Figure 5.2 A).

5.2.3 Dopamine et ségrégation D1/D2

Dans chaque modèle, nous avons simulé l'effet de différents niveaux de dopamine tonique sur la sélection de l'action. Dans les modèles *CBG* et *GPR*, l'effet de la dopamine est modélisé par l'augmentation (respectivement diminution) de l'activité dans les *MSN D1* (respectivement *D2*), changeant ainsi l'équilibre entre chemins direct et indirect du modèle.

Cette modélisation est plus délicate à implémenter dans le modèle *rBCBG* car il fait initialement l'hypothèse d'un total recouvrement de ces chemins, et à ce titre ne fait pas la distinction entre *MSN D1* et *D2*. Dans un premier temps nous modélisons donc l'effet de la dopamine sur les autres noyaux du modèle. Les neurones du *STN* ainsi que du *GP* possèdent des récepteurs dopaminergiques de type *D2* (Smith et Kieval 2000) qui sont négligés dans les modèles *CBG* et *GPR*. De plus, il a été montré que les connexions synaptiques du *STN* et du *GP* sont renforcées chez les patients parkinsoniens (Fan et al. 2012), indiquant qu'une diminution chronique du niveau dopaminergique entraîne un renforcement de la connectivité entre *STN* et *GP*. En conséquence, l'état parkinsonien du modèle *BCBG* se traduit par un renforcement des connections synaptiques de ces noyaux. Nous modélisons cet effet en augmentant conjointement l'amplitude A_n et la demie-vie D_n dans le *STN* et le *GP* (Liénard et Girard 2014). Les poids synaptiques utilisés dans le

modèle réduit étant proportionnels à ces deux valeurs, nous modélisons le renforcement de la connectivité par une augmentation/diminution des poids synaptiques de ±20%.

Afin de modéliser une possible asymétrie dans les chemins direct et indirect, nous testons différents niveaux de ségrégations qui restent en accord avec la connectivité globale du modèle. Nous considérons le chemin indirect absent puisque tous les *MSN* projettent vers le *GPe* dans le modèle *BCBG*. Ainsi en considérant une proportion identique de neurones ayant des récepteurs *D1* et *D2*, l'effet de la dopamine sera compensée dans le chemin indirect. Cependant nous pouvons jouer sur la proportion de *MSN D1, D2* et colocalisés projetant vers le *GPi* tout en respectant le fait que $\mathcal{P}_{MSN \to GPi} = 82\%$ (voir Figure 5.1). Nous avons testé différentes configurations en augmentant progressivement la proportion de *MSN D1* dans le chemin direct (voir Tableau 5.3).

En effet, on peut exprimer la proportion de *MSN* projetant vers *GPi* par :

$$\mathcal{P}_{MSN \to GPi} = p_{MSND1 \to GPi} + p_{MSND2 \to GPi} + p_{MSND1/D2 \to GPi} = 82\%$$

Ne connaissant a priori pas les proportions de *MSN D1*, $p_{MSND1 \to GPi}$, *MSN D2*, - $p_{MSND2 \to GPi}$ et *MSN* avec une colocalisation des deux types de récepteurs $p_{MSND1/D2 \to GPi}$, on peut imaginer avoir une proportion de *MSN D1* plus importante que de *MSN D1* et ainsi avoir un renforcement positif de ce chemin avec une augmentation du niveau dopaminergique.

Notons, que si la proportion exacte de neurones de type *D1/D2* et de neurones ayant une colocalisation des deux types de récepteurs n'est pas connue avec précision, des études rapportent que la proportion de *MSN D1* est proche de la proportion de *MSN D2* (Gerfen et Surmeier 2012). Cependant la proportion de *MSN* présentant une colocalisation des deux types de récepteurs est floue dans la littérature puisqu'elle peut aller de 2-5% (Aubert et al. 2000) à près de 60% (Nadjar et al. 2006).

Aussi, si certaines configurations sont plausibles d'un point de vue biologique, d'autres le sont moins. Une configuration peut être considérée plausible si la proportion de *MSN D1* et *MSN D2* dans le striatum est similaire. Or, lorsque $p_{MSND1 \to GPi} > 0.5$ on a forcément une proportion de *MSN D1* > *MSN D2*. Dans la littérature, on observe plutôt un nombre similaire de *MSN D1* et *D2*. Typiquement on considérera une configuration plausible si $p_{MSND1 \to GPi} \leq 0.40$.

5.2.4 Sortie du *GPi* et probabilité de sélection

Afin de mesurer l'impact de différents niveaux dopaminergiques sur la sélection, nous avons tout d'abord converti le signal de sortie des ganglions de la base (e.g. l'activité du *GPi/SNr*) en une mesure de probabilité de sélection de l'action. Dans Humphries et al. 2012, la probabilité est extraite de l'activité de *GPi* par :

$$p(a_i | GPi^i) = \frac{1 - GPi^i}{\sum_j 1 - GPi^j}$$

Ce calcul de probabilité repose sur le fait que l'activité maximale des noyaux du *GPR* est de 1. L'activité maximale du *GPi* dans le modèle *BCBG* est de 400Hz. Cependant

prendre 400Hz comme activité maximale a pour effet d'écraser la densité de probabilité. Nous avons décidé de conserver pour le modèle *rBCBG* le même rapport, r, entre activité maximale et l'activité au repos du noyau de sortie considérée dans l'étude originale avec le *GPR* pour le modèle *rBCBG*.

Nous avons donc pour le modèle *GPR* :

$$r = \frac{\phi_{rest}^{GPi}}{\phi_{max}^{GPi}} = \frac{0.185}{1.}$$

On en déduit pour le *rBCBG* :

$$r = \frac{\phi_{rest}^{GPi}}{M} \implies M = \frac{\phi_{rest}^{GPi}}{r} = \frac{65.84}{0.185} = 355.9$$

Ainsi nous calculons la probabilité de choisir une action dans le *rBCBG* par :

$$p(a_i|GPi^i) = \frac{355.9 - GPi^i}{\sum_j (355.9 - GPi^j)}$$

À partir de cette mesure de probabilité déduite de la sortie du *GPi*, nous calculons l'entropie associée :

$$H = -\sum_i^n p(a_i) log_2(p(a_i))$$

L'entropie est une mesure de répartition de l'information liée à la deuxième loi de la thermodynamique. Elle a été utilisée dans Humphries et al. 2012 afin d'évaluer le compromis exploration/exploitation du modèle. Si la densité de probabilité sur les actions est proche d'une densité uniforme, l'entropie est maximale et la sélection sera proche d'une sélection aléatoire et donc hautement exploratoire. Au contraire si la densité de probabilité n'est forte que pour une seule action, l'entropie sera minimale et la sélection déduite sera alors proche d'une sélection totale et portée sur l'exploitation des connaissances.

5.2.5 Parkinson et oscillations β

Le modèle *BCBG* permet d'obtenir des oscillations β (oscillations dont la fréquence est comprise entre 13 et 30Hz) en modélisant un état parkinsonien. La maladie de Parkinson se traduit notamment par la mort des neurones dopaminergiques de la *SNc* et de la *VTA* entraînant une diminution du niveau dopaminergique dans les ganglions de la base. Cette diminution se traduit par une augmentation des oscillations β dans le noyau subthalamique (Weinberger et al. 2006). Afin de tester la possible apparition d'oscillations β lors de la modélisation d'un état parkinsonien dans le modèle *rBCBG*, nous avons modélisé une diminution du niveau dopaminergique dans le *STN* et *GP* en augmentant progressivement la connectivité de ces noyaux de 0 à 50%.

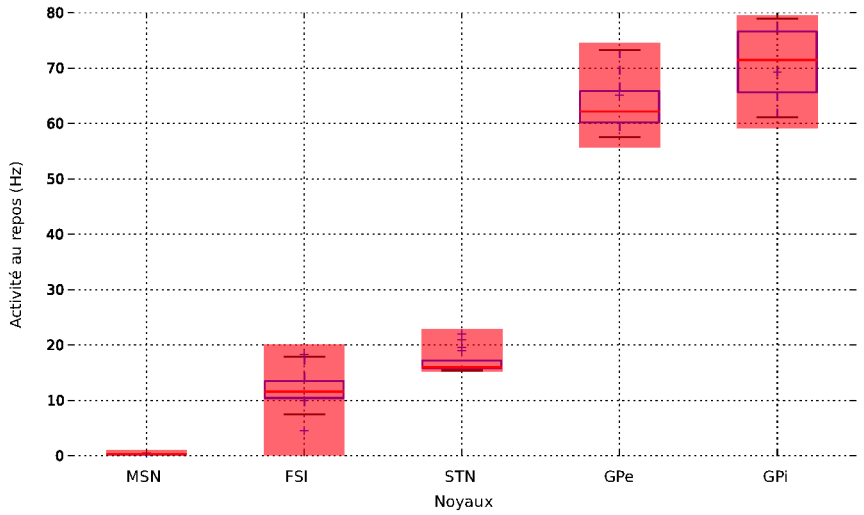

Figure 5.3 – *Activité au repos, des différents noyaux, calculée par le modèle* rBCBG *avec les différentes paramétrisations du modèle* BCBG *obtenues par* Liénard *et* Girard *2014. En rouge sont indiquées les bornes de plausibilité d'activité de chaque noyau.*

5.3 Résultats

5.3.1 Des taux de décharge biologiquement plausibles

Notre volonté est de garder la plausibilité biologique du modèle tout en simplifiant la dynamique du système et ainsi son utilisation dans des tâches plus complexes. Les différentes paramétrisations du modèle trouvées par Liénard et Girard permettent d'optimiser deux critères, l'un anatomique et l'autre physiologique. Cette optimisation multicritère a permis de construire multiples paramétrisations du modèle, optimales au sens de Pareto et capables de sélection.

Ainsi la première des propriétés que le modèle simplifié se doit de conserver est l'activité biologiquement plausible des différents noyaux au repos. À cette fin nous avons simulé notre modèle réduit en utilisant les 1468 paramétrisations du *BCBG* pendant 1000 ms. Il s'avère que ce modèle réduit reproduit fidèlement les taux de décharges au repos enregistrés chez le singe et utilisés pour l'optimisation du modèle originel (voir Figure 5.3). En effet, l'activité des moyennes des *MSN* au repos est de $0.25Hz$ (écart type : 0.018, maximum : 0.28, minimum : 0.21). L'activité des *MSN* est donc bien comprise dans l'intervalle $0.5Hz \pm 0.5$. De même, les *FSI* ont une activité au repos moyenne de $12.21Hz$ sur l'ensemble des solutions testées (écart type : 3.63, minimum : 4.465, maximum : 18.26), ce qui est bien compris sur l'intervalle $0-20Hz$. Le *STN* montre une activité moyenne de $16.87Hz$ (écart type : 1.76, minimum : 15.37, maximum : 21.93) comprise dans l'intervalle $19 \pm 3.8Hz$. De la même façon nous avons le *GPe* qui a une activité au repos moyenne de $63.56Hz$ (écart type : 4.77, minimum : 57.48, maximum : 73.22) et *GPi* de $70.53Hz$ (écart type : 5.69, minimum : 61.07, maximum : 78.93) (voir Figure 5.3).

Ces résultats montrent que la réduction proposée par le modèle *rBCBG* permet de

garder une activité au repos compatible avec les critères de plausibilité biologique définis dans Liénard et Girard 2014, basés sur l'activité au repos des noyaux des ganglions de la base du primate.

5.3.2 Un modèle de sélection de l'action

Deux canaux en compétition

Liénard et Girard ont également montré que le modèle *BCBG* a des capacités de sélection de l'action et que celles-ci émergent de la construction anatomique du modèle. Aussi nous avons souhaité étudier la capacité de sélection de l'action du modèle réduit. Nous avons implémenté deux autres modèles de la littérature : les modèles *CBG* (Girard et al. 2008) et *GPR* (Gurney et al. 2001b). Cela nous permet de comparer trois modèles de la littérature, présentant des différences importantes tant par leur construction que par leur anatomie (voir Chapitre III), sur leur capacité de sélection de l'action.

Nous avons réalisé plusieurs tests de sélection variant la salience des deux canaux en compétitions. Nous testons ainsi la capacité des systèmes à sélectionner une action lorsque deux canaux ont une salience plus ou moins proche. Les résultats sont illustrés dans la Figure 5.4. Le modèle *rBCBG* est capable de sélectionner correctement [3] le canal associé à la plus haute salience uniquement lorsque le deuxième canal n'a pas une salience trop importante. Notamment, lorsque la salience la plus faible dépasse $9Hz$, le système commence à perdre sa capacité à sélectionner le canal associé à la plus haute salience lorsque les deux canaux ont une salience relativement proche (voir Figure 5.4 A). Ce phénomène s'accentue lorsque l'activité des deux canaux dépasse $12Hz$. En effet, lorsque la salience des deux canaux dépasse $12Hz$, le contraste entre l'activité de sortie des deux canaux devient très faible rendant la sélection d'un unique canal difficile.

Les modèles *GPR* et *CBG* quant à eux ont plus de difficultés à départager les saliences faibles en ayant une frontière assez nette (de 0.28 pour le *CBG* et 0.16 pour le *GPR* ; voir Figure 5.4 B et C), en dessous de laquelle un canal ne peut être sélectionné et ce même si le canal en compétition a une salience nulle. En contrepartie, le modèle *GPR* n'a pas de problème à départager deux canaux ayant des saliences supérieures à ce seuil. Le modèle *CBG* a, tout comme le modèle *rBCBG*, une limite à partir de laquelle il ne peut distinguer entre deux saliences trop proches. En effet, si les deux saliences sont au-dessus de ce seuil de 0.7 (voir Figure 5.4 B gauche et droite), les deux canaux peuvent être sélectionnés simultanément. On notera que le modèle *CBG* arrive à avoir un contraste assez élevé même pour des saliences assez proches lorsque celles-ci sont comprises entre 0.3 et 0.7. Comparativement, le modèle *GPR* a un contraste moins élevé pour des saliences proches, ce qui se traduit sur la Figure 5.4 B droite par une diagonale moins nette et marquée pour le modèle *GPR* que pour le modèle *CBG* (voir Figure 5.4 C droite) ; et a plus forte raison que pour le modèle *rBCBG* où la diagonale disparaît pour des saliences trop grandes.

De ces tests se dégagent des limites nettes dans les capacités de sélection des différents

3. Rappelons que l'on considère un seuil arbitraire de 1% de l'activité au repos du noyau en dessous duquel la différence d'activité des canaux en compétition est considérée comme trop faible pour permettre la sélection (voir Méthode).

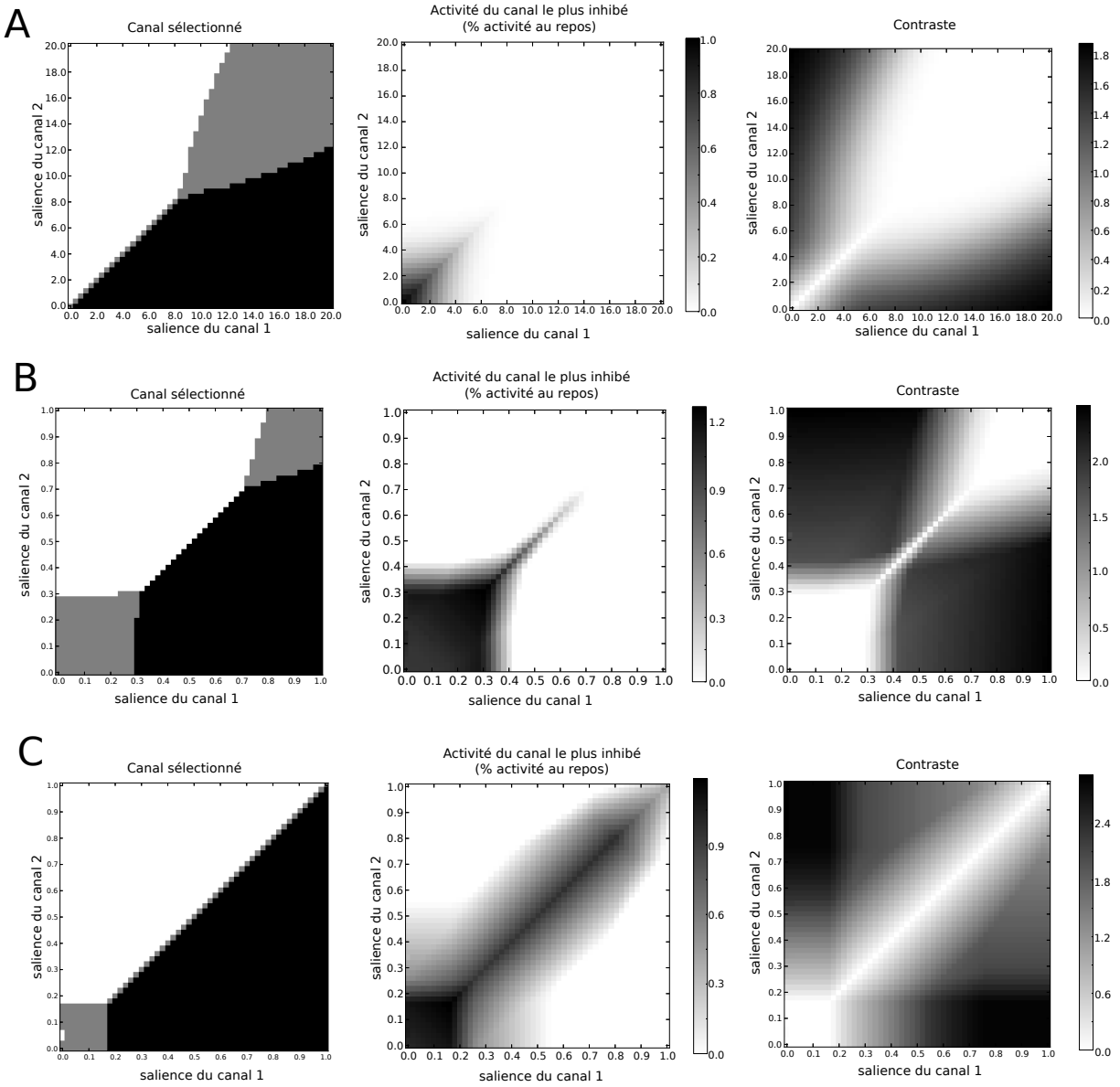

Figure 5.4 – *Résultats de la sélection avec deux canaux pour les différents modèles. A : modèle* rBCBG*. B : modèle* CBG*. C : modèle* GPR*. Les figures de gauche montrent quel canal a été choisi (un canal est choisi s'il est plus inhibé que le deuxième canal d'au moins 1% le taux de décharge au repos du noyau de sortie). Les figures du milieu montrent l'activité du canal le plus inhibé en fonction de la salience de chaque canal. Les figures de droite montrent l'évolution du contraste entre les deux canaux au niveau du noyau de sortie du modèle.*

modèles, que ce soit pour des saliences faibles ne permettant pas la sélection ou des saliences élevées mais trop proches pour que la différences permettent au modèle de faire la sélection. D'un point de vue théorique, cette capacité limitée à choisir entre deux options ayant une forte salience, est dommageable à la sélection. Cependant, plusieurs études ont montré dans le cadre des saccades oculaires que lorsque deux cibles intéressantes sont présentées, il n'est pas rare d'observer des sélections partielles se traduisant par des saccades moyennes (Becker et Jürgens 1979).

Six canaux en compétition

Connaissant les limites des capacités de sélection des différents modèles, nous avons testé le comportement des différents noyaux lorsque six canaux sont en compétition. Ce test se décompose en cinq phases (voir Méthode). Les phases 2 à 5 permettent d'évaluer les capacités à sélectionner et désélectionner un canal parmi d'autres.

Le modèle *rBCBG* est capable de sélectionner le canal 1 lors de la deuxième phase (voir Figure 5.5). En effet, l'activité du premier canal passe de $65.85Hz$ au repos à $38.93Hz$, en sortie du *GPi*, montrant une nette sélection par rapport aux autres canaux qui ont alors une activité de $67.76Hz$. On note, que l'augmentation de l'activité des canaux 2 à 6 est faible comparativement à la diminution de l'activité du premier canal. Dans la troisième phase la salience du deuxième canal est augmentée, ce qui se traduit par la diminution de l'activité de sortie du deuxième canal à $26.84Hz$. L'activité du premier s'en trouve légèrement augmentée passant à $41.41Hz$. De même, l'activité des autres canaux est augmentée à $70.93Hz$. Ainsi le canal 2 est seulement partiellement sélectionné. Dans la quatrième phase, les canaux 1 et 2 ont tous deux la même salience et ont la même activité de sortie de $27.58Hz$. Les deux canaux sont donc similairement sélectionnés. Les autres canaux ont une activité de $72.12Hz$. Dans la cinquième phase, l'activité des différents canaux est identique à celle observée dans la troisième phase et le canal 2 est partiellement sélectionné.

Le modèle *CBG* a une activité au repos de 0.10. Durant la phase 2, le canal 1 est sélectionné avec une activité de 0.014 et les autres canaux voient leur activité augmentée à 0.15 (voir Figure 5.6). Dans la phase 3, le canal 2 est sélectionné seul avec une activité nulle et les autres canaux ont tous une activité de 0.19. On peut observer ici que malgré la salience du canal 1 non nulle et supérieure aux autres canaux, son activité est identique à ces derniers. Dans la 4ème phase, les canaux 1 et 2 sont similairement sélectionnés avec une activité de 0.029 et les autres canaux ont une activité de 0.26. La phase 5 est identique à la phase 3.

Le modèle *GPR* a une activité au repos de 0.18. Lors de la phase 2 il sélectionne correctement le premier canal (voir Figure 5.7), dont l'activité est de 0.048, tandis que les autres canaux ont une activité de 0.31. Dans la 3ème phase, le canal 2 est sélectionné seul avec une activité de 0.07. Le canal 1 voit son activité augmentée à 0.27 et les autres canaux ont une activité encore plus grande à 0.53. Dans la phase 4, les canaux 1 et 2 sont similairement sélectionnés avec une activité de 0.17 et les autres canaux ont une activité de 0.62. De nouveau la phase 5 montre des résultats identiques à la phase 3.

On peut donc noter que les capacités de sélection du modèle *rBCBG* se trouvent légère-

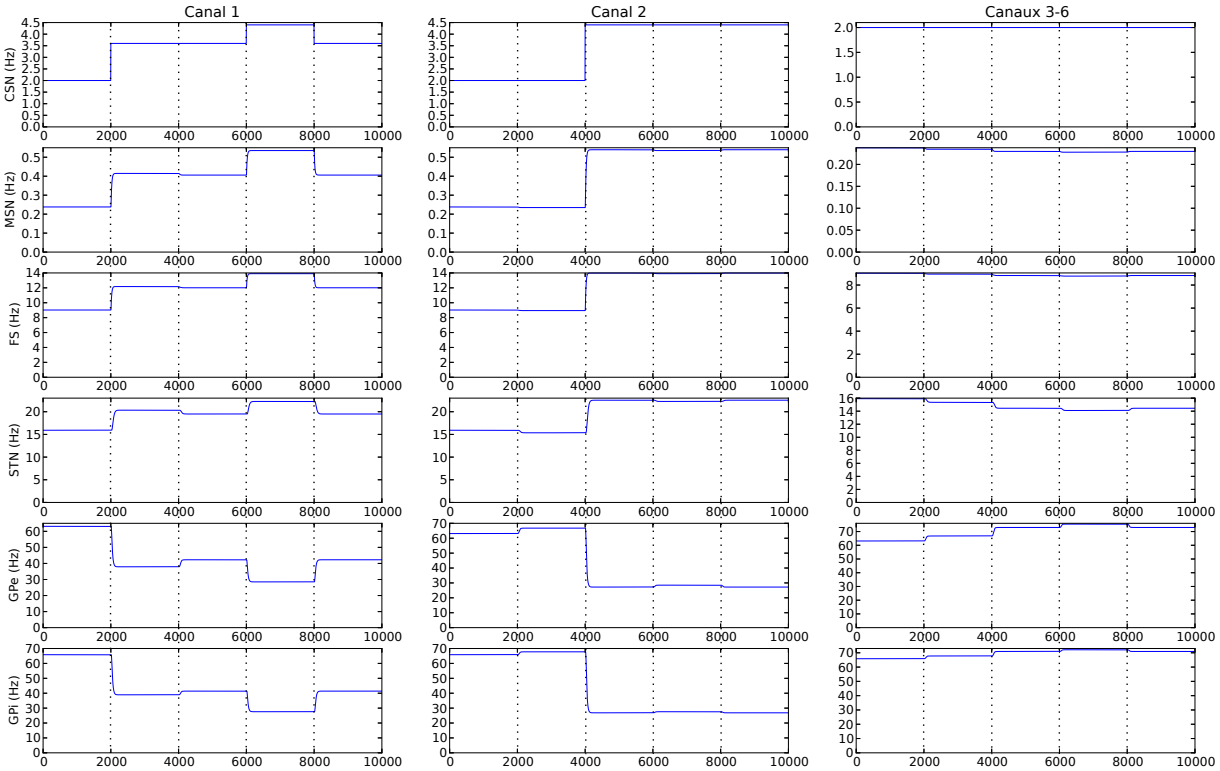

Figure 5.5 – *Activité des différents noyaux du modèle réduit* rBCBG.

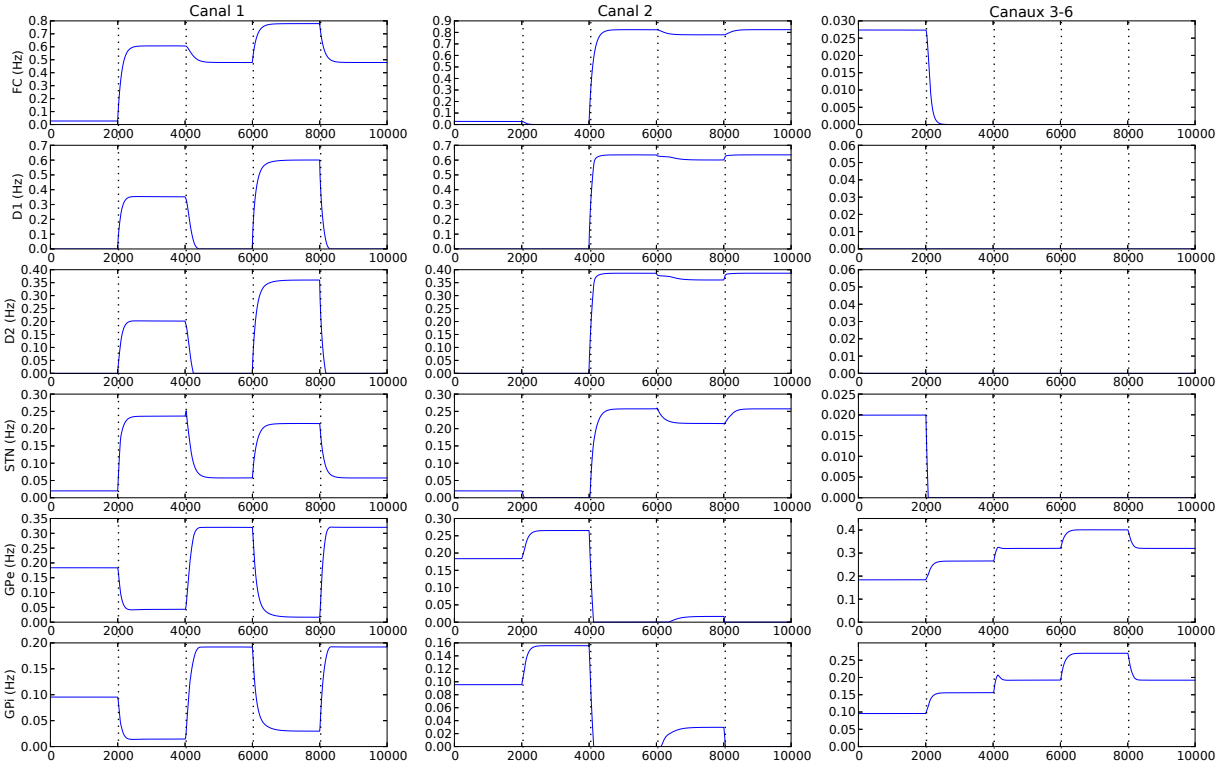

Figure 5.6 – *Activité des différents noyaux du modèle* CBG.

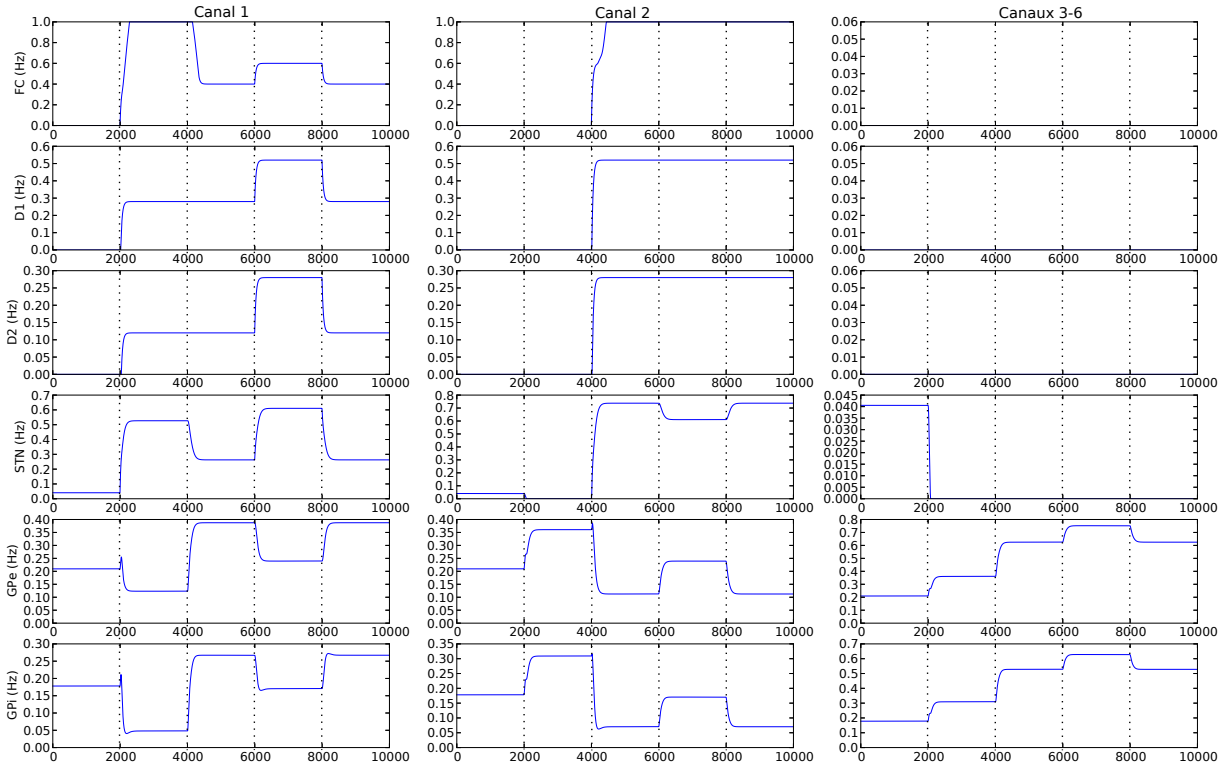

Figure 5.7 – *Activité des différents noyaux du modèle* GPR.

ment en deçà des modèles *GPR* et *CBG*. Ceci est dû au fait que la sélection du canal 2 est partielle dans la 3ème et 5ème phase. Cela semble être dû à sa capacité limitée à exciter les canaux ayant une salience moins grande en sortie du *GPi*. La sélection se fait donc plus sur l'inhibition des canaux ayant une salience positive que par l'excitation des canaux en compétition, ce qui empêcherait leur sélection. Ainsi, avec le modèle *rBCBG*, plusieurs canaux peuvent être en même temps sélectionnés par le système. On notera que, dans le modèle *rBCBG*, l'activité des neurones du faisceau pyramidal (*PTN*) est un signal constant, indépendant de la salience des canaux en compétition. Or, il est probable que ce soit une simplification du modèle et la prise en compte de la salience des différents canaux au niveau des *PTN* devrait permettre une meilleur sélection.

Toutefois, si la sélection simultanée de plusieurs options n'est pas forcément souhaitable dans un modèle de sélection, des exemples de sélections partielles existent dans la littérature. En effet, certaines sélections peuvent ne pas être totales et résulter d'un mélange de deux ou plusieurs actions ; c'est notamment le cas des saccades oculaires moyennes (Becker et Jürgens 1979). Ainsi la sélection simultanée de plusieurs actions peut être plausible d'un point de vue biologique à défaut d'optimale d'un point de vue de l'efficacité de la sélection.

