Coordination de systèmes de mémoire : modèles théoriques du comportement animal et humain

Guillaume Viejo

Thèse de doctorat
de l'Université Pierre et Marie Curie

École doctorale nº 158 : Cerveau, Cognition, Comportement

Présentée par : **Guillaume Viejo**

pour obtenir le grade de
Docteur de l'Université Pierre et Marie Curie

28 Novembre 2016

Coordination de systèmes de mémoire : modèles théoriques du comportement animal et humain

Jury

M. Nicolas Rougier,	INRIA	Rapporteur
M. Jerome Sallet,	Univ. of Oxford	Rapporteur
Mme Laure Buhry,	LORIA - Univ. de Lorraine	Examinatrice
M. Emmanuel Procyk,	Univ. de Lyon	Examinateur
M. Mathias Pessiglione,	ICM	Examinateur
M. Benoît Girard,	ISIR - UPMC - CNRS	Directeur de thèse
M. Mehdi Khamassi,	ISIR - UPMC - CNRS	Directeur de thèse

Institut des Systèmes Intelligents et de Robotique
Université Pierre et Marie Curie
Pyramide - T55/65, 4 Place Jussieu 75005 Paris

THESE

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 Contexte

La mémoire est considérée aujourd'hui comme émergeant de l'activité des milliards de neurones qui constituent notre cerveau. Elle permet de se reconnaître entre nous, de savoir qui nous sommes et de pouvoir réaliser une multitude de tâches. Si la mémoire est un terme utilisé dans le langage courant, sa signification dans le champ scientifique est souvent multiforme. Si l'on pose la question à deux neuroscientifiques différents de définir la mémoire, les réponses s'accorderont sans doute en premier sur quelques principes basiques, divergeront sur la manière de l'étudier en fonction des sujets d'études personnels et s'accorderont enfin sur une compréhension limitée. Si l'on demande à ces deux neuroscientifiques de débattre, il est alors fort probable qu'au cours de leur discussion la mémoire devienne les mémoires, c'est-à-dire un ensemble de fonctions servant différentes finalités et auxquelles nous avons accès librement ou en fonction des circonstances. C'est précisément cette question que nous allons aborder d'un point de vue théorique dans ce manuscrit. Est-il possible de modéliser la coordination de systèmes de mémoires, ce qui fait qu'ils coopèrent par moments, ou au contraire qu'ils entrent parfois en compétition pour la réalisation d'une tâche ? Si oui, quels sont les principes qui gouvernent les variations temporelles de cette coordination ? Plus précisément, quels sont les principes qui font qu'un système de mémoire prime par rapport à l'autre dans certains types de tâches ou à certains moments particuliers de la tâche, et que l'autre système de mémoire soit impliqué davantage à d'autres moments ?

En premier, il est nécessaire de discuter du contexte dans lequel se place le travail développé dans ce manuscrit. Bien évidemment, la délimitation du cadre dans laquelle se meut l'objet d'étude détermine entièrement ce que l'on peut produire ou penser. Toutefois et dans la majorité des cas, ce cadre s'impose de lui-même au gré de l'évolution des idées. Un exemple simple concernant le sujet de la mémoire est celui de la métaphore de l'entrepôt [KORIAT et GOLDSMITH, 1996] qui a dominé l'étude de la mémoire et qui a été explicitée par H. Roediger :

> Les processus mentaux sont souvent décrits dans des termes qui s'appliquent à des comportements dans un espace physique. Nous parlons de *stocker* des souvenirs, de les *chercher* et de les *retrouver*. Nous *organisons* nos pensées ; nous *cherchons* nos souvenirs *perdus*, et si nous sommes chanceux, nous les *trouvons*. (ROEDIGER [1980], en anglais dans le texte)

Les métaphores sont courantes en science car elles permettent de penser un phénomène en le ramenant à quelque chose de connu. Dans notre cas, la métaphore de l'entrepôt permet de concevoir la mémoire comme un lieu où l'information est stockée. Il est d'ailleurs fort probable que cette métaphore constitue le faîte de toutes les métaphores en science comme le montre ROEDIGER [1980] en listant tous les objets ayant servi de comparaison avec la mémoire depuis la tablette d'argile d'Aristote.

En opposition à cette vision *structurale* de la mémoire, on peut citer la vision *procédurale* et la vision *fonctionnelle* [NEATH et SURPRENANT, 2003]. Dans le premier cas, l'accent est mis sur le processus qui crée et re-crée la mémoire en continu plutôt que sur sa localisation. Dans le deuxième cas, le but est d'énumérer des principes généraux à la mémoire qui permettent de répondre à la question : quelle est la fonction ou le but d'un tel acte comportemental ? Il est évident qu'aucune vision seule ne peut expliquer la totalité du processus mnésique. Néanmoins, la vision structurelle à travers la métaphore de l'entrepôt est actuellement dominante.

Pour compléter le cadre de ce manuscrit, cette vision de la mémoire par la métaphore de l'entrepôt est complémentaire d'une autre métaphore dominante actuelle : le cerveau

comme un outil traitant l'information. Cette approche de l'étude du cerveau, catapultée entre autres par D. Marr et son influent «Vision» [MARR, 1982], est résumable ainsi :

> La plupart des phénomènes qui sont centraux pour nous en tant qu'êtres humains - le mystère de la vie et de l'évolution, de la perception et des sentiments et des pensées - sont principalement des phénomènes de traitement de l'information, et si nous voulons les comprendre pleinement, notre réflexion sur ces sujets doit inclure cette perspective. (MARR [1982] en anglais dans le texte)

De manière plus générale, la volonté de D. Marr était ainsi la compréhension des systèmes complexes manipulant de l'information. Cet objectif l'a ainsi conduit à poser trois niveaux d'analyse d'un tel système. Le premier niveau doit produire une théorie computationnelle permettant de donner du sens aux règles de manipulation de l'information (dans notre cas, l'information est équivalente au contenu mnésique). Le deuxième niveau pose la question de la mise en oeuvre algorithmique permettant une simulation avec une définition des entrées et des sorties de l'algorithme. Le troisième niveau concerne la compréhension de la réalisation physique de l'algorithme (dans notre cas par le substrat neuronal).

C'est précisément la simulation, par l'utilisation de l'outil informatique, qui rend aujourd'hui cette approche à la fois bénéfique et incontournable. Les résultats présentés dans ce manuscrit sont des résultats de simulation de modèles computationnels du comportement. En comparant un comportement simulé grâce à un algorithme à un comportement réel dans le cadre d'une tâche étudiant l'interaction de mémoires, il nous est ainsi possible d'inférer et de discuter les processus en jeu lorsque des systèmes de mémoire interagissent.

Pour finir, l'étude et la modélisation de la coordination des systèmes de mémoire est aussi motivée par des considérations techniques. Le transfert des connaissances issues de l'étude du cerveau (ou plus généralement de l'étude du vivant) permet aujourd'hui des avancées majeures en robotique. La capacité de doter un robot d'une ou plusieurs stratégies identifiables à des systèmes de mémoire permet ainsi une plus grande autonomie comportementale.

1.2 Problématique

Un rapide survol de la littérature sur la mémoire humaine et animale permet de mesurer la complexité du concept. La mémoire n'est pas un phénomène simple et unifié. Dans la plupart des cas, ce phénomène dépend de l'intérêt et du paradigme de recherche de celui qui l'étudie. Il semble ainsi plus juste de parler de phénomènes différents qu'il faut expliquer et ceci à travers des paradigmes parfois radicalement distincts.

En psychologie, une expérience classique est l'apprentissage de listes variant entre 10, 20 ou 30 mots. Il a ainsi été montré que si les sujets étaient interrogés pour restituer la liste juste après la mémorisation, les derniers mots de la liste étaient restitués en priorité. Connue sous le nom d'effet de mémoire récent, cette observation disparaît si le test s'effectue après 15 ou 30 secondes tandis que la restitution des premiers mots des listes est peu affectée. Il fut ainsi naturel de postuler l'existence de deux types de mémoire et de les appeler mémoire à court-terme et mémoire à long-terme [ATKINSON et SHIFFRIN, 1968]. C'est une distinction qui a été reprise et étendue dans beaucoup de directions comme nous le verrons au cours du deuxième chapitre. La mémoire à court-terme est ainsi invoquée pour expliquer le fait que les éléments à la fin de la liste sont bien récités juste après

que la liste est présentée. L'échelle de temps de la mémoire à court-terme peut ainsi varier de quelques secondes à quelques minutes. Au contraire, la capacité de rétention de la mémoire à long-terme peut s'étendre d'une minute à une heure jusqu'à la vie entière.

Dans cet exemple simple, le but est d'évaluer la capacité de la mémoire tout en réduisant sensiblement ses différentes fonctionnalités. Dans certains cas, la mémoire peut aussi servir à prendre une décision en fonction d'un contexte. En ayant accumulé de l'expérience, il est ainsi possible de choisir la meilleure option. On parlera plus couramment de stratégie de décision. Chaque stratégie associée à un système de mémoire traite et stocke l'information différemment. L'exemple traditionnellement invoqué est celui du trajet de son lieu de travail à son domicile. Effectuées quotidiennement, les actions deviennent automatiques et inconscientes. Dans le cadre d'une visite chez le médecin, le trajet va nécessiter une modification du comportement impliquant le souvenir de la cartographie de l'environnement. Dans les deux cas, les décisions sont guidées différemment et, comme nous le verrons au chapitre 2, il est possible de les relier à différents substrats neuronaux.

1.3 Plan

Ce manuscrit se compose de 7 chapitres.

— Le chapitre 2 présente les expériences de neurobiologie ayant permis le développement de la théorie des systèmes de mémoire parallèles. Nous verrons notamment les études de cas de dommages focaux dans certaines parties du cerveau chez l'humain induisant des amnésies spécifiques ainsi que leur tentative de reproduction par des lésions dans les modèles animaux. Dans une deuxième partie, nous présenterons les expériences de séparation des systèmes de mémoire chez le rongeur par lésions de substrats neuronaux spécifiques. Pour finir, nous parlerons des processus d'interaction révélés dans certaines conditions par ces mêmes études de lésions.