On a ainsi pu montrer que notre simplification du modèle *BCBG*, qui consiste à négliger la durée de l'intégration temporelle synaptique du signal entrant par les récepteurs AMPA, NMDA et GABA au cours du temps, permet de garder des taux de décharges ainsi que des capacités de sélection de l'action similaires au modèle complet. Cependant cette

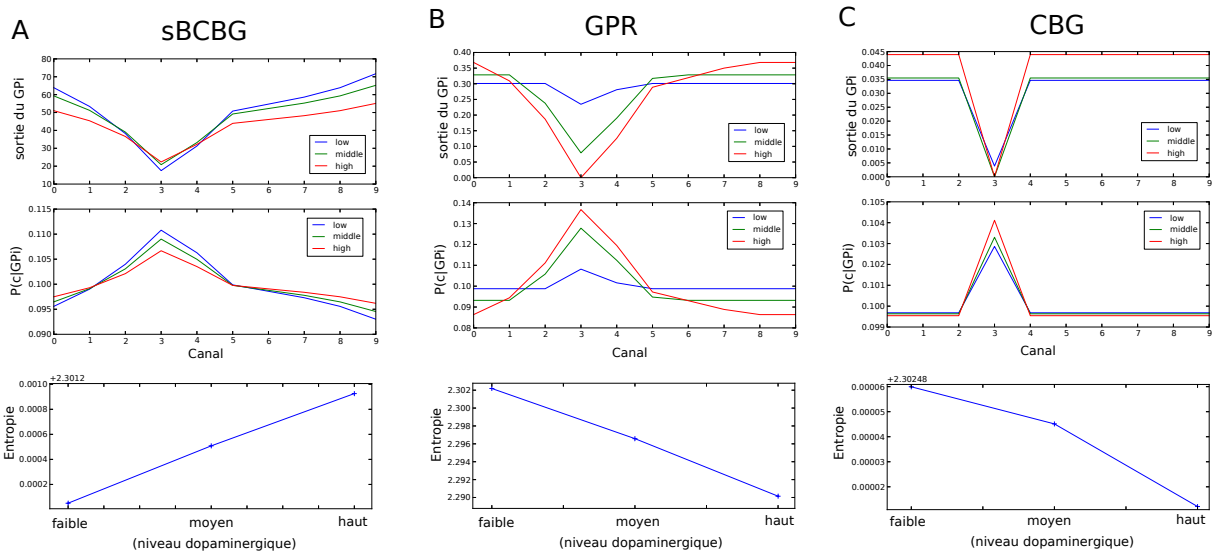

Figure 5.8 – *Reproduction des résultats obtenus dans* Humphries *et al. 2012 avec les modèles : A.* rBCBG*, B.* GPR *et C.* CBG*. Les courbes du haut représentent les sorties du* GPi *des différents modèles pour les différents canaux, sous différents niveaux de dopamine tonique. Les courbes du milieu représentent la fonction de probabilité de choisir un canal calculé en fonction de la sortie du* GPi*. Les courbes du bas représentent l'entropie de la fonction de probabilité sur les canaux en fonction de trois niveaux dopaminergiques : bas, moyen et haut.*

simplification peut avoir un effet sur la dynamique du système dans les phases transitoires.

5.3.3 Effet de différents niveaux de DA tonique sur la sélection

Les neurones dopmaminergiques de la *SNc* et de *VTA* projettent massivement vers les ganglions de la base et affectent fortement son activité. En effet, un dérèglement du niveau dopaminergique peut avoir un effet important sur l'activité des ganglions de la base (voir Chapitre II et Chapitre III) et ainsi sur la sélection de l'action. Une étude récente a montré que dans un modèle de type *GPR*, la dopamine tonique contrôle le compromis exploration/exploitation de la sélection de l'action des ganglions de la base (Humphries et al. 2012).

Dans le modèle *GPR* ainsi que le modèle *CBG*, le niveau dopaminergique contrôle l'activité des *MSN* via la modulation des poids synaptiques entre le cortex et le striatum. Les différents types de récepteurs à la dopamine des MSN D1 et D2 créent un système asymétrique en chemin direct et chemin indirect (voir Chapitre III). Le modèle *rBCBG* ne différenciant pas les MSN selon leur type de récepteurs, nous avons dans un premier temps modélisé l'effet d'une diminution (respectivement augmentation) de l'activité dopaminergique via l'augmentation (respectivement diminution) des forces de connexions synaptiques dans le *GP* et le *STN*. En effet, la dopamine affecte également les autres noyaux des ganglions, dans lesquelles l'effet des récepteurs D2 semble dominé (Hassani et al. 1996 ; Lindvall et Björklund 1979 ; Smith et Kieval 2000). Cependant cette

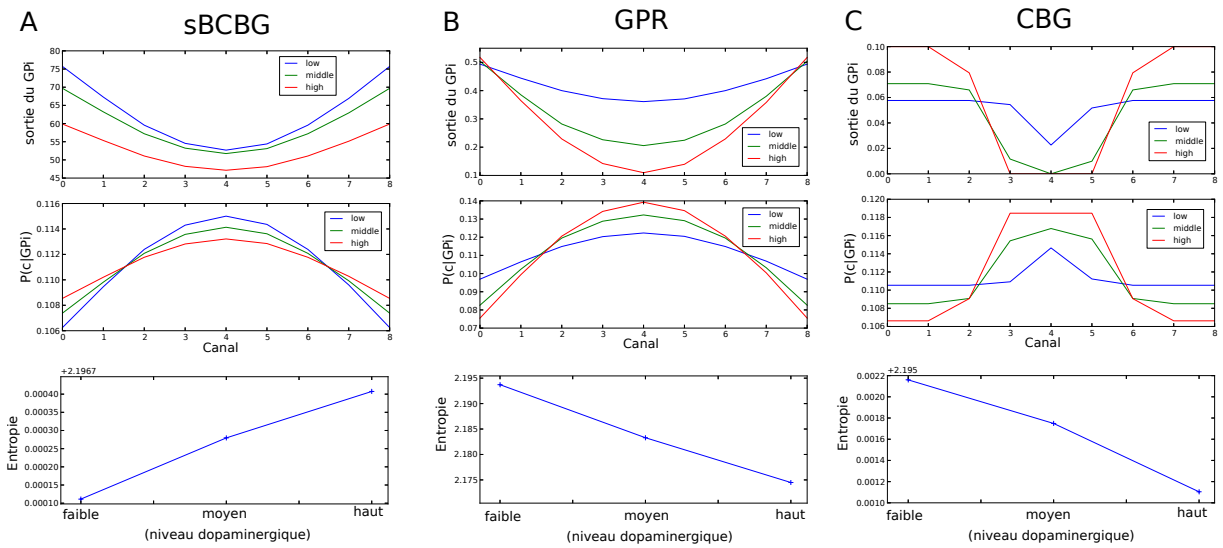

Figure 5.9 – *Test de l'effet de différents niveaux de dopamine tonique sur une salience compatible avec les résultats de* Georgopoulos *et al. 1982 avec les modèles : A.* sBCBG, *B.* GPR *et C.* CBG. *Les courbes du haut représentent les sorties du* GPi *des différents modèles pour les différents canaux, sous différents niveaux de dopamine tonique. Les courbes du milieu représentent la fonction de probabilité de choisir un canal calculé en fonction de la sortie du* GPi. *Les courbes du bas représentent l'entropie de la fonction de probabilité sur les canaux en fonction de trois niveaux dopaminergiques : bas, moyen et haut.*

hypothèse sera à vérifier compte tenu de la présence de récepteurs de type D1 chez le *STN* et *GP* (Rommelfanger et Wichmann 2010).

Ainsi après avoir étudié dans les parties précédentes les capacités de sélection de l'action des modèles avec un niveau de dopamine tonique constant, nous nous sommes intéressés à l'effet d'une diminution/augmentation de ce niveau sur la sélection.

Nous avons dans un premier temps reproduit, avec le modèle *GPR*, les résultats de Humphries et al. 2012 (voir Figure 5.8 B), puis observé comment se comportent les modèles *rBCBG* et *CBG* (voir Figure 5.8 A et B) dans les mêmes conditions afin de tester les prédictions de cette étude. Nous avons donc soumis les modèles à un vecteur de salience basé sur une distribution $\Gamma(2, 0.1)$ (voir Méthode) et observé l'activité de sortie des différents noyaux afin d'évaluer si une augmentation du niveau dopaminergique affecte la sélection. Afin de mesurer cet effet, l'activité du *GPi* est transformée en densité de probabilité sur les canaux (voir Méthode). Enfin, l'entropie liée à la fonction de probabilité est calculée pour les différents niveaux dopaminergiques. Nous testons 3 niveaux dopaminergiques : un moyen où les modèles sont en condition standard de fonctionnement, un deuxième niveau où la dopamine tonique est plus haute que la normale et enfin un troisième niveau où elle est plus basse.

Les modèles *GPR* et *CBG* semblent être affectés de la même façon par la dopamine. En effet une augmentation du niveau dopaminergique entraîne une diminution de l'entropie de la fonction de probabilité et donc favorise la sélection du canal associé à la plus forte salience. Cela se traduit, dans l'activité de sortie des modèles et leur densité de probabilité, par un pic plus marqué au niveau du canal associé à la plus haute salience. Dans le modèle *CBG* cela se traduit plus par l'augmentation de l'activité des canaux en compétition que par une inhibition plus forte du canal sélectionné (voir Figure 5.8 C). En effet, même avec un niveau moyen de dopamine, le système sélectionne complètement le canal optimal. Chez le *GPR*, au contraire, cela se traduit plus par une inhibition plus franche du canal sélectionné. Ainsi une augmentation du niveau dopaminergique favorise l'exploitation de la fonction de salience et une diminution favorise plus d'exploration et une sélection plus aléatoire du canal. Le *GPR* ayant une activité quasi-constante sur l'ensemble des canaux lorsque le niveau dopaminergique est faible.

Le modèle *rBCBG* se comporte de façon radicalement différente de ces deux modèles. En effet une augmentation du niveau dopaminergique entraîne au contraire une augmentation de l'entropie (voir Figure 5.8 A), favorisant ainsi l'exploration. Une diminution ayant globalement l'effet inverse en augmentant le contraste entre les canaux sélectionnés et non sélectionnés.

Nous avons également reproduit ce test avec une salience issue de la littérature du contrôle moteur afin de tester ces prédictions sur un signal cortical plausible (Georgopoulos et al. 1982 ; Kalaska et al. 1989 ; voir Figure 5.9). Ce changement de salience montre des effets comparables sur l'effet du niveau dopaminergique sur la sélection du système. On notera toutefois que chez les trois modèles, une augmentation du niveau dopaminergique tend à sélectionner plusieurs canaux simultanément. C'est notamment le cas pour le modèle *CBG* (voir Figure 5.9 C) pour lequel 3 canaux sont complètement desinhibés en sortie du *GPi* avec un haut niveau dopaminergique. Dans une moindre mesure, c'est également le cas pour les modèles *rBCBG* et *GPR* (voir Figure 5.9 A et B), où le niveau

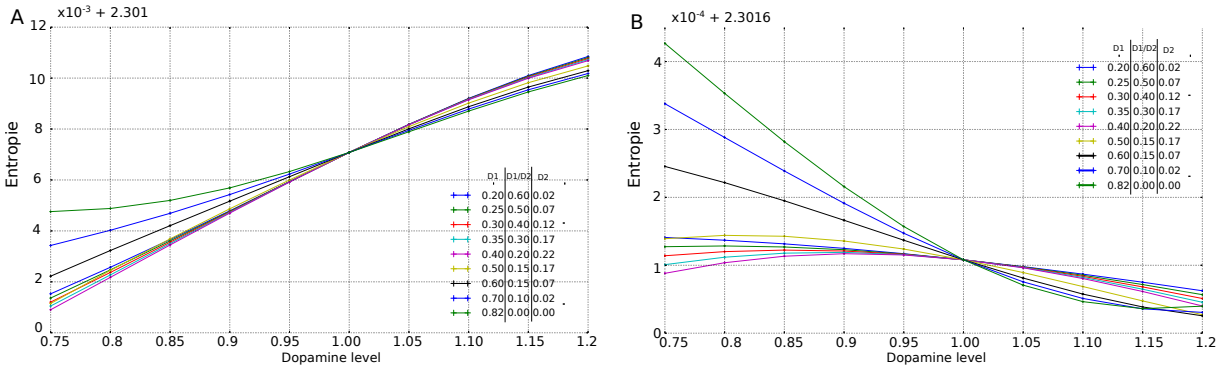

Figure 5.10 – *Évolution de l'entropie du signal en fonction du niveau dopaminergique pour différentes ségrégations du chemin direct. A. Résultats avec la modélisation de l'impact de la dopamine sur le complexe* STN/GP *et B. sans.*

d'activité globale des canaux tend à être plus faible en sortie du *GPi*.

Les différences sur l'effet du niveau de dopamine tonique sur les différents modèles, peuvent s'expliquer par l'absence de différenciation *D1/D2* des *MSN* qui est présente dans les modèles *GPR* et *CBG* mais absente dans le modèle *rBCBG*. En effet, le modèle *rBCBG* suppose que la proportion de neurones projetant à la fois vers les chemins direct et indirect ne permet pas une ségrégation franche de ces chemins. Cependant il est possible de supposer, tout en gardant les contraintes anatomiques du modèle *BCBG*, la présence partielle du chemin direct. En effet le modèle suppose que 82% des *MSN* du striatum projettent vers *GPi* (voir Figure 5.1). Il est donc possible de supposer que parmi ces neurones il y a une majorité de *MSN D1* (voir Méthode).

Nous avons simulé plusieurs variations du modèle *rBCBG* soumis à la salience de Humphries et al. 2012 dont la proportion de *MSN D1* projetant vers *GPi* varie (voir Tableau 5.3). Ces différents modèles ont été soumis à plusieurs niveaux dopaminergiques afin de voir l'évolution de l'entropie dérivée du signal de sortie du modèle. Nous avons testé ces modèles avec la modélisation de l'impact du niveau dopaminergique sur les noyaux *STN* et *GP* (voir Figure 5.10 A), ce qui est biologiquement plausible ou sans cette modélisation (voir Figure 5.10 B), ce qui est moins plausible mais plus proche de la modélisation du *GPR* et *CBG*.

Nous pouvons observer que la séparation partielle du chemin direct ne permet pas, en gardant l'impact du niveau dopaminergique sur les noyaux *GP/STN* de changer l'effet de la dopamine sur la sélection (voir Figure 5.10 A). En effet un plus haut niveau dopaminergique continue d'augmenter l'entropie de la sélection prônant ainsi une politique exploratoire. Cependant lorsque l'on ne prend pas en compte l'impact de la dopamine sur *GP/STN*, on peut observer une légère diminution de l'entropie avec l'augmentation du niveau dopaminergique. Cette effet est relativement faible mais présent pour des ségrégations plausibles. Avec une proportion plus importante de *MSN D1* dans le chemin direct on obtient l'effet observé chez les modèles *CBG* et *GPR*. Ceci montre que l'effet initial inversé sur l'entropie entre le modèle *rBCBG* et ces deux modèles n'est pas un défaut

du premier, mais s'explique bel et bien par la séparation partielle entre chemins direct et indirect.

Dans notre modélisation, l'effet de la dopamine est de la même force dans le striatum que dans les noyaux *GP* et *STN*. Si nous avions diminué l'effet de la dopamine au niveau de *GP* et *STN* tout en renforçant son effet sur le striatum, il est possible que nous aurions pu observer le même biais dans la sélection dans notre modèle que dans les modèles *GPR* et *CBG*, tout en modélisant l'effet de la dopamine sur l'ensemble des noyaux. Néanmoins, il faut supposer une prédominance des *MSN D1* dans le chemin indirect suffisamment importante.

Nous avons pu mettre en évidence qu'un modèle anatomiquement plausible des ganglions de la base du primate prédit un effet du niveau dopaminergique radicalement différent des modèles supposant une séparation importante des chemins direct et indirect. De plus la présence partielle de ces chemins ne permet pas de retrouver les résultats obtenus avec une ségrégation forte lorsque l'on prend en considération l'impact de la dopamine sur tous les noyaux du modèle *rBCBG*. Nous avons également testé de moduler l'activité du *STN* et *GP* par la dopamine dans les modèles *CBG* et *GPR*. Cependant, dans ces modèles, l'effet de la dopamine sur ces noyaux est secondaire par rapport à son effet sur les chemins direct et indirect. La modulation de *STN* et *GP* par la dopamine ne change ainsi pas l'effet de la dopamine sur l'entropie de la sélection dans les modèles *CBG* et *GPR*

5.3.4 Maladies de Parkinson et Oscillations β

La maladie de Parkinson se traduit par la mort des neurones dopaminergiques dans la *SNc* entraînant une diminution du niveau dopaminergique dans les ganglions de la base, affectant le fonctionnement normal du système (Bartels et Leenders 2009 ; Israel et Bergman 2008 ; Obeso et al. 2000). Cette diminution du niveau dopaminergique entraîne également l'apparition d'oscillations β (oscillation allant de 13 à $30Hz$) dans l'activité des noyaux des ganglions de la base pouvant expliquer les troubles moteurs observés chez les sujets atteints de Parkinson (Brown 2003 ; Gale et al. 2008 ; Hutchison et al. 2004). Liénard et Girard ont montré que la modélisation d'une diminution du niveau dopaminergique, mimant un état Parkinsonien, dans le modèle *BCBG* donne lieu à l'apparition d'oscillations β observées chez des patients parkinsoniens dans les noyaux *STN* et *GP*. L'apparition de ces oscillations est, au moins en partie, due aux délais axonaux pris en compte dans le modèle (voir Tableau 5.1). Cependant, la prise en compte d'une dynamique d'intégration fine du signal synaptique, et notamment des différences d'intégration du signal entre récepteurs *AMPA* et *NMDA* peut être clé dans l'apparition de ces oscillations.

Aussi, bien que nous ayons montré que le *rBCBG* permet de garder les propriétés de sélection de l'action du modèle ainsi qu'une activité au repos plausible, il n'est pas trivial que notre réduction ne prévienne pas l'apparition de ces oscillations. En effet, la réduction joue sur la vitesse d'intégration synaptique du signal et peut ainsi affecter la dynamique fine du système. En testant l'apparition d'oscillations dans le *rBCBG*, nous testons si les différences d'intégration synaptique dépendant du type de récepteurs est clé ou si la seule

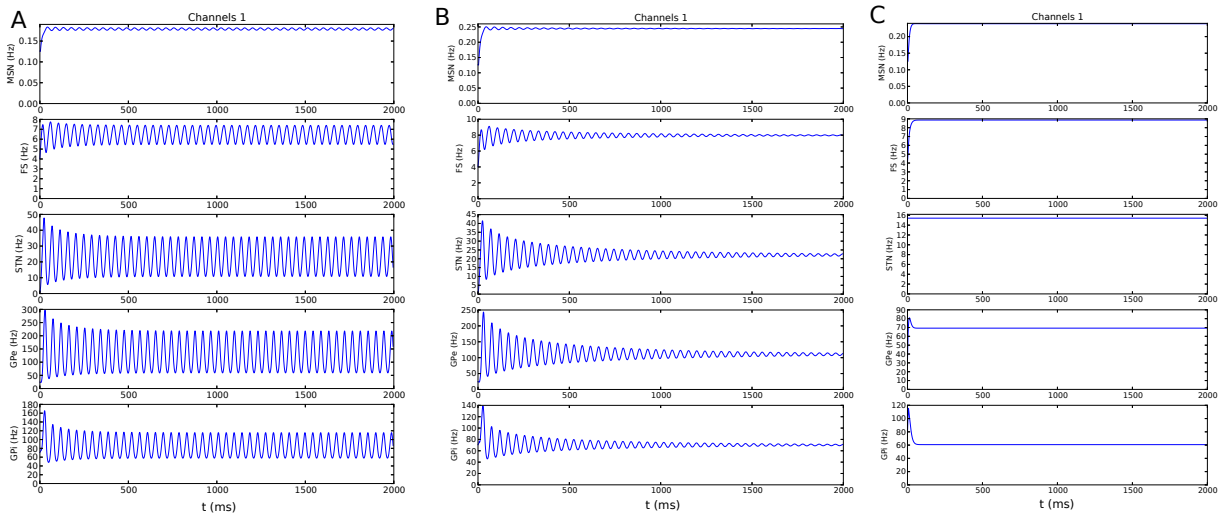

Figure 5.11 – *Apparition d'oscillations avec un faible niveau dopaminergique dans les différents noyaux du modèle* sBCBG. *A. Oscillations avec un renforcement de 30% de la connectivité dans* GP *et* STN *(22 oscillations par seconde). B. Oscillations avec un renforcement de 25% de la connectivité. C. Activité avec* STN *constant et 30% d'augmentation de la connectivité.*

prise en compte du délai axonal est suffisante.

En augmentant la force synaptique des signaux entrant de 30% dans les noyaux *GP* et *STN*, simulant une diminution du niveau dopaminergique agissant sur des récepteurs dopaminergique D2, nous pouvons observer l'apparition d'oscillations β dans les noyaux des ganglions de la base (voir 5.11 A), avec 22 oscillations par seconde. Avec une augmentation de 25% de la connectivité, on observe un début de système oscillatoire mais non persistant (voir 5.11 B) et une augmentation plus faible ne donne lieu à aucune oscillation dans le système. Ainsi l'apparition d'oscillations β dans le système n'est pas due à la différence d'intégration temporelle du signal par les récepteurs *AMPA* et *NMDA* ; la prise en compte des délais axonaux entre les différents noyaux des ganglions de la base (voir Tableau 5.1) est suffisante pour permettre l'apparition de ces oscillations. Cependant dans le modèle original du *BCBG*, tous les paramétrages conservés du modèles présentent des oscillations lors d'une augmentation de 20% de la connectivité dans *GP/STN*. Ainsi, si la réduction proposée permet l'apparition de ces oscillations, elle en change légèrement les conditions d'apparition.

Nous avons ensuite cherché à observer le rôle des différents noyaux dans l'apparition de ces oscillations. Des études ont montré qu'une stimulation du *STN* chez les patients parkinsoniens entraîne une disparition nette des troubles moteurs (Anderson et al. 2005), permettant de réduire la médication des patients. Ainsi il semble que le *STN* soit au coeur de l'apparition des oscillations. Nous avons donc simulé une diminution du niveau dopaminergique tout en gardant l'activité du *STN* constante au cours de la simulation. Nous avons pu observer la disparition totale d'oscillations dans les autres noyaux (voir Figure 5.11). Nous avons également testé une diminution de l'activité du *STN* au cours de la simulation et avons montré que de diminuer d'un facteur 2 l'activité du *STN* prévient

également l'apparition d'oscillations. Ce test permet de simuler la disparition d'oscillations lors d'une lésion du *STN* (Alvarez et al. 2005).

Nous avons, de la même façon, testé le rôle du *GPe* et nous avons également pu observer la disparition des oscillations lors d'une stabilisation de son activité. Une simple diminution de son niveau d'activité permet la disparition des oscillations. Cependant la perte de l'inhibition que le *GPe* exerce sur le *GPi* entraîne une forte augmentation du niveau d'activité de *GPi* contrairement à une diminution de l'activité du *STN* qui a un impact plus limité sur l'activité du *GPi*.

Ces résultats montrent l'importance des noyaux *GPe* et *STN* dans l'apparition d'oscillations β dans les ganglions de la base lors d'une diminution du niveau dopaminergique. De plus nous avons montré que la différenciation de l'intégration synaptique du signal par les récepteurs *NMDA* ou *AMPA* n'est pas clé dans l'apparition de ces oscillations puisque la prise en compte du délai axonal entre les différents noyaux est suffisante pour l'apparition d'oscillations.

5.4 Discussion

Nous avons dans cette partie proposé un nouveau modèle computationel, le *rBCBG*, des ganglions de la base issus du modèle *BCBG* (Liénard et Girard 2014). Ce modèle a été obtenu par une réduction de la dynamique d'intégration synaptique du signal entrant présente dans le modèle *BCBG*. Cette réduction a été faite dans le but de pouvoir ensuite appliquer le modèle à des données comportementales impliquant des prises de décision successives et de l'apprentissage modulé par la dopamine phasique (voir chapitre suivant). Nous avons pu déterminer que la réduction proposée permet de garder les propriétés du modèle original sur les critères d'activité au repos, de sélection de l'action et permet de reproduire les oscillations β dans les noyaux des ganglions de la base en modélisant un état parkinsonien par une diminution du niveau dopaminergique dans le *GP* et le *STN*. Nous avons donc pu montrer que le modèles *rBCBG* permet de garder une plausibilité importante avec une complexité computationelle réduite.

De plus nous avons étudié l'effet d'une diminution ou augmentation du niveau tonique de dopamine dans trois modèles des ganglions de la base en comparant son effet sur le modèle *rBCBG* ainsi que de deux modèles de la littérature : le *GPR* (Gurney et al. 2001b) et le *CBG* (Girard et al. 2008). Nous avons pu observer la singularité du modèle *rBCBG* pour qui une augmentation/diminution du niveau dopaminergique module le compromis exploration-exploitation (comme dans les deux autres modèles) mais donne une direction des effets contraires aux deux autres modèles testés. Ceci donne lieu à une prédiction expérimentale qui permettrait de tester la validité des différents modèles.

5.4.1 La sélection de l'action

Depuis notamment les travaux de Mink 1996 et Redgrave et al. 1999b, les ganglions de la base ont intensivement été étudié avec le spectre de la sélection de l'action (voir Chapitre 3). Depuis, plusieurs contrôleurs robotiques ont été créés sur la base de modèles

de ganglions de la base (Girard et al. 2003 ; Gurney et al. 2004 ; Khamassi et al. 2005), montrant les capacités de ces modèles à effectuer la sélection de l'action non pas seulement dans le cadre simplifié de simulations discrètes et parfaites, mais également dans un cadre robotique, impliquant une dynamique continue et plus complexe d'interaction avec l'environnement.

Afin de tester les capacités de sélection de notre modèle, nous avons comparé ses capacités de sélection à deux autres modèles de la littérature. Le but étant d'observer comment trois modélisations très différentes des ganglions de la base permettent la sélection et quelles sont leurs limites.

Nous avons tout d'abord pu observer que les trois modèles sont capables d'une sélection correcte de l'entrée corticale la plus forte. En effet chacun d'eux permet de satisfaire la majorité des conditions du test de Gurney et al. 2001b. Seul le *rBCBG* ne permet pas une sélection totale lorsque deux canaux avec une salience positive sont présentés. En effet, ce test a mis en évidence une spécificité du modèle *rBCBG* dans sa capacité limitée à exciter les canaux non désirés en sortie du *GPi* lorsque ceux-ci reçoivent un input cortical fort ; cela se traduit par une sélection partielle de plusieurs canaux simultanément.

Cette capacité réduite à exciter les canaux les moins pertinents peut être en partie due au fait que, dans le modèle *rBCBG*, le *STN* reçoit en entrée corticale le signal des *PTN* qui est modélisé comme un signal constant indépendant du signal de salience intégré dans les *CSN*. Ainsi il est possible que le signal excitateur du *STN* ne permette pas d'augmenter l'excitation des options les moins désirées aussi efficacement que les autres modèles dans lesquels le *STN* reçoit un signal intégrant la salience. Ainsi il est possible qu'une modélisation différente du *rBCBG* dans laquelle les *PTN* intègrent ce signal de salience soit à tester.

Nous avons également mis en évidence les limites de chaque modèle dans leur capacité à sélectionner des canaux ayant des saliences trop proches. Chaque modèle a en effet des difficultés à sélectionner la meilleure option parmi deux saliences proches. On a pu observer que *GPR* et *CBG* ont tous deux une salience limite minimale en deçà de laquelle ils ne peuvent sélectionner une option même si les options concurrentes ont une salience nulle. Cela n'a pas été trouvé chez le *rBCBG*. Par contre, nous avons également pu établir une salience limite maximale au-dessus de laquelle deux options peuvent être sélectionnées simultanément si elles ont une salience relativement proche et supérieure à cette limite. Le *GPR* n'a pas montré de limitation de ce type contrairement aux deux autres modèles. Il semble que cette absence de limite supérieure démontre une capacité supérieure à exciter les options non désirées. Bien que la salience de l'option non désirée soit forte, cette capacité d'excitation des options concurrentes permet tout de même une sélection totale de l'option associée à la salience optimale.

Nous avons donc pu mettre en évidence des différences majeures dans la capacité de sélection de l'action de ces trois modèles. Ces différences peuvent résulter de leurs différences de construction anatomiques ainsi que dans leur implémentation. Il est notamment probable que la capacité forte du modèle *GPR* à exciter les options concurrentes provienne de sa ségrégation plus forte en chemin direct et indirect, avec un chemin indirect plus marqué que dans les modèles *CBG* et a plus forte raison *rBCBG*. Le rôle du chemin indirect pouvant être interprété comme un *No-Go*, car il exerce globalement une

excitation sur les actions les empêchant d'être sélectionnées (Frank et al. 2004). Aussi si la présence de la ségrégation des chemins direct et indirect semble être bénéfique pour la sélection, elle n'est plus vue aujourd'hui comme biologiquement plausible chez le rongeur ou le primate (Calabresi et al. 2014 ; Lévesque et Parent 2005 ; Wu et al. 2000).

Lorsque les saliences en compétition sont dans le champ d'action du modèle *CBG* (Girard et al. 2008), ce dernier a également des capacités de sélection de l'action remarquables. Ces capacités à sélectionner une unique action peuvent être le résultat de la boucle thalamo-corticale présente dans ce modèle et peut permettre une intégration dans le temps de la sélection du modèle permettant un contraste plus marqué. Les deux autres modèles n'intègrent pas cette boucle. Cependant le modèle *GPR* a déjà été implémentée et étudié avec cette boucle thalamo-corticale. Humphries et Gurney 2002 ont en effet montré que l'ajout de la boucle thalamo-corticale dans le modèle *GPR* permet d'améliorer les capacités de sélection en augmentant la capacité du modèle à sélectionner des canaux avec une salience faible et augmentant également le contraste dans le *GPi* entre les canaux en compétition.

Ainsi, nous pouvons voir que la présence des chemins direct et indirect peut permettre une sélection plus nette et que l'intégration de la boucle thalamo-corticale permet d'augmenter le contraste et ainsi d'améliorer la sélection.

On peut également noter que dans le test de sélection de l'action de Gurney et al. 2001b, un meilleur score est donné au modèle lorsque des canaux ayant une salience similaire sont inhibés de façon similaire au niveau du *GPi*. Il n'est pas évident qu'une telle propriété soit optimale pour un processus de sélection de l'action. En effet, si le but final des ganglions de la base est de sélectionner une seule action, alors la sélection partielle et simultanée de deux actions peut être vue comme une propriété non désirable. Notamment dans un contexte robotique l'ajout d'une option concurrente lors d'une tâche ne doit pas modifier le cours de l'action entamée par le robot. Aussi considérer le fait de sélectionner de façon égale deux canaux ayant une salience similaire comme positif est un parti pris qui peut être discuté. Néanmoins, considérer que c'est négatif est également un parti pris étant donné qu'il existe des situations dans lesquelles le déclenchement de deux actions simultanées peut être souhaitable (e.g. marcher et parler ou lire en même temps). Enfin, il existe des données expérimentales dans lesquelles le comportement observé (e.g. saccade entre deux cibles) s'explique parcimonieusement comme la sélection simultanée de deux actions (e.g. saccade vers cible gauche et saccade vers cible droite), résultant en un comportement intermédiaire (e.g. saccade vers le milieu des deux cibles). Les propriétés du *rBCBG* peuvent donc être vues comme positives ou négatives selon le contexte.

Nous avons également mis en évidence les limites de sélection de notre modèle *rBCBG* et nous avons montré que des fréquences trop élevées (à partir de $8Hz$) du *CSN* sur plusieurs canaux entraînent une sélection totale et simultanée de ces canaux. Cependant l'entrée corticale relevée dans Georgopoulos et al. 1982 montre des taux de décharge bien plus élevés et pouvant monter jusqu'à plus de $60Hz$. Ces résultats montrent donc une limitation du modèle *rBCBG* dans sa capacité à gérer des entrées corticales trop grandes, pourtant plausibles d'un point de vue biologique. Une piste pour que le modèle puisse gérer ce genre d'entrée corticale est la prise en compte de la salience au niveau des *PTN*. Ces derniers projettent en particulier vers le *STN* qui permet le contrôle de

l'activité dans les différents noyaux du *GP*. Ainsi, la révision du modèle en intégrant un signal non stationnaire chez les *PTN* pourrait permettre l'utilisation d'entrées corticales plausibles au niveau du striatum.