— Le chapitre 3 introduit la théorie de l'apprentissage par renforcement. Issue de l'intelligence artificielle, cette théorie nous donne les outils formels permettant de modéliser un comportement dont le but est de maximiser l'obtention d'une récompense. Nous verrons comment cette théorie s'avère pertinente pour décrire comment un sujet apprend à mémoriser les valeurs de ses actions de façon à prendre de mieux en mieux les décisions qui lui permettent d'obtenir une récompense. Dans nos modèles, elle sera la base principale des processus de décision utilisant les différentes formes de mémoire.

— Le chapitre 4 présente un ensemble de modèles théoriques de la coordination de systèmes de mémoire. Certains de ces modèles constitueront ainsi une source d'inspiration pour la contribution de ce manuscrit.

Les chapitres suivants présentent notre contribution.

— Dans le chapitre 5, nous présentons les résultats de la modélisation d'une interaction entre une mémoire de travail et une mémoire procédurale. Différents modèles de coordination issus de la littérature ainsi qu'une proposition nouvelle ont été testés. Dans un premier temps, ces modèles ont été confrontés à la reproduction d'un comportement (choix et temps de réaction) observé chez des sujets humains dans une tâche d'association visuo-motrice. Dans un deuxième temps, c'est le comportement observé chez le singe dans une tâche similaire qui a été reproduit.

— Dans le chapitre 6, nous présentons un travail de modélisation complémentaire qui vise à reproduire l'apprentissage de séquences d'actions motrices chez la souris dans un labyrinthe sans indice visuel. Nous avons testé plusieurs systèmes de mémoire, chacun contenant une stratégie possible de résolution du problème.

Pour finir, nous discuterons dans le dernier chapitre des différents choix de modélisation effectués ainsi que des limites possibles de notre approche et des pistes pour les dépasser.

Chapitre 2

Systèmes de mémoire parallèles

Sommaire

2.1 Introduction

La première définition de «système de mémoire» est donnée par SHERRY et SCHACTER [1987] en essayant d'intégrer la théorie des systèmes de mémoire parallèles d'un point de vue évolutionniste. Ainsi, le terme *système de mémoire* fait référence à l'interaction entre des processus d'acquisition, de rétention et de récupération caractérisés par certaines règles d'opérations. L'expression *systèmes de mémoire multiples* est donc l'idée de systèmes qui cohabitent et fonctionnent selon des règles d'opérations fondamentalement différentes.

De manière plus élaborée, une autre définition d'un système de mémoire a été donnée par SCHACTER et TULVING [1994] en énonçant d'abord ce qu'il n'est pas. Un système de mémoire n'est pas une forme de mémoire, un processus mnésique ou une tâche quantifiant un processus mnésique. Ces termes sont importants pour l'étude de la mémoire mais ne suffisent pas à définir un système de mémoire. Une distinction existe entre des systèmes de mémoire et des formes de mémoire comme la mémoire olfactive, la mémoire verbale, la mémoire sensorielle qui sont utiles pour organiser des faits empiriques. Un système de mémoire exprime une forme de mémoire mais l'inverse n'est pas vrai selon ces auteurs. La même distinction est possible pour les processus mnésiques couramment utilisés pour parler de mémoire. L'encodage, la rétention ou la restitution d'information au service de la résolution d'une tâche mnésique sont des processus génériques à un système de mémoire mais ne le définissent pas. En conséquence, ces auteurs proposent une liste ouverte de trois critères.

1. Le premier critère est catégoriel. Pour un système de mémoire, il existe un ensemble de tâches que ce système peut résoudre. Ces tâches se différencient par leur contenu informationnel mais se rejoignent sur le type d'information retenu (des mots pour la mémoire verbale ou des souvenirs pour la mémoire autobiographique). Cette catégorisation fonctionnelle de systèmes de mémoire permet d'introduire l'expérience de dissociation. Si un changement d'état dans le système nerveux (de préférence localisable ou identifiable) tel qu'une lésion, une administration de drogues ou un manque de sommeil induit la suppression totale d'une seule catégorie fonctionnelle tout en épargnant d'autres catégories, alors le substrat neuronal altéré est impliqué dans tout ou partie de la catégorie fonctionnelle. Cette expérience de dissociation est très importante pour la théorie des systèmes de mémoire parallèles car elle permet de révéler un système de mémoire. Plusieurs de ces expériences seront exposées au cours de ce chapitre.
2. L'isolation d'un système de mémoire doit permettre de définir une liste de propriétés et fonctions spécifiques. Une liste doit ainsi contenir des règles d'inférence, un type d'information manipulée et un substrat neuronal. De plus, une finalité du système en relation avec une utilité biologique doit pouvoir être énoncée. Un exemple est donné dans O'KEEFE et NADEL [1978] avec la description d'un système de codage de l'espace sous forme de route (*taxon*) ou sous forme de carte (*locale*).
3. Le troisième critère est la convergence des dissociations. En effet, une critique portée notamment par ROEDIGER et collab. [1990] est le risque de prolifération de systèmes de mémoire. Pour contenir une telle épidémie, la convergence de plusieurs expériences de dissociation est nécessaire pour établir, sur des bases empiriques solides, l'existence d'un système de mémoire.

Cette liste, bien que volontairement peu précise, permet néanmoins de limiter l'utilisation de l'expression *système de mémoire*. Cela permet aussi de se détacher, principalement pour les travaux en psychologie cognitive humaine, de genres de mémoire liés le

plus souvent à une théorie. Une liste non exhaustive est ainsi donnée par NEATH et SURPRENANT [2003] dans le tableau 2.1.

TABLEAU 2.1 – Différentes formes de mémoire selon leur appartenance à une théorie. Tableau reproduit de NEATH et SURPRENANT [2003]

Type d'information	Terme neutre	Terme lié à une théorie
Sensations	Mémoire sensorielle	Mémoire iconique
		Mémoire résonante
		Mémoire acoustique
		precatégorielle
Information retenue brièvement	Mémoire immédiate	Stockage à court-terme
		Mémoire à court-terme
		Mémoire primaire
		Mémoire de travail
Information retenue indéfiniment	Mémoire générique	Stockage à long-terme
		Mémoire à long-terme
		Mémoire secondaire
Histoire personnelle	Mémoire autobiographique	Mémoire épisodique
Connaissance	Mémoire générique	Mémoire sémantique

Par exemple, le terme mémoire de travail fait implicitement référence à la proposition de BADDELEY et HITCH [1974] pour décrire un système mêlant boucle phonologique et calepin visuo-spatial. Le terme mémoire immédiate est l'équivalent neutre. Le but de ce chapitre n'est donc pas d'exposer la liste complète des formes de mémoire mais les études qui nous informent sur la théorie des systèmes de mémoire parallèles.

Comme nous l'avons vu dans la liste des critères, l'expérience de dissociation est la pierre angulaire de la théorie des systèmes de mémoire parallèles. Une grande partie de ce chapitre est donc dédiée à exposer une vue d'ensemble sur les travaux de dissociation tels que présentés dans la plupart des revues sur la question [GOLD, 2004; HARTLEY et BURGESS, 2005; KIM et BAXTER, 2001; PACKARD et GOODMAN, 2013; POLDRACK et PACKARD, 2003; SQUIRE, 2004; WHITE et collab., 2013; YIN et KNOWLTON, 2006].

D'un point de vue empirique, on peut dater le début de l'histoire des systèmes de mémoire avec le célèbre patient H.M. [SCOVILLE et MILNER, 1957]. Celui-ci était incapable d'encoder ou de restituer des évènements ou des faits tout en pouvant apprendre une tâche motrice. Cependant, la préhistoire de cette théorie est riche en débats philosophiques [POLSTER et collab., 1991]. L'exemple immanquable est le débat qui opposait les théoriciens de l'apprentissage Hull et Thorndike [HULL, 1943; THORNDIKE, 1933] à Tolman [TOLMAN, 1948]. Les premiers défendaient la théorie *behavioriste*. Le comportement est explicable à travers l'apprentissage d'associations stimulus-réponse. Leur but était d'éliminer du débat des concepts non scientifiques tels que l'intentionnalité, l'attente ou les représentations internes. Le second défendait la théorie *cognitiviste*. L'animal acquiert une connaissance de la conséquence de ses actions. Au final, ces débats sont apparus comme solubles dans la théorie des systèmes de mémoire au fur et à mesure de l'accumulation des données empiriques dont une partie sera exposée dans ce chapitre [YIN et KNOWLTON, 2006].

Ce chapitre va donc parcourir l'histoire d'un champ scientifique en partant du milieu du XX^{e} siècle, c'est-à-dire avec le patient H.M. La première partie sera consacrée aux observations de dissociation chez l'humain et aux tentatives de reproduction chez le singe. Cela débouchera sur une classification des mémoires chez l'humain au début des années

2000. Dans une seconde partie, nous parlerons des expériences de dissociation chez le rongeur qui ont aussi largement contribué à fournir des preuves empiriques sur l'existence des systèmes de mémoire. La dernière partie sera consacrée aux études, principalement en conditionnement instrumental ou en navigation chez le rat, qui nous informent sur les processus sous-tendant l'interaction entre systèmes de mémoire.

2.2 Amnésie et dissociation chez l'humain

2.2.1 Le cas H.M.

Le cas le plus cité d'amnésie chez l'homme est celui du patient H.M. [SCOVILLE et MILNER, 1957]. En 1953, à l'âge de 27 ans, H.M. subit une intervention chirurgicale pour lui retirer le lobe temporal médian dans le but de soigner une épilepsie chronique. Le résultat notable fut une perte de la capacité à créer de nouveaux souvenirs.

Pendant 14 ans, les neuropsychologues Suzanne Corkin et Brenda Milner, entre autres, vont suivre le patient H.M. en le soumettant à des tests en laboratoire. Dans le contexte des connaissances de l'époque, la question était de délimiter les fonctions cognitives qui étaient affectées. Pour l'essentiel, il fut montré que les capacités intellectuelles, motrices, perceptuelles et langagières étaient préservées [CORKIN, 1968, 1984; MILNER et collab., 1968]. La conclusion était donc que la perte du lobe temporal médian chez le patient induisait exclusivement des problèmes mnésiques.

Globalement, le cas H.M. et d'autres cas d'amnésies ont été utilisés pour appuyer des théories sur (1) le rôle de l'hippocampe pour la mémoire [SCOVILLE et MILNER, 1957] (2) la distinction entre mémoire à court-terme et long-terme [ATKINSON et SHIFFRIN, 1968; WICKELGREN, 1968], (3) la distinction entre mémoire procédurale et déclarative [COHEN et collab., 1985; SQUIRE et ZOLA-MORGAN, 1991] et (4) la distinction entre tâche implicite et explicite [CERMAK et collab., 1995]. Ces distinctions seront discutées ci-dessous.