5.4.2 Effet de la dopamine sur la sélection de l'action

En suivant la méthode mise en place dans Humphries et al. 2012, nous avons testé l'effet de différents niveaux de dopamine tonique sur la sélection en comparant son effet sur les trois modèles des ganglions de la base. Humphries et al. 2012 ont émis l'hypothèse que la dopamine tonique joue un rôle de facteur d'exploration/exploitation dans la sélection des ganglions de la base. Ils ont validé cette hypothèse avec le modèle *GPR*, également utilisé ici, en montrant que la sélection en sortie du *GPi* est modifiée par le niveau de dopamine tonique. Lorsque le niveau de dopamine augmente, la sélection se trouve plus largement portée par le canal ayant la plus forte entrée corticale, favorisant donc l'exploitation de l'information corticale. Au contraire une diminution du niveau de dopamine tonique se traduit par une diminution du contraste entre les différentes options en sortie du *GPi* et donc par une plus forte exploration des différentes options. Des effets similaires de la dopamine tonique ont pu être observés dans le modèle *CBG* (Girard et al. 2008). Ces deux modèles, bien que différents, supposent en grande partie une ségrégation des *MSN D1* et *D2* en chemin direct et indirect projetant respectivement vers le *GPi* et *GPe*. Notamment, les deux modèles supposent la présence du chemin direct total. Le modèle *rBCBG* prédit un effet inverse de l'effet de la dopamine avec une augmentation du niveau dopaminergique favorisant l'exploration et une diminution, l'exploitation.

Cependant la modélisation de la dopamine dans le modèle *rBCBG* est différente de celles des deux autres modèles. En effet, dans le *GPR* et *CBG*, nous avons modélisé l'effet de la dopamine tonique uniquement au niveau du striatum. Dans le modèle *rBCBG* il n'y a au départ aucune distinction entre *MSN D1* et *D2* ce qui empêche une modélisation de l'effet de la dopamine à ce niveau. Nous avons donc choisi dans un premier temps de modéliser la dopamine tonique uniquement au niveau du *STN* et du *GP* (négligée dans les deux autres modèles). Nous avons alors observé un effet de la dopamine tonique à l'opposé de celui trouvé dans le *GPR* et *CBG* en modélisant la dopamine uniquement au niveau du striatum. Nous avons donc considéré la possibilité d'un chemin direct partiel dans le *rBCBG* en supposant qu'une majorité des 82% des *MSN* projetant vers le *GPi* ait des récepteurs de types *D1*. Nous avons testé plusieurs types de ségrégations. Cela nous a permis de montrer qu'il est possible d'obtenir un effet de la dopamine tonique comparable à celui obtenue avec le *CBG* et *GPR* avec une hypothèse de forte ségrégation du chemin direct, montrant ainsi que c'est bien cette propriété de ces modèles qui est clé pour reproduire ces résultats. Cependant cette ségrégation forte du chemin direct n'est pas biologiquement plausible chez le primate et le rat (Calabresi et al. 2014 ; Lévesque et Parent 2005). De plus, dans le modèle *rBCBG*, cet effet ne s'obtient qu'en négligeant le possible impact de la dopamine tonique sur les autres noyaux du modèle.

Nous observons, dans le modèle *rBCBG*, une prédominance de l'effet de la dopamine tonique au niveau du *STN* et *GP* par opposition aux modèles *GPR* et *CBG* où la modélisation de la dopamine au niveau du *STN* et *GP* n'a qu'un effet marginal par rapport à une

modélisation dans le striatum. Nous avons donc mis en évidence différentes prédictions expérimentales sur le rôle de la dopamine, qui pourront être testées ultérieurement pour valider ou réfuter le modèle. Une modélisation au niveau du striatum avec l'hypothèse d'une séparation des *MSN* en chemin direct et indirect semble être en accord avec les résultats de Humphries et al. 2012, par opposition à la modélisation au niveau du *GP* et *STN* qui tend à prédire un effet opposé.

La littérature actuelle ne permet pas de déterminer avec certitude si le niveau de dopamine tonique a un effet direct sur l'aspect exploration/exploitation de la sélection de l'action. Cependant, une étude utilisant des souris ayant un haut niveau dopaminergique a montré qu'elles sont capables d'un apprentissage normal mais ont une capacité moindre à exploiter cet apprentissage (Beeler et al. 2010 ; Marchand et al. 2014). Ces études suggèrent donc qu'un haut niveau dopaminergique favorise l'exploration plutôt que l'exploitation. Ceci va donc dans le sens des résultats obtenus avec le modèle *rBCBG*. On notera toutefois que nous nous concentrons ici sur l'effet de la dopamine sur la sélection en sortie du *GPi*. Or, dans Humphries et al. 2012, les auteurs argumentent qu'au niveau d'un noyau recevant le signal de *GPi* (noyau *target* dans l'étude, s'apparentant au thalamus), l'effet de la dopamine peut être inversé en fonction de la force de connectivité entre le *GPi* et le noyau *target*.

Dans une autre étude récente (Fountas et Shanahan 2014), les auteurs ont montré, dans un modèle avec des neurones à spike de type Izhikevich, que sous certaines fréquences faibles d'oscillations, un faible niveau dopaminergique permet de favoriser la capacité des ganglions de la base à choisir l'option associée à l'entrée corticale la plus forte, favorisant donc l'exploitation. Cette étude bien qu'utilisant un modèle complètement différent tant par son anatomie que par la finesse de modélisation de la dynamique des neurones permet de trouver des résultats comparables aux résultats obtenus avec le modèle *rBCBG* sur l'effet de la dopamine sur la sélection. Il semble donc que plusieurs résultats tendent à montrer qu'un haut niveau de dopamine entraîne une plus forte exploration alors qu'un niveau plus faible entraîne une plus forte exploitation du signal cortical.

L'activité tonique des neurones dopaminergiques a été dans la littérature plus souvent associée à un aspect motivationnel lié à la vigueur ou au *wanting* (Berridge 2007 ; Mingote et al. 2005 ; Niv et al. 2007). D'autres études voient le niveau tonique de dopamine comme la somme des récompenses précédemment obtenues (Daw et al. 2002 ; McClure et al. 2003 ; Niv et al. 2007). Aussi nous pouvons voir ce problème via le spectre de ces deux hypothèses principales sur le rôle de la dopamine dans l'apprentissage et le comportement qui s'affrontent et coexistent aujourd'hui : l'apprentissage via l'encodage d'erreur de prédiction (Schultz et al. 1997) et la vigueur ('*wanting*' ; Berridge 2007 ; Mingote et al. 2005 ; Niv et al. 2007) de la sélection. En considérant le spectre de l'apprentissage, si la dopamine corrèle avec une erreur de prédiction, alors il est naturel de penser que l'augmentation progressive du niveau de dopamine tonique signalerait un environnement récompensant mais générant de nombreuses erreurs de prédiction. Dans un environnement encore relativement inconnu ou générant de nombreuses erreurs, faire de l'exploration peut savérer payant. Ainsi cette interprétation va dans le sens des résultats obtenus avec le modèle *rBCBG*. Cependant si on considère que la dopamine corrèle avec

la vigueur du choix, il est plus plausible de penser qu'une augmentation du niveau de dopamine devrait entraîner une plus forte exploitation telle que prédit par les modèles *GPR* et *CBG*.

Il est donc aujourd'hui difficile de déterminer quel est le véritable effet de la dopamine tonique sur la sélection et des travaux testant ces hypothèses devront être menés pour déterminer quel est le rôle exact de ce neurotransmetteur dans le compromis exploration/exploitation.

5.4.3 Oscillations et dynamique temporelle

La maladie de Parkinson entraîne une diminution du niveau dopaminergique dans les ganglions de la base. De nombreuses études ont reporté, chez des patient parkinsoniens ou des primates traités au MPTP [4] que la perte des neurones dopaminergiques entraîne une augmentation des oscillations dans le *STN* et dans le *GP* (Bergman et al. 1994 ; Brown et al. 2001).

Cependant, les paramètres liés à l'apparition de ces oscillations restent flous. Le fait que le modèle *rBCBG* permet la reproduction des oscillations en modélisant un état parkinsonien dans le modèle suggère que la réduction effectuée sur le modèle original n'affecte que marginalement l'apparition de ces oscillations. En effet, si on a pu observer des oscillations β, l'apparition de ces oscillations n'arrive qu'avec une augmentation de 30% de la connectivité du *STN* et du *GP* (avec le modèle original, une augmentation de 20% est suffisante). Cela suggère donc que la prise en compte de la différence d'intégration du signal synaptique en fonction du type de récepteurs n'est pas nécessaire à l'apparition de ces oscillations et que la seule prise en compte du délai axonal permet leur apparition.

Pour aller plus loin, nous avons montré que la simulation d'une ablation totale ou partielle du *STN* permet de supprimer ces oscillations. Ces résultats font écho aux travaux qui montrent qu'une subthalamotomie permet de réduire les symptômes moteurs chez des personnes fortement atteintes de Parkinson (Alvarez et al. 2005).

De même, nous avons montré que de fixer une activité constante sur le *STN* permet également de prévenir l'apparition d'oscillations dans d'autres noyaux. Cela montre l'importance du *STN* sur l'apparition des oscillations β dans la maladie de Parkinson. Bien que nous ne modélisons pas l'effet d'une stimulation des noyaux profonds sur le *STN*, ces résultats sont cohérents avec les études qui montrent que d'appliquer un contrôle de l'activité du *STN*, via une stimulation haute fréquence, permet d'améliorer de façon significative les symptômes des patients atteints de Parkinson, sans recourir à une ablation complète du noyau (Benabid et al. 2009).

Nous avons également vu des résultats relativement similaires sur le *GPe*, mais enlever ce noyau entraîne une modification de l'activité du *GPi* bien plus importante qu'avec le *STN*. En effet, en enlevant le signal inhibiteur du *GPe* sur le *GPi*, l'activité de ce dernier

4. MPTP : 1-mthyl-4-phenyl-1,2,3,6-tetrahydropyridine ; les singes traités au MPTP présentent les mêmes symptômes que les patients parkinsoniens. Le MPTP détruit en effet les neurones dopaminergiques dans la substance noire, simulant la maladie de Parkinson.

s'en trouve augmentée de façon importante. Cela prédirait un état apathique chez des sujets ayant une ablation du *GPe*.

Conclusion

Le modèle *rBCBG* développé dans ce travail s'est montré capable de reproduire des taux de décharge biologiquement plausibles, a montré des capacités de sélection de l'action comparables à deux autres modèles de la littérature. Nous avons également discuté du rôle possible de la dopamine tonique sur la sélection en explicitant les prédictions faites par les différents modèles. Le modèle *rBCBG* a alors montré sa singularité par rapport aux autres modèles en prédisant que plus de dopamine favorise l'exploration et non pas l'exploitation. De plus, nous avons démontré que la réduction de l'intégration synaptique du *rBCBG* n'affecte que marginalement l'apparition d'oscillations β lors de la modélisation d'un état parkinsonien.

Dans le dernier chapitre de résultats de cette thèse, nous intégrerons le signal de dopamine phasique au modèle des ganglions de la base pour tester ses capacités d'apprentissage.

Chapitre 6

Modélisation du rôle de la dopamine dans l'apprentissage dans les ganglions de la base

Sommaire

6.1 Introduction

Dans le chapitre précédent, nous avons analysé l'effet de la dopamine tonique sur la sélection de l'action et ainsi observé comment le système de sélection, contrôlé par les ganglions de la base, est modifié par ce type d'activité dopaminergique. Dans ce chapitre, nous étudierons le rôle de la composante phasique de la dopamine dans l'apprentissage et

testerons les prédictions de notre modèle sur une tâche de prise de décision de la littérature introduite par Frank et al. 2004, afin de confronter les prédictions de notre modèle avec des résultats expérimentaux.

Comme présenté dans le chapitre 2, le rôle de la dopamine phasique est associé dans la littérature à un signal d'apprentissage et de renforcement en encodant une erreur de prédiction de la récompense (Schultz 1998 ; Schultz et al. 1997 ; voir Chapitre II). La dopamine a de plus un rôle modulateur sur la connectivité entre les neurones du striatum et l'entrée corticale. Ainsi, de nombreuses études font l'hypothèse que le signal phasique de la dopamine guide la sélection de l'action des ganglions de la base en modulant les poids synaptiques cortico-striataux (Berthet et al. 2012 ; Frank et al. 2004 ; Redgrave et al. 2011).

La maladie de Parkinson est associée à une dégénérescence de l'innervation dopaminergique dans les ganglions de la base (Obeso et al. 2000). Un traitement utilisé actuellement consiste à prescrire aux patients parkinsoniens de la Levo-DOPA (L-DOPA), qui peut par la suite être transformée en dopamine, palliant le déficit dopaminergique entraîné par la mort des neurones de la *SNc*. Ainsi, les sujets parkinsoniens présentent un niveau faible de dopamine en l'absence de médication et un niveau fort avec la médication. Cette particularité fait qu'ils permettent d'étudier l'effet de la dopamine sur la sélection de l'action et l'apprentissage.

Plusieurs études montrent en effet, qu'en plus de leurs troubles moteurs, les sujets parkinsoniens présentent un déficit cognitif entraînant un changement dans la sensibilité du sujet à apprendre via des retours positifs ou négatifs (Cox et al. 2015 ; Frank et al. 2004 ; Maril et al. 2013 ; Shiner et al. 2012 ; Smittenaar et al. 2012). Cette sensibilité est modifié lorsque le patient est sous un traitement de remplacement de la dopamine. Cela montre l'effet prédominant de la dopamine pour la gestion de la sensibilité de l'apprentissage par la récompense et la punition.

Ces études reposent sur la même tâche expérimentale et observent des résultats différents.

6.1.1 Présentation de la tâche

Nous avons modélisé la tâche expérimentale utilisée dans Frank et al. 2004 et reproduite dans différentes autres études (Shiner et al. 2012 ; Smittenaar et al. 2012). Cette tâche a pour but de comparer les capacités d'apprentissage et les performances de sujets atteints de la maladie de Parkinson, avec (ON) ou sans (OFF) médication, avec des sujets sains (*Senior* ou contrôle [1]). Le but est d'observer si les différents sujets apprennent mieux à choisir un stimulus récompensant (*choose A*) ou à éviter un stimulus non récompensant (*avoid B*).

Dans cette tâche, les sujets sont confrontés à différentes paires de stimuli et chaque stimulus est associé à une probabilité de récompense entre 20% et 80% (voir Figure 6.1).

1. Ces sujets contrôle sont appelé *Senior* dans l'expérience de Frank et al. 2004 à cause de leur l'âge. Ils jouent cependant le rôle de contrôle

Le sujet doit alors choisir un stimulus de la paire qui lui est présentée. Dans l'expérience originale, chacun des 6 stimuli visuels est représenté par un hiragana (caractère japonais). Le but de l'utilisation de ce genre de symboles est d'avoir des stimuli neutres à la fois d'un point de vue symbolique et esthétique, et ainsi d'éviter que les sujets n'aient un *a priori* ou une préférence envers l'un ou l'autre stimulus. Par souci de simplicité nous considérerons les stimuli comme étant des lettres de l'alphabet de A à F.

La tâche comprend deux phases distinctes. La première phase permet l'apprentissage de l'association entre stimulus et probabilité de récompense. Au cours de cette phase, uniquement certaines paires de stimuli sont présentées (voir Figure 6.1). Durant un essai, le sujet doit choisir un stimulus et reçoit un retour positif ou négatif lui permettant d'évaluer son choix. Lors de cette phase les paires de stimuli présentés sont AB (80% 20%), CD (70% 30%) et EF (60% 40%) [2]. La difficulté de la sélection augmentant ainsi avec les paires CD et EF où la différence de probabilité de récompense entre les stimuli diminue. La phase d'apprentissage consiste en 360 essais répartis équitablement sur chaque paire. Ainsi, lorsque le sujet est confronté aux stimuli A et B, il apprend à choisir le plus souvent le stimulus A et apprend à éviter de choisir B. De même que pour les couples de stimuli (C,D) et (E,F), les sujets apprennent à choisir C et E le plus fréquemment et à éviter D et F. Cet apprentissage est contrôlé dans Frank et al. 2004 par les performances d'apprentissage qui doivent être supérieures à 65% dans le cas (A,B), de 60% dans le cas (C,D) et de 50% dans le cas (E,F). Le dernier cas permet de contrôler que le sujet n'a pas de préférences esthétiques pour le stimulus F a priori moins récompensant que le stimulus E.

La deuxième phase consiste en une phase de test où l'on évalue les performances du sujet sur de nouvelles paires de stimuli. En effet, les stimuli A et B sont couplés aux stimuli C,D,E et F. La phase de test consiste en 60 essais au total, répartis équitablement en essais avec A et avec B. Notons qu'alors, aucune information de retour n'est délivrée au sujet sur le choix lors de la phase de test, et ce afin de prévenir tout apprentissage supplémentaire. On évalue ainsi les performances de chaque sujet comme leur capacité à choisir dans ces nouvelles paires le stimulus associé à la plus forte probabilité de récompense. On appellera ainsi les performances dans les essais avec A *choose A* et les essais avec B *avoid B*. Ainsi, si le sujet a des performances importantes pour choisir A cela révèle une sensitivité importante à la récompense. S'il est plus apte à éviter le stimulus B, il est plus sensible à la punition. Si les performances sont égales sur les deux conditions, alors le sujet n'a pas de différence de sensibilité à la récompense et à la punition.

Cette tâche permet donc d'évaluer la capacité des différents sujets à apprendre par la récompense ou par la punition et de tester l'influence de la dopamine sur cet apprentissage.

Pour chaque expérience, nous avons simulé 100 occurrences de notre modèle *rBCBG* (voir Chapitre 4) sur cette tâche afin d'observer les performances comportementales en fonction du niveau de ségrégation des chemins direct et indirect ainsi que du type de sujet modélisé.

2. Les pourcentages présentés représentent la probabilité d'obtenir un retour positif.

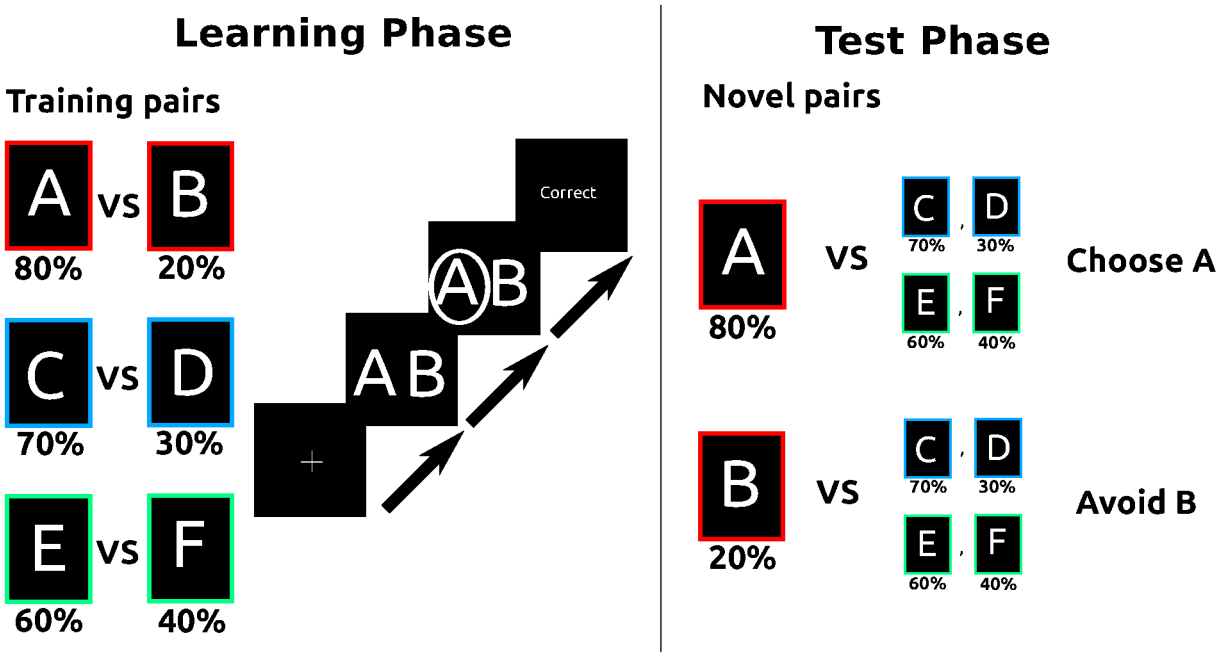

Figure 6.1 – *Illustration de la tâche utilisée dans* Frank *et al. 2004 (figure modifié de* Shiner *et al. 2012). Pour la description de la tâche, se référer au texte.*

6.1.2 Description des résultats expérimentaux

Frank et al. 2004 ont observé que les sujets sous médication ont une plus grande capacité à apprendre via la récompense, *choose A*, que les sujets contrôles et que leur capacité à apprendre via la punition, *avoid B*, est diminuée, bien que statistiquement non différente des contrôles (voir Figure6.2 A). Par opposition, les sujets sans médication ont des capacités d'apprentissage via la punition augmentées – ils sont meilleurs que les contrôles dans cette condition – et leur capacité à apprendre de la récompense diminue – sans être statistiquement moins bonne que les contrôles (voir Figure 6.2 A). Frank et al. 2004 font l'hypothèse que cette asymétrie dans l'apprentissage par la punition et la récompense

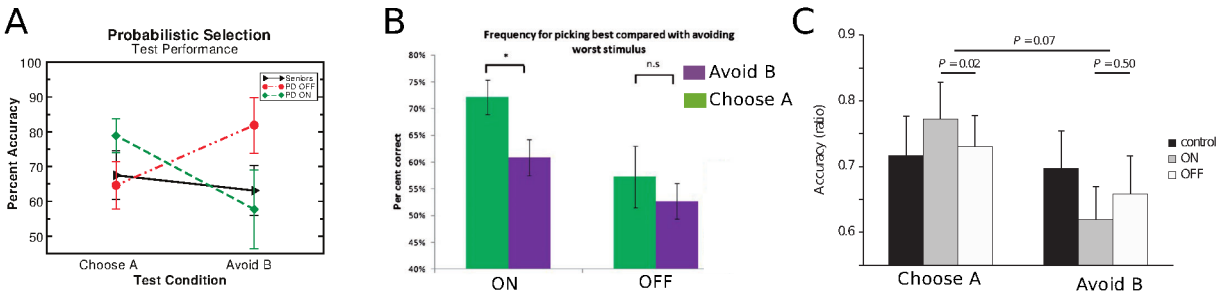

Figure 6.2 – *Résultats expérimentaux obtenus par A.*Frank *et al. 2004, B.*Shiner *et al. 2012 et C.*Smittenaar *et al. 2012 en test pour les patients parkinsonien avec ou sans médication.*

est portée par l'influence asymétrique de la dopamine sur le chemin direct, *Go*, favorisant l'approche, et le chemin indirect, *NoGo*, favorisant l'évitement. Ils proposent une règle d'apprentissage liée à la dopamine phasique qui suppose un apprentissage asymétrique sur les *MSN* en fonction de leur type de récepteurs dopaminergique. Cette règle d'apprentisage leur a permis, avec un modèle supposant la présence des chemins direct et indirect, de reproduire leurs résultats expérimentaux. Notre modèle ne faisant pas l'hypothèse d'une ségrégation franche des sites de projections des *MSN* en fonction de leurs récepteurs dopaminergiques, nous souhaitons tester ces prédictions sur cette tâche.

Plusieurs autres études ont utilisé cette tâche expérimentale pour observer ces différences comportementales (Shiner et al. 2012 ; Smittenaar et al. 2012). Cependant, les résultats obtenus ne sont que partiellement en accord avec les résultats observés initialement par Frank et al. 2004. En effet, ces études montrent bien une aptitude accrue, des patients sous L-DOPA, à avoir de meilleures performances pour aller vers la récompense que les autres groupes (voir Figures 6.2 B,C). Cependant, aucune de ces deux études n'a trouvé d'augmentation des performances sur l'évitement de la punition chez les patients parkinsoniens non traités. On observe plutôt une diminution – bien que non statistiquement significative – de ces capacités (voir Figures 6.2 B,C).

Shiner et al. 2012 ont de plus testé plus loin les différences d'apprentissage des patients sous et sans médication, en changeant l'état de médication des sujets entre la phase d'apprentissage et la phase de test. Un sujet peut avoir eu l'expérience de la phase d'apprentissage sans médication et passer la phase de test avec. Ces expériences ont montré que tous les sujets ont des performances similaires en apprentissage mais que leur performance en test dépend de leur état de médication. En particulier, les patients sous médication pendant la phase de test ont de meilleures performances que les patients sans médication, et ce indépendamment de leur état de médication durant la phase d'apprentissage.

Bien qu'en désaccord sur certains points, ces études permettent de mettre en évidence des différences comportementales liées au niveau de médication de patients par rapport à des patients sains. Par ce biais elles mettent en évidence une influence directe de la dopamine dans l'apprentissage par la récompense et la punition.

Dans ce chapitre, nous nous proposons de tester les prédictions de notre modèle sur la tâche utilisée dans ces études afin de les confronter à leurs différents résultats comportementaux. Pour ce faire nous utiliserons dans un premier temps la règle d'apprentissage proposée par Frank et al. 2004, sur notre modèle des ganglions de la base. De plus, comme dans le chapitre précédent, nous testons différents niveaux de ségrégation des chemins direct et indirect – aujourd'hui remis en question par de nombreuses études (voir Chapitre 3) – afin de mieux comprendre l'implication spécifique des récepteurs dopaminergiques dans la prise de décision. Nous verrons que la règle d'apprentissage proposée par Frank et al. 2004 ne permet pas d'atteindre de bonnes performances lorsque le degré de ségrégation D1/D2 est faible. Nous avons donc proposé, en accord avec différentes études de la littérature (Grace 2000 ; Grieder et al. 2012), que l'information phasique de la dopamine n'est interprétée que par les neurones ayant des récepteurs D1. Nous montrons

que, dans ces conditions, un modèle avec une faible ségrégation D1/D2 permet d'avoir de bonnes performances en apprentissage.

6.2 Méthode

6.2.1 Modélisation de l'apprentissage guidé par la dopamine dans le modèle *rBCBG*

Les neurones dopaminergiques de la *SNc* et de *VTA*, projettent massivement vers le striatum et la dopamine a un effet modulateur sur la connectivité synaptique entre le cortex et le striatum (Redgrave et al. 2011 ; voir Figure 6.4). Cette modulation est supposée guider et faciliter la sélection de l'action réalisée par les ganglions de la base.

Ainsi, dans les modèles de Frank et collègues, l'apprentissage via la récompense se fait principalement via l'influence de la dopamine sur les poids synaptiques cortico-striataux (Frank et Claus 2006 ; Frank et al. 2004 ; Frank et al. 2007 ; Frank 2005). Leur modèle repose sur le *framework* Leabra, développé par O'Reilly 1998. Nous avons repris le formalisme de cet apprentissage via l'effet de la dopamine dans les poids corticaux-striataux.

Afin de calculer le changement de connectivité dans les poids corticaux-striataux, le système est simulé durant deux phases distinctes de 200ms au cours de chaque essai de la tâche. Lors de la première phase, la phase *minus*, aucun signal dopaminergique n'est envoyé au striatum, cette phase permet d'obtenir l'activité de chaque canal du système avant apprentissage. L'entrée corticale représente les signaux indiquant les options accessibles (0 si l'option n'est pas accessible et 1 sinon).

À la fin de cette première phase, le système sélectionne un canal, associé à un stimulus de la tâche présenté précédemment (il y a donc 6 canaux en compétition). Cette sélection se fait à partir de l'activité de sortie du *GPi* (voir Chapitre 5). Cette sélection se fait d'après la densité de probabilité $P(a_i|GPi_i)$ défini comme :

$$P(a_i|GPi_i) = 1 - \frac{GPi_i}{\sum_j GPi_j}$$

où a_i est l'action/stimulus représenté par le canal i, GPi_i l'activité de sortie du *GPi* à la fin de la phase de sélection. On définit ainsi la probabilité de choisir l'action représentée par un canal [3].

Dans la deuxième phase de l'essai, phase *plus*, le choix a été fait et le sujet a perçu un signal de retour positif ou négatif. En cas de retour positif, le canal choisi reçoit une stimulation dopaminergique au niveau des *MSN*, en ayant un effet excitateur sur les neurones ayant des récepteurs de type D1, et un effet inhibiteur sur les neurones ayant des récepteurs de type D2 (voir équation ci-dessous). En cas de retour négatif, on simule

3. Cette définition est légèrement différente de celle utilisée dans le Chapitre 5 mais est sensiblement équivalente.

une diminution du niveau dopaminergique par un effet inhibiteur sur les *MSN D1* et excitateur sur les *MSN D2*. Ainsi, on a :

$$\Delta V_{MSND1}(t) = \sum_{(y,n)} \psi_x^n(t) V_0^n(t) C(x,y) + DA \tag{6.1}$$

$$\Delta V_{MSND2}(t) = \sum_{(y,n)} \psi_x^n(t) V_0^n(t) C(x,y) - DA \tag{6.2}$$

avec DA représentant le signal phasique de la dopamine. Ce signal est toujours nul dans la phase *minus* (voir partie Méthode du Chapitre 5 pour la description du reste du modèle). Dans la phase *plus*, il est positif si la récompense a été obtenue et négatif sinon. Seul le signe de ce paramètre change et son intensité est fixé à 10 afin de prévenir une convergence trop rapide des poids de connexion.

Notons que l'effet d'excitation/inhibition de la dopamine, ainsi simulée sur les neurones, permet uniquement de calculer le renforcement des poids de connexion, et son action a, *in fine*, un effet modulateur. Le calcul de cette modulation se fait en calculant la différence entre l'activité des *MSN* calculée lors de la phase 2 avec stimulation dopaminergique et calculée lors de la phase 1 sans stimulation dopaminergique.

Règle d'apprentissage :

$$\Delta w_i = x_i^+ y_i^+ - x_i^- y_i^-$$

Où w_i est le poids de connexion entre l'entrée corticale et le striatum (avec des poids différents pour les *MSN D1*, *MSN D2* et *MSN D1/D2*), x_i^+,x_i^- l'activité du cortex lors de la phase plus et minus respectivement et y_i^+,y_i^- l'activité des *MSN* lors de la phase plus et minus respectivement. En pratique, dans notre simulation, l'activité corticale lors des deux phases est identique. En revanche, au travers de l'influence dopaminergique lors de la phase *plus*, l'activité des *MSN* est différente lors des deux phases.

On met par la suite à jour le poids de connexion $w_i = w_i + \Delta w_i$.

Ainsi, lorsque l'option choisie mène à une récompense, la connexion entre le canal cortical et striatal de cette option se trouvera renforcée chez les *MSN D1*, diminuée chez les *MSN D2* et inchangée chez les *MSN D1/D2*. Si l'on considère une ségrégation des chemins direct et indirect, cela renforce cette action dans le chemin direct et diminue celle du chemin indirect ayant pour effet de favoriser le choix de cette action au prochain essai similaire rencontré.

Au contraire, si un retour négatif est perçu, la connexion cortex-*MSN D1* se trouve diminuée et la connexion cortex-*MSN D2* se trouve renforcée. Dans l'hypothèse des chemins direct et indirect, cela revient à diminuer la probabilité de rechoisir cette action à la prochaine occurrence du même type d'essai (i.e. présentation des mêmes stimuli).

Les différents types de sujets sont modélisés par différents niveaux de dopamine et nous supposons un effet sur la dopamine phasique. Ainsi, lors de la phase *plus*, les *MSN* seront plus ou moins excités/inhibés, en fonction du niveau de dopamine du sujet (voir section suivante).

On notera que dans cette modélisation, l'apprentissage se fait exclusivement via l'influence de la dopamine sur les poids corticaux-striataux, tels que décrits en tant qu'apprentissage intrinsèque dans Redgrave et al. 2011 (voir Figure 6.4). Cependant, le fait que le cortex n'envoie que des informations sensorielles sans aucune valuation des options peut être une simplification étant donné le rôle du cortex préfrontal dans l'évaluation des préférences (Miller et Cohen 2001).