2.2.2 Etudes de lésions humaines

D'autres patients furent décrits après H.M. On peut notamment citer le patient R.B. [ZOLA-MORGAN et collab., 1986] qui environ 20 ans après H.M. fut décrit comme atteint d'une amnésie antérograde. Le patient est incapable de se souvenir d'évènements vécus *après* le début de l'amnésie. La zone atteinte après une ischémie était le CA1, c'est-à-dire une sous-structure de l'hippocampe. La lésion de R.B. était beaucoup plus circonscrite que H.M.ce qui suggérait l'idée que l'hippocampe était le principal acteur du système de mémoire atteint. Cette idée sera renforcée dans les études de lésions chez le singe et le rongeur et nous y reviendrons dans les sections suivantes.

En revenant aux années 60, des doutes sur l'étude des patients amnésiques, c'est-à-dire leur incapacité de se souvenir d'évènements, furent apportés par WARRINGTON et WEISKRANTZ [1968, 1970]. Les auteurs ont fait apprendre des listes de mots courants de différentes formes en commençant d'abord par une version fragmentée vers une version complète à des patients atteints d'amnésie antérograde suite à une dégradation du lobe temporal médian et à un groupe contrôle. Le taux d'erreurs est ainsi défini par le nombre d'étapes de complétion du mot avant reconnaissance. La restitution d'une liste de mots (1 min après apprentissage) est ensuite mesurée sous trois conditions : rappel verbal , reconnaissance visuelle ou présentation d'une version fragmentée. Le résultat fut que les patients amnésiques échouaient aux deux premières conditions mais égalaient les sujets

contrôles dans la troisième condition. La conclusion était donc que les patients amnésiques pouvaient restituer une information si un indice était fourni.

Le phénomène d'amorçage a été décrit à la même époque [MEYER et SCHVANEVELDT, 1971] mais ce phénomène ne sera décrit comme mémoire implicite que plus tard. L'amorçage est ainsi défini comme l'influence d'un premier stimulus pour une décision liée à un second stimulus. Suite à l'observation de la préservation de l'amorçage chez les patients amnésiques, la vision dominante de l'époque était de parler de déficit de restitution dans un certain contexte. Il semblerait qu'il existât une certaine réfraction à parler de systèmes de mémoire multiples, le principe de simplicité y contribuant.

Les premiers apports de la mémoire implicite comme intégrable dans la théorie de systèmes de mémoire sont donnés dans SCHACTER et BUCKNER [1998]; TULVING et SCHACTER [1990]. Ainsi cette mémoire implicite est mesurée à travers des tâches qui ne requièrent pas de rappel conscient d'information encodée. Selon SCHACTER et BUCKNER [1998], il est aussi possible de distinguer l'amorçage conceptuel de l'amorçage perceptuel. Le premier induit une similarité sémantique et le second une similarité de forme. Comme énoncé dans l'introduction, un substrat neuronal sous-tendant la fonction est nécessaire pour acquérir une dénomination de système de mémoire. Une première réponse a été donnée dans GABRIELI et collab. [1995] par une expérience de dissociation. Le patient M.S. a subi à 16 ans une ablation du lobe occipital droit dans le but de soigner une épilepsie. A l'issue de la chirurgie, le patient souffrait d'hémianopsie, c'est-à-dire la perte d'une moitié de son champ visuel. Néanmoins, aucune autre perte de facultés mentales n'a pu être décelée par des tests de cognition. Les auteurs ont soumis M.S. à des tests de reconnaissance (faisant appel à une mémoire explicite) et des tests d'amorçage (utilisant donc une mémoire implicite). Il apparut que M.S. montrait une capacité normale dans le rappel verbal et l'amorçage conceptuel mais était déficient dans l'amorçage visuel. Les travaux de BUCKNER et collab. [1995] sont venus consolider ces résultats primaires. Utilisant les techniques de neuro-imagerie de l'époque, les auteurs ont enregistré les variations de flux sanguins de sujets réalisant une tâche d'amorçage. Il apparut que, durant la phase de complétion de l'indice verbal, le flux sanguin diminuait dans le cortex occipital, région associée à un traitement perceptif selon les auteurs. Cette observation a été reproduite ultérieurement dans plusieurs études [SCHACTER et BUCKNER, 1998; SCHACTER et collab., 1996].

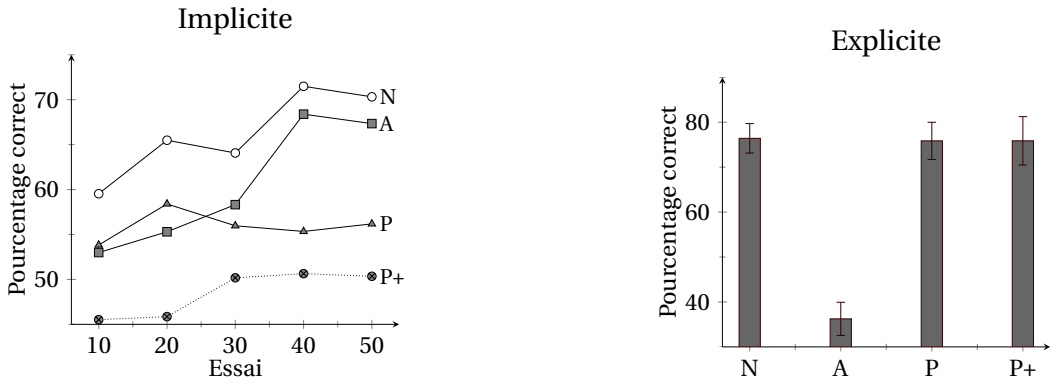

FIGURE 2.1 – *Gauche* : Performance sur la tâche de classification probabiliste par un groupe contrôle (N, n=15), un groupe de patients amnésiques (A, n=12), un groupe de patients parkinsoniens (P, n=20) et un groupe de patients parkinsoniens avec les symptômes aggravés (P+, n=10). *Droite* : Performance sur la tâche de mémoire déclarative des mêmes groupes. (Adapté de KNOWLTON et collab. [1996])

Pour finir, la première description d'un système sous-tendant l'apprentissage d'habitudes chez l'humain a été apporté par KNOWLTON et collab. [1996]. Le striatum avait déjà été lié à la consolidation progressive d'associations stimulus-réponses chez le rongeur [PACKARD et collab., 1989] mais l'équivalent d'une expérience de dissociation n'avait pas été démontré chez l'humain. Dans KNOWLTON et collab. [1996], les auteurs ont testé une expérience de dissociation entre mémoire déclarative et mémoire d'habitudes entre 12 patients souffrant de lésions du lobe temporal et 20 patients atteints de la maladie de Parkinson. Cette maladie cause une dégénérescence de la substance noire, noyau du système nerveux situé au niveau du mésencéphale. Dans une première tâche, les sujets apprennent ainsi l'association entre un ensemble d'indices et une conséquence. La combinaison probabiliste des indices rend impossible la mémorisation explicite des associations. Dans une deuxième tâche, les sujets doivent répondre à un questionnaire à choix multiples sur les indices, la forme de l'écran d'ordinateur ou le type d'essai qu'ils ont expérimenté. La dissociation apparaît ainsi clairement puisque les patients parkinsoniens réalisent la tâche d'association au niveau de la chance et les patients amnésiques échouent à la tâche déclarative. Ce résultat est reproduit dans la figure 2.1.

2.2.3 Modèles primates de l'amnésie

Les tentatives de reproduction de l'amnésie antérograde de H.M. ont commencé dans les années 80 et le modèle animal a rendu possible la recherche systématique des structures spécialisées dans la mémoire. Le plan est simple. Les singes sont lésés bilatéralement pour une région ou combinaison de régions. L'effet de la lésion est ensuite évalué quantitativement. La performance d'un singe est mesurée sur des tâches similaires à celles utilisées pour détecter les amnésies chez les patients humains.

La première preuve d'une réplication complète d'amnésie antérograde chez le singe a été fournie par MISHKIN [1978]. La problématique de l'époque était surtout de circonscrire les régions impliquées pour l'obtention d'une amnésie antérograde complète.

Des troubles de la mémoire spatiale avaient déjà été observés pour des lésions de l'hippocampe sur le modèle animal sans toutefois parvenir au désordre décrit pour des cas d'amnésie humaine [O'KEEFE et NADEL, 1978]. En élargissant la lésion à l'amygdale et en testant sur une tâche de reconnaissance d'objet (les singes doivent se souvenir de quel objet ils ont vu, pour que, à la seconde présentation de l'objet apparié avec un nouvel objet, ils choisissent l'objet nouveau), l'auteur montre ainsi qu'une combinaison de lésion de l'hippocampe et de l'amygdale est nécessaire pour induire un désordre mnésique profond. Cette observation est en concordance avec le fait que les patients ayant des troubles mnésiques similaires à ceux de H.M. ont des lésions dans les deux régions.

Ces expériences ont été reproduites au cours de la décennie suivante pour répondre à diverses questions. Une de ces questions était, par exemple, le rôle de la tige temporale dans ces expériences d'amnésie induite. En abrégé, la tige temporale est un ensemble de fibres de matière blanche passant à travers le lobe temporal et connectant diverses structures notamment l'insula, l'amygdale ou le noyau caudé [CHOI et collab., 2010]. Dans ZOLA-MORGAN et collab. [1982], les auteurs ont reproduit l'expérience de MISHKIN [1978] tout en y incluant une lésion de la tige temporale. L'ensemble de ces résultats est représenté dans la figure 2.2. La nécessité d'une lésion de l'hippocampe et de l'amygdale (H+A+) apparaît ainsi clairement lorsque l'on compare les différents groupes de singes chacun ayant une lésion particulière à ceux n'ayant pas de lésion (N). Aussi il apparut qu'une lésion de la tige temporale (Ts) n'était pas requise pour l'induction d'une amnésie profonde.

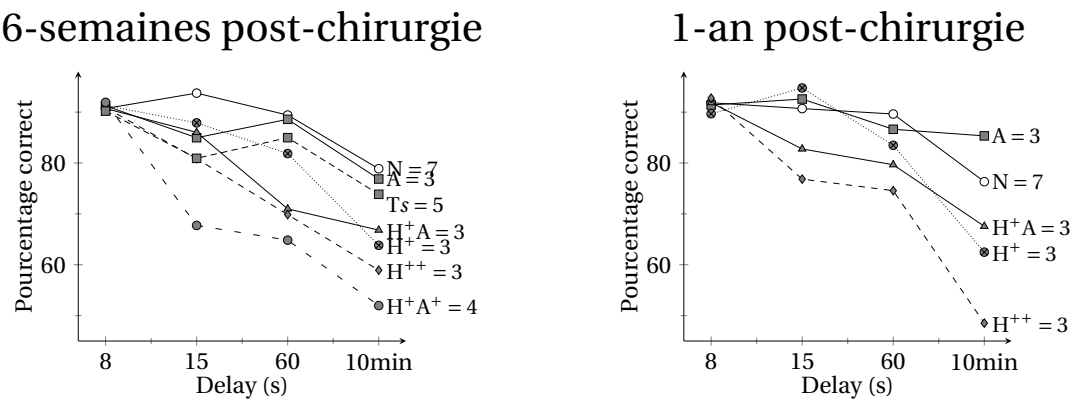

FIGURE 2.2 – *Gauche* : performance sur une tâche différée (delay) de reconnaissance de forme pour sept singes normaux (N) et 5 groupes de singes avec lésions 6 à 8 semaines après la chirurgie : A, lésion circonscrite à l'amygdale épargnant le cortex alentour ; H^+, lésion de l'hippocampe et cortex para-hippocampique ; H^+A, une combinaison de ces deux lésions ; H^+A^+, lésions de l'hippocampe, de l'amygdale et de régions corticales adjacentes (cortex perirhinal, entorhinal et parahippocampique), H^{++}, lésions de l'hippocampe et des régions adjacentes ; T*s*, lésion de la tige temporale. *Droite* : performance sur la même tâche 1 à 2 ans après chirurgie pour 4 des 5 groupes de singes. (Adapté de SQUIRE et ZOLA-MORGAN [1991]; ZOLA-MORGAN et collab. [1982])

Comme cela a été expliqué dans l'introduction, un système de mémoire est établi comme tel s'il montre un ensemble de dissociations. Un déficit dans un système de mémoire doit s'accompagner d'une baisse de performance dans une tâche *A* tout en gardant des performances intactes pour la tâche *B* sachant que les tâches *A* et *B* exploitent deux types de mémoire fondamentalement distincts.

Il était donc naturel que, lors de l'établissement du modèle d'amnésie antérograde chez le singe, la spécificité et les limites du déficit soient établies, tout comme elles le furent pour le patient H.M. Ainsi, la même tâche de reconnaissance d'objet différée a été reproduite tout en incluant un distracteur entre la présentation et le choix ou en augmentant la durée du retard de présentation [ZOLA-MORGAN et SQUIRE, 1985]. De même, l'influence de la modalité sensorielle a été évaluée dans MURRAY et MISHKIN [1984]. La tâche de discrimination d'objets a été portée dans une condition tactile : les singes doivent palper dans le noir un objet pour l'encoder en mémoire ou le discriminer. Il apparut que le déficit mnésique induit par la double lésion hippocampe-amygdale n'était pas dépendant de la modalité sensorielle. Les singes lésés échouent autant à la tâche en version tactile qu'en version visuelle.

La question de la sauvegarde de l'apprentissage de fonctions motrices (*skills*) a été résolue dans ZOLA-MORGAN et SQUIRE [1984]. Une première tâche consistait à attraper un gressin en dehors de la cage et à le ramener dans la cage en le faisant traverser trois rangées de barreaux espacés de 3cm. La progression de l'apprentissage du *skill* (qui s'étale sur 4 jours) est équivalente pour des singes normaux et des singes avec lésions de l'hippocampe et de l'amygdale. La deuxième tâche consistait à manoeuvrer un bonbon attaché sur une barre de fer pour le récupérer dans un laps de temps fini. La diminution du temps de manoeuvre du bonbon pour les deux groupes montre encore une fois l'équivalence de l'apprentissage moteur.

La mémoire immédiate a été évaluée dans OVERMAN et collab. [1990]. Néanmoins, les résultats de la figure 2.2 montraient déjà des performances équivalentes à celles du groupe contrôle pour un retard de 8s. Cela suggérait que la mémoire immédiate est préservée pour une lésion hippocampe-amygdale. Dans OVERMAN et collab. [1990], les singes

doivent effectuer une tâche de reconnaissance immédiate et le résultat est congruent avec les résultats de la figure 2.2. Ainsi, il semblerait que le seuil critique au-delà duquel l'information est perdue soit de 10 secondes.

Une expérience de lésion intéressante est celle de SALMON et collab. [1987]. Le déficit mnésique de H.M. est qualifié d'amnésie antérograde. Dans SALMON et collab. [1987], les singes montrent une amnésie rétrograde, c'est-à-dire l'incapacité à se souvenir d'évènements survenus *avant* le début de l'amnésie. Les singes furent ainsi entraînés 32, 16, 8, 4 ou 2 semaines avant la chirurgie dans une tâche de discrimination d'objets, tâche dans laquelle ils excellent comme le montre le groupe contrôle. C'est un groupe de 100 objets appariés deux par deux qui est donc familier au singe. Dans toutes les conditions de période d'entraînement, les singes sont incapables de résoudre la tâche après la chirurgie. Ils montrent donc une amnésie rétrograde. Les auteurs ont aussi évalué la mémoire motrice et les singes lésés ont une restitution parfaite d'un *skill* moteur appris avant la lésion. Cette étude montre donc qu'une lésion de l'hippocampe et de l'amygdale induit une amnésie rétrograde sévère pour des faits ayant été appris entre 2 semaines et 8 mois avant la chirurgie.

Au début des années 90, le consensus était donc qu'une lésion combinée de l'amygdale et de l'hippocampe était nécessaire pour reproduire l'amnésie antérograde de H.M. Une remise en cause de ces résultats a été apportée par SQUIRE et ZOLA-MORGAN [1991]. Un examen des coupes histologiques minimisait ainsi le rôle de l'amygdale. Lors de la procédure de lésion, un ensemble de régions adjacentes à l'amygdale était lésé par défaut. Ces régions étaient le cortex entorhinal et perirhinal qui sont des voies d'entrées majeures pour l'hippocampe. Comme montré dans la figure 2.2, les piètres performances du groupe présentant une lésion de l'hippocampe, du cortex entorhinal antérieur et du cortex perirhinal (H^{++}) (donc épargnant l'amygdale) sont équivalentes à celles du groupe $H^{+}A^{+}$. Il fut ainsi conclu que l'amygdale n'était pas une composante du lobe temporal médian et même que cette structure pouvait constituer un système de mémoire propre. Cela sera abordé à nouveau dans la section sur l'étude de dissociations chez le rongeur.

2.2.4 Classification des mémoires chez l'humain

Au début des années 2000, l'accumulation de données empiriques a permis de clarifier la situation. En effet, la dichotomie entre mémoire déclarative et mémoire non-déclarative est apparue comme trop simple. Si le système de mémoire déclarative existe, le terme mémoire non-déclarative fait référence à un ensemble de systèmes de mémoire (habitudes, amorçage, *skills*) qui ont parfois peu de choses en commun. Dans SQUIRE [2004], l'auteur propose la classification reportée dans la figure 2.3.

L'amnésie antérograde comme celle de H.M. est donc une atteinte à la mémoire déclarative. Cette mémoire encode les faits et les évènements et son fonctionnement est lié au lobe temporal médian.

En opposition, il existe des mémoires non déclaratives qui peuvent être subdivisées en plusieurs catégories : mémoire des *skills* et des habitudes, mémoire d'amorçage, mémoire de conditionnement classique. Cette subdivision en capacités inconscientes distinctes est liée à d'autres courants de recherche sur d'autres formes de mémoire. Par exemple, l'apprentissage d'habitudes a été étudié chez le rongeur [PACKARD et collab., 1989] et nous y reviendrons dans la section suivante. Néanmoins, ces capacités mnésiques sont liées par une incapacité à fournir un rappel explicite conscient et une certaine inflexibilité.

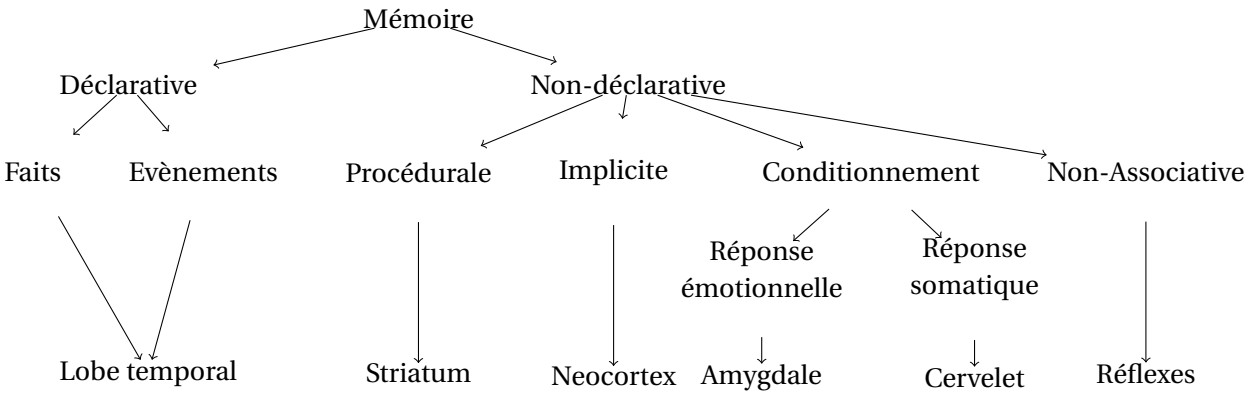

FIGURE 2.3 – Taxonomie des systèmes de mémoire chez les mammifères. Chaque système est associé à la structure qui supposément le sous-tend. Adapté de SQUIRE [2004]

2.3 Dissociation chez le rongeur

2.3.1 Dissociation double

Les effets de lésions de l'hippocampe sur la mémoire chez le rongeur ont été étudiés à partir des années 70 [HIRSH, 1974; O'KEEFE et NADEL, 1978; OLTON et collab., 1978]. Ces premières études ont ainsi mis en évidence l'importance de l'hippocampe pour des tâches de navigation. Néanmoins, il apparut qu'une lésion de l'hippocampe pouvait être sans conséquence pour certaines tâches [HARLEY, 1972; SAMUELS, 1972; SILVEIRA et KIMBLE, 1968]. Dans la plupart de ces études, le facteur déterminant pour la réussite de la tâche est une réponse appropriée après la discrimination d'un indice sensoriel. De même, plusieurs études montraient que le striatum était important pour résoudre ce type de tâche. Par exemple, dans PRADO-ALCALA et collab. [1975], les animaux avec une lésion du noyau caudé échouent à une tâche d'évitement passif. Dans WHISHAW et collab. [1987], les animaux avec une lésion du noyau caudé et du putamen ne peuvent utiliser une stratégie de suivi d'indices dans une piscine de Morris.