Dopamine phasique et récepteurs D2

Dans l'apprentissage décrit ci-dessus, ont suppose une symétrie dans la modulation du signal phasique dopaminergique sur les poids synaptiques chez les *MSN* D1 et D2 (à l'image des travaux de Frank et collègues). Or, dans la littérature certaines études tendent à montrer que le signal phasique dopaminergique est trop bref et de trop forte amplitude pour que les récepteurs D2 puissent l'intégrer ; contrairement aux récepteurs D1 ((Grace 2000 ; Grieder et al. 2012)). Ainsi nous avons également testé un apprentissage dépendant uniquement des récepteurs D1 afin de tester cette hypothèse. Dans cette modélisation, les récepteurs D2 ne sont sensibles qu'à la partie tonique du signal dopaminergique.

6.2.2 Modélisation de l'apprentissage dans la maladie de Parkinson

La maladie de Parkinson se traduit par la mort de nombreux neurones dopaminergiques dans la *SNc* entraînant un certain nombre de dérèglements dans les ganglions de la base et les processus d'apprentissage mis en jeux dans le modèle (voir Chapitre 5 ; nous rappelons les effets principaux ci-dessous).

Une diminution du niveau dopaminergique se traduit notamment par :

- Renforcement de la connectivité cortico-striatale chez les MSN ayant des récepteurs dopaminergiques D1 et diminution de la connectivité cortico-striatale chez les MSN présentant des récepteurs dopaminergiques de type D2. Dans le modèle, on suppose qu'une partie des neurones présente les deux types de récepteurs, notés *MSN D1/D2*. Nous supposons que la dopamine n'affecte pas l'activité de ces neurones qui ne sont donc pas affectés par différents niveaux de dopamine (voir Figure 6.3).
- Une diminution tonique du niveau dopaminergique traduit par un renforcement de la connectivité du *Globus Pallidus* (GP) et du noyau subthalamique (*STN* ; voir Chapitre 5).
- Nous supposons que la perte de neurones dans *SNc* affecte également l'activité phasique de la dopamine, et donc l'apprentissage possiblement lié à ce neurotransmetteur.

La perte de neurones dopaminergiques dans les aires A10 et A9 (VTA et SNc) affecte ainsi fortement le fonctionnement des ganglions de la base à la fois dans la sélection de l'action (voir Chapitre 5) ainsi que, nous le supposons, dans l'apprentissage, via son effet

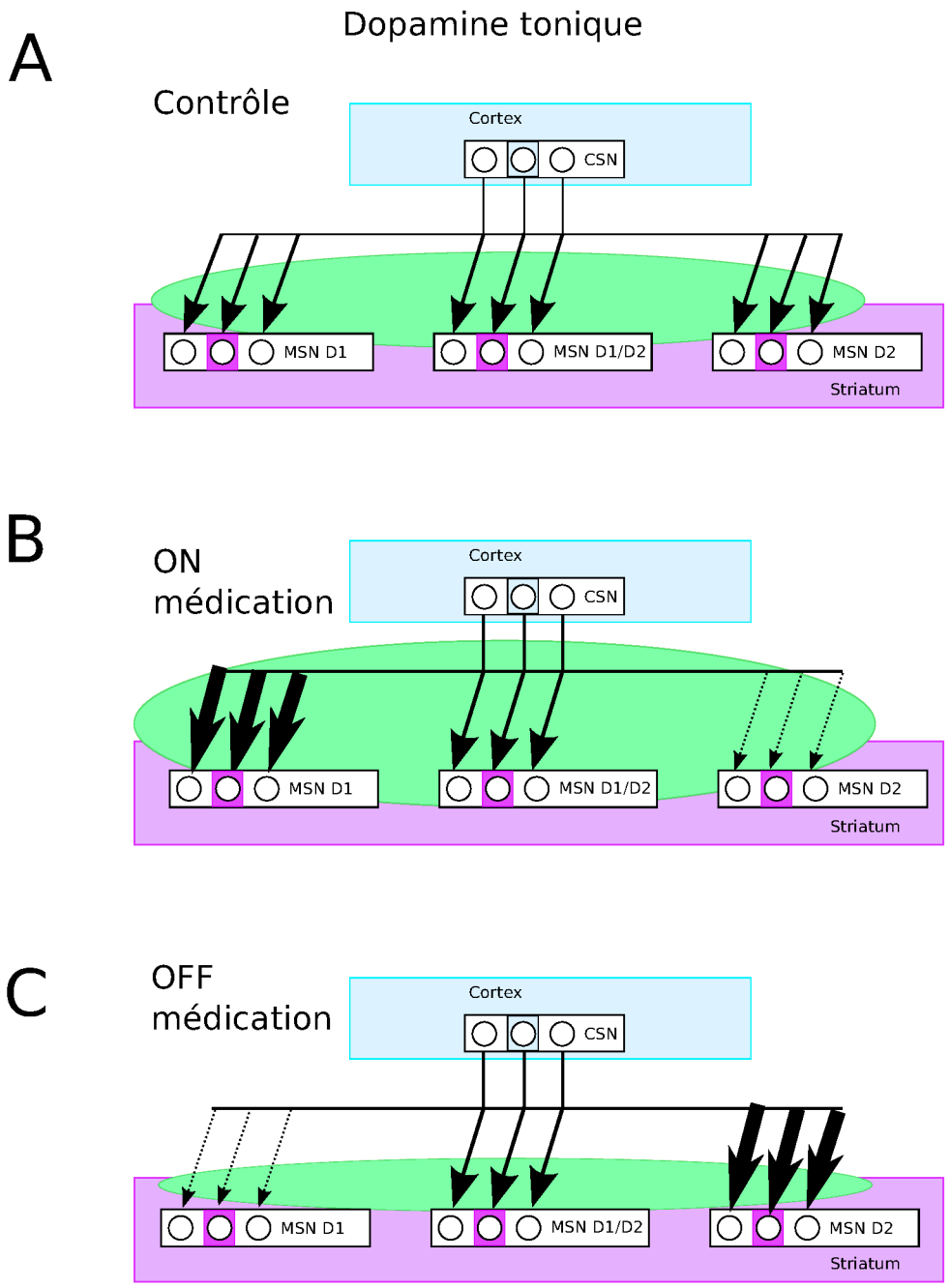

Figure 6.3 – *Illustration de l'effet de la dopamine tonique dans le modèle* rBCBG *pour les différents niveaux de dopamine modélisés. On suppose 3 actions concurrentes et une situation où l'action numéro 2 a été sélectionnée. A. Modèle contrôle ayant un niveau de dopamine tonique de base (e.g. ne modifiant pas l'activité du modèle* BCBG *de* Liénard *et* Girard *2014). B. Modèle ON médication avec une augmentation du niveau de dopamine se traduisant par une augmentation de la connexion Cortex→*MSN D1 *et une diminution de la connexion Cortex→*MSN D2*. On suppose aucun changement chez les* MSN D1/D2*. C. Modèle OFF médication avec un niveau faible de dopamine se traduisant par un renforcement des* MSN D2 *et une diminution des* MSN D1*.*

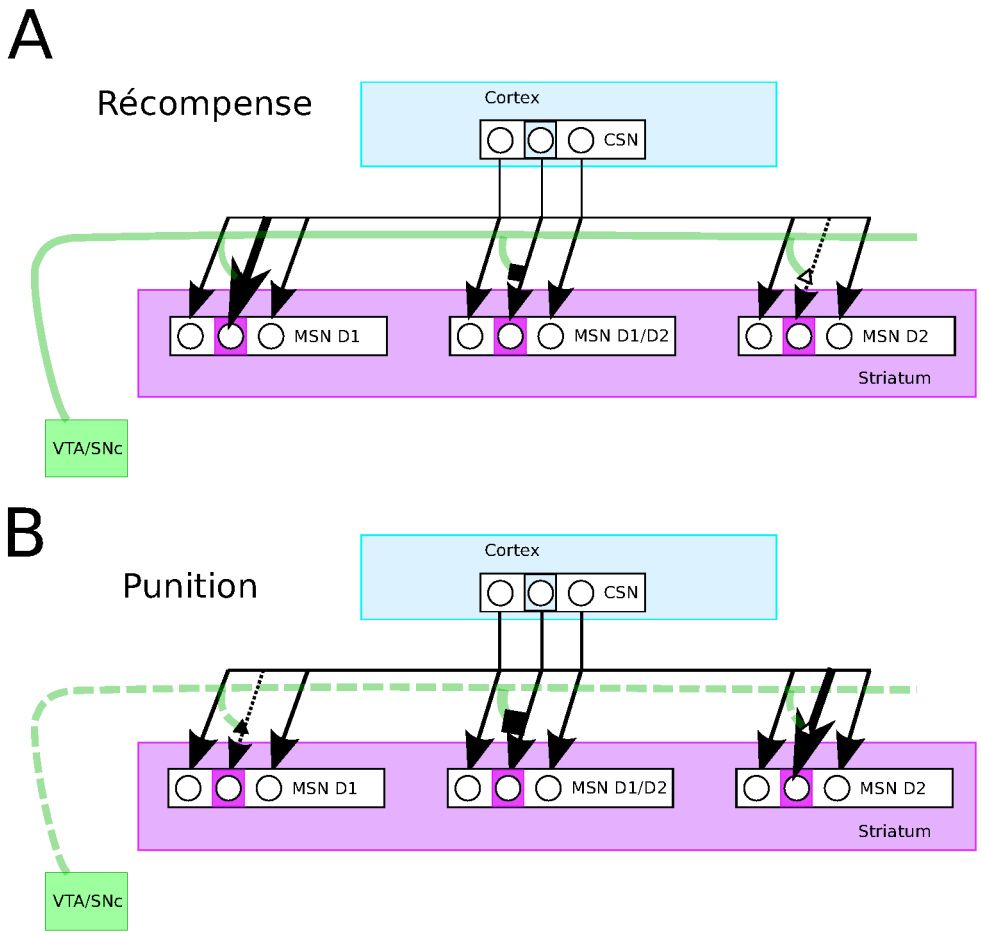

Figure 6.4 – *Illustration de l'effet de la dopamine phasique dans le modèle* rBCBG. *On suppose 3 actions concurrentes et une situation où l'action numéro 2 a été sélectionnée. A. Effet de l'augmentation phasique de la dopamine lors de la réception d'une récompense. L'excitation phasique a pour effet de renforcer la connexion entre le cortex et les* MSN D1 *dans le canal sélectionné et de diminuer celle entre le cortex et les* MSN D2 *dans ce même canal. B. Effet d'une inhibition phasique de la dopamine lors de la réception d'un retour négatif. Cela a pour effet de diminuer la connexion cortex→*MSN D1 *et d'augmenter la connexion cortex→*MSN D2 *dans le canal sélectionné.*

sur la dopamine phasique.

Modélisation des sujets sous médication

Les sujets ON médication ont un niveau de dopamine important ce qui renforce globalement la connexion entre les *MSN D1* et le cortex et diminue les poids de connexions entre le cortex et les *MSN D2*. De plus cela diminue globalement l'activité dans le *STN* ainsi que dans le *GP*. De plus nous supposons, à l'image de la règle d'apprentissage de Frank et collègues, que le niveau élevé de dopamine renforce lexcitation phasique exercée par la dopamine sur les *MSN* d'un facteur 1.3, lorsqu'une récompense est perçue. Par contre on suppose que cela diminue d'un facteur 1.3 l'effet de la dopamine phasique lors de la perception d'une récompense négative.

On a donc une asymétrie en supposant que les individus sous médication seront plus sensibles à une augmentation qu'une diminution du niveau dopaminergique. L'idée est que la LEVO-Dopa donnée à ces patients a pour conséquence une élévation globale du niveau dopaminergique rendant l'inhibition phasique de ces neurones peu perceptible lors d'une omission de récompense.

Parkinson sans médication

Les sujets OFF médication ont un niveau dopaminergique plus faible que les contrôles. Cela entraîne une diminution des poids de connexion entre le cortex et les *MSN D1* et un renforcement des poids de connexion entre le cortex et les *MSN D2*. Cela a également pour effet de renforcer la connectivité du *STN* ainsi que du *GP*. Coté apprentissage, on suppose que cela diminue à la fois l'excitation, mais égalementt l'inhibition phasique de la dopamine. Cela aura donc pour effet de diminuer l'effet de ce signal phasique sur l'apprentissage d'un facteur 1.3.

Ici on suppose que l'excitation phasique est diminuée à cause de la disparition de neurones dopaminergiques chez les sujets parkinsoniens. De plus, le faible niveau dopaminergique fait qu'une inhibition phasique n'entraînera qu'un changement faible de l'activité de base du sujet, rendant ce signal moins perceptible (Frank et al. 2004 ; Frank 2005).

Ségrégation D1/D2

Dans notre modèle *rBCBG*, nous considérons *a priori* une faible ségrégation des *MSN D1/D2* (voir Chapitre 5). En effet, dans le modèle *BCBG* complet, aucune ségrégation n'est considérée, si bien que la différenciation entre *MSN D1* et *D2* est inexistante. Cependant, comme nous l'avons vu précédemment, le degré de ségrégation de ces deux populations est encore sujet à discussion.

En l'absence de consensus sur ce thème dans la littérature, nous avons testé plusieurs niveaux de ségrégations des chemins direct et indirect (voir Tableau 6.1). Nous supposons, d'après les résultats de Lévesque et Parent 2005, que la ségrégation est relativement faible compte tenu que leurs résultats suggèrent que tous les neurones ont des terminaisons dans le *GPe* (voir Chapitre 3). Aussi, certaines ségrégations testées nous semblent plus

	Vers *GPi*			Vers *GPe*		
Modèle	*MSN D1*	*MSN D1/D2*	*MSN D2*	*MSN D1*	*MSN D1/D2*	*MSN D2*
1	42	20	20	33	33	33
2	52	10	10	33	33	33
3	62	15	5	33	33	33
4	72	5	5	33	33	33
5	82	0	0	33	33	33
6	82	0	0	0	0	100

Table 6.1 – *Niveaux de ségrégations D1/D2 dans le chemin direct. Chaque proportion est en rapport du nombre total de* MSN *présent dans le striatum. Nous gardons constante la proportion de 82% de* MSN *projetant vers le* GPi*. Nous avons testé trois versions du modèle avec uniquement des* MSN D1 *dans le chemin direct (modèle* rBCBG6*) : (a) avec un chemin indirect avec autant de* MSN *de chaque type ; (b) avec uniquement des* MSN D2 *(permet de se rapprocher au plus près de la modélisation de* Frank *et al. 2004) ; (c) uniquement des* MSN D2 *dans le chemin indirect et une connexion focalisée entre* GPe *et* GPi *(toutes les autres simulations considèrent une connexion diffuse).*

ou moins plausibles mais permettent d'étudier différents cas extrêmes du système afin d'en comprendre les propriétés et de générer des conditions expérimentales. Nous avons également testé des ségrégations totales des chemins direct et indirect, incompatibles avec plusieurs études récentes (voir Chapitre 3), mais permettant de se rapprocher du modèle de Frank et collègues et ainsi une meilleur comparaison des résultats. Nous supposons également que le nombre de *MSN D1* est comparable au nombre de *MSN D2*. Aussi, toutes les ségrégations testées ne sont pas forcément biologiquement plausibles, nous discuterons donc de ce point au regard des résultats.

6.3 Résultats

6.3.1 Modélisation d'un niveau de ségrégation faible

Performance dans la phase d'apprentissage

Dans un premier temps, nous avons testé les capacités d'apprentissage et les performances en test du modèle avec une faible ségrégation des chemins direct et indirect. La ségrégation que nous avons testée repose sur plusieurs hypothèses : (1) tous les *MSN* projettent vers le *GPe* ; (2) 82% des *MSN* projettent vers *GPi* (Lévesque et Parent 2005) ; (3) la proportion de *MSN D1* et *MSN D2* est comparable ; (4) une proportion non négligeable des *MSN* exprime les deux types de récepteurs dopaminergiques (Nadjar et al. 2006).

Ces hypothèses nous ont amené à considérer une ségrégation faible des chemins direct et indirect. Nous avons ainsi testé une ségrégation dans laquelle le chemin direct est composé

de 42% de *MSN D1*, de 20% de *MSN D1/D2* et de 20% de *MSN D2*[4](i.e. modèle rBCBG 1 dans le Tableau 6.1). Ce modèle a ensuite été simulé sous trois niveaux de dopamine simulant les trois groupes de sujets présents dans l'expérience de Frank et al. 2004 : contrôles ; ON L-DOPA ; OFF L-DOPA (voir Méthode).

Les résultats obtenus avec cette faible ségrégation des chemins direct et indirect montrent une influence importante du niveau dopaminergique dans les performances en apprentissage (voir Figure 6.5A-F). Chez les modèles contrôles, les performances sont assez faibles avec peu de sessions qui atteignent le critère de performance défini dans l'expérience de Frank et al. 2004.Seuls 21% des modèles simulés choisissent le stimulus A plus de 65% du temps et ont des performances moyennes de 57%. Dans les conditions C vs D, seules 30% des sessions atteignent le niveau de performance de 60%, avec une performance moyenne de 54% et uniquement 67% de modèles atteignent le critère de performance de 50% pour E vs F avec des performances moyennes de 51%.

Dans les modèles ON médication, les performances sont en moyenne significativement meilleures que celles des contrôles sur l'ensemble de la phase d'apprentissage. En effet, la majorité des modèles simulés atteignent les critères de performance de Frank et al. 2004 (voir Figure 6.5C,D). Sur la paire A vs B, 89% des modèles choisissent A plus de 65% du temps et en moyenne les modèles choisissent A 76% du temps. Pour C vs D, 71% des modèles ont plus de 60% de performances, avec des performances moyennes de 66%. Pour E vs F, 84% des modèles ont des performances supérieures à 50% et des performances moyennes de 58%.

Au contraire, les performances des modèles OFF médication sont en moyenne inférieures à celles des modèles contrôles sur la phase d'apprentissage (voir Figure 6.5E,F). Pour A vs B, seuls 3% des modèles ont des performances supérieures à 65% pour une performance moyenne de 49%. Pour C vs D seuls 17% des modèles ont des performances supérieures à 55% avec une performance moyenne de 49% et pour la paire E vs F, 59% des modèles ont des performances supérieures à 50% avec une moyenne de 51% de performance.

Ces résultats montrent plusieurs choses. Tout d'abord, on observe que la dopamine favorise dans le modèle de meilleures performances et un meilleur apprentissage ; ce qui semble logique compte tenu de la place centrale qu'elle a dans le modèle. On peut ainsi observer que chez les modèles ON médication, l'apprentissage est plus rapide et les performances finales sont meilleures que les contrôles (voir Figure 6.5). Au contraire en modélisant un niveau faible de dopamine, on observe un apprentissage quasi-nul avec des performances globalement constantes autour de 50%.

Ces différences dans le niveau d'apprentissage peuvent être observées au niveau de l'évolution de l'activité des différents groupes de *MSN*, ainsi que du *GPi*. Au cours de l'apprentissage, l'activité du *GPi* reflète la probabilité de choisir un stimulus plutôt qu'un autre. Ces activités sont illustrées dans la Figure 6.6 pour les différents types de modèles (ON, OFF médication et contrôle). Dans les modèles contrôles, les *MSN D1* et *D2* ont des activités comparables et de même intensité en début d'apprentissage (voir Figure 6.6). Cependant la dopamine renforce l'activité des *MSN D1* et diminue l'activité des *MSN D2*

4. Les pourcentages sont ici donnés en fonction de la proportion totale de *MSN* et non pas en proportion de neurones projetant vers *GPi*.

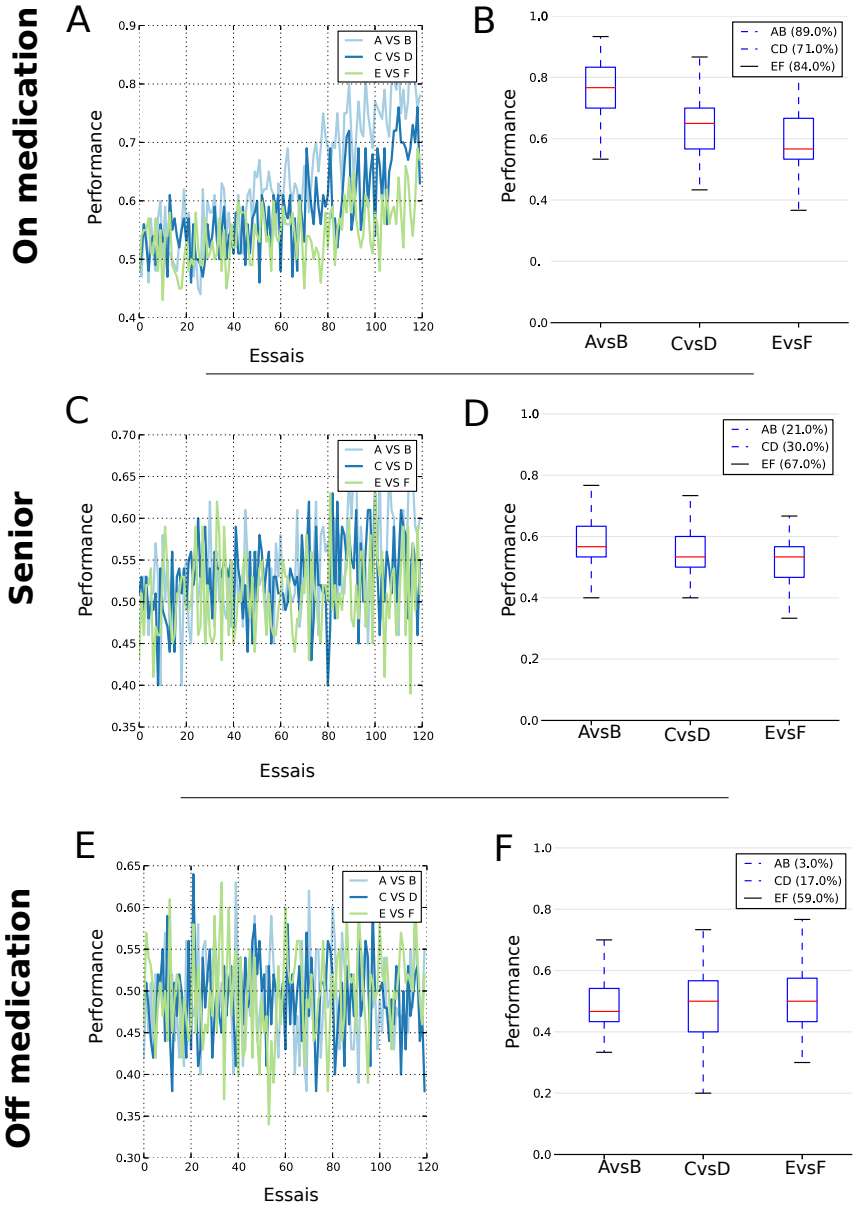

Figure 6.5 – *Résultats comportementaux obtenus en apprentissage avec une ségrégation faible du chemin direct (42%* MSN D1*, 20%* MSN D1/D2 *et 20%* MSN D2*) et indirect (modèle rBCBG 1). A,C,E. Évolution des performances au cours de la phase d'apprentissage pour les différentes paires AB, CD et EF. B,D,F. Performances finales calculées sur les 30 derniers essais de la phase d'apprentissage. Les pourcentages indiquent la proportion de modèles atteignant le critère de performance de* Frank *et al. 2004.*

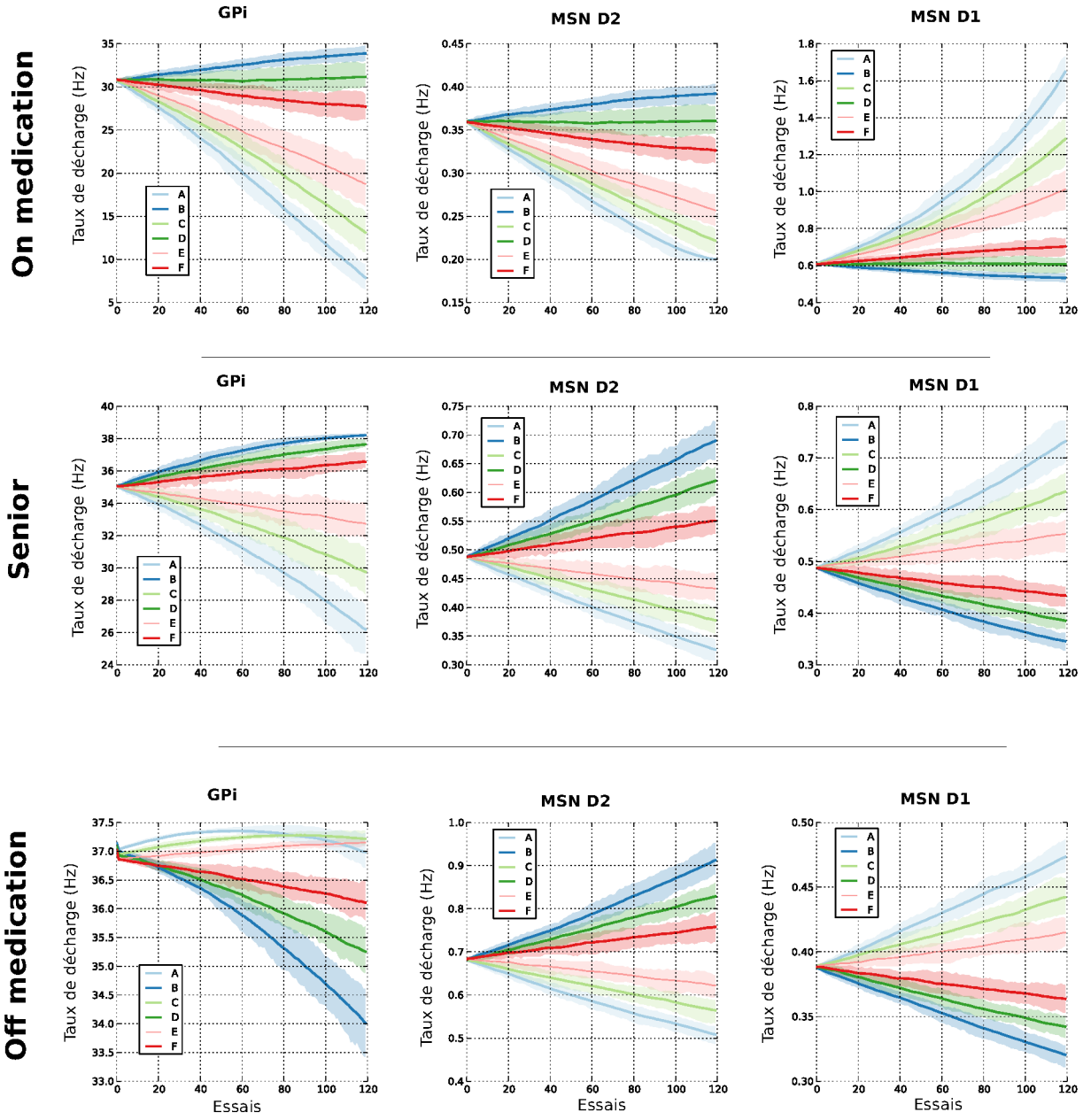

Figure 6.6 – *Évolution de l'activité enregistrée dans les noyaux de sortie et les* MSN *durant la phase d'apprentissage avec un modèle présentant une ségrégation faible des chemins direct et indirect.*

induisant cette évolution opposée de l'activité des différentes options dans ces noyaux. En sortie du *GPi* on observe ainsi que le stimulus A est assez fortement inhibé de même que, dans une moindre mesure, les stimuli C et E alors que les autres stimuli ont des activités plus grandes. Le modèle choisira donc plus facilement les stimuli A,C,E face aux stimuli B,D,F.

Dans les modèles ON médication, les *MSN D2* ont une activité réduite autour de 0.35Hz au début de la simulation et les *MSN D1* ont une activité renforcée, autour de 0.6Hz en début de simulation. Cela est dû au niveau tonique élevé de dopamine dans le modèle (voir Méthode). Lors de l'apprentissage, ont peut remarquer que les stimuli A,C,E sont plus fortement inhibés dans la condition ON que chez les contrôles au niveau des *MSN D2* et ils sont plus fortement excités au niveau des *MSN D1*, ce qui traduit la plus forte influence de la dopamine phasique sur l'apprentissage. Il en résulte que les stimuli les plus attractifs (A,C et E) sont globalement plus inhibés au niveau du *GPi* en fin d'apprentissage, permettant l'apparition de meilleures performances en ON médication qu'en contrôle.

Dans les modèles OFF médication, on a une symétrie opposée par rapport au modèle ON médication en ce qui concerne la force d'activité des *MSN*. En effet, les *MSN D2* ont une activité plus importante que les *MSN D1* (voir Figure 6.6). De plus le changement d'activité au cours de l'apprentissage est plus faible que chez les autres modèles (ON médication et contrôle). Ce qui est plus surprenant est l'évolution de l'activité du *GPi*. En effet, on peut observer que le stimulus B devient, au cours de l'apprentissage, le plus inhibé en sortie du *GPi* et donc, pour le modèle, celui qui sera choisi le plus facilement – bien qu'il soit le moins attractif. Ce biais dans l'apprentissage vient du renforcement de l'activité des *MSN D2* présents dans le chemin direct. En effet, bien que le chemin direct soit composé en majorité de *MSN D1*, le renforcement des *MSN D2* fait qu'ils prennent le pas sur les *MSN D1*. Il en résulte qu'une inhibition phasique de dopamine vient renforcer les *MSN D2* et ainsi le chemin direct. Ainsi la présence même minoritaire de *MSN D2* dans le chemin direct, peut entraîner des choix éronnés ce qui n'est pas le cas dans les différentes expériences menées sur cette tâche (Frank et al. 2004 ; Shiner et al. 2012 ; Smittenaar et al. 2012). Il est possible que la modélisation de la baisse du niveau dopaminergique soit en partie surévaluée, il est ainsi possible de s'approcher des performances des sujets contrôle en diminuant cet effet. Cependant, la présence de *MSN D2* dans le chemin direct créé un signal apprenant à aller vers des actions non récompensantes.

Ce résultat montre l'effet de l'ajout de projections des *MSN D2* dans le chemin direct qui peut amener, lors de la simulation de patients parkinsoniens, à un apprentissage biaisé en renforçant des stimuli non désirés. De plus, la dopamine a ici pour rôle d'augmenter les capacités d'apprentissage du modèle.

Performances dans la phase de test

Nous avons également observé les performances des modèles à choisir le stimulus A (*choose A*) et à éviter le stimulus B (*avoid B*) pendant la phase de test. Dans le cas contrôle, le modèle choisit le stimulus A 56% du temps et évite le stimulus B 54% du

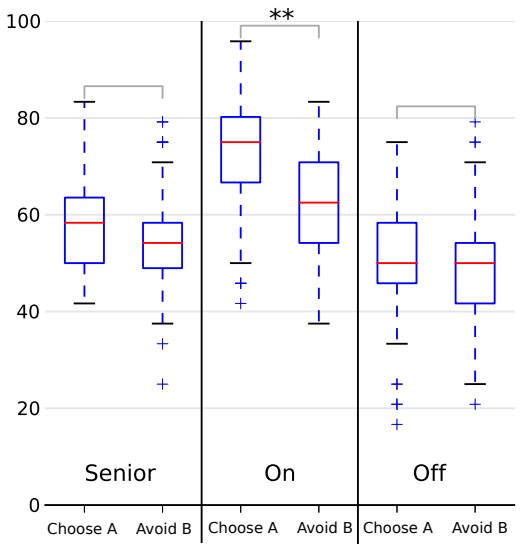

Figure 6.7 – *Score sur* choose A *et* avoid B *en phase de test sous différents niveaux de dopamine avec une ségrégation faible des chemins direct et indirect.*

temps (voir Figure 6.7). Le modèle n'est pas statistiquement meilleur sur l'un des deux critères (test Mann-Whitney, p = 0.07), ce qui est similaire au groupe contrôle dans Frank et al. 2004. Dans le cas *ON* médication, le modèle choisit le stimulus A 72.8% du temps et évite le stimulus B 61% du temps en moyenne. Dans le cas ON médication, le modèle est statistiquement plus apte à choisir A qu'à éviter B (test Mann-Whitney , $p << 0.001$). Ce résultat est comparable à ceux obtenus par Frank et collègues. Ils ont en effet montré que des patients ON médication avait une capacité accrue à choisir le stimulus A. On notera toutefois que la capacité à éviter le stimulus B a ici été légèrement renforcée en moyenne (test Mann-Whitney, $p << 0.001$).