La première double dissociation à partir d'une seule tâche a été établie dans PACKARD et collab. [1989]. En effet, le problème dans les études précédemment citées est l'absence d'homogénéité. Il apparaît ainsi difficile de dissocier et comparer des systèmes de mémoire si les tâches comportementales sont fondamentalement différentes. Dans PACKARD et collab. [1989], les auteurs ont utilisé le même labyrinthe radial de 8 bras dans les deux tâches.

La première tâche est appelée *win-stay*. A chaque essai, des récompenses sont placées dans 4 bras choisis aléatoirement. Chaque bras est éclairé comme indice visuel. Après qu'un rat a visité un bras éclairé, celui-ci est muni d'un appât une seconde fois. Lors de la seconde visite, la lumière s'éteint. Le rat peut ainsi gagner 8 récompenses pendant un essai en moins de 10 min. Une erreur est définie comme une entrée dans un bras non-éclairé. L'apprentissage s'effectue pendant 15 jours et le test de rétention est effectué des jours 16 à 19. Durant ces tests, un seul bras choisi aléatoirement est éclairé et muni d'un appât à la fois. L'essai s'arrête quand 8 bras ont été visités. La seconde tâche est appelée *win-shift*. A chaque essai, chaque bras est muni d'un appât une seule fois. Le rat peut donc obtenir 8 récompenses en moins de 10 min. Le nombre d'erreurs est mesuré en comptant le nombre de fois où le rat rentre une nouvelle fois dans un bras sans récompense. La phase d'apprentissage s'interrompt quand le rat est capable de visiter 7 bras différents une seule fois. Dans les deux tâches, les rats ont été motivés à explorer leur environnement en réduisant leur quantité de nourriture.

Deux groupes de rats ont reçu une lésion du fornix qui est une voie de sortie majeure de l'hippocampe (groupe H) et deux groupes de rats ont reçu une lésion bilatérale étendue du noyau caudé (groupe C). Dans chaque cas, un groupe de rat est entraîné sur une seule tâche. Le résultat est une incapacité du groupe H à résoudre la tâche *win-shift* tandis que le groupe C est équivalent au groupe contrôle. Inversement, le groupe C fait beaucoup plus d'erreurs que le groupe contrôle pour la tâche *win-stay*. Mais il apparut aussi clairement que le groupe H était significativement supérieur au groupe contrôle pour cette tâche. Le résultat est donc une double dissociation entre l'hippocampe et le noyau caudé chez le rongeur.

A l'époque, la dissociation entre noyau caudé et hippocampe était déjà suggérée par les études de lésion chez l'homme et le primate [ZOLA-MORGAN et collab., 1982; ZOLA-MORGAN et SQUIRE, 1985]. L'un des ajouts de cette étude chez le rat était donc de généraliser à travers l'ensemble des mammifères la possibilité de systèmes de mémoire distincts. La deuxième observation majeure était la facilitation de l'apprentissage pour le groupe H sur la tâche *win-stay*. Les auteurs ont suggéré pour la première fois le processus de compétition entre les systèmes de mémoire.

Chez le rongeur, d'autres doubles dissociations ont été étudiées à la même époque. Dans SUTHERLAND et MCDONALD [1990], la dissociation entre l'hippocampe et l'amygdale a été testée dans

1. une tâche de reconnaissance d'indices visuels et olfactifs appariés
2. une tâche de néophobie gustative (un simple test de mémoire qui détermine des changements dans les habitudes de consommation de nouveaux aliments)
3. une tâche de discrimination entre des indices seuls (amenant une récompense) ou appariés (n'amenant pas de récompense)
4. une tâche de navigation allocentrique dans une piscine de Morris
5. une tâche de conditionnement à la peur.

La performance des rats avec lésion de l'hippocampe est normale dans 1 si les deux stimuli sont dans la même modalité sensorielle, et dans 2. Ce groupe est inférieur au groupe normal dans la tâche 1 si les deux stimuli sont dans deux modalités différentes, dans 3, dans 4 et dans 5. Pour le groupe avec lésion de l'amygdale, les performances sont altérées dans les tâches 2 et 3 et sont *grosso-modo* identiques au groupe contrôle pour les tâches 1, 4 et 5. Le rôle de l'hippocampe est donc toujours impliqué dans des tâches nécessitant l'apprentissage et la rétention des relations entre stimuli ou indices mais aussi dans la reconnaissance d'un contexte dans le cas de la tâche 5. Le rôle de l'amygdale, bien qu'évident dans la deuxième tâche, est beaucoup plus controversé selon les auteurs. Pour conclure et dans le langage des systèmes de mémoire parallèles, la dissociation entre amygdale et hippocampe apparaît pour les tâches 2, 4 et 5.

Dans le cas d'une tâche de conditionnement, la dissociation entre hippocampe et amygdale est plus compliquée comme le montre PHILLIPS et LEDOUX [1992] à la même époque. Le rôle du contexte conditionné en opposition au stimulus conditionné a ainsi été examiné dans une chambre de conditionnement classique. L'hypothèse de base étant : l'hippocampe encode l'association contexte-réponse émotionnelle et l'amygdale encode l'association stimulus-réponse émotionnelle. En comparant un groupe avec lésion de l'amygdale (A) et un groupe avec lésion de l'hippocampe (H) sur la durée de réponse émotionnelle, il apparaît que seul le groupe A montre une absence de réaction dans les deux conditions. Le groupe H montre une durée de réaction fortement diminuée dans le cas contextuel mais une réaction normale avec stimulus conditionné. Selon les auteurs, les stimuli complexes, particulièrement ceux dont la configuration spatiale est importante

et formant un contexte sont traités par l'hippocampe avant d'être envoyés par projection avec le subiculum vers l'amygdale. Ainsi l'hippocampe constituerait, dans ce cas, un relai vers l'amygdale, tel le cortex sensoriel, au lieu d'un système à part entière. Cette étude constitue notre premier exemple d'une interdépendance entre les systèmes de mémoire.

2.3.2 Dissociation triple

Dans McDonald et White [1993], les auteurs présentent pour la première fois une triple dissociation entre l'hippocampe, le noyau caudé et l'amygdale. Cette étude est une continuation de Packard et collab. [1989] décrite dans la section précédente. On retrouve ainsi les tâches *win-stay* et *win-shift* plus une tâche de conditionnement. Selon les données de l'époque, le rôle de l'amygdale dans le conditionnement était déjà bien établi. Néanmoins, il apparaissait incertain que l'association entre un stimulus neutre et une réponse non-conditionnée fût sous-tendue que par un seul système de mémoire.

Dans la tâche de conditionnement adaptée au labyrinthe radial de Packard et collab. [1989], deux bras non-adjacents sont choisis aléatoirement et les autres bras sont fermés. Pendant un essai, l'un des deux bras est allumé tandis que l'autre reste éteint. Le rat est donc autorisé à naviguer entre un bras éclairé, un bras éteint et la plateforme centrale. Durant la phase de conditionnement, la moitié des rats d'un groupe a accès à 70 pièces de nourriture tout en étant confinée dans le bras éclairé tandis que l'autre moitié est nourrie tout en étant confinée dans le bras non-éclairé. Durant la session de test, les deux bras sont ouverts et le temps passé dans chaque bras est mesuré.

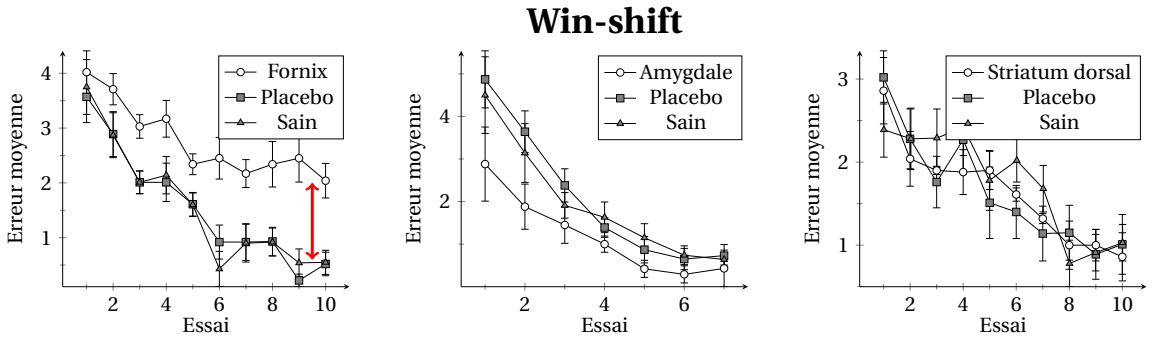

FIGURE 2.4 – Erreur moyenne (± SE) sur la tâche *win-shift* du groupe de rats avec lésion du fornix (*gauche*), avec lésion de l'amygdale latérale (*centre*) ou avec lésion du striatum dorsal (*droite*). Chaque groupe est représenté avec les performances des groupes placebo et sain respectifs. Adapté de McDonald et White [1993]

Au total, ce sont trois cent rats qui furent inclus dans l'étude. Pour apprendre chaque tâche séparément, trois groupes sont formés pour chaque type de lésion. Chaque groupe est subdivisé en un sous-groupe lésé, un sous-groupe placebo et un sous-groupe sain. L'amygdale latérale, le striatum dorsal et le fornix ont ainsi été lésés séparément. Les résultats sont présentés dans la figure 2.4 pour le *win-shift*, 2.5 pour la tâche de conditionnement et 2.6 pour le *win-stay*. Dans chaque figure, le principal résultat est souligné par une flèche rouge. Tout comme dans l'étude précédente [Packard et collab., 1989], les mêmes effets sont observés pour l'hippocampe et le striatum dorsal. De plus, une lésion de l'amygdale n'a d'effet que dans la tâche de conditionnement. L'amygdale n'est donc pas nécessaire pour apprendre le *win-shift* ou le *win-stay*.