Dans le cas *OFF* médication, les modèles ont une capacité réduite à choisir A, ainsi qu'une capacité à éviter B qui n'augmente pas par rapport au cas contrôle. En effet, les modèles OFF médication n'évitent A et ne choisissent B qu'environ 50% du temps ce qui indique un choix aléatoire (voir Figure 6.7). Cela montre une première différence importante par rapport aux résultats de Frank et collègues qui ont mis en évidence une capacité plus grande des patients OFF médication à éviter le stimulus B, ce qui n'a pas été reproduit par Shiner et al. 2012 et n'est pas non plus observé ici. Par contre nous avons bien une diminution de la capacité à choisir A – qui semble cependant due à une diminution globale des performances.

Si l'on observe un effet net de la dopamine dans l'apprentissage et une capacité accrue à choisir le stimulus A en test chez les modèles ON médication, on voit également que la capacité d'apprentissage est globalement faible sur l'ensemble des modèles, se traduisant par des performances médiocres en test et en apprentissage.

On peut se demander si le manque d'apprentissage observé ici n'est pas dû à une influence trop faible du paramètre DA dans les équations 6.1 et 6.2. À ces fins, nous avons également simulé le modèle avec un paramètre DA plus élevé que précédemment (DA = 25). Les résultats sont montrés en Annexes 8. Ils mettent en évidence que, si l'augmentation de la vitesse d'apprentissage est bénéfique sur les performances globales des modèles *senior* et *OFF* médication, elle entraîne un renforcement du biais sur l'apprentissage par la punition chez les patients OFF médication. Ces résultats confirment ainsi que la règle d'apprentissage utilisé par Frank et collègues ne supporte pas l'absence d'une franche ségrégation D1/D2. Cependant cette augmentation drastique de l'effet de la dopamine phasique entraîne un changement important de l'activité des *MSN* ; ce qui n'est pas forcément souhaitable.

Il est également probable que le fait de trouver des résultats différents de Frank et al. 2004, malgré une règle d'apprentissage proche, soit dû à cette faible ségrégation. Aussi par la suite nous avons testé les autres modèles *rBCBG* présentant une ségrégation D1/D2 plus importante.

6.3.2 Différents niveaux de ségrégations

Nous avons testé plusieurs niveaux de ségrégation en augmentant progressivement le pourcentage de *MSN D1* dans le chemin direct (rBCBG 2 à 6 dans le Tableau 6.1). Nous avons également testé plusieurs ségrégations dans lesquelles le chemin indirect est composé uniquement de *MSN D2* (rBCBG 5 et 6 dans le Tableau 6.1). Nous avons donc testé une ségrégation totale des chemins direct et indirect (modèle 6 dans le Tableau 6.1). Nous illustrons également une ségrégation dans laquelle le chemin direct n'est composé que de *MSN D1* et le chemin indirect contient tous les *MSN* (Figures 6.8,6.9,6.10 ; rBCBG 5). Ce dernier cas n'est donc pas une ségrégation totale.

Modèle	AvsB	CvsD	EF	*Choose A*	*Avoid B*
rBCBG1	76.03	65.86	57.93	72.87	61.08
rBCBG2	54.83	51.55	50.52	51.33	54.08
rBCBG3	88.34	76.45	65.66	81.07	65.03
rBCBG4	92.41	84.52	67.97	71.44	60.73
rBCBG5	94.76	86.76	69.31	91.87	71.54
rBCBG6	95.62	87.00	70.17	92.75	72.62
rBCBG6 focalisé	95.83	87.10	73.69	92.25	72.25

Table 6.2 – *Performance en apprentissage et en test des différentes ségrégations testés avec des modèles ON médication. Les trois premières colonnes indiquent les proportions de* MSN *projetant vers* GPi*, par rapport à la quantité totale des* MSN *du modèle. Les trois colonnes suivantes concernent les performances en apprentissage sur les différentes paires et les deux dernière les performances en test* choose A *et* avoid B*. La ligne en rose clair et rose vif indique que seuls les* MSN D2 *projettent vers le* GPe*. La ligne rose vif indique également que la projection entre* GPe *et* GPi *est focalisée afin de favoriser l'influence du chemin indirect sur le choix.*

Modèle	AvsB	CvsD	EvsF	*Choose A*	*Avoid B*
rBCBG1	57.86	54.48	51.66	56.96	54.25
rBCBG2	52.45	49.62	49.86	51.96	52.25
rBCBG3	69.66	62.34	56.86	62.49	57.03
rBCBG4	73.76	65.90	57.38	61.08	55.74
rBCBG5	78.10	68.72	59.17	71.67	60.63
rBCBG6	77.66	68.03	59.24	72.08	62.67
rBCBG6 focalisé	79.93	70.79	60.62	75.71	61.87

Table 6.3 – *Performance en apprentissage et en test des différentes ségrégations testés avec des modèles contrôles. Le format est identique au tableau 6.2.*

Modèle	AvsB	CvsD	EF	*Choose A*	*Avoid B*
rBCBG1	48.62	49.31	51.07	50.67	50.04
rBCBG2	48.69	49.34	49.62	50.46	49.17
rBCBG3	57.48	53.86	51.10	52.96	51.58
rBCBG4	60.59	57.72	53.59	55.06	53.61
rBCBG5	66.48	58.79	53.28	58.21	56.33
rBCBG6	63.55	58.55	55.07	58.75	55.67
rBCBG6 focalisé	68.48	63.14	55.48	61.67	57.92

Table 6.4 – *Performance en apprentissage et en test des différentes ségrégations testés avec des modèles OFF médication. Le format est identique au tableau 6.2.*

Performances dans la phase apprentissage

Comme prédit précédemment, l'augmentation du niveau de ségrégation dans le chemin direct permet d'obtenir un apprentissage plus rapide et efficace et ainsi d'obtenir des performances finales meilleures qu'avec une ségrégation faible et ce pour tous les types de modèles (voir Tableaux 6.2,6.4,6.3 pour les résultats concernant tous les niveaux de ségrégation testés).

On peut en effet noter que quel que soit le type de sujet modélisé, l'augmentation du pourcentage de *MSN D1* dans le chemin direct permet d'améliorer significativement l'apprentissage et les performances lors de la phase d'apprentissage sur les trois paires présentées. Par exemple, sur les modèles contrôles, les performances pour la paire AB passent de 57% avec une ségrégation faible (42% *MSN D1*, 20% *MSN D1/D2* et 20% *MSN D2*) à près de 80% avec une ségrégation totale sur le chemin direct. Cette effet se reproduit chez toutes les paires et chez les trois types de sujets modélisés dans cette tâche : une augmentation du niveau de ségrégation entraîne une augmentation globale des capacités d'apprentissage chez les trois groupes de sujets modélisés. On notera toutefois que la présence ou non du chemin indirect n'affecte qu'assez peu les performances.

On peut noter qu'avec uniquement des *MSN D1* dans le chemin direct, une très grande majorité des modèles ON médication et contrôles ont des performances suffisantes pour

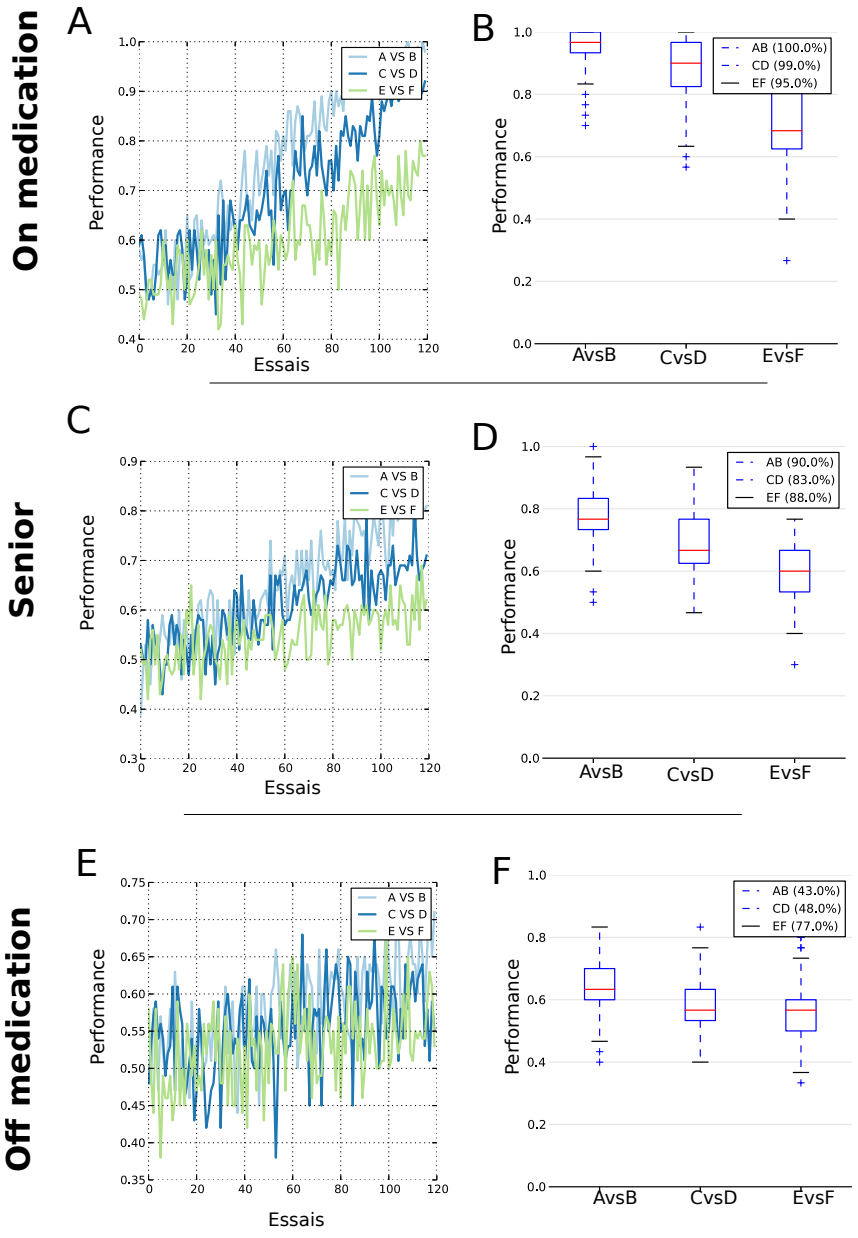

Figure 6.8 – *Résultats comportementaux obtenues avec une ségrégation totale des chemins direct et indirect (*rBCBG 6*).*

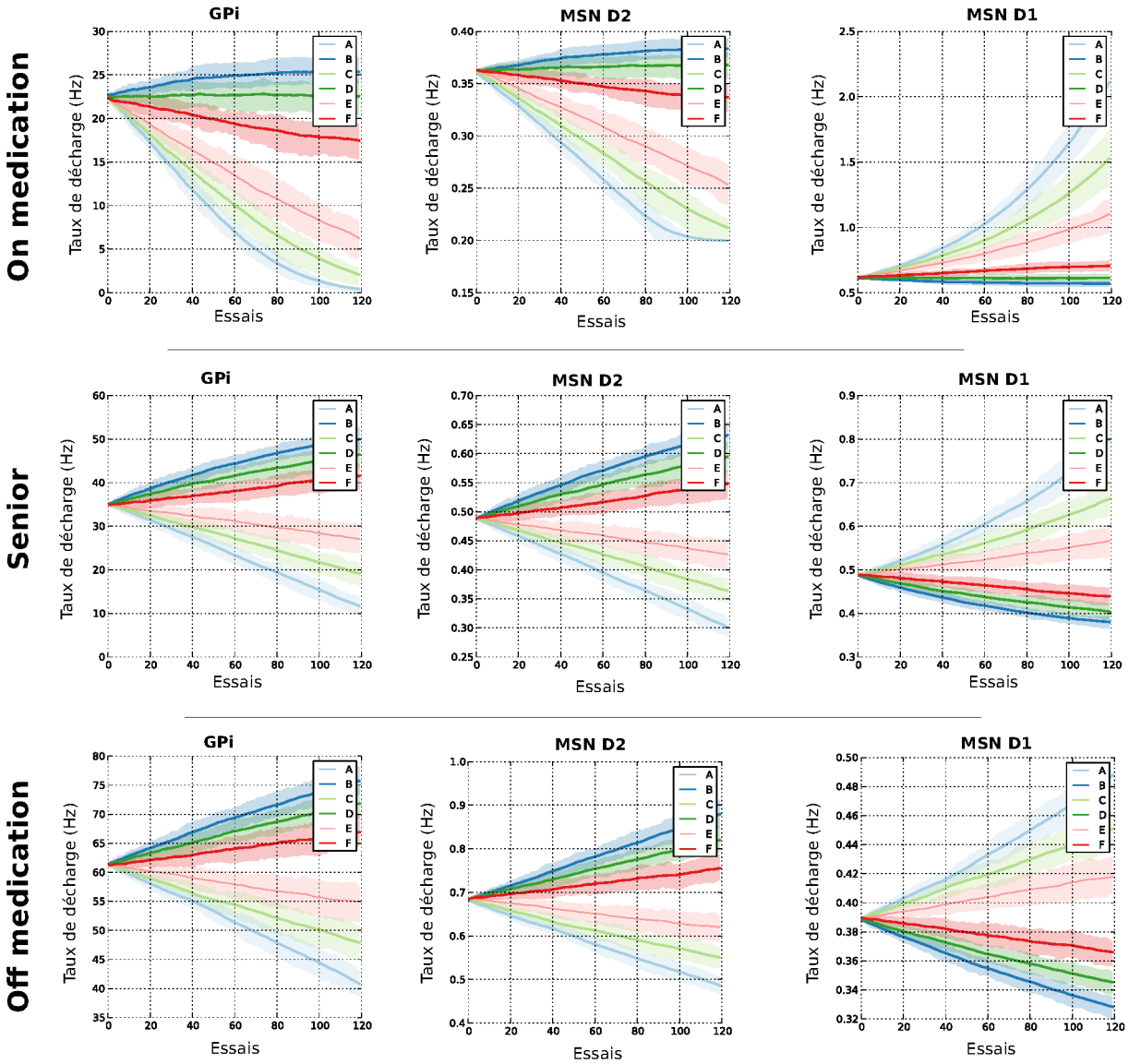

Figure 6.9 – *Évolution de l'activité enregistrée dans le noyaux de sortie et les* MSN *durant la phase d'apprentissage avec un modèle présentant une ségrégation totale des chemins direct et indirect,* rBCBG 6.

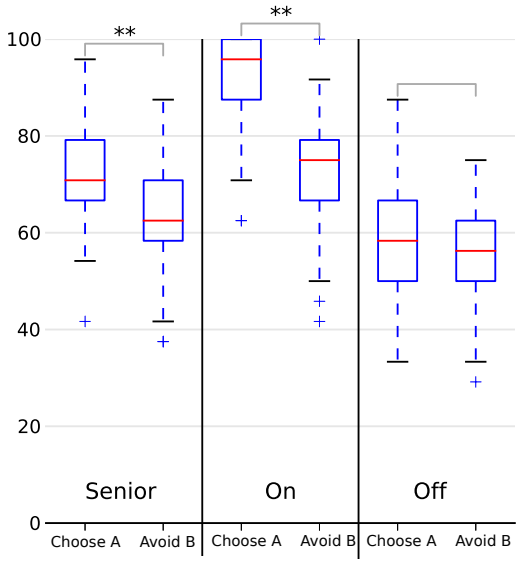

Figure 6.10 – *Score sur* choose A *et* avoid B *sous différents niveau de dopamine avec une ségrégation totale des chemins direct et indirect (*rBCBG 6*).*

satisfaire les conditions de Frank et al. 2004 (voir Figure 6.8). Cependant, seule une minorité des modèles ont des performances suffisantes pour le groupe OFF médication (moins de 50% pour A vs B et C vs D). Cette performance est néanmoins bien meilleure que les OFF medication du modèle rBCBG1 ayant une faible ségrégation dans le chemin direct.

Nous avons également observé l'activité des *MSN* et de *GPi* (voir Figure 6.9). À l'image des résultats présentés précédemment, on peut observer que l'activité des *MSN D1* chez les patients ON médication est augmentée de façon plus importante que chez les modèles contrôles et *OFF* médication. De plus le fait qu'il n'y ait plus de *MSN D2* dans le chemin direct permet, chez les modèles OFF médication, d'éviter le biais dans l'apprentissage observé avec une faible ségrégation, qui entraînait une inhibition trop importante du stimulus B au niveau du *GPi*.

Mis bout à bout, ces résultats valident qu'une ségrégation forte permet un meilleur apprentissage avec la règle d'apprentissage implémentée dans nos modèles et inspirée de Frank et al. 2004 et que le niveau de dopamine du modèle affecte grandement les performances d'apprentissage.

Performances dans la phase de test

Les résultats sur la phase de test montrent également qu'une ségrégation plus prononcée ainsi qu'un plus haut niveau dopaminergique augmente les capacités d'apprentissage du modèle et ses performances finales. Cependant avec ce niveau de ségrégation, les per-

formances du groupes OFF médication sont au delà de la chance (autour de 60% de performance). De plus les modèles contrôles montrent une préférence plus grande à la récompense que précédemment. Toutefois, à part ces quelques différences, on observe globalement les mêmes tendances principales qu'avec le modèle à faible niveau de ségrégation sur les résultats en test. En effet, on voit une augmentation de la capacité à choisir le stimulus A dans les modèle ON médication par rapport aux contrôles (voir Tableau 6.2, Figure 6.10) et il n'y a pas d'augmentation significative de la capacité des modèles OFF médication à éviter le stimulus B (voir Tableau 6.4).

Notamment, en considérant uniquement des *MSN D1* dans le chemin direct (mais pas de ségrégation dans le chemin indirect), les modèles contrôles choisissent A 71% du temps et évitent B 60% du temps (voir Figure 6.10). Les modèles ON médication ont de meilleures performances en choisissant A 91% du temps et évitent B 71% du temps – montrant une augmentation globale de performance et pas uniquement la capacité à apprendre de la récompense. Les modèles OFF médication ont toujours des performances moindres avec 58% de choix A et 56% de choix B.

Cette tendance est la même avec les deux ségrégations totales que nous avons testées (voir Tableaux 6.2,6.4,6.3). Cela suggère que le fait de ne pas observer une meilleure capacité à éviter B chez les modèles OFF médication n'est pas due à l'absence de ségrégation dans le chemin indirect. Afin de tester toutes les configurations, nous avons également testé une projection focalisée plutôt que diffuse de *GPe* vers *GPi*, mais cela n'a pas changé la dynamique du système dans l'apprentissage par la punition.

Notre modèle prédit donc qu'une augmentation du niveau dopaminergique affecte bien les performances à choisir un stimulus associé à une forte probabilité de récompenses, conformément aux résultats obtenus par différentes études (Frank et al. 2004 ; Shiner et al. 2012 ; Smittenaar et al. 2012). Cependant, nous ne prédisons aucune augmentation des performances à éviter un stimulus aversif chez les modèles OFF médication, ce qui est en accord avec les résultats de Shiner et al. 2012 et Smittenaar et al. 2012, mais en contradiction avec les résultats de Frank et al. 2004 – malgré l'utilisation d'une règle d'apprentissage inspirée de cette étude.

6.3.3 Une nouvelle règle d'apprentissage

Un point de franc désaccord entre nos résultats et les données de la littérature sont les performances des modèles (notamment OFF médication) dans la phase d'apprentissage ne permettant que rarement d'atteindre les performances seuil définies par Frank et collègues dans l'expérience originale et notamment dans le cas où l'on considère une faible ségrégation. Comme nous avons pu le constater deux raisons principales sont en cause : 1) la règle d'apprentissage utilisée fait l'hypothèse que moins de dopamine entraîne moins d'apprentissage ; 2) la présence des deux types de récepteurs dans les chemins direct et indirect, entraînent une compensation du changement d'activité – inverse sur les deux types de récepteurs, étant donné l'hypothèse d'apprentissage symétrique chez les *MSN D1* et *MSN D2* – empêchant un apprentissage unilatéral dans chaque chemin. Or, certaines études suggèrent que les récepteurs dopaminergiques D2 sont plus sensibles à un influx

constant et tonique de la dopamine. Les récepteurs dopaminergiques D1 sont quant à eux plus sensible à la composante phasique de la dopamine. Ainsi, il parait raisonnable de ne pas modéliser un apprentissage lié à la dopamine phasique chez les neurones présentant des récepteurs de type D2 (voir Méthode).

Nous avons donc testé un apprentissage unilatéral chez les neurones ayant des récepteurs D1 sur le modèle rBCBG 1 qui intègre un recouvrement plus important. Nous avons pu observer que cette nouvelle modélisation de l'apprentissage permet, dans ce modèle, d'obtenir des performances bien meilleures et comparables au modèle rBCBG 6 supposant une ségrégation forte des chemins direct et indirect (voir Figures 6.8 et 6.11).

En phase d'apprentissage une majorité des modèles simulés ont des performances suffisantes pour satisfaire les conditions de performances introduite par Frank et al. 2004 (voir Méthode). Chez les modèles ON médication 90% des modèles en moyenne satisfont ces conditions pour les trois paires d'apprentissage et 75% en moyenne pour les modèles contrôles (voir Figures 6.11 A-D). Seuls 50% modèles OFF médication satisfont ces critères de performances pour les paires AB et CD et 76% pour la paire EF nécessitant un apprentissage moins important.

Dans la phase de test on notera des performances sur le critère *choose A* toujours plus élevées que pour *avoid B* chez tous les modèles, bien que dans le cas OFF médication cette différence soit minime (voir Figure 6.12). On observe une augmentation significative des performances à choisir A et éviter B chez les patients ON médication, en comparaison des modèles contrôles. Toutefois, l'augmentation de ces performances est bien plus importante sur le critère *choose A*. Chez les modèles OFF médication on observe une diminution des performances sur le critère *choose A*. Le critère *avoid B*, n'est que légèrement diminué.

La modélisation de l'apprentissage uniquement sur les *MSN D1* permet de prévenir une détérioration de l'apprentissage due à la présence de *MSN D2* dans le chemin direct, et ainsi d'avoir un modèle d'apprentissage viable sur des modèles des ganglions de la base avec une faible ségrégation des *MSN D1* et *D2*.

6.4 Discussion

Dans ce travail, nous avons testé les prédictions de nos modèles des ganglions de la base, ayant différents degrés de ségrégation des *MSN D1* et *D2* du striatum, sur la tâche de Frank et al. 2004 en modélisant des patients parkinsoniens avec ou sans médication. Nous avons supposé que le traitement influerait à la fois sur la composante tonique et phasique de la dopamine, impliquant un effet sur la sélection de l'action (voir Chapitre 5) et l'apprentissage. Notre modélisation de l'apprentissage se base sur deux hypothèses fortes : (1) l'apprentissage se fait via la modulation des poids cortico-striataux par la dopamine (Frank et al. 2004 ; Redgrave et al. 2011), (2) l'entrée corticale est un signal sensoriel indiquant la présence ou absence de chaque stimulus.

Les résultats obtenus dans notre étude sont partiellement en accord avec les résultats expérimentaux de la littérature. Nous avons en effet montré que la modélisation d'un haut niveau de dopamine permet d'augmenter la capacité du système à choisir le stimulus récompensant, ce qui a été démontré par de nombreuses études (Frank et al. 2004 ;

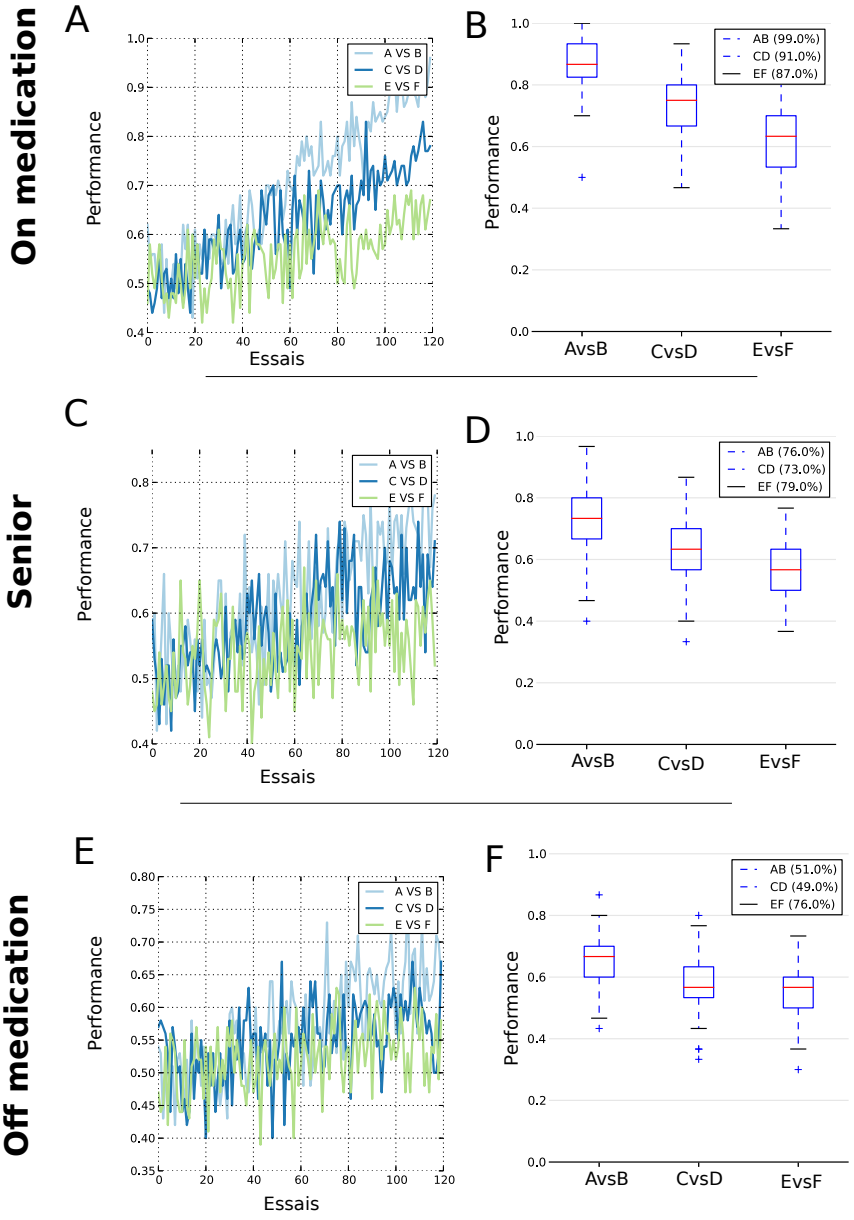

Figure 6.11 – *Résultats comportementaux obtenus avec le modèle rBCBG1 sans apprentissage sur les* MSN D2.

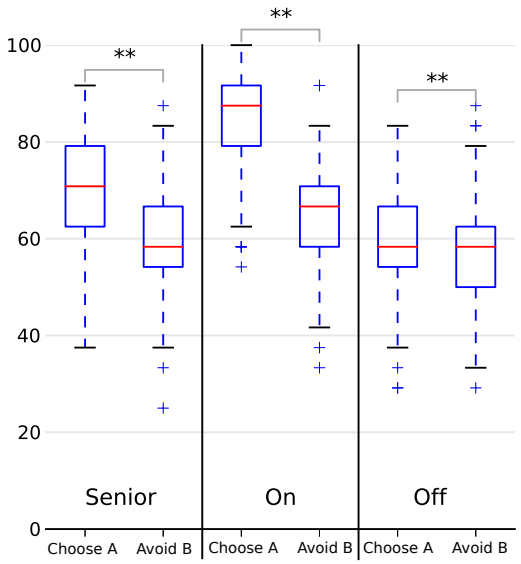

Figure 6.12 – *Résultats obtenus en test par le modèle rBCBG1 sans apprentissage sur les* MSN D2*.*

Maril et al. 2013 ; Shiner et al. 2012 ; Smittenaar et al. 2012). Un plus faible niveau dopaminergique a, au contraire, eu tendance à diminuer à la fois les performances sur la capacité à choisir le stimulus récompensant mais également à éviter le stimulus non récompensant. Le fait qu'un faible niveau dopaminergique n'augmente pas les performances du modèle à éviter la punition est en désaccord avec certaines études (Cox et al. 2015 ; Frank et al. 2004 ; Maril et al. 2013), mais en accord avec d'autres (Shiner et al. 2012 ; Smittenaar et al. 2012). Nos résultats ont également montré un lien fort entre performances et niveau dopaminergique suggérant que plus de dopamine implique de meilleures performances. Cependant comme le montre notamment les résultats de Shiner et al. 2012, le niveau de médication n'affecte pas les performances en apprentissage des sujets.

6.4.1 Faible ségrégation et apprentissage

Nos travaux ont permis de mettre en évidence que la règle d'apprentissage proposée par Frank et collègues – supposant un apprentissage asymétrique des poids corticaux-striataux entre les *MSN D1* et *MSN D2* – ne permet pas l'expression d'une asymétrie dans l'apprentissage par la récompense et la punition dans nos modèles. De plus, nous avons observé que cette règle d'apprentissage ne permet pas d'obtenir un apprentissage valide dans les modèles ayant un faible degré de ségrégation, qui sont pourtant plus proche de l'anatomie observé chez le primate (Lévesque et Parent 2005 ; voir Chapitre 3). Ainsi en accord avec des études analysant l'affinité des récepteurs dopaminergiques à la dopamine, nous avons proposé que l'apprentissage guidé par la dopamine phasique ne se répercute que sur les neurones ayant des récepteurs D1 (Grace 2000 ; Grieder

et al. 2012). Nous montrons que l'absence de ségrégation des chemins est viable dans ces conditions.

La présence de *MSN D2* dans la connexion striatum→*GPi* a un effet négatif sur l'apprentissage, si on utilise une règle d'apprentissage tel que proposée dans Frank et al. 2004. Si on considère un apprentissage symétrique par les récepteurs D1 et D2, un renforcement de l'un entraîne une diminution de l'autre. Le fait que les projections ne soit pas ségréguées entraîne une compétition entre les deux apprentissages. Or, lors d'une diminution du niveau dopaminergique on suppose un renforcement des connexions entre l'entrée corticale et les *MSN D2*, ce qui peut entraîner un biais important des *MSN D2* dans le chemin direct, ce qui est non souhaitable. Nous avons donc supposé que l'interaction entre la dopamine et les récepteurs n'est pas symétrique mais dépend de la sensibilité du récepteur à la dopamine. Les résultats expérimentaux ayant permis de montrer que la dopamine phasique exerce son influence sur les récepteurs D1 et moins sur les D2 qui, ayant une plus grande sensibilité à la dopamine, ne peuvent intégrer ces pics importants de dopamine (Grace 2000 ; Grieder et al. 2012).

Cette modélisation de l'apprentissage uniquement sur les récepteurs D1 se veut plus plausible d'un point de vue biologique et la prise en compte des différences d'affinité des récepteurs peut être clé pour une meilleure compréhension du rôle de la dopamine dans les ganglions de la base. Il faudra certainement par la suite tester plus loin cette hypothèse et cela montre que les règles d'apprentissage doivent prendre en considération ces résultats biologiques car ils ont un effet non négligeable sur les prédictions des modèles.

6.4.2 Dopamine, apprentissage et performances

Un point d'importante différence entre les résultats expérimentaux et nos résultats concerne les performances lors de la phase d'apprentissage. Le lien entre dopamine, apprentissage et performances est difficile à déterminer avec précision. Si l'on s'en tient à une interprétation classique du rôle de la dopamine dans l'apprentissage alors, comme nous l'avons observé dans nos simulations, un niveau important de dopamine doit entraîner plus d'apprentissage se traduisant par des performances accrues. Cependant, les différentes études expérimentales ne montrent pas une augmentation (respectivement diminution) unilatérale des performances chez des patients parkinsonien avec (respectivement sans) médication. Les résultats de Shiner et al. 2012 montrent que les différents niveaux de médications n'affectent pas les performances lors de la phase d'apprentissage de la tâche, mais affectent les capacités des sujets à choisir en test le stimulus attractif, ou a éviter le stimulus punitif. Cela suggère qu'un haut niveau de dopamine n'est pas nécessaire à l'apprentissage de la tâche, mais est nécessaire pour maximiser la performance en test.

Ce résultat de Shiner et collègues est pour le moins surprenant car il remet en perspective le rôle de signal d'apprentissage attribué à la dopamine. Le fait que le niveau de médication n'influence pas les performances en phase d'apprentissage suggère en effet que le niveau de dopamine n'a pas d'effet sur l'apprentissage. Cependant, le niveau de médication influe sur les performances lors de la phase de test ce qui suggère un effet de la dopamine dans la capacité des sujets à généraliser l'apprentissage précédent sur les nouvelles paires de stimuli.