Cette triple dissociation suggère donc que chaque système soutient un traitement spécifique de l'information. L'hippocampe permettrait un encodage de la relation entre

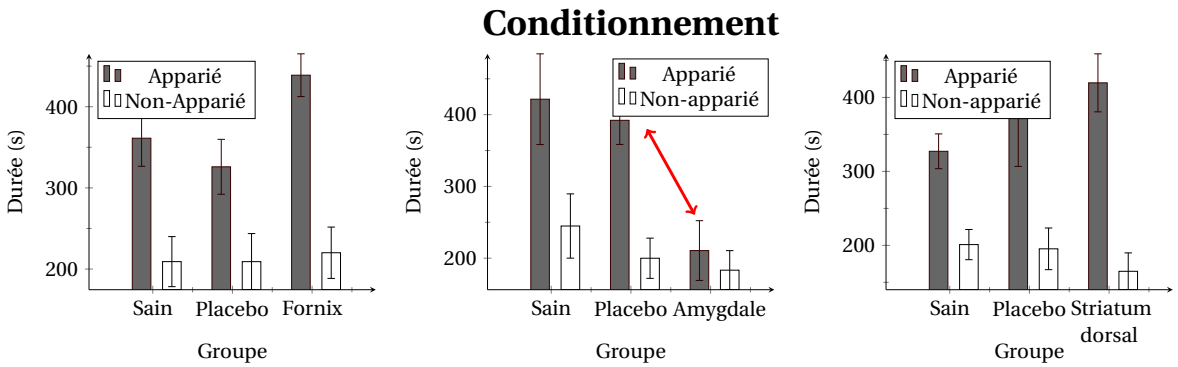

FIGURE 2.5 – Temps moyen (± SE ; secondes) passé dans le bras apparié et le bras non-apparié du groupe de rats avec lésion du fornix (*gauche*), avec lésion de l'amygdale latérale (*centre*) ou avec lésion du striatum dorsal (*droite*). Chaque groupe est représenté avec les performances des groupes placebo et sain respectifs. Adapté de MCDONALD et WHITE [1993]

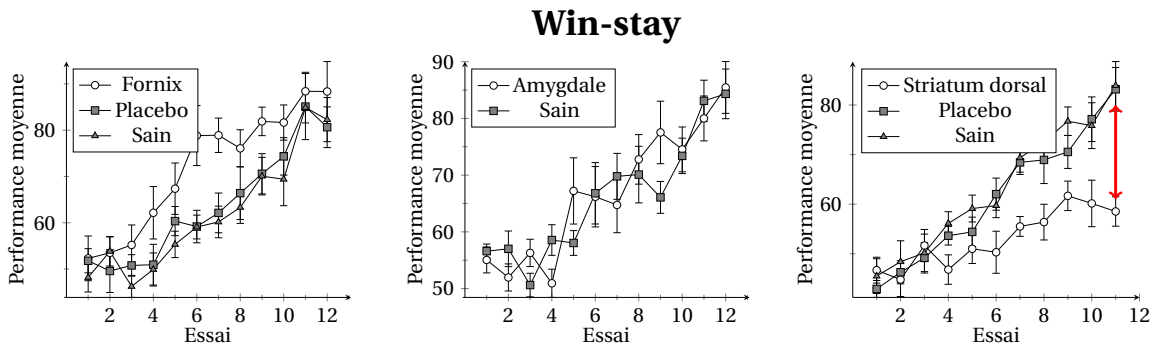

FIGURE 2.6 – Performance moyenne (± SE) sur la tâche *win-stay* du groupe de rats avec lésion du fornix (*gauche*), avec lésion de l'amygdale latérale (*centre*) ou avec lésion du striatum dorsal (*droite*). Chaque groupe est représenté avec les performances des groupes placebo et sain respectifs. Adapté de MCDONALD et WHITE [1993]

les stimuli, l'amygdale traiterait les associations entre stimuli et récompense dans des tâches ne nécessitant pas de réponse et le striatum dorsal renforcerait des associations stimulus-réponse.

2.3.3 Modèle conceptuel

Tout comme L. Squire au début des années 2000, les auteurs N. White et R. McDonald ont proposé une théorie des systèmes de mémoire parallèles [WHITE et MCDONALD, 2002] sur la base, entre autres, des résultats de MCDONALD et WHITE [1993]. Chaque système de mémoire a ainsi une structure centrale : hippocampe, striatum ou amygdale comme representé dans la figure 2.7.

Les auteurs postulent ainsi une direction de l'information de l'entrée sensorielle vers le comportement. Cette information neuronale est égalitairement reçue par chaque système et contient les relations entre les stimuli ou les différents évènements du monde extérieur. Tous les systèmes de mémoire ont ainsi accès aux mêmes informations sur la tâche à résoudre. Néanmoins, chaque système possède une architecture neuronale distincte et invariable. Selon les auteurs, la conséquence est une spécialisation dans le traitement de l'information reçue, et cela de manière indépendante.

De plus, la sélection d'un système pour encoder une situation est dépendante de la concordance entre l'information reçue et le style de traitement du système. Une telle si-

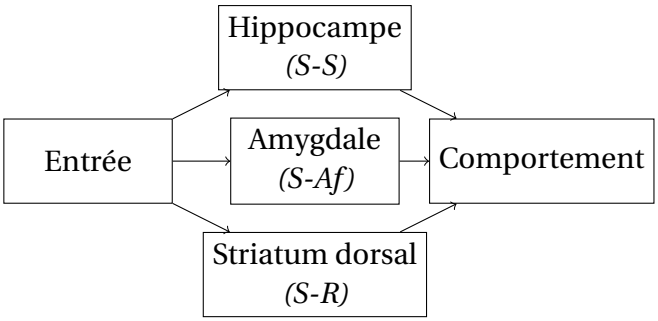

FIGURE 2.7 – Le concept de traitement parallèle (S : stimulus, R : réponse, Af : affect). Adapté de WHITE et MCDONALD [2002]

.

milarité va ainsi induire une activité neuronale cohérente dans le système concerné. Si un système possède un style de traitement moins similaire, l'encodage d'information sera moins élevé dans le système empêchant ainsi une influence sur la même situation future.

L'interaction entre les systèmes de mémoire est un sujet que nous aborderons dans la section suivante. Néanmoins, nous pouvons déjà énoncer les propositions des auteurs : coopération ou compétition entre les comportements induits par chaque système. Un exemple de coopération entre l'amygdale et l'hippocampe a déjà été décrit dans la section 2.3.1 avec une expérience de dissociation selon la présence d'un stimulus conditionné ou d'un contexte conditionné [PHILLIPS et LEDOUX, 1992]. A l'opposé, une compétition se produit quand deux systèmes proposent deux comportements mnésiques différents. Ainsi, un système peut produire un comportement défini comme bonne réponse selon l'expérimentateur et un autre système peut engendrer un comportement erroné, résultant en une interférence et une baisse possible des performances de l'animal. Une lésion du système produisant le comportement incorrect, supprimant l'interférence, fera augmenter les performances de l'animal au-dessus de celles du groupe sain [MCDONALD et WHITE, 1993; PACKARD et collab., 1989].

2.4 Interaction

2.4.1 Temporalité

La notion de temporalité entre systèmes de mémoire est apparue pour la première fois dans PACKARD et MCGAUGH [1996]. Les auteurs ont entraîné des rats à trouver une récompense dans un labyrinthe en croix tel qu'introduit dans TOLMAN [1948]. Le point de départ se situe au bras nord ou sud et la récompense est posée au bras est ou ouest. Cette étude a été réalisée, à l'époque, pour répondre au débat entre Tolman et Hull/Thorndike tel qu'énoncé dans l'introduction. Les connaissances grâce aux expériences de dissociation avaient progressé comme nous l'avons vu tout au long de ce chapitre. L'hippocampe apparaissait ainsi comme le siège possible de la théorie *cognitiviste* et le noyau caudé comme le siège possible de la théorie *behavioriste*.

Dans PACKARD et MCGAUGH [1996], les rats ont donc été entraînés pendant 7 jours à parcourir le chemin bras sud vers bras ouest. Au 8ème jour, les rats ont été testés en étant placés au départ dans le bras nord. Selon les auteurs, les animaux faisant le choix d'aller vers le bras ouest usent d'une stratégie de lieux tandis que les animaux allant dans le bras est usent d'une stratégie de réponse. Après le 8ème jour, l'entraînement continue avec départ du bras sud jusqu'au second test au jour 16 avec départ du bras nord. La position

de la récompense reste inchangée tout au long de l'expérience. Deux groupes reçoivent une infusion de lidocaïne le jour du test permettant une inactivation de l'hippocampe ou du noyau caudé. Deux groupes contrôles reçoivent une injection de solution saline dans l'une des deux structures. Les résultats sont présentés dans la figure 2.8.

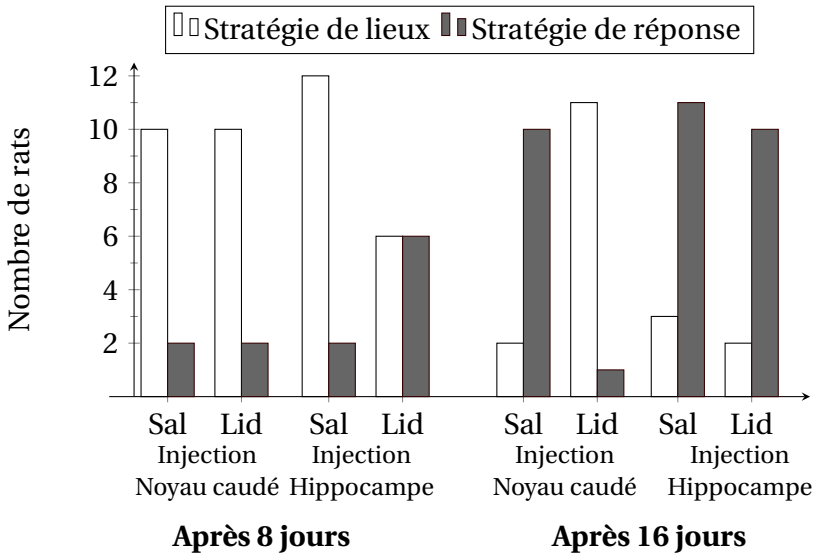

FIGURE 2.8 – Nombre de rats exprimant une stratégie de lieux ou une stratégie de réponse après inactivation de l'hippocampe ou du noyau caudé par une solution de lidocaïne (lid). Le groupe contrôle reçoit une injection saline (sal). Les tests s'effectuent après 8 jours ou 16 jours d'entraînement. Adapté de PACKARD et MCGAUGH [1996]

En comparant entre groupes recevant une injection saline et groupe recevant une injection de lidocaïne, on peut donc conclure que l'hippocampe sous-tend une stratégie de lieux et que le noyau caudé sous-tend une stratégie de réponse. Ce résultat est peu surprenant comme le remarquent les auteurs. L'autre observation est la diminution du nombre de rats exprimant une stratégie de lieux et inversement l'augmentation du nombre de rats exprimant une stratégie de réponse entre les jours 8 et 16. L'effet de surentraînement était déjà connu [HICKS, 1964; RITCHIE et collab., 1950] mais cette étude établit le lien avec deux structures neurales et dans les termes de la théorie des systèmes de mémoire parallèles. Plus important encore, une inactivation du noyau caudé montre une préservation de la stratégie de lieux après 16 jours d'entraînement. La prédominance de la stratégie de réponse après surentraînement n'est donc pas due à une disparition de la stratégie de lieux encodée dans l'hippocampe.