Computationnellement parlant, cela signifierait que la dopamine joue un rôle important dans la représentation des états pour la sélection de l'action. Elle pourrait avoir un rôle important dans la factorisation des états permettant une plus grande généralisation de l'apprentissage. La représentation différente des états pourrait par exemple faire partie d'un apprentissage cortical non pris en compte dans cette étude. Le cortex orbito-frontal notamment est supposé permettre d'encoder les valeurs des différents stimuli (Gottfried et al. 2003 ; Hare et al. 2008 ; Padoa-Schioppa et Assad 2006) et également de différencier différents état ou différents contextes (Schoenbaum et al. 2011 ; Takahashi et al. 2011). On peut supposer que la variation du niveau dopaminergique aura également une influence sur la représentation des états et l'évaluation de la valeur au niveau cortical qui pourrait expliquer ces capacités de généralisation. Cependant cette hypothèse devra être testée dans de futures études.

De façon plus étonnante encore, le fait que les performances en apprentissage ne soient pas dépendantes du niveau dopaminergique suggère que ce dernier n'influence pas les capacités d'apprentissage à l'aide d'une information de retour (ou *feedback*), rôle normalement attribué à la dopamine phasique. Il est possible que le traitement de remplacement de la dopamine dans la maladie de Parkinson n'affecte que la composante tonique de la dopamine, n'influençant pas la composante phasique. On peut donc supposer que l'élévation artificielle du niveaux dopaminergique n'affecte pas l'apprentissage directement.

6.4.3 Évitement de la punition et chemin indirect

Un des enjeux de cette étude est de mieux comprendre comment la dopamine peut influer sur l'apprentissage par la récompense et la punition. Dans notre modèle, la dopamine est étroitement liée à la performance et ne permet pas un apprentissage asymétrique par la récompense ou la punition. Notamment aucun de nos modèles n'a permis d'observer une augmentation des performances en test dans l'évitement du stimulus B. Dans Frank et al. 2004, ils font l'hypothèse que le chemin indirect (ici supposé être Striatum$\rightarrow$*GPe*$\rightarrow$*GPi*) joue le rôle de No-Go. Or, le modèle *BCBG* suggère que la connexion entre les noyaux *GPe* et *GPi* est diffuse. Cependant cette connexion a une importance particulière sur l'interprétation du chemin indirect en No-Go dans les ganglions de la base. En effet, si cette connexion est diffuse, alors les canaux du *GPi* sont tous inhibés de façon uniforme par les canaux du *GPe* donnant lieu à une forme de No-Go globale ; se rapprochant plus d'un rôle de régulation de l'activité proche de celui donné au *STN*. Le rôle de cette connexion serait donc de prévenir une sélection trop rapide lorsque le niveau moyen des canaux du chemin indirect est trop élevé indiquant un choix compliqué entre multiples options à éviter.

Comme le type de connectivité entre *GPe* et *GPi* n'est pas connu de façon certaine, nous avons également autorisé une projection localisée permettant de se rapprocher de la modélisation de Frank et al. 2004. Le fait que nous n'ayons pas trouvé d'augmentation des performances en test sur le critère *avoid B* suggère que la règle d'apprentissage utilisée dans le modèle de Frank et al. 2004 n'est pas le seul facteur permettant de modéliser ces résultats.

6.4.4 Conclusion

Ce travail semble poser plus de question qu'il n'y répond. L'utilisation de notre modèle des ganglions de la base incorporant une plausibilité biologique intrinsèque a permis de reproduire une partie des résultats issus de la littérature, tels que l'augmentation des performances sur la partie de la tâche basée sur la récompense. Cependant, les hypothèses de modélisation de l'apprentissage par la dopamine phasique ne permettent pas de rendre compte de la relation entre niveau dopaminergique et performance. Cela montre la nécessité de revoir notre vision du rôle de la dopamine dans la représentation des états et la généralisation d'un contexte à l'autre, possiblement via un apprentissage cortical.

Chapitre 7

Conclusions

7.1 Résumé du travail de thèse

Durant ce travail de thèse, nous nous sommes intéressés à la place de la dopamine dans l'apprentissage et la prise de décision. Nous nous sommes concentrés sur son effet dans l'apprentissage et la sélection de l'action dans les ganglions de la base. Nous avons commencé par analyser l'activité dopaminergique dans une tâche à choix multiples en testant la similarité entre ce signal dopaminergique et le signal d'erreur de prédiction calculé par différents algorithmes d'apprentissage par renforcement (voir Chapitre 4). Nous avons mis en évidence que le signal dopaminergique enregistré était mieux reproduit avec un signal mixte composé de valeur et d'erreur de prédiction de la récompense. Lors de ce travail, nous avons de plus remarqué une possible dissociation entre l'évolution comportementale des animaux et l'activité des neurones dopaminergiques, suggérant que la convergence du signal dopaminergique n'est pas nécessaire pour la convergence comportementale. Cette étude a utilisé des modèles mathématiques éloignés de la réalité biologique. Cette méthode nous a permis de conceptualiser facilement l'information encodée par l'activité dopaminergique, mais manque de base biologique.

Aussi, dans un second temps nous avons utilisé un modèle des ganglions de la base développé précédemment par Liénard et Girard 2014, ayant une plausibilité biologique forte (voir Chapitre 5 et 6). Ce modèle est en effet basé sur un grand nombre de données anatomiques et physiologiques du primate, issues de la littérature des ganglions de la base. Il a montré que la seule prise en compte de ces paramètres biologiques permet l'émergence de capacité de sélection de l'action, et que l'apparition d'oscillations dans le *STN* résulte de la connectivité interne des ganglions de la base. De plus, contrairement à la plupart des modèles actuels, le modèle *BCBG* fait l'hypothèse d'une ségrégation faible des *MSN D1/D2* (Lévesque et Parent 2005).

Après avoir proposé une réduction de ce modèle, nous avons modélisé l'effet de la dopamine sur les différents noyaux des ganglions de la base et sur les forces synaptiques cortico-striatales afin d'observer les conséquences d'une perturbation du signal dopaminergique sur l'activité des différents noyaux, sur la sélection de l'action ainsi que sur l'ap-

prentissage. Nous avons vu que la dopamine tonique peut influer sur la sélection de l'action en terme de compromis exploration/exploitation. Cependant, selon le type de modèle utilisé, nous avons observé des effets contradictoires. Les modèles faisant une hypothèse de ségrégation forte des sites de projections des *MSN* en fonction de leurs récepteurs prédisent que la composante tonique de la dopamine favorise l'exploitation des connaissances, comme précédemment trouvé par Humphries et al. 2012. Au contraire dans notre modèle avec peu ou pas de ségrégation D1/D2, nous avons montré que la dopamine tonique favorise l'exploration (voir Chapitre 5 pour discussion).

Enfin, nous avons testé le rôle de la dopamine dans la modulation des poids cortico-striataux sur l'apprentissage dans une tâche impliquant des punitions et récompenses (voir Chapitre 6). Nous avons simulé différents sujets parkinsoniens en jouant sur le niveau de dopamine présent dans nos modèles des ganglions de la base. Nous avons dans un premier temps fait l'hypothèse que l'apprentissage induit par la composante phasique du signal dopaminergique se fait de façon symétrique chez les neurones ayant des récepteurs D1 ou D2, tel que supposé dans les modèles précédents (Frank 2005). Nous avons observé que ces hypothèses classiques de la modélisation de l'apprentissage sur les poids synaptiques cortico-striataux ne sont pas compatibles avec une faible ségrégation D1/D2 telle que celle présente dans notre modèle initial et pourtant telle que les données anatomiques chez le primate le suggèrent (Lévesque et Parent 2005 ; Nadjar et al. 2006). Nous avons ainsi proposé que l'apprentissage via la dopamine phasique n'influence que les récepteurs D1, ce qui est dû à leur plus faible affinité à la dopamine et permet d'intégrer le signal phasique de la dopamine. Au contraire, comme le suggère des études, les récepteurs D2 sont trop rapidement saturés pour intégrer ce signal intense (Grace 2000 ; Grieder et al. 2012 ; voir la partie discussion du Chapitre 6). La prise en compte de l'affinité des récepteurs D1 et D2 à la dopamine a permis d'obtenir de bonnes performances d'apprentissage sur le modèle ayant une faible ségrégation D1/D2.

Ces différents travaux nous ont permis d'explorer le rôle de la dopamine dans l'apprentissage et la sélection de l'action par l'analyse de son activité et de son effet sur un modèle des ganglions de la base. Ils nous permettent de répondre, du moins partiellement, aux différentes questions posées dans l'introduction de cette thèse concernant le lien entre la dopamine et l'apprentissage/le comportement.

Quel type de signal est encodé par l'activité dopaminergique au cours d'une tâche impliquant un choix parmi plusieurs actions possibles ?

Dans le chapitre 3, nous avons cherché à déterminer le type de signal, supposé d'erreur de prédiction, encodé par l'activité dopaminergique dans une tâche impliquant un choix de la part de l'animal. Nos résultats ont montré que l'activité dopaminergique reflète un signal mixte entre valeur et erreur de prédiction de la récompense. Ce résultat nous a empêché d'établir avec précision quel algorithme d'apprentissage par renforcement est compatible avec ce type de signal. Nous avons toutefois confirmé les suppositions de Roesch et al. 2007 en montrant que Q-learning semble le meilleur candidat. Nous avons néanmoins

montré que, contrairement à l'interprétation initiale des auteurs, Actor-Critic ne peut être mis de côté sur la seule base de ces données. Cela supporte globalement l'idée que le signal d'erreur dopaminergique ne prend pas en compte l'action future que l'animal sapprête à réaliser. Nous avons supposé que la présence explicite ou non des différents choix par différents stimuli de l'environnement pourrait être clé dans l'intégration par le signal dopaminergique du choix futur de l'animal, expliquant les résultats apparemment incompatibles de Roesch et al. 2007 et Morris et al. 2006.

Quelle est l'influence du niveau tonique de la dopamine sur la sélection de l'action ? Comment change-t-elle l'activité des noyaux des ganglions de la base ?

Dans le Chapitre 4, nous avons testé l'implication de la dopamine tonique dans la sélection de l'action. La réponse à cette question est toutefois complexe compte tenu des prédictions différentes des différents modèles de la littérature testés. De plus, si certaines études permettent de valider les prédictions de notre modèle (Beeler et al. 2010 ; Fountas et Shanahan 2014), force est de constater que nous n'avons pu apporter une réponse définitive à cette question d'un point de vue computationnel. Le rôle et l'effet exact de la composante tonique de la dopamine sur la sélection des ganglions de la base restent ainsi à découvrir.

Nous avons également pu observer comment un changement tonique de la dopamine modifie la dynamique des noyaux. Nous avons ainsi pu reproduire l'apparition d'oscillations β dans le *STN* et *GP* lorsque l'on modélise une diminution du niveau de dopamine. Cela montre que ces types d'oscillations résultent principalement de l'interaction entre *GPe* et *STN* et que l'intégration de délais de connexion plausibles dans les ganglions de la base est suffisante pour observer ces oscillations.

Ces résultats reposent toutefois sur la présence de récepteurs D2 dans les noyaux extrastriataux des ganglions de la base. Or, certains travaux montrent également la présence de récepteurs de type D5, appartenant à la famille D1, au niveau du *STN* (Rommelfanger et Wichmann 2010). Ainsi il faudra, par la suite, vérifier la dominance de l'effet des récepteurs D2 dans ces noyaux afin de valider ou non notre hypothèse.

Comment différents niveaux de ségrégation des neurones D1 et D2 affectent l'effet de la dopamine sur un modèle des ganglions de la base ?

La principale singularité du modèle *BCBG*, ainsi que de sa version réduite, est de supposer que la majorité des *MSN* projettent à la fois vers le *GPe* et le *GPi*. Ce modèle tranche donc avec un grand nombre de modèles de la littérature et nous avons souhaité observer combien cette particularité change notre vision de l'effet de la dopamine sur les ganglions de la base. Dans un premier temps, nous avons vu que le degré de ségrégation D1/D2 change l'effet de la dopamine tonique sur le facteur d'exploration et d'exploitation des connaissances dans la sélection de l'action (voir Chapitre 5). Enfin, nous avons montré que la supposition d'une faible ségrégation D1/D2 doit s'accompagner d'une réévaluation de la modélisation de l'apprentissage via la dopamine phasique dans le système (voir Chapitre 6).

La prise en compte de projections de *MSN D1* vers le *GPe* et de *MSN D2* vers le *GPi* nous oblige donc à revoir notre interprétation du fonctionnement global des ganglions de la base et de l'influence de la dopamine sur le système.

Comment modéliser les processus d'apprentissage liés à la dopamine sur un modèle reproduisant un état parkinsonien sous différents niveaux de traitement médical (tel que la LEVODOPA) ?

Comme mentionné précédemment, le niveau de ségrégation D1/D2 nous a amené à modifier la modélisation de l'apprentissage via la dopamine dans les ganglions de la base. Cependant, nous avons vu la supposition selon laquelle le niveau de médication influence directement l'apprentissage peut être une interprétation erronée étant donné la faible différence de performance des différents sujets dans les phases d'apprentissage (Shiner et al. 2012). Il semble que les différences de performances soient plus importantes dans les phases de test, dans lesquelles les sujets doivent exploiter leurs connaissances issues de la phase d'apprentissage afin de choisir les stimuli les plus récompensants dans des nouvelles paires. On peut ainsi supposer que les capacités intrinsèques d'apprentissage des sujets parkinsoniens ne sont pas directement influencées par le traitement mais que cela touche des capacités de généralisation des connaissances. Cela supposerait que le traitement n'influence pas la composante phasique de la dopamine. Cependant, la composante tonique seule ne peut expliquer les différences de performances des sujets ; du moins si celle-ci n'est modélisée qu'au niveau sous-cortical. Il faudrait ainsi explorer la possibilité que les traitements de remplacement de la dopamine, comme la LEVODOPA, jouent sur la représentation des connaissances, possiblement au niveau cortical.

7.2 Les multiples chemins de la dopamine

Nous avons observé les capacités d'apprentissage de notre modèle des ganglions de la base sur la tâche de Frank et al. 2004. Il en est ressorti que malgré l'utilisation d'une règle d'apprentissage proche de celle utilisée dans l'expérience originale, nous n'avons pu obtenir des résultats similaires dans la gestion de la punition. Il est probable que les différences observées soient intrinsèques à la construction du modèle et pas seulement imputables à la présence ou à l'absence de ségrégation ainsi qu'à la règle d'apprentissage.

Un point d'importance relevé dans notre étude est l'absence d'une influence importante du chemin indirect à permettre l'évitement et la prédominance du chemin direct dans la sélection de l'action. Dans le modèle BCBG (et dans sa forme réduite), la fonction du chemin indirect n'a ainsi plus un effet No-Go comme imaginé par Frank et collègues (Collins et Frank 2014 ; Frank et al. 2004), mais joue un rôle régulateur notamment par son interaction avec le *STN*. De plus, le fait de ne plus considérer la ségrégation D1/D2 dans les ganglions de la base, participe un peu plus à une remise en question de l'interprétation en Go-NoGo. La question est alors comment la punition et la récompense sont-elles prises en compte dans les boucles striato-nigrales ? Comment le système gère le compromis entre évitement de la punition et approche de la récompense, et quel est la

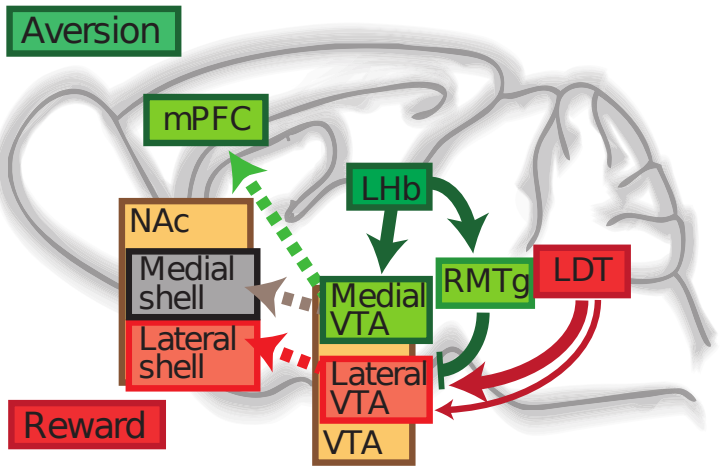

Figure 7.1 – *Les chemins parallèles de la punition et de la récompense. Illustration des chemins observés par Lammel et collègues contrôlant les comportements d'approche liés à la récompense et d'évitement liés à la punition. LhB : habénula latérale ; LDT : tegmentum latéro-dorsal ; RMTg : noyau tegmental rostro-médial. Figure reprise de* Lammel *et al. 2012* .

place de la dopamine dans le processus ?

La vision Go-NoGo des ganglions de la base permettait en effet de répondre de façon élégante à ces questions. Les travaux présentés dans Collins et Frank 2014 se fondant sur cette séparation entre apprentissage de la valeur positive des actions et apprentissage de la valeur négative de celle-ci permettent en effet d'expliquer de nombreux résultats notamment issus des expériences d'optogénétique. Cependant, ces études ont été réalisées sur des rongeurs et en majorité sur des souris, chez qui la ségrégation D1/D2 est plus forte que chez le primate (Wu et al. 2000). De nombreuses études se sont en effet penchées sur le lien entre la stimulation des différents types de récepteurs dopaminergiques et la punition/récompense. Les résultats d'optogénétique sont globalement en accord avec l'interprétation Go-NoGo des ganglions de la base de Frank et collègues en montrant qu'une stimulation des *MSN D1* permet de renforcer des associations et facilite un comportement d'approche de la part de l'animal (Kravitz et al. 2012 ; Paton et Louie 2012). On notera toutefois que des résultats indiquent que lors du mouvement les *MSN D1* et *D2* doivent être actifs, remettant en perspective la séparation franche entre D1→mouvement, D2→inhibition du mouvement (Cui et al. 2013 ; Friend et Kravitz 2014). Cela peut indiquer que l'activation des neurones D2 est nécessaire pour prévenir la sélection non désirée.

Toutefois la vision Go-NoGo semble remise en cause chez le primate, il faut donc trouver une nouvelle interprétation sur la gestion de la récompense et la punition permettant de reproduire ces résultats. Une possibilité est de considérer les résultats de Lammel et collègues (Lammel et al. 2008, 2011, 2012) montrant que différents chemins impliquant différents sous-groupes de neurones dopaminergiques sont liés à la punition ou la récompense (voir Figure 7.1). En effet, ils montrent que la stimulation via optogénétique de l'habénula latéral (LHb) entraîne un comportement d'évitement dans une tâche de con-

ditionnement par le lieu (*CPP* : *conditioned place preference*). Au contraire, l'activation du tegmentum latéro-dorsal (LDT) entraîne une préférence de l'animal pour la pièce dans laquelle la stimulation a lieu, suggérant un effet d'approche.

Les auteurs ont montré que *LHb* et *LDT* projettent vers des populations distinctes de neurones dopaminergiques de la *VTA* mettant en évidence l'implication de sous-populations de neurones dopaminergiques dans l'approche et lévitement (voir Figure 7.1). La question est de savoir si les neurones dopaminergiques liés à la récompense ou la punition projettent respectivement vers les *MSN D1* et les *MSN D2* (voir Figure 7.2B) ? Ou encore si les neurones dopaminergiques excités par la punition (Brischoux et al. 2009 ; Matsumoto et Hikosaka 2009 ; voir Chapitre 2) appartiennent au chemin lié à la punition (voir Figure 7.2) ?

Ces chemins parallèles liés à la punition et la récompense, impliquant différents groupes de neurones dopaminergiques, pourraient permettre une réinterprétation Go-NoGo de la prise de décision en lien avec les ganglions de la base.

La réévaluation de l'organisation des ganglions de la base devra de plus prendre en considération de récents travaux qui ont montré l'existence d'une connexion directe entre le *GPi* et le cortex, ce qui remet en perspective le rôle des différents chemins des ganglions de la base sur le cortex (Saunders et al. 2015). Théoriquement, la sortie du *GPi* était supposée inhiber le thalamus qui à son tour excite les régions corticales, permettant au final de faciliter le mouvement. Si certaines boucles des ganglions de la base court-circuitent le thalamus alors le contrôle de l'activité corticale sera direct au lieu de ne jouer que sur l'amplification de la réverbération du signal thalamo-cortical (Girard et al. 2006 ; voir Figure7.2 B). Cette nouvelle connexion sera donc à prendre en compte.

Dopamine dans les boucles cortico-basales

La dopamine intervient dans les différentes boucles cortico-striatales. Nous avons vu que la connectivité entre les ganglions de la base peut être décomposée en différentes boucles parallèles : limbique, associative et motrice (Haber et al. 2000 ; Joel et Weiner 2000). On a donc au moins trois chemins différents impliquant la dopamine : VTA→cortex, VTA→ventral striatum, SNc→dorsal striatum. Les différentes boucles étant liées à différents types d'évaluation, il est raisonnable de supposer que selon le chemin dans lequel les neurones dopaminergiques sont impliqués, ils encoderont une information différente. Cet aspect n'a pas été pris en compte dans ce travail de thèse. Notre analyse a tout d'abord porté sur l'étude d'un signal dopaminergique possiblement proche du système limbique, puisqu'enregistré en majorité dans VTA. Mais qu'en est-il de l'activité des neurones impliqués dans les autres boucles ?

Matsumoto 2015 fait l'hypothèse que les neurones dopaminergiques de *VTA*, associés préférentiellement aux régions limbiques et associatives, encodent un signal lié à la valeur de la récompense (possiblement de type *RPE*) et que les neurones dopaminergiques de *SNc*, associés préférentiellement aux régions motrices, encodent un signal de salience. Ainsi, les neurones de *SNc* encoderaient un signal d'attention et de motivation alors que

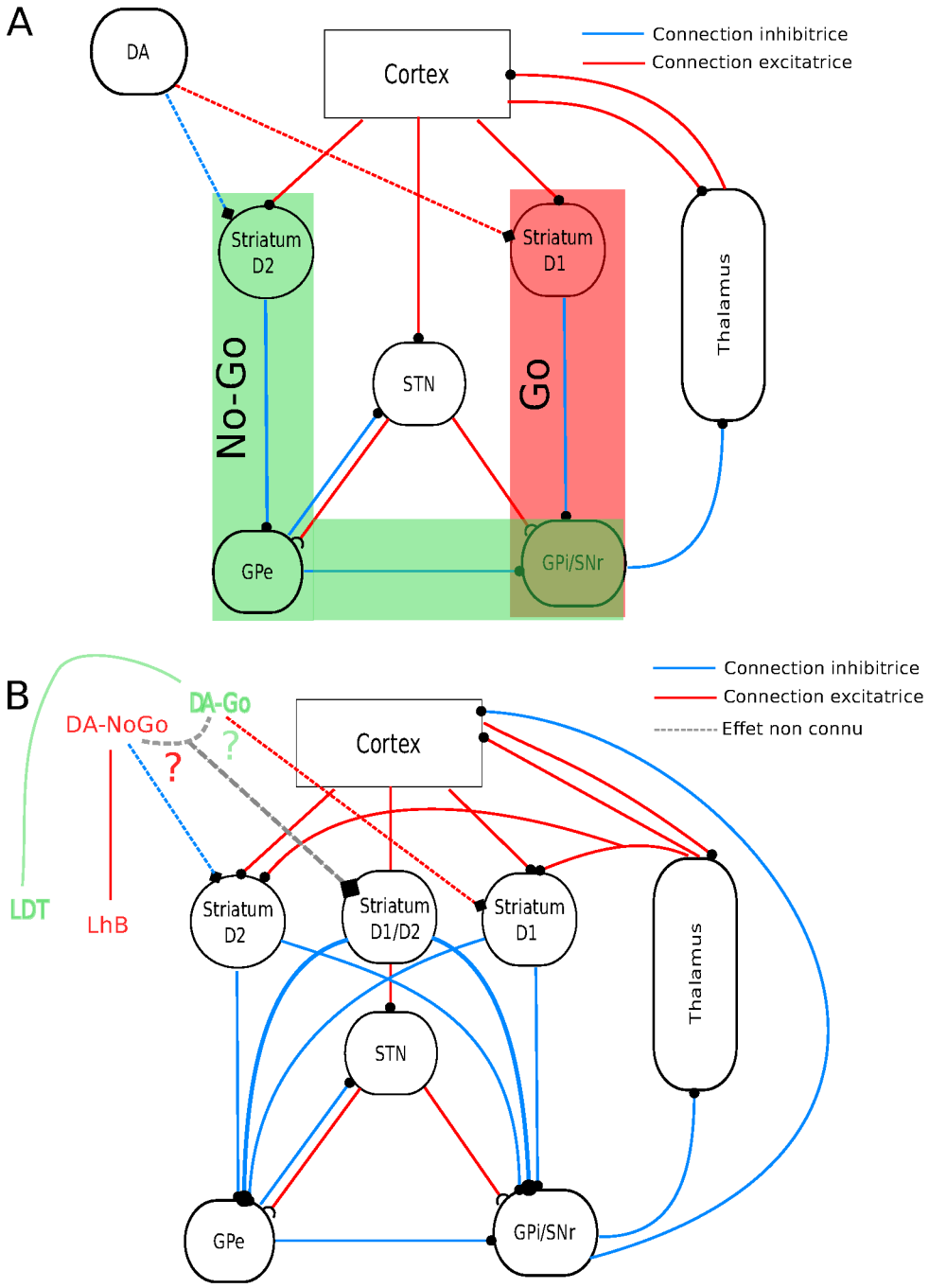

Figure 7.2 – *Illustration de la connectivité des ganglions de la base A. avec l'hypothèse Go-NoGo de Frank et collègues, B. en ajoutant les connexions probables entre les différentes populations de* MSN *vers* GPe *et* GPi. *DA GO correspond aux neurones dopaminergiques appartenant au chemin lié à la récompense et DA NoGo les neurones dopaminergiques appartenant au chemin lié à la punition.*

les neurones de *VTA* encoderaient plutôt un signal lié à la valeur de la récompense. Cette interprétation se fonde notamment sur les travaux montrant que les neurones qui sont excités à la fois par la punition et la récompense sont dans la *SNc* (Matsumoto et Hikosaka 2009). Cette interprétation est certainement simpliste, notamment car certains neurones de *VTA* sont également excités par la punition (Brischoux et al. 2009), mais permet de différencier le rôle computationnel de la dopamine dans les différentes boucles cortico-striatales.

La compréhension du rôle de la dopamine dans les différentes boucles passe également par l'établissement de modèles permettant de décrire le rôle de ces différentes boucles dans le comportement. Un modèle influant de la littérature, Daw et al. 2005, fait l'hypothèse que différentes boucles permettraient un apprentissage basé sur un modèle du monde ou non. Ils font ainsi l'hypothèse que plusieurs systèmes d'apprentissage fonctionnent en parallèle. Un système apprend à associer un état avec une action (système *model free*), quand l'autre système apprend à prédire les conséquences des actions permettant d'anticiper et de planifier à l'avance des séquences d'action (système *model based*).

Cette vision *model based/model free*, a été utilisée dans de nombreux modèles (Bornstein et Daw 2011 ; Dollé et al. 2010 ; Keramati et al. 2011 ; Lesaint et al. 2014). Cette hypothèse suppose que la partie ventrale du striatum (le noyau accumbens) est découpée en *shell model based* et *core model free* et la partie dorsale est décomposée en *DLS model free* et *DMS model based* (Bornstein et Daw 2011). L'implication de l'information dopaminergique sur ces différentes régions striatales n'a pas été étudiée avec précision. On sait toutefois que différentes sous population dopaminergiques projettent vers ces différentes régions striatales (Haber 2003 ; Haber et al. 2000). Chaque population encoderait donc différentes informations liées à cette séparation en régions *model based* et *model free* (Khamassi et Humphries 2012). Il est ainsi possible que les régions *model based* reçoivent une information de type erreur de prédiction sur les états, possiblement des neurones dopaminergiques. Ce type d'erreur étant insensible à la punition ou la récompense, les neurones excités par ces deux résultats pourraient en réalité refléter cette information permettant de mettre à jour le système *model based*.

7.3 Limitations et perspectives

7.3.1 Compléter le modèle *rBCBG*

Nous avons analysé le modèle *rBCBG* dans sa capacité à sélectionner une action et à reproduire un état parkinsonien. Cela a permis de valider le fait que la réduction du modèle *BCBG* permet de conserver les propriétés importantes du modèle. Cette réduction nous a notamment permis d'ajouter la composante apprentissage sur le modèle en ajoutant la modulation des poids synaptiques cortico-striataux par la dopamine.

La simplicité du *rBCBG* devrait également nous permettre de compléter ce modèle en ajoutant la boucle thalamocorticale. Cela nécessiterait ainsi d'ajouter deux sous-régions du noyau thalamique : le noyau ventro-latéral (*VL*) et le noyau réticulé du thalamus (*TRN* ;

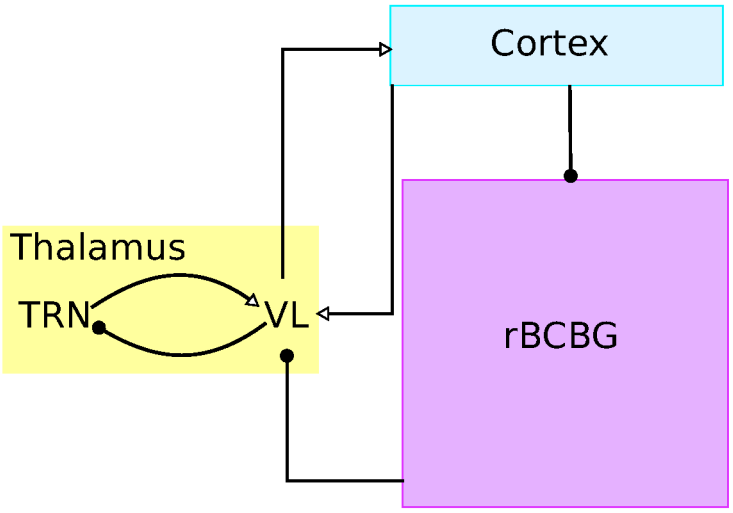

Figure 7.3 – *Illustration de la boucle thalamocorticale complétant le modèle* rBCBG.

voir Figure 7.3). Chaque connexion à rajouter ne nécessite l'optimisation que de deux paramètres : la force de connexion et le délai de transmission du signal. L'ajout de cette boucle dans le modèle original est d'une tout autre complexité puisque chaque connexion dépend de nombreux paramètres biophysiques, qui ne sont pas tous nécessairement fournis dans la littérature.

Ce que l'on peut attendre de l'ajout de cette boucle thalamo-corticale est l'amélioration des capacités de sélection (Humphries et Gurney 2002) et éventuellement l'explication des oscillations θ (3-8Hz) observées dans le cadre de Parkinson car la boucle thalamocorticale implique des délais de transmission plus longs.

Une autre limite de notre modèle est l'utilisation d'un tirage aléatoire pour finaliser la sélection. En effet, dans notre modèle des ganglions de la base, la sélection de l'action finale passe par un tirage aléatoire en fonction de la densité de probabilité déduite de la sortie du *GPi*, comme c'est classiquement fait pour les modèles d'apprentissage par renforcement (Sutton et Barto 1998). Ainsi la sélection finale n'est pas réalisée par un processus biomimétique mais mathématique. L'intégration d'un système plus plausible d'un point de vue biologique pourra s'avérer cruciale dans l'étude future des questions posées dans cette thèse.

7.3.2 Dopamine et comportement

Nos résultats nous ont amené à supposer une dissociation entre comportement et activité dopaminergique (voir Chapitre 3). Or, de nombreuses études ont également montré que la dopamine n'est pas toujours nécessaire à l'apprentissage. Cela suggère que le comportement est guidé par plusieurs processus, certains guidés par la dopamine et d'autres non. Lesaint et al. 2014 ont proposé un modèle se basant sur les travaux de Daw et al. 2005 et supposent que la dopamine n'est nécessaire que pour la mise à jour du système

model free, la partie *model based* du modèle étant indépendante de la dopamine.