Est-il possible de modifier la transition entre stratégie de lieux et stratégie de réponse en stimulant un système de mémoire ? La réponse est donnée dans PACKARD [1999] en remplaçant la lidocaïne par du glutamate pour la même tâche dans la perspective d'activer un système de mémoire. Deux groupes (hippocampe et noyau caudé) ont ainsi reçu plusieurs jours avant les jours de test (8 et 16) une dose de glutamate juste après l'entraînement. Le résultat est une prédominance de la stratégie de lieux aux jours 8 et 16 pour le groupe stimulé dans l'hippocampe et une prédominance de la stratégie de réponse aux jours 8 et 16 pour le groupe stimulé dans le noyau caudé. Il est donc possible d'avancer le moment de transition en stimulant avec un neuro-transmetteur le noyau caudé. Un résultat similaire a été obtenu dans CHANG et GOLD [2003] en inactivant seulement l'hippocampe par de la lidocaïne sur la même tâche de labyrinthe en croix. Le résultat est une accélération significative des performances des rats pour apprendre la stratégie de réponse. L'inactivation de l'hippocampe facilite donc l'acquisition de la stratégie de

réponse et ce résultat est similaire à l'effet de compétition décrit dans PACKARD et collab. [1989].

Le rôle possible du cortex préfrontal dans la temporalité entre systèmes de mémoire a été étudié dans une série d'études [COUTUREAU et KILLCROSS, 2003; KILLCROSS et COUTUREAU, 2003]. Mais, ces études ne sont pas formulées dans le language des systèmes de mémoire parallèles et nécessitent un certain contexte. Ce courant de recherche sur le comportement en conditionnement instrumental sera abordé dans la section suivante.

2.4.2 Conditionnement instrumental

Contexte historique

L'origine du courant de recherche sur le comportement instrumental, et plus particulièrement la distinction entre comportement dirigé vers un but et comportement habituel remonte encore une fois à l'opposition entre *behavioriste* et *cognitiviste* [DICKINSON, 1985]. Une description plus détaillée de ce débat est ici nécessaire. La théorie *behavioriste* prônait l'explication de tout comportement par l'apparition d'un stimulus qui déclenchait une réponse. Si l'on pose la question de «pourquoi un tel comportement est produit», la réponse est simple : «parce qu'un stimulus en est la cause». Cette théorie a été avancée pour contrer la notion de télos ou finalité. Le télos ou finalité appartient à l'une des quatres catégories causales : cause matérielle, cause formelle, cause motrice, et cause finale. Par exemple, la question «Pourquoi une maison ?» peut être répondue par un «parce que» dans les 4 catégories : «parce qu'elle est constituée de briques» comme cause matérielle, «parce qu'un architecte a dessiné un plan» comme cause formelle, «parce que des maçons se sont activés pour la construire» comme cause motrice et enfin, «parce que quelqu'un a l'intention d'y habiter» comme cause finale. Dans la dernière réponse, l'effet précède la cause, le futur affecte le présent et c'est cette apparente violation de la continuité du temps qui rend ce concept si dur à cerner pour une théorie scientifique (le premier à évacuer le problème avec succès fut d'ailleurs Newton, l'un des fondateurs de la science moderne).

Dans le cas présent, discerner les causes d'un comportement est facile pour les trois premières catégories (stimulus, substrat neuronale, activités électriques des neurones, etc). La quatrième catégorie fait appel à une intentionnalité de la part de l'agent montrant un tel comportement. Dans les termes de la théorie *cognitiviste*, l'animal a une connaissance des conséquences liées à ses actions, d'où la notion de comportement lié à un but. Au contraire, la théorie *behavioriste* n'assigne aucun rôle causal à la connaissance de la conséquence d'une action. Dans cette théorie, seul le lien entre stimulus et réponse conduit le comportement. Comme le font remarquer les auteurs d'une revue sur le sujet [YIN et KNOWLTON, 2006], il convient de noter que cette dernière théorie, bien qu'apparemment radicale, continue d'influencer le travail des neurosciences contemporaines. La plupart de l'activité neuronale enregistrée est ensuite interprétée par les chercheurs en fonction des stimuli qui l'ont précédée.

Tâche de conditionnement instrumental

Ainsi, plusieurs études ont montré de manière concluante que les animaux encodent effectivement la relation causale entre leurs actions et les conséquences et possèdent un contrôle de leur choix en fonction de leur désir ou de leur anticipation de l'effet. [BALLEINE et DICKINSON, 1998; DICKINSON, 1985; DICKINSON et BALLEINE, 1994; DICKINSON et collab., 1995; HAMMOND, 1980; KRIECKHAUS et WOLF, 1968]. Ces études sont basées

sur une tâche d'appui de levier dans une chambre de conditionnement. Pendant une première phase d'entraînement, le rat, motivé par une privation d'eau ou de nourriture, apprend à appuyer sur un levier pour obtenir une récompense. Après un temps d'entraînement variable, la valeur de la conséquence de l'action est soit augmentée, soit diminuée et une phase d'extinction mesure le nombre d'appuis sur le levier. Dans le contexte de l'époque, la raison de la prééminence de cette tâche comme seule permettant d'étudier le comportement lié à un but est longuement discutée dans DICKINSON et BALLEINE [1994]. Très brièvement, le premier critère est de type contingent. L'animal connaît la relation entre l'action d'appuyer sur le levier et l'apparition de la récompense. Si le comportement est lié à un but, la dégradation de cette contingence entraîne une diminution du nombre d'appuis. Cela permet de différencier des études sur le conditionnement pavlovien qui associe stimulus conditionné et comportement automatique. Le deuxième critère est motivationnel. La performance d'un animal doit dépendre de la valeur qu'il associe à la conséquence de son action. Cette valeur est donc fonction de l'état motivationnel et influence le comportement. Pour conclure, les tâches de navigation apparaissaient comme ambigües pour ces auteurs car un apprentissage de type pavlovien pouvait soustendre l'association entre indices visuels et récompense au lieu d'un apprentissage des contingences action-conséquence.

Le lien avec la théorie des systèmes de mémoire parallèles apparaît si on considère qu'il existe des comportements liés à un but et des comportements habituels et qu'il est possible d'induire une transition de l'un vers l'autre par surentraînement [DICKINSON, 1985]. Si l'animal encode et réutilise des associations ou contingences diverses, le terme mémoire est approprié. Et si ces contingences sont de nature variée, on peut aussi supposer qu'elles sont encodées dans des structures neuronales différentes.

Des réponses ont ainsi été apportées dans PACKARD et MCGAUGH [1996] et PACKARD [1999], décrits précédemment. Pour rappel, le rat doit trouver une récompense dans un labyrinthe en croix. Si la stratégie de lieux ne peut être considérée comme instrumentale, il est communément admis que la stratégie de réponse est basée sur un comportement habituel tel que décrit dans la théorie *behavioriste* [YIN et KNOWLTON, 2006]. De même, la tâche de *win-stay* décrite dans MCDONALD et WHITE [1993] correspond aussi au développement d'une stratégie d'habitudes. Dans tous les cas, l'animal renforce l'association entre un stimulus et une réponse et les lésions pratiquées ont rendu possible l'identification des structures sous-jacentes.

Dissociation du striatum

Dans PACKARD et collab. [1989]; PACKARD et MCGAUGH [1996], la structure lésée ou infusée est le noyau caudé. Dans PACKARD [1999], l'auteur indique qu'il infuse la structure double noyau caudé et putamen. Dans MCDONALD et WHITE [1993], les auteurs parlent de lésion du striatum dorsal. Il y avait sûrement de la place pour plus de précision.

En quittant brièvement la sphère du conditionnement instrumental, les résultats présentés dans DEVAN et collab. [1999] et DEVAN et WHITE [1999] ont ainsi indiqué une différence entre le striatum dorso-latéral et le striatum dorso-médian chez le rongeur dans une tâche de navigation. Dans les deux études, les animaux ont été entraînés à nager dans une piscine de Morris pour localiser une plateforme qui était soit visible (stratégie de réponse), soit submergée (stratégie de lieux), ou les deux en même temps. Les auteurs ont ensuite lésé séparément le striatum dorso-médian, le striatum dorso-latéral et le fornix. L'effet le plus significatif dans DEVAN et collab. [1999] est la tendance des rats avec lésion du striatum dorso-médian à s'engager dans une stratégie de réponse quand les deux

types de plateforme sont présentés concurrentiellement. Les autres groupes de rats ne montrent pas de préférence. Pour DEVAN et WHITE [1999], le principal résultat est l'effet de facilitation de la stratégie de réponse quand les rats sont lésés pour le striatum dorso-médian ou le fornix. Un effet de compétition entre systèmes de mémoire a déjà été décrit dans 2.3.3 avec MCDONALD et WHITE [1993]. Pour conclure, les auteurs proposent que le striatum dorso-médian est impliqué dans le même système fonctionnel que l'hippocampe.

Toujours dans le domaine de la navigation, la différence entre les deux régions du striatum a aussi été montrée pour le labyrinthe en croix dans YIN et KNOWLTON [2004]. Néanmoins, les auteurs, pour des raisons principalement anatomiques, ont lésé trois parties du striatum : dorso-médian antérieur, dorso-médian postérieur et dorso-latéral. Cette distinction est pertinente puisque seuls les rats avec lésion du striatum dorso-médian postérieur exhibent en majorité une stratégie de réponse après 7 ou 14 jours d'entraînement. Les rats avec lésion du striatum dorso-médian antérieur ne montrent pas de différence dans la stratégie choisie tandis que les rats avec lésion du striatum dorso-latéral utilisent principalement une stratégie de lieux.

Pour revenir à la problématique du conditionnement instrumental, la tentation est donc grande de considérer le striatum dorso-médian comme siège du comportement lié à un but et le striatum dorso-latéral comme siège du comportement habituel. Cette hypothèse a été confirmée dans une série d'articles [YIN et collab., 2004, 2005, 2006].

Très brièvement, la première étude montre que les rats avec lésion du striatum dorso-latéral appuient significativement moins sur le levier que le groupe contrôle en phase d'extinction. Après dévaluation de la conséquence de l'action, une persistance d'appui sur le levier est considérée comme un comportement habituel.

Dans la deuxième étude, les rats ont été entraînés à choisir entre deux leviers amenant deux récompenses différentes. Chaque groupe de rats est ensuite dévalué pour une récompense. Le groupe contrôle montre ainsi une préférence pour le levier non dévalué tandis que le groupe infusé avec un bloqueur des récepteurs NMDA dans le striatum dorso-médian postérieur ne montre pas de préférence. Selon les auteurs, ce résultat indique que cette région est impliquée dans le comportement lié à un but puisque les animaux ne font plus de différence dans la valeur associée à chaque action.

Dans la dernière étude, les rats, après surentraînement, peuvent augmenter leur taux de récompense en s'abstenant d'appuyer sur le levier. Durant cette phase d'omission, les rats avec inactivation du striatum dorso-latéral se retiennent très vite d'appuyer sur le levier contrairement au groupe contrôle. En phase d'extinction, le groupe avec inactivation appuie significativement moins sur le levier que le groupe contrôle. Les auteurs en concluent donc que le striatum dorso-latéral sous-tend le comportement habituel.

A cet instant, le rôle du striatum dans la théorie des systèmes de mémoire a été reconsidéré. Le striatum est fonctionnellement hétérogène et ne peut être la structure unique sous-tendant l'apprentissage d'habitudes sous la forme d'association stimulus-réponse. Ce rôle semble être dévolu uniquement à la partie latérale. Et puisqu'une perturbation du striatum dorso-médian a des effets dans les tâches de navigation et de conditionnement instrumental, la stratégie de lieux est confondable avec le comportement lié à un but malgré les différences dans les actions motrices engagées par l'animal.

2.4.3 Transfert de contrôle et cortex préfrontal

Après ce long détour, nous allons donc revenir à la question de l'interaction entre la stratégie liée à un but et la stratégie habituelle dans une tâche de conditionnement ins-

trumental. Comme nous l'avons vu avec les résultats précédents, il est plus que probable que ces stratégies soient sous-tendues par des systèmes de mémoire tels que définis dans l'introduction de ce chapitre. Ainsi, les études qui posent la question des processus de transition entre stratégies sont donc informatives pour la théorie des systèmes de mémoire parallèles.

L'implication du cortex préfrontal dans le comportement lié à un but a été mis en avant par plusieurs études avec des rôles différents au début des années 2000. Par exemple, il fut montré que le cortex orbitofrontal est impliqué dans l'établissement d'une valeur motivationnelle de la conséquence d'une action dans TREMBLAY et SCHULTZ [1999]. Dans GEHRING et KNIGHT [2000], les auteurs enregistrent sous imagerie cérébrale des patients avec des lésions du cortex préfrontal latéral exhibant un déficit de supervision d'une action volontaire. Chez les rats, le cortex préfrontal médian a été associé à la capacité d'apprendre les contingences entre l'action et la valeur de la conséquence d'une action [BALLEINE et DICKINSON, 1998].

Dans KILLCROSS et COUTUREAU [2003], les auteurs sont les premiers à poser la question de l'interaction entre un comportement lié à un but et un comportement habituel. Certains résultats publiés après cette étude sur le rôle du striatum dans le comportement lié à un but ont déjà été décrits dans cette section. Au nom de la clarté de l'évolution des idées, il convient de noter que l'hypothèse des auteurs était de considérer le cortex préfrontal comme sous-tendant à la fois le comportement lié à un but et la responsabilité de la transition vers un comportement habituel. Les auteurs ont donc comparé un groupe de rats avec lésion du cortex préfrontal infralimbique, un groupe de rats avec lésion du cortex préfrontal prélimbique et un groupe contrôle. Pendant la phase d'entraînement, tous les rats ont été entraînés modérément et extensivement dans deux chambres de conditionnement différentes avec deux récompenses différentes. La phase d'extinction mesure le nombre d'appuis sur le levier dans chaque chambre avec récompense dévaluée ou non dévaluée. Les résultats sont présentés dans la figure 2.9

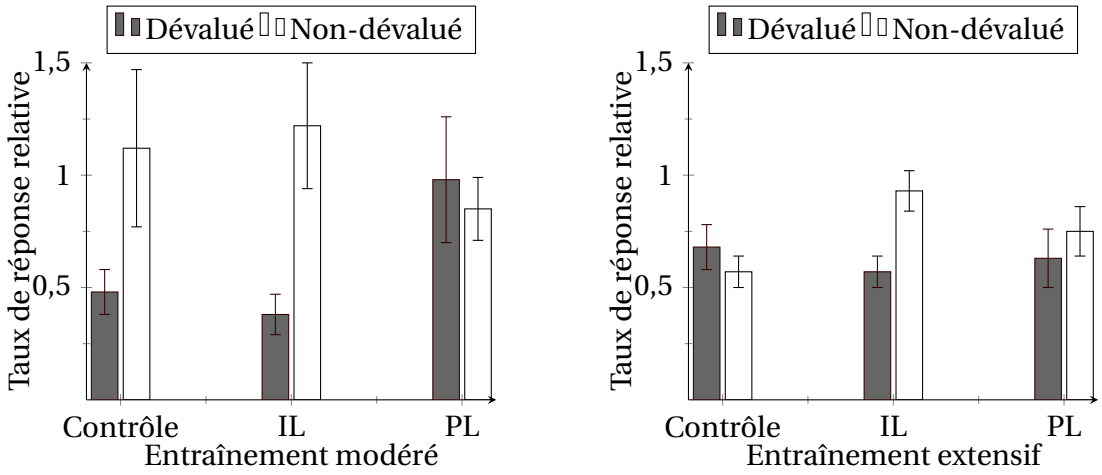

FIGURE 2.9 – (Taux d'appui sur le levier (± SE) pour les groupes de rats étant testés dans la chambre de conditionnement avec entraînement modéré (*gauche*) et les groupes de rats étant testés dans la chambre de conditionnement avec entraînement extensif (*droite*). IL : groupe de rats avec lésion du cortex préfrontal infralimbique, PL : groupe de rats avec lésion du cortex préfrontal prélimbique. Adapté de KILLCROSS et COUTUREAU [2003]

Tout comme PACKARD et MCGAUGH [1996], la transition entre comportement lié à un but et comportement habituel apparaît clairement après un entraînement extensif pour le groupe contrôle. Les rats normaux deviennent insensibles à la valeur associée à

la conséquence d'une action. Outre ce résultat, les rats avec lésion du cortex préfrontal pré-limbique ne montrent pas de sensibilités à la dévaluation en entraînement modéré au contraire des deux autres groupes de rats. Le deuxième résultat est la sensibilité à la dévaluation en entraînement extensif pour le groupe avec lésion du cortex préfrontal infralimbique. Très brièvement, les auteurs ont proposé que la prévalence progressive du comportement habituel est explicable par un contrôle inhibiteur venant de la région infralimbique vers la région prélimbique. Cet agencement expliquerait ainsi les différences de réponses entre entraînement modéré et entraînement extensif pour chaque groupe de rat.

Etant donné ces résultats, COUTUREAU et KILLCROSS [2003] ont posé la question suivante sur la transition vers un comportement habituel : est-ce l'influence de l'association entre action et conséquence qui décroît ou est-ce l'association qui disparaît ? Si son influence décroît, une inactivation du cortex préfrontal infralimbique, en tant qu'inhibiteur, en phase d'extinction devrait faire revenir la sensibilité à la valeur et donc le comportement lié à un but. Si l'association disparaît, le rat devrait conserver son comportement habituel. La réponse est donnée dans la figure 2.10. Après inactivation du cortex préfrontal infralimbique, la capacité des rats à préférer le levier non-dévalué reste intacte comparé au groupe contrôle. Les auteurs en concluent donc que l'association action-conséquence reste intacte après que le comportement habituel devient prévalent comme le démontre le retour de la sensibilité à la valeur de l'action.

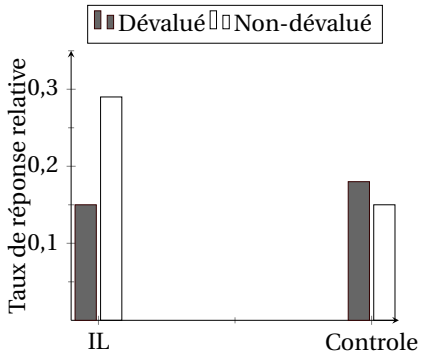

FIGURE 2.10 – (Taux d'appui sur le levier en phase d'extinction après dévaluation ou sans dévaluation. Le groupe IL a été infusé dans le cortex préfrontal infralimbique avec du Muscimol. Adapté de COUTUREAU et KILLCROSS [2003]

Pour élargir sur le rôle du cortex préfrontal, il existe une littérature qui a testé la capacité des rats à changer de stratégie [FLORESCO et collab., 2006; RAGOZZINO, 2002, 2007; RAGOZZINO et collab., 2003; STEFANI et collab., 2003]. Néanmoins, ces études sont différentes de celles précédemment citées car le surentraînement n'est pas la cause du changement de stratégie. Typiquement, les rats doivent apprendre, dans une première phase, à discriminer entre deux stimuli dans une modalité sensorielle puis, dans une deuxième phase, à discriminer entre deux stimuli dans une modalité sensorielle différente tout en inhibant la première stratégie. Cela nécessite bien sûr un contrôle cognitif important mais il n'est pas précisé par les auteurs si cela requiert