Cette hypothèse se fonde sur l'étude du comportement de rats montrant différents types de comportements dans une tâche pavlovienne. Certains rats, dit *goal trackers*, apprennent à aller directement vers la récompense sans se préoccuper d'un levier indiquant l'imminence de la récompense. D'autres rats, nommés *sign trackers* vont toujours aller mordre et actionner le levier avant d'aller vers la récompense bien que l'actionnement du levier ne soit pas requis pour l'obtention de la récompense (Flagel et al. 2010). Les résultats montrent que l'activité dopaminergique des *goal trackers* ne semble pas diminuer au moment de la récompense, contrairement aux *sign trackers* pour qui l'activité dopaminergique suit le schéma d'activité observé par Schultz et collègues.

Le modèle de Lesaint et al. 2014 a permis de reproduire ce type de données et pourrait dans une certaine mesure expliquer pourquoi nous trouvons une dissociation entre comportement et activité dopaminergique.

L'analyse précise de l'évolution conjointe de la dopamine et du comportement semble aujourd'hui un point de grande importance pour résoudre les différentes interprétations liées à son activité. L'observation conjointe et précise de l'adaptation comportementale et de l'activité dopaminergique devrait nous permettre de mieux comprendre son rôle dans l'apprentissage. Il semble également important de pouvoir distinguer à quelle boucle le signal dopaminergique appartient (mésocortical, mésolimbique, nigro striatal), ce qui permettrait certainement d'observer une hétérogénéité de l'activité de la dopamine.

À ces fins, il serait intéressant d'étudier l'évolution d'enregistrements dopaminergiques de tôt jusque tard dans l'apprentissage dans une tâche impliquant un choix de la part de l'animal. En ce sens, les données de Roesch et al. 2007 ne sont pas idéales puisque la tâche contient de nombreux changements de la contingence empêchant l'apprentissage de l'animal de converger complètement.

7.3.3 Traitement à Parkinson et généralisation de l'apprentissage

Nous avons vu que le traitement à la maladie de Parkinson semble, selon certaines études, ne pas influer sur la capacité d'apprentissage des sujets, mais semble jouer sur leurs capacités à généraliser leur apprentissage sur de nouvelles paires de stimuli (Shiner et al. 2012). Cela suggère un effet du traitement de remplacement de la dopamine totalement différent de ce que l'on pourrait attendre d'une augmentation du niveau phasique ou même tonique dans les ganglions de la base.

Une hypothèse parmi d'autres serait d'étudier la possible implication d'une augmentation du niveau tonique de la dopamine dans la représentation des états. Il est possible qu'un haut niveau de dopamine tonique permette aux sujets de faire plus rapidement le lien entre la valeur d'un stimulus appris dans une paire donnée avec le même stimulus dans une autre paire. L'idée serait que les sujets perçoivent dans un premier temps les deux stimuli comme un contexte particulier, et apprennent la valeur de ce contexte $V(AB)$, $V(CD)$ et $V(EF)$. Le fait de retrouver le stimulus A face à un stimulus auquel il n'a

jamais été associé (C,D,E ou F) créera ainsi un nouveau contexte $V(AC)$ inconnu. La capacité à déduire la valeur de ce nouveau contexte à partir des autres valeurs apprises pourrait dépendre du niveau de dopamine dans le cortex préfrontal, ou par exemple dans le cortex orbitofrontal (Takahashi et al. 2011), qui permettrait de déduire de $V(AB)$ et $V(CD)$ la valeur $V(AE)$ et ainsi d'aider à la prise de décision.

Cette question est d'autant plus intéressante qu'elle nous obligera à sortir de la modélisation classique de l'effet de la médication chez ces patients. Elle pourrait également permettre de tester la relation entre dopamine et représentation de l'environnement. Cette réflexion pourrait rejoindre le fait que, lorsque deux choix sont représentés par un unique stimulus, le signal phasique de la dopamine n'encode pas le choix futur (Roesch et al. 2007), tandis que quand chaque choix est explicitement représenté par un stimulus qui lui est propre, l'activité dopaminergique reflète le choix futur (Morris et al. 2006).

Pour ce type de travaux, notre modèle des ganglions de la base ne semble toutefois pas adapté et un modèle de plus haut niveau, tel que les modèles d'apprentissage par renforcement factorisé, pourrait être approprié pour répondre à cette question (Lesaint et al. 2014).

Chapitre 8

Annexes

Actor-Critic avec deux paramètres d'apprentissage

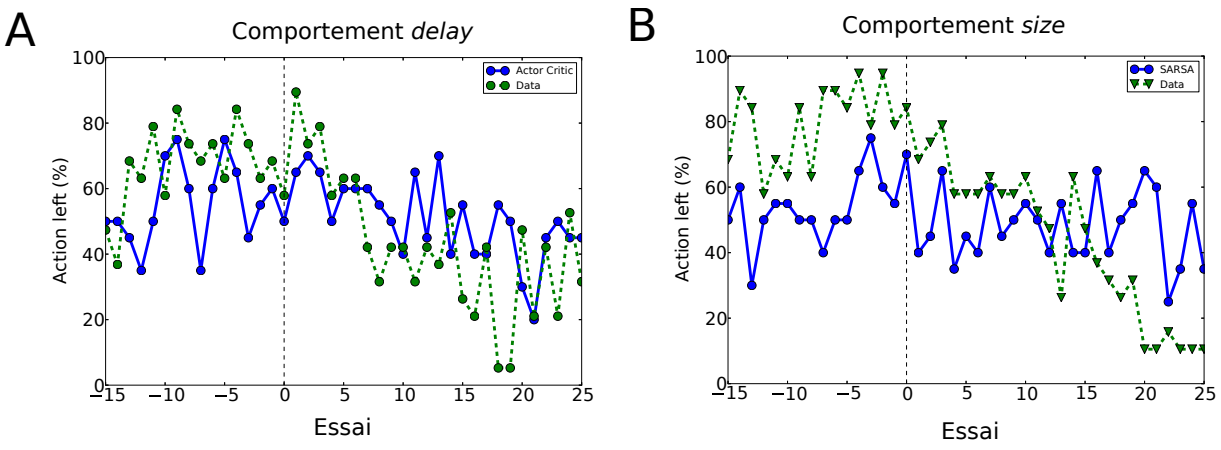

Figure 8.1 – *Meilleure reproduction du comportement avec le modèle* Actor-Critic *à 4 paramètres. Les paramètres de ce modèle sont : (*$\alpha 1 = 1$,$\alpha 2 = 0.05$,$\beta = 0.15$,$\gamma = 0.95$*). A. Reproduction du comportement durant le changement entre deux blocs* delay *et B. entre deux blocs* size*. Le faible facteur d'apprentissage* $\alpha 2$ *donne un comportement hautement exploratoire.*

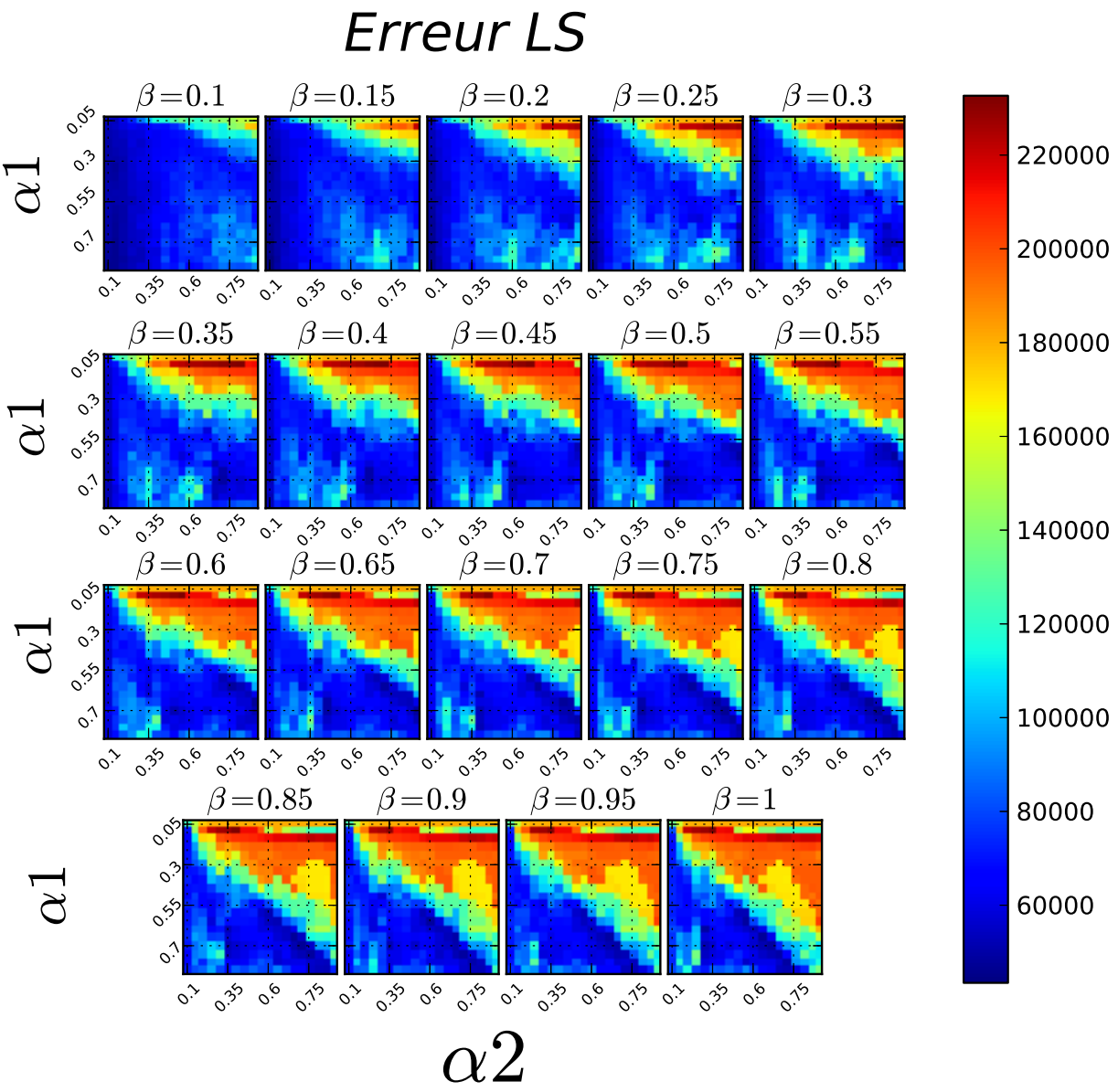

Figure 8.2 – *Erreurs* LS *obtenues avec* Actor-Critic *ayant un paramètre d'apprentissage différents pour la mise à jour du critique* ($\alpha 1$) *et le critique* ($\alpha 2$). *Ces résultats ont été obtenus pour* $\gamma = 0.95$.

Facteur d'apprentissage élevé dans le modèle *rBCBG1*

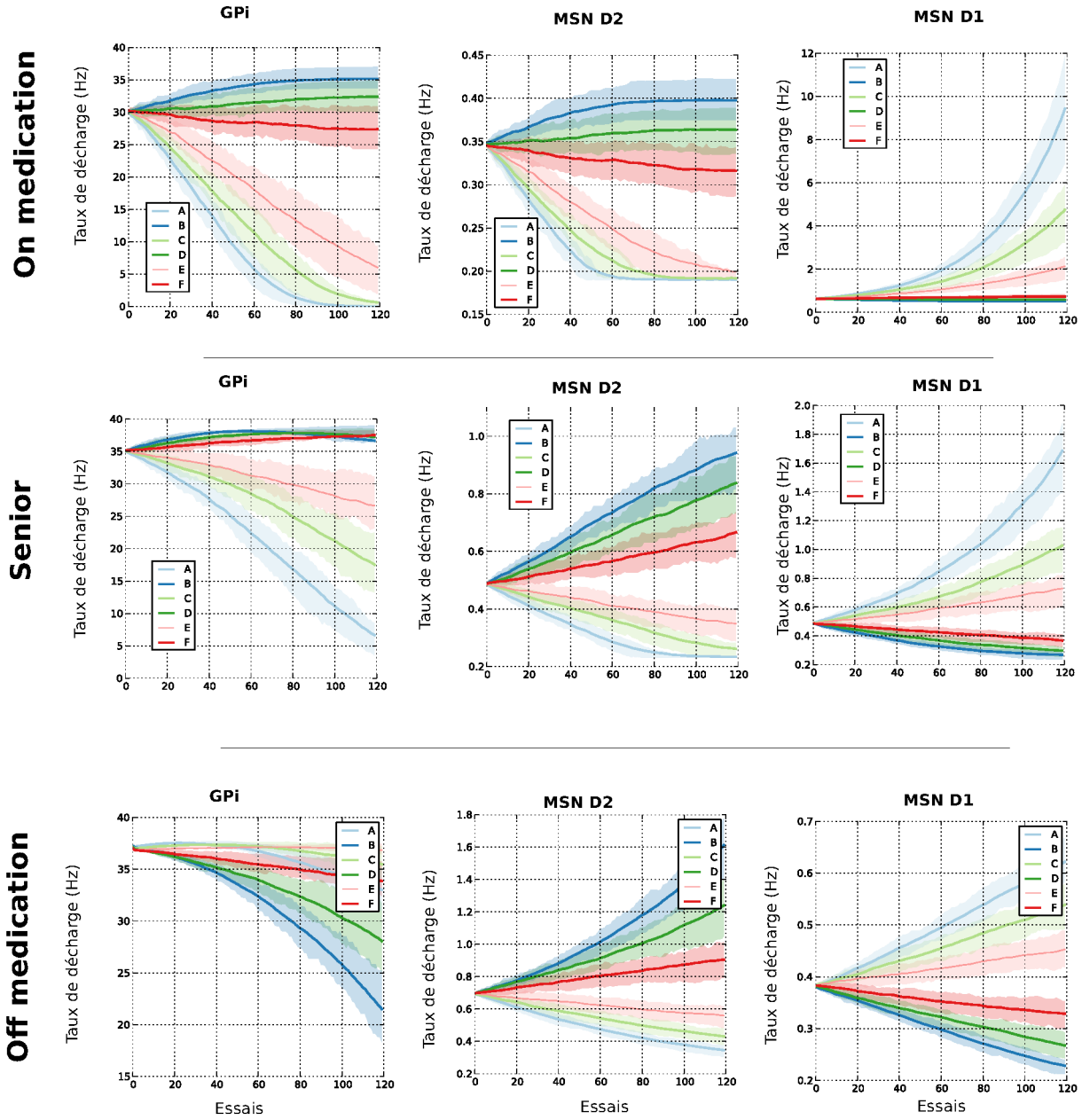

Figure 8.3 – *Résultats comportement obtenus en phase d'apprentissage avec le modèle* rBCBG1 *et un facteur DA = 25.*

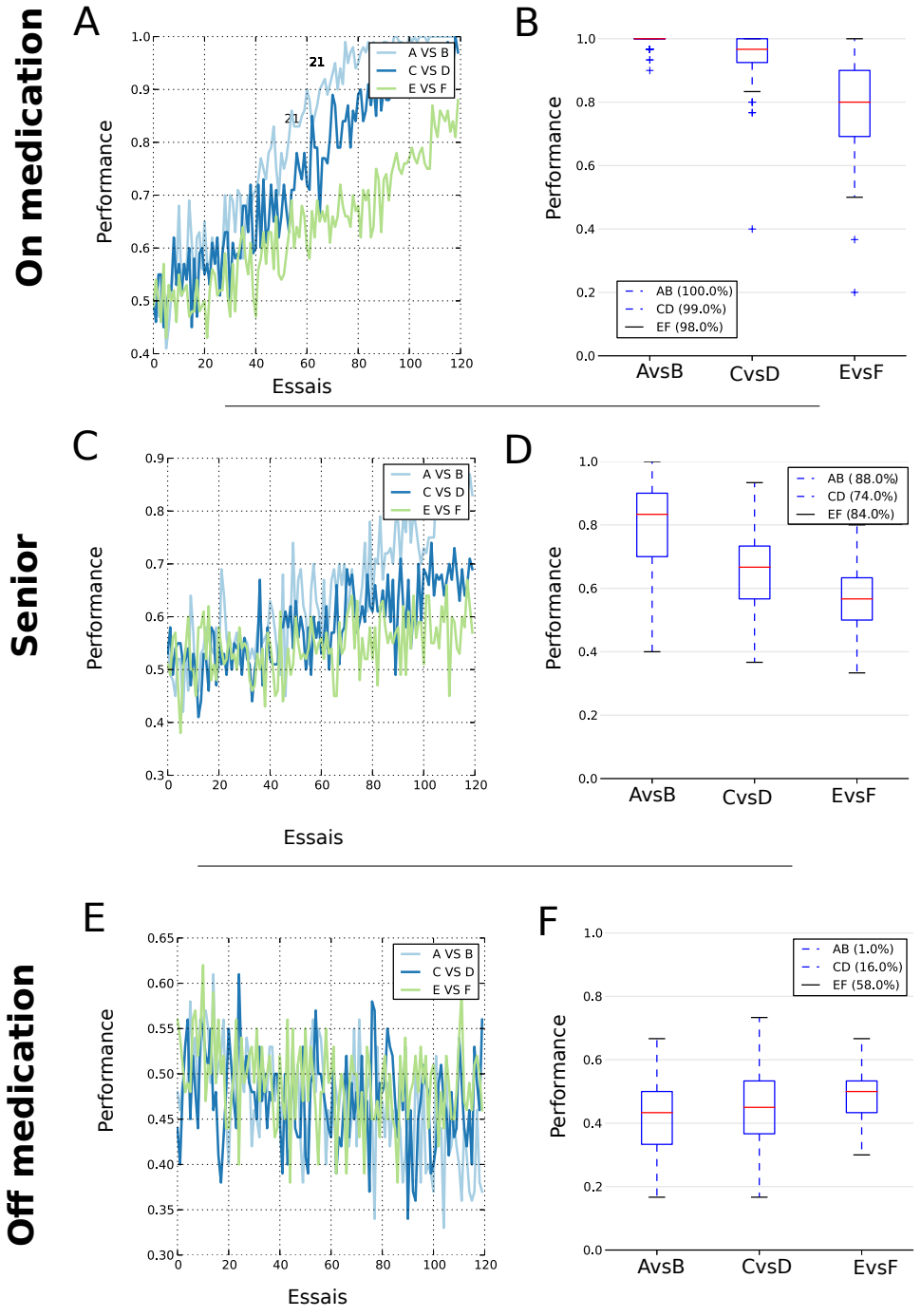

Figure 8.4 – *Évolution de l'activité des différents noyaux aux cours de la phase d'apprentissage avec le modèle* rBCBG1 *avec un facteur DA = 25.*

Bibliographie

Albada, S. J van et P. a Robinson (avr. 2009). « Mean-field modeling of the basal ganglia-thalamocortical system. I Firing rates in healthy and parkinsonian states. » Dans : *Journal of theoretical biology* 257.4, p. 642–63. issn : 1095-8541.

Albada, S. J van, R. T Gray, P. M Drysdale et P. a Robinson (avr. 2009). « Mean-field modeling of the basal ganglia-thalamocortical system. II Dynamics of parkinsonian oscillations. » Dans : *Journal of theoretical biology* 257.4, p. 664–88. issn : 1095-8541.

Albin, R L, a B Young et J B Penney (oct. 1989). « The functional anatomy of basal ganglia disorders. » Dans : *Trends in neurosciences* 12.10, p. 366–75.

Alexander, G. E et M. D Crutcher (juil. 1990). « Functional architecture of basal ganglia circuits : neural substrates of parallel processing. » Dans : *Trends in neurosciences* 13.7, p. 266–71.

Alexander, G. E, M. R DeLong et P. L Strick (1986). « Parallel organization of functionally segregated circuits linking basal ganglia and cortex ». Dans : *Annual review of neuroscience* 9.1, p. 357–381.

Alvarez, L, R Macias, G Lopez, E Alvarez, N Pavon, M C Rodriguez-Oroz, J L Juncos, C Maragoto, J Guridi, I Litvan, E S Tolosa, W Koller, J Vitek, M R DeLong et J a Obeso (mar. 2005). « Bilateral subthalamotomy in Parkinson's disease : initial and long-term response. » Dans : *Brain : a journal of neurology* 128.Pt 3, p. 570–83.

Anderson, V. C, K. J Burchiel, P. Hogarth, J. Favre et J. P Hammerstad (2005). « Pallidal vs subthalamic nucleus deep brain stimulation in Parkinson disease ». Dans : *Archives of neurology* 62.4, p. 554–560.

Atallah, H. E, D. Lopez-Paniagua, J. W Rudy et R. C O'Reilly (jan. 2007). « Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. » Dans : *Nature neuroscience* 10.1, p. 126–31. issn : 1097-6256.

Aubert, I., I. Ghorayeb, E. Normand et B. Bloch (2000). « Phenotypical Characterization of the Neurons Expressing the D1 and D2 Dopamine Receptors in the Monkey Striatum ». Dans : 32.November 1999, p. 22–32.

Balleine, B. W et J. P O'Doherty (jan. 2010). « Human and rodent homologies in action control : corticostriatal determinants of goal-directed and habitual action. » Dans : *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 35.1, p. 48–69. issn : 1740-634X.

Balleine, B W, M R Delgado et O Hikosaka (2007). « The role of the dorsal striatum in reward and decision-making ». Dans : *The Journal of Neuroscience* 27.31, p. 8161–8165.

Bar-Gad, I et H Bergman (2001). « Stepping out of the box : information processing in the neural networks of the basal ganglia ». Dans : *Current opinion in neurobiology* 11, p. 689–695.

Bartels, A. L et K. L Leenders (sept. 2009). « Parkinson's disease : the syndrome, the pathogenesis and pathophysiology. » Dans : *Cortex ; a journal devoted to the study of the nervous system and behavior* 45.8, p. 915–21.

Barto, A G (1995). « Adaptive Critics and the Basal Ganglia ». Dans :

Bayer, H. M et P. W Glimcher (2005). « Midbrain dopamine neurons encode a quantitative reward prediction error signal ». Dans : *Neuron* 47.1, p. 129–141.

Bayer, H M, B Lau et P W Glimcher (2007). « Statistics of midbrain dopamine neuron spike trains in the awake primate ». Dans : *Journal of Neurophysiology* 98.3, p. 1428–1439.

Becker, W et R Jürgens (1979). « An analysis of the saccadic system by means of double step stimuli ». Dans : *Vision research* 19.9, p. 967–983.

Beeler, J. A, N. Daw, C. RM Frazier et X. Zhuang (2010). « Tonic dopamine modulates exploitation of reward learning ». Dans : *Frontiers in behavioral neuroscience* 4.

Benabid, A. L, S. Chabardes, J. Mitrofanis et P. Pollak (jan. 2009). « Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. » Dans : *The Lancet. Neurology* 8.1, p. 67–81.

Bergman, H, T Wichmann, B Karmon et MR DeLong (1994). « The primate subthalamic nucleus. II. Neuronal activity in the MPTP model of parkinsonism ». Dans : *Journal of neurophysiology* 72.2, p. 507–520.

Berns, G. S et T. J Sejnowski (1998). « A Computational Model of How the Basal Ganglia Produce Sequences ». Dans : p. 108–121.

Berridge, K. C (avr. 2007). « The debate over dopamine's role in reward : the case for incentive salience. » Dans : *Psychopharmacology* 191.3, p. 391–431.

Berridge, K C (avr. 2012). « From prediction error to incentive salience : mesolimbic computation of reward motivation. » Dans : *The European journal of neuroscience* 35.7, p. 1124–43.

Berthet, P., J. Hellgren-Kotaleski et A. Lansner (jan. 2012). « Action selection performance of a reconfigurable basal ganglia inspired model with Hebbian-Bayesian Go-NoGo connectivity. » Dans : *Frontiers in behavioral neuroscience* 6.October, p. 65.

Bertsekas, D. P et J. N Tsitsiklis (1995). « Neuro-dynamic programming : An overview ». Dans : *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on.* T. 1. IEEE, p. 560–564.

Bhatnagar, S., M. Ghavamzadeh, M. Lee et R. S Sutton (2007). « Incremental natural actor-critic algorithms ». Dans : *Advances in neural information processing systems*, p. 105–112.

Bolam, J P, J J Hanley, P a Booth et M D Bevan (mai 2000). « Synaptic organisation of the basal ganglia. » Dans : *Journal of anatomy* 196 (Pt 4, p. 527–42.

Bornstein, A. M et N. D Daw (juin 2011). « Multiplicity of control in the basal ganglia : computational roles of striatal subregions. » Dans : *Current opinion in neurobiology* 21.3, p. 374–80. issn : 1873-6882.

Brischoux, F., S. Chakraborty, D. I Brierley et M. a Ungless (mar. 2009). « Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. » Dans : *Proceedings of the National Academy of Sciences of the United States of America* 106.12, p. 4894–9. issn : 1091-6490.

Brown, P. (2003). « Oscillatory nature of human basal ganglia activity : relationship to the pathophysiology of Parkinson's disease ». Dans : *Movement Disorders* 18.4, p. 357–363.

Brown, P, A Oliviero, P Mazzone, A Insola, P Tonali et V Di Lazzaro (2001). « Dopamine dependency of oscillations between subthalamic nucleus and pallidum in Parkinson's disease ». Dans : *The Journal of neuroscience* 21.3, p. 1033–1038.

Bunzeck, N et E Düzel (2006). « Absolute coding of stimulus novelty in the human substantia nigra/VTA ». Dans : *Neuron* 51.3, p. 369–379.

Calabresi, P., B. Picconi, Al. Tozzi, V. Ghiglieri et Ma. Di Filippo (juil. 2014). « Direct and indirect pathways of basal ganglia : a critical reappraisal ». Dans : *Nature Neuroscience* 17.8, p. 1022–1030. issn : 1097-6256.

Cannon, C M et R D Palmiter (2003). « Reward without dopamine ». Dans : *The Journal of neuroscience* 23.34, p. 10827–10831.

Chevalier, G. et J.M Deniau (1990). « Disinhibition as a basic process in the expression of striatal functions ». Dans : *Trends in neurosciences* 13.7, p. 277–280.

Christian, A (2003). « The Nature of Learning - A study of Reinforcement Learning Methodology ». Dans :

Cohen, J D, S M McClure et A J Yu (mai 2007). « Should I stay or should I go ? How the human brain manages the trade-off between exploitation and exploration. » Dans : *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 362.1481, p. 933–42.

Cohen, J. Y, S. Haesler, L. Vong, B. B Lowell et N. Uchida (fév. 2012). « Neuron-type-specific signals for reward and punishment in the ventral tegmental area. » Dans : *Nature* 482.7383, p. 85–8. issn : 1476-4687.

Collins, A. (2010). « Apprentissage et contrôle cognitif : une théorie computationnelle de la fonction exécutive préfontale humaine ». Thèse de doct. UPMC - Université Pierre et Marie Curie - Paris VI.

Collins, A G E et M J Frank (juil. 2014). « Opponent actor learning (OpAL) : Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. » Dans : *Psychological review* 121.3, p. 337–66.

Colombo, M (mar. 2014). « Deep and beautiful. The reward prediction error hypothesis of dopamine. » Dans : *Studies in history and philosophy of biological and biomedical sciences* 45, p. 57–67. issn : 1879-2499.

Corbit, L H et B W Balleine (août 2011). « The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31.33, p. 11786–94.

Cox, S M L, M J Frank, K Larcher, L K Fellows, C a Clark, M Leyton et A Dagher (jan. 2015). « Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. » Dans : *NeuroImage* 109, p. 95–101.

Cromwell, H C et W Schultz (2003). « Effects of expectations for different reward magnitudes on neuronal activity in primate striatum ». Dans : *Journal of Neurophysiology* 89.5, p. 2823–2838.

Cui, G., S. B. Jun, X. Jin, M. D Pham, S. S Vogel, D. M Lovinger et R. M Costa (2013). « Concurrent activation of striatal direct and indirect pathways during action initiation ». Dans : *Nature*.

Daw, N. D (déc. 2007). « Dopamine : at the intersection of reward and action ». Dans : *Nat Neurosci* 10.12, p. 1505–1507. issn : 1097-6256.

Daw, N D (2011). « Trial-by-trial data analysis using computational models ». Dans : *Decision making, affect, and learning : Attention and performance XXIII* 23, p. 3–38.

Daw, N D (2013). « Advanced Reinforcement Learning ». Dans : Academic Press. Chap. 13.

Daw, N D et K Doya (avr. 2006). « The computational neurobiology of learning and reward. » Dans : *Current opinion in neurobiology* 16.2, p. 199–204.

Daw, N. D, S. Kakade et P. Dayan (juin 2002). « Opponent interactions between serotonin and dopamine ». Dans : *Neural Networks* 15.4-6, p. 603–616. issn : 08936080.

Daw, N. D, Y. Niv et P. Dayan (déc. 2005). « Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control ». Dans : *Nat Neurosci* 8.12, p. 1704–1711. issn : 1097-6256.

Day, J. J, J. L Jones, R M. Wightman et R. M Carelli (août 2010). « Phasic nucleus accumbens dopamine release encodes effort- and delay-related costs. » Dans : *Biological psychiatry* 68.3, p. 306–9. issn : 1873-2402.

Deco, G., V. K Jirsa, P.a Robinson, M. Breakspear et K. Friston (jan. 2008). « The dynamic brain : from spiking neurons to neural masses and cortical fields. » Dans : *PLoS computational biology* 4.8, e1000092. issn : 1553-7358.

DeLong, M R (1990). « Primate models of movement disorders of basal ganglia origin ». Dans : *Trends in neurosciences* 13.7, p. 281–285.

Dollé, L., D. Sheynikhovich, B. Girard, R. Chavarriaga et A. Guillot (oct. 2010). « Path planning versus cue responding : a bio-inspired model of switching between navigation strategies. » Dans : *Biological cybernetics* 103.4, p. 299–317. issn : 1432-0770.

Doya, K. (oct. 1999). « What are the computations of the cerebellum, the basal ganglia and the cerebral cortex ? » Dans : *Neural Networks* 12.7-8, p. 961–974.

Doya, K (2002). « Metalearning and neuromodulation ». Dans : 15, p. 495–506.

Doya, K. (2007). « Reinforcement learning : Computational theory and biological mechanisms ». Dans : *HFSP Journal* 1.1, p. 30. issn : 19552068.

Enomoto, K., N. Matsumoto, S. Nakai, T. Satoh, T. K Sato, Y. Ueda, H. Inokawa, M. Haruno et M. Kimura (2011). « Dopamine neurons learn to encode the long-term value of multiple future rewards ». Dans : *PNAS*.

Fan, K. Y, J. Baufreton, D J. Surmeier, C S. Chan et M. D Bevan (oct. 2012). « Proliferation of external globus pallidus-subthalamic nucleus synapses following de-

generation of midbrain dopamine neurons. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32.40, p. 13718–28. ISSN : 1529-2401.

Faure, A, U Haberland, F Condé et N El Massioui (2005). « Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation ». Dans : *The Journal of Neuroscience* 25.11, p. 2771–2780.

Fino, E. (2007). « Transmission et plasticité activité-dépendante au niveau des synapses cortico-striatales ». Thèse de doct. Université Pierre et Marie Curie.

Fiorillo, C. D (août 2013). « Two dimensions of value : dopamine neurons represent reward but not aversiveness. » Dans : *Science (New York, N.Y.)* 341.6145, p. 546–9. ISSN : 1095-9203.

Fiorillo, C. D, P. N Tobler et W. Schultz (2003). « Discrete coding of reward probability and uncertainty by dopamine neurons ». Dans : *Science* 299.5614, p. 1898.

Fiorillo, C D, S R Yun et M R Song (mar. 2013a). « Diversity and homogeneity in responses of midbrain dopamine neurons. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33.11, p. 4693–709.

Fiorillo, C D, M R Song et S R Yun (mar. 2013b). « Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33.11, p. 4710–25.

Flagel, S. B, J. J Clark, T. E Robinson, L. Mayo, A. Czuj, I. Willuhn, C. A Akers, S. M Clinton, P. E.M Phillips et H. Akil (2010). « A selective role for dopamine in stimulus-reward learning ». Dans : *Nature* 469.7328, 5357.

Fountas, Z. et M. Shanahan (2014). « Phase Offset Between Slow Oscillatory Cortical Inputs Influences Competition in a Model of the Basal Ganglia ». Dans :

Frank, M J et E D Claus (avr. 2006). « Anatomy of a decision : striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. » Dans : *Psychological review* 113.2, p. 300–26.

Frank, M. J, L. C Seeberger et R. C O'reilly (déc. 2004). « By carrot or by stick : cognitive reinforcement learning in parkinsonism. » Dans : *Science (New York, N.Y.)* 306.5703, p. 1940–3. ISSN : 1095-9203.

Frank, M J, A a Moustafa, H M Haughey, T Curran et K E Hutchison (oct. 2007). « Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. » Dans : *Proceedings of the National Academy of Sciences of the United States of America* 104.41, p. 16311–6.

Frank, Michael J (jan. 2005). « Dynamic dopamine modulation in the basal ganglia : a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. » Dans : *Journal of cognitive neuroscience* 17.1, p. 51–72.

— (oct. 2006). « Hold your horses : a dynamic computational role for the subthalamic nucleus in decision making. » Dans : *Neural networks : the official journal of the International Neural Network Society* 19.8, p. 1120–36.

Freeze, B S, A V Kravitz, N Hammack, J D Berke et A C Kreitzer (2013). « Control of basal ganglia output by direct and indirect pathway projection neurons ». Dans : *The Journal of Neuroscience* 33.47, p. 18531–18539.

Friend, D M et A V Kravitz (mai 2014). « Working together : basal ganglia pathways in action selection. » Dans : *Trends in neurosciences* 37.6, p. 301–303. issn : 1878-108X.

Fujiyama, F, J Sohn, T Nakano, T Furuta, K C Nakamura, W Matsuda et T Kaneko (fév. 2011). « Exclusive and common targets of neostriatofugal projections of rat striosome neurons : a single neuron-tracing study using a viral vector. » Dans : *The European journal of neuroscience* 33.4, p. 668–77.

Furuyashiki, T (2012). « Roles of Dopamine and Inflammation-Related Molecules in Behavioral Alterations Caused by Repeated Stress ». Dans : *Journal of Pharmacological Sciences* 120.2, p. 63–69.

Gale, J. T, R. Amirnovin, Z. M Williams, A. W Flaherty et E. N Eskandar (jan. 2008). « From symphony to cacophony : pathophysiology of the human basal ganglia in Parkinson disease. » Dans : *Neuroscience and biobehavioral reviews* 32.3, p. 378–87.

Gan, J. O, M. E Walton et P. E M Phillips (jan. 2010). « Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. » Dans : *Nature neuroscience* 13.1, p. 25–7. issn : 1546-1726.

Georgopoulos, A. P, J. F Kalaska, R. Caminiti et J. T Massey (1982). « On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex ». Dans : *The Journal of neuroscience* 2.11, p. 1527–1537.

Gerfen, C. R et D J. Surmeier (2012). « Modulation of striatal projection systems by dopamine ». Dans : 2, p. 441–466.

Gerfen, C. R et C. J Wilson (1996). « Chapter II The basal ganglia ». Dans : *Handbook of chemical neuroanatomy* 12, p. 371–468.

Gershman, S J (2013). « Dopamine ramps are a consequence of reward prediction errors ». Dans : p. 1–7.

Gillies, A et G Arbuthnott (sept. 2000). « Computational models of the basal ganglia ». Dans : *Movement Disorders* 15.5, p. 762–770.

Girard, B, V Cuzin, A Guillot, K Gurney et T Prescott (2003). « A basal ganglia inspired model of action selection evaluated in a robotic survival task ». Dans : *Journal of Integrative Neuroscience* 2, p. 179–200.

Girard, B, N Tabareau, A Berthoz et J Slotine (2006). « Selective amplification using a contracting model of the basal ganglia ». Dans : *NeuroComp 2006*, p. 30–33.

Girard, B., N. Tabareau, Q. C Pham, A. Berthoz et J-J. Slotine (mai 2008). « Where neuroscience and dynamic system theory meet autonomous robotics : a contracting basal ganglia model for action selection. » Dans : *Neural networks : the official journal of the International Neural Network Society* 21.4, p. 628–41. issn : 0893-6080.

Gittis, A. H et A. C Kreitzer (2012). « Striatal microcircuitry and movement disorders ». Dans : *Trends in neurosciences* 35.9, p. 557–564.

Gläscher, J., N. Daw, P. Dayan et J. P O'Doherty (mai 2010). « States versus rewards : dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. » Dans : *Neuron* 66.4, p. 585–95. issn : 1097-4199.

Glimcher, P W (2011). « Understanding dopamine and reinforcement learning : the dopamine reward prediction error hypothesis ». Dans : *Proceedings of the National Academy of Sciences* 108.Supplement 3, p. 15647–15654.

Gottfried, J A, J O'Doherty et R J Dolan (2003). « Encoding predictive reward value in human amygdala and orbitofrontal cortex ». Dans : *Science* 301.5636, p. 1104–1107.

Grace, A A (2000). « The tonic/phasic model of dopamine system regulation and its implications for understanding alcohol and psychostimulant craving ». Dans : *Addiction* 95.8s2, p. 119–128.

Grieder, T E, O George, H Tan, S R George, B Le Foll, S R Laviolette et D van der Kooy (2012). « Phasic D1 and tonic D2 dopamine receptor signaling double dissociate the motivational effects of acute nicotine and chronic nicotine withdrawal ». Dans : *Proceedings of the National Academy of Sciences* 109.8, p. 3101–3106.

Guillot, A et J Meyer (2003). « La contribution de l'approche animat aux sciences cognitives ». Dans : *Cognito* 1.1, p. 1–26.

Gurney, K, T J Prescott et P Redgrave (juin 2001a). « A computational model of action selection in the basal ganglia. I. A new functional anatomy. » Dans : *Biological cybernetics* 84.6, p. 401–10.

Gurney, K., T. J Prescott et P. Redgrave (juin 2001b). « A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. » Dans : *Biological cybernetics* 84.6, p. 411–23. issn : 0340-1200.

Gurney, K., T. J Prescott, J. R Wickens et P. Redgrave (août 2004). « Computational models of the basal ganglia : from robots to membranes. » Dans : *Trends in neurosciences* 27.8, p. 453–9.

Haber, S. N. (déc. 2003). « The primate basal ganglia : parallel and integrative networks ». Dans : *Journal of Chemical Neuroanatomy* 26.4, p. 317–330. issn : 08910618.

Haber, S. N et R. Calzavara (fév. 2009). « The cortico-basal ganglia integrative network : the role of the thalamus. » Dans : *Brain research bulletin* 78.2-3, p. 69–74. issn : 1873-2747.

Haber, S. N, J. L Fudge et N. R McFarland (mar. 2000). « Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 20.6, p. 2369–82. issn : 1529-2401.

Hardman, C. D, J. M Henderson, D. I Finkelstein, M. K Horne, G. Paxinos et G. M Halliday (2002). « The journal of comparative neurology 445 :238 255 (2002) ». Dans : 255.November 2000, p. 238–255.

Hare, T A, J O'Doherty, C F Camerer, W Schultz et A Rangel (2008). « Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors ». Dans : *The Journal of Neuroscience* 28.22, p. 5623–5630.

Hart, A S, R B Rutledge, P W Glimcher et P E M Phillips (jan. 2014). « Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 34.3, p. 698–704.

Hassani, O-K, M Mouroux et J Feger (1996). « Increased subthalamic neuronal activity after nigral dopaminergic lesion independent of disinhibition via the globus pallidus ». Dans : *Neuroscience* 72.1, p. 105–115.

Haynes, W. I. A. et S. N. Haber (2013). « The organization of prefrontal-subthalamic inputs in primates provides an anatomical substrate for both functional specificity and integration : implications for basal ganglia models and deep brain stimulation ». Dans : *J Neuroscience* 33.11, p. 4804–4814.

Hebb, Donald Olding (1949). *The organization of behavior : A neuropsychological approach*. John Wiley & Sons.

Hollerman, J. R et W. Schultz (1998). « Dopamine neurons report an error in the temporal prediction of reward during learning ». Dans : *Nat Neurosci* 1.4, 304309.

Houk, J C. et S P. Wise (1995). « Distributed modular architectures linking Basal Ganglia, Cerebellum, and Cerebral Cortex : their role in planning and controlling action ». Dans : *Cerebral cortex*.

Howe, M. W, P. L Tierney, S. G Sandberg, P. EM Phillips et A. M Graybiel (2013). « Prolonged dopamine signalling in striatum signals proximity and value of distant rewards ». Dans : *Nature* 500.7464, p. 575–579.

Humphries, M D et T J Prescott (avr. 2010). « The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. » Dans : *Progress in neurobiology* 90.4, p. 385–417.

Humphries, M. D, R. D Stewart et K. N Gurney (2006). « A physiologically plausible model of action selection and oscillatory activity in the basal ganglia ». Dans : *The Journal of neuroscience* 26.50, p. 12921–12942.

Humphries, M. D, M. Khamassi et K. Gurney (jan. 2012). « Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. » Dans : *Frontiers in neuroscience* 6.February, p. 9. issn : 1662-453X.

Humphries, M.D et K.N Gurney (2002). « The role of intra-thalamic and thalamocortical circuits in action selection ». Dans : *Network : Computation in Neural Systems* 13.1, p. 131–156.

Hutchison, W. D, J. O Dostrovsky, J. R Walters, R. Courtemanche, T. Boraud, J. Goldberg et P. Brown (2004). « Neuronal oscillations in the basal ganglia and movement disorders : evidence from whole animal and human recordings ». Dans : *The Journal of Neuroscience* 24.42, p. 9240–9243.

Israel, Z. et H. Bergman (jan. 2008). « Pathophysiology of the basal ganglia and movement disorders : from animal models to human clinical applications. » Dans : *Neuroscience and biobehavioral reviews* 32.3, p. 367–77.

Ito, M. et K. Doya (2009). « Validation of decision-making models and analysis of decision variables in the rat basal ganglia ». Dans : *The Journal of neuroscience* 29.31, p. 9861–9874.

— (juin 2011). « Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. » Dans : *Current opinion in neurobiology* 21.3, p. 368–73. issn : 1873-6882.

Jaakkola, T, M I Jordan et S P Singh (1994). « On the convergence of stochastic iterative dynamic programming algorithms ». Dans : *Neural computation* 6.6, p. 1185–1201.

Joel, D. et I. Weiner (2000). « Commentary the connections of the dopaminergic system with the striatum in rats and primates : an analysis with respect to the functional and compartmental organization of the striatum ». Dans : *Neuroscience* 96.3, p. 451–474.
Joel, D., Y. Niv et E. Ruppin (2002). « Actor-critic models of the basal ganglia : new anatomical and computational perspectives ». Dans : *Neural Networks* 15.4-6, 535547.
Kalaska, J. F, D. A D Cohen et L. Hyde (1989). « Centre de recherche en sciences neurologiques, Departemente de physiologie, Faculte de medecine, Universite de Montreal, Montreal, Quebec, Canada, H3C 3J7, and *Department of Veterinary and Comparative Anatomy, Pharmacology and Physiology, College of Veterinary Medicine, Washington State University, Pullman, Washington 99164 ». Dans : June, p. 2080–2102.
Kawaguchi, Y., C. J Wilson et P. C Emson (1990). « Projection subtypes of rat neostriatal matrix cells revealed by intracellular injection of biocytin ». Dans : *The Journal of Neuroscience* 10.10, p. 3421–3438.
Kawaguchi, Y., C. J Wilson, S. J Augood et P. C Emson (1995). « Striatal interneurones : chemical, physiological and morphological characterization ». Dans : *Trends in neurosciences* 18.12, p. 527–535.
Keeler, J F, D O Pretsell et T W Robbins (juil. 2014). « Functional implications of dopamine D1 vs D2 receptors : A 'Prepare and Select' model of the striatal direct vs. indirect pathways. » Dans : *Neuroscience* 282, p. 156–175.
Keramati, M et B Gutkin (2013). « Imbalanced decision hierarchy in addicts emerging from drug-hijacked dopamine spiraling circuit ». Dans : *PloS one* 8.4, e61489.
— (2014). « Collecting reward to defend homeostasis : A homeostatic reinforcement learning theory ». Dans : *bioRxiv*, p. 005140.
Keramati, M., A. Dezfouli et P. Piray (mai 2011). « Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes ». Dans : *PLoS Comput Biol* 7.5, e1002055.
Khamassi, M. et M. D Humphries (2012). « Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies ». Dans : *Frontiers in behavioral neuroscience* 6.
Khamassi, M, L Lachèze, B Girard, A Berthoz et A Guillot (2005). « Actor–Critic models of reinforcement learning in the basal ganglia : from natural to artificial rats ». Dans : *Adaptive Behavior* 13.2, p. 131–148.
Khamassi, M, A B Mulder, E Tabuchi, V Douchamps et S I Wiener (2008). « Anticipatory reward signals in ventral striatal neurons of behaving rats ». Dans : *European journal of neuroscience* 28.9, p. 1849–1866.
Konda, V. R et J. N Tsitsiklis (1999). « Actor-Critic Algorithms. » Dans : *NIPS*. Citeseer, p. 1008–1014.
Kravitz, A V, L D Tye et A C Kreitzer (juin 2012). « Distinct roles for direct and indirect pathway striatal neurons in reinforcement. » Dans : *Nature neuroscience* 15.6, p. 816–8.
Kravitz, E A et R Huber (2003). « Aggression in invertebrates ». Dans : *Current opinion in neurobiology* 13.6, p. 736–743.

Lammel, S., A. Hetzel, O. Häckel, I. Jones, B. Liss et J. Roeper (mar. 2008). « Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. » Dans : *Neuron* 57.5, p. 760–73. issn : 1097-4199.

Lammel, S., D. I Ion, J. Roeper et R. C Malenka (juin 2011). « Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. » Dans : *Neuron* 70.5, p. 855–62. issn : 1097-4199.

Lammel, S., B. K. Lim, C. Ran, K. W. Huang, M. J Betley, K. M Tye, K. Deisseroth et R. C Malenka (oct. 2012). « Input-specific control of reward and aversion in the ventral tegmental area. » Dans : *Nature*. issn : 1476-4687.

Lardeux, S, D Paleressompoulle, R Pernaud, M Cador et C Baunez (oct. 2013). « Different populations of subthalamic neurons encode cocaine vs. sucrose reward and predict future error. » Dans : *Journal of neurophysiology* 110.7, p. 1497–510.

Leblois, A, T Boraud, W Meissner, H Bergman et D Hansel (2006). « Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia ». Dans : *The Journal of Neuroscience* 26.13, p. 3567–3583.

Lesaint, F., O. Sigaud, S. B. Flagel, T. E. Robinson et M. Khamassi (2014). « Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses : A Dual Learning Systems Approach with Factored Representations. » Dans : *PLoS Computational Biology*.

Lévesque, M. et A. Parent (août 2005). « The striatofugal fiber system in primates : a reevaluation of its organization based on single-axon tracing studies. » Dans : *Proceedings of the National Academy of Sciences of the United States of America* 102.33, p. 11888–93. issn : 0027-8424.

Liénard, J. et B. Girard (sept. 2014). « A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection. » Dans : *Journal of computational neuroscience*. issn : 1573-6873.

Lindvall, O. et A. Björklund (1979). « Dopaminergic innervation of the globus pallidus by collaterals from the nigrostriatal pathway ». Dans : *Brain research* 172.1, p. 169–173.

Liénard, J. (2013). « A Biologically Constrained Basal Ganglia Model ». Thèse de doct. Université Pierre et Marie Curie.

Ljungberg, T., P. Apicella et W. Schultz (jan. 1992). « Responses of monkey dopamine neurons during learning of behavioral reactions ». Dans : *Journal of Neurophysiology* 67.1, p. 145 –163.

Lloyd, K., N. Becker, M. W Jones et R. Bogacz (jan. 2012). « Learning to use working memory : a reinforcement learning gating model of rule acquisition in rats. » Dans : *Frontiers in computational neuroscience* 6.October, p. 87. issn : 1662-5188.

Maia, T V et M J Frank (fév. 2011). « From reinforcement learning models to psychiatric and neurological disorders. » Dans : *Nature neuroscience* 14.2, p. 154–62.

Marchand, A, V Fresno, M Khamassi et Coutureau E. (2014). « Dopaminergic modulation of the exploration level in a non-stationary probabilistic task ». Dans : *FENS Abstract Milan Italy*.

Maril, S, S Hassin-Baer, O S Cohen et R Tomer (avr. 2013). « Effects of asymmetric dopamine depletion on sensitivity to rewarding and aversive stimuli in Parkinson's disease. » Dans : *Neuropsychologia* 51.5, p. 818–24.

Masana, M, A Bortolozzi et F Artigas (fév. 2011). « Selective enhancement of mesocortical dopaminergic transmission by noradrenergic drugs : therapeutic opportunities in schizophrenia. » Dans : *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)* 14.1, p. 53–68.

Matsumoto, M (mar. 2015). « Dopamine signals and physiological origin of cognitive dysfunction in Parkinson's disease. » Dans : *Movement disorders : official journal of the Movement Disorder Society* 30.4, p. 472–483.

Matsumoto, M. et O. Hikosaka (2009). « Two types of dopamine neuron distinctly convey positive and negative motivational signals ». Dans : *Nature* 459.7248, 837841.

McClure, S M, N D Daw et P R Montague (2003). « A computational substrate for incentive salience ». Dans : *Trends in neurosciences* 26.8, p. 423–428.

Miller, E K et J D Cohen (2001). « An integrative theory of prefrontal cortex function ». Dans : *Annual review of neuroscience* 24.1, p. 167–202.

Mingote, S., S. M Weber, K. Ishiwari, M. Correa et J. D Salamone (2005). « Ratio and time requirements on operant schedules : effort-related effects of nucleus accumbens dopamine depletions ». Dans : *European Journal of Neuroscience* 21.6, p. 1749–1757.

Mink, J. W (nov. 1996). « The basal ganglia : focused selection and inhibition of competing motor programs. » Dans : *Progress in neurobiology* 50.4, p. 381–425. issn : 0301-0082.

Mink, JW et WT Thach (1991). « Basal ganglia motor control. III. Pallidal ablation : normal reaction time, muscle cocontraction, and slow movement ». Dans : *J Neurophysiol* 65.2, p. 330–351.

Mirenowicz, J. et W. Schultz (1994). « Importance of unpredictability for reward responses in primate dopamine neurons ». Dans : *Journal of Neurophysiology* 72.2, p. 1024 –1027.

Morris, G., A. Nevet, D. Arkadir, E. Vaadia et H. Bergman (2006). « Midbrain dopamine neurons encode decisions for future action ». Dans : *Nat Neurosci* 9.8, p. 1057–1063. issn : 1097-6256.

Nadjar, A., J. M Brotchie, C. Guigoni, Q. Li, S. Zhou, G. Wang, P. Ravenscroft, F. Georges, A. R Crossman et E. Bezard (août 2006). « Phenotype of striatofugal medium spiny neurons in parkinsonian and dyskinetic nonhuman primates : a call for a reappraisal of the functional organization of the basal ganglia. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26.34, p. 8653–61. issn : 1529-2401.

Nakanishi, S, T Hikida et S Yawata (avr. 2014). « Distinct dopaminergic control of the direct and indirect pathways in reward-based and avoidance learning behaviors. » Dans : *Neuroscience*.

Nakano, K, T Kayahara, T Tsutsumi et H Ushiro (2000). « Neural circuits and functional organization of the striatum ». Dans : *Journal of Neurology* 247, p. 1–15.

Nambu, A (déc. 2008). « Seven problems on the basal ganglia. » Dans : *Current opinion in neurobiology* 18.6, p. 595–604.

Nambu, A, H Tokuno et M Takada (2002). « Functional significance of the cortico-subthalamo-pallidal hyperdirect pathway ». Dans : *Neuroscience Research* 43, p. 111–117.

Nestler, E J (2001). « Molecular basis of long-term plasticity underlying addiction ». Dans : *Nature reviews neuroscience* 2.2, p. 119–128.

N'Guyen, S, C Thurat et B Girard (2014). « Saccade learning with concurrent cortical and subcortical basal ganglia loops ». Dans : *Frontiers in computational neuroscience* 8.

Niv, Y., M.O. Duff et P. Dayan (2005). « Dopamine, uncertainty and TD learning ». Dans : *Behavioral and Brain Functions* 1 :6, p. 1–9.

Niv, Y., N. D Daw et P. Dayan (2006). « Choice values ». Dans : *Nature neuroscience* 9.8, 987988.

Niv, Y., N. D Daw, D. Joel et P. Dayan (2007). « Tonic dopamine : opportunity costs and the control of response vigor ». Dans : *Psychopharmacology* 191.3, p. 507–520.

Obeso, J. A, M. C Rodriguez-Oroz, M. Rodriguez, J. L Lanciego, J. Artieda, N. Gonzalo et C. W Olanow (2000). « Pathophysiology of the basal ganglia in Parkinson's disease ». Dans : *Trends in neurosciences* 23, S8–S19.

O'Doherty, J, P Dayan, J Schultz, R Deichmann, K Friston et R J Dolan (2004). « Dissociable roles of ventral and dorsal striatum in instrumental conditioning ». Dans : *Science* 304.5669, p. 452–454.

O'Reilly, R C (nov. 1998). « Six principles for biologically based computational models of cortical cognition. » Dans : *Trends in cognitive sciences* 2.11, p. 455–62.

Padoa-Schioppa, C et J A Assad (2006). « Neurons in the orbitofrontal cortex encode economic value ». Dans : *Nature* 441.7090, p. 223–226.

Parent, A. (1994). « Extrinsic connections of the basal ganglia ». Dans : *Trends in neurosciences* 7, p. 7980–7984.

Parent, A. et L. Hazrati (1995). « Functional anatomy of the basal ganglia. » Dans : 20, p. 91–127.

Parent, A, F Sato, Y Wu, J Gauthier, M Lévesque et M Parent (oct. 2000). « Organization of the basal ganglia : the importance of axonal collateralization. » Dans : *Trends in neurosciences* 23.10 Suppl, S20–7.

Paton, J J et K Louie (juin 2012). « Reward and punishment illuminated. » Dans : *Nature neuroscience* 15.6, p. 807–9.

Pessiglione, M., B. Seymour, G. Flandin, R. J Dolan et C. D Frith (2006). « Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans ». Dans : *Nature* 442.7106, p. 1042–1045.

Prescott, T J, F M Montes González, K Gurney, M D Humphries et P Redgrave (jan. 2006). « A robot model of the basal ganglia : behavior and intrinsic processing. » Dans : *Neural networks : the official journal of the International Neural Network Society* 19.1, p. 31–61.

Redgrave, P, T J Prescott et K Gurney (1999a). « Is the short-latency dopamine response too short to signal reward error ? » Dans : *Trends in neurosciences* 22.4, p. 146–151.

Redgrave, P., T. J Prescott et K. Gurney (1999b). « The Basal Ganglia : a vertabrate solution to the selection problem ? » Dans : *Neuroscience* 89.4, p. 1009–1023.

Redgrave, P, N Vautrelle et JNJ Reynolds (2011). « Functional properties of the basal ganglia's re-entrant loop architecture : selection and reinforcement ». Dans : *Neuroscience* 198, p. 138–151.

Redish, A D (2004). « Addiction as a computational process gone awry ». Dans : *Science* 306.5703, p. 1944–1947.

Reiner, A (2009). « The Basal Ganglia IX ». Dans : Advances in Behavioral Biology 58. Sous la dir. d'Hendrik Jan Groenewegen, Pieter Voorn, Henk W. Berendse, Antonius B. Mulder et Alexander R. Cools, p. 3–24.

Rescorla, R A et A R Wagner (1972). « A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement ». Dans : *Classical conditioning : current research and theory*.

Robinson, M J F et K C Berridge (fév. 2013). « Instant transformation of learned repulsion into motivational "wanting". » Dans : *Current biology : CB* 23.4, p. 282–9.

Roesch, M. R, D. J Calu et G. Schoenbaum (déc. 2007). « Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards ». Dans : *Nat Neurosci* 10.12, p. 1615–1624. issn : 1097-6256.

Rommelfanger, K S et T Wichmann (2010). « Extrastriatal dopaminergic circuits of the basal ganglia ». Dans : *Frontiers in neuroanatomy* 4.

Rosenblatt, F (1958). *Two theorems of statistical separability in the perceptron*. United States Department of Commerce.

Samejima, K. et K. Doya (mai 2007). « Multiple representations of belief states and action values in corticobasal ganglia loops. » Dans : *Annals of the New York Academy of Sciences* 1104, p. 213–28. issn : 0077-8923.

Samejima, K., Y. Ueda, K. Doya et M. Kimura (2005). « Representation of action-specific reward values in the striatum ». Dans : *Science* 310.5752, p. 1337.

Sano, H, S Chiken, T Hikida, K Kobayashi et A Nambu (avr. 2013). « Signals through the striatopallidal indirect pathway stop movements by phasic excitation in the substantia nigra. » Dans : *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33.17, p. 7583–94.

Saunders, A, I A Oldenburg, V K Berezovskii, C A Johnson, N D Kingery, H L Elliott, T Xie, C R Gerfen et B L Sabatini (2015). « A direct GABAergic output from the basal ganglia to frontal cortex ». Dans : *Nature*.

Schoenbaum, G, Y Takahashi, T Liu et M A McDannald (2011). « Does the orbitofrontal cortex signal value ? » Dans : *Annals of the New York Academy of Sciences* 1239.1, p. 87–99.

Schultz, W. (juil. 1998). « Predictive Reward Signal of Dopamine Neurons ». Dans : *Journal of Neurophysiology* 80.1, p. 1 –27.

Schultz, W (août 2001). « Reward signaling by dopamine neurons. » Dans : *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry* 7.4, p. 293–302.

Schultz, W. et R. Romo (1987). « Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey ». Dans : *Journal of Neurophysiology* 57.1, p. 201–217.

Schultz, W, P Apicella et T Ljunberg (1993). « Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task ». Dans : *Journal of Neuroscience* 13.March.

Schultz, W., P. Dayan et P. R. Montague (mar. 1997). « A Neural Substrate of Prediction and Reward ». Dans : *Science* 275.5306, p. 1593 –1599.

Shen, W., M. Flajolet, P. Greengard et D J Surmeier (août 2008). « Dichotomous dopaminergic control of striatal synaptic plasticity. » Dans : *Science (New York, N.Y.)* 321.5890, p. 848–51. issn : 1095-9203.

Shiner, T., B. Seymour, K. Wunderlich, C. Hill, K. P Bhatia, P. Dayan et R. J Dolan (juin 2012). « Dopamine and performance in a reinforcement learning task : evidence from Parkinson's disease. » Dans : *Brain : a journal of neurology* 135.Pt 6, p. 1871–83.

Sigaud, O et O Buffet (2008). *Processus décisionnels de Markov en intelligence artificielle*. T. 1. Lavoisier-Hermes Science Publications.

Smith, A, M Li, S Becker et S Kapur (2006). « Dopamine, prediction error and associative learning : a model-based account ». Dans : *Network : Computation in Neural Systems* 17.1, p. 61–84.

Smith, Y. et J. Z Kieval (2000). « Anatomy of the dopamine system in the basal ganglia ». Dans : *Trends in neurosciences* 23, S28–S33.

Smittenaar, P, H W Chase, E Aarts, B Nusselein, B R Bloem et R Cools (avr. 2012). « Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection–learning or performance ? » Dans : *The European journal of neuroscience* 35.7, p. 1144–51.

Steinberg, E E, R Keiflin, J R Boivin, I B Witten, K Deisseroth et P H Janak (2013). « A causal link between prediction errors, dopamine neurons and learning ». Dans : *Nature neuroscience* 16.7, p. 966–973.

Stephenson-Jones, M, E Samuelsson, J Ericsson, B Robertson et S Grillner (2011). « Evolutionary conservation of the basal ganglia as a common vertebrate mechanism for action selection ». Dans : *Current Biology* 21.13, p. 1081–1091.

Sutton, R. S. et A. G. Barto (mar. 1998). *Reinforcement Learning : An Introduction*. The MIT Press. isbn : 9780262193986.

Takahashi, Y. K, M. R Roesch, R. C Wilson, K. Toreson, P. O'Donnell, Y. Niv et G. Schoenbaum (2011). « Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex ». Dans : *Nature neuroscience* 14.12, p. 1590–1597.

Tanaka, S. C, K. Doya, G. Okada, K. Ueda, Y. Okamoto et S. Yamawaki (2004). « Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops ». Dans : *Nature Neuroscience* 7.8, 887893.

Tepper, J. M, C. J Wilson et T. Koós (2008). « Feedforward and feedback inhibition in neostriatal GABAergic spiny neurons ». Dans : *Brain research reviews* 58.2, p. 272–281.

Thorndike, E L (1927). « The law of effect ». Dans : *The American Journal of Psychology*, p. 212–222.

Thurat, C, S NGuyen et B Girard (2015). « Biomimetic race model of the loop between the Superior Colliculus and the Basal Ganglia : Subcortical selection of saccade targets ». Dans : *Neural Networks*.

Tobler, P N, C D Fiorillo et W Schultz (2005). « Adaptive coding of reward value by dopamine neurons ». Dans : *Science* 307.5715, p. 1642–1645.

Tsirogiannis, G. L, G. a Tagaris, D. Sakas et K. S Nikita (fév. 2010). « A population level computational model of the basal ganglia that generates parkinsonian Local Field Potential activity. » Dans : *Biological cybernetics* 102.2, p. 155–76.

Utter, A. et M. Basso (jan. 2008). « The basal ganglia : an overview of circuits and function. » Dans : *Neuroscience and biobehavioral reviews* 32.3, p. 333–42.

Voorn, P, L JMJ Vanderschuren, H J Groenewegen, T W Robbins et C MA Pennartz (2004). « Putting a spin on the dorsal–ventral divide of the striatum ». Dans : *Trends in neurosciences* 27.8, p. 468–474.

Waelti, P, A Dickinson et W Schultz (2001). « Dopamine responses comply with basic assumptions of formal learning theory ». Dans : *Nature* 412.6842, p. 43–48.

Wang, D. V et J. Z Tsien (jan. 2011). « Convergent processing of both positive and negative motivational signals by the VTA dopamine neuronal populations. » Dans : *PloS one* 6.2, e17047. issn : 1932-6203.

Weinberger, M., N. Mahant, W. D Hutchison, A. M Lozano, E. Moro, M. Hodaie, A. E Lang et J. O Dostrovsky (déc. 2006). « Beta oscillatory activity in the subthalamic nucleus and its relation to dopaminergic response in Parkinson's disease. » Dans : *Journal of neurophysiology* 96.6, p. 3248–56.

White, N. M et R. J McDonald (2002). « Multiple parallel memory systems in the brain of the rat ». Dans : *Neurobiology of learning and memory* 77.2, p. 125–184.

Wilson, S W (1991). « The animat path to AI ». Dans :

Wilson, SA K. (1928). « Modern problems in neurology ». Dans : *British medical journal* 2.3541, p. 914.

Wise, R A (1985). « The anhedonia hypothesis : Mark III ». Dans : *Behavioral and Brain Sciences* 8.01, p. 178–186.

Wise, R A et P Rompré (1989). « Brain dopamine and reward ». Dans : *Annual review of psychology* 40.1, p. 191–225.

Witjas, T, C Baunez, J M Henry, M Delfini, J Regis, A A Cherif, J C Peragut et J P Azulay (août 2005). « Addiction in Parkinson's disease : impact of subthalamic nucleus deep brain stimulation. » Dans : *Movement disorders : official journal of the Movement Disorder Society* 20.8, p. 1052–5.

Wu, Y., S. Richard et A. Parent (2000). « The organization of the striatal output system : a single-cell juxtacellular labeling study in the rat ». Dans : *Neuroscience Research* 38.1, p. 49–62.

Yelnik, J., C. Francis, G. Percheron et D. Tandéa (1991). « Morphological taxonomy of the neurons of the primate striatum ». Dans : *Journal of Comparative Neurology* 313.2, p. 273–294.

Yelnik, J., C. François, G. Percheron et D. Tandé (1996). « A spatial and quantitative study of the striatopallidal connection in the monkey ». Dans : *Neuroreport* 7.5, p. 985–988.

Yin, H. H et B. J Knowlton (juin 2006). « The role of the basal ganglia in habit formation. » Dans : *Nature reviews. Neuroscience* 7.6, p. 464–76. issn : 1471-003X.

Yin, H. H, B. J Knowlton et B. W Balleine (2004). « Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning ». Dans : *European Journal of Neuroscience* 19.August 2003, p. 181–189.

Yin, H. H, S. B Ostlund, B. J Knowlton et B. W Balleine (juil. 2005). « The role of the dorsomedial striatum in instrumental conditioning. » Dans : *The European journal of neuroscience* 22.2, p. 513–23. issn : 0953-816X.

Yin, H H, X Zhuang et B W Balleine (2006). « Instrumental learning in hyperdopaminergic mice ». Dans : *Neurobiology of learning and memory* 85.3, p. 283–288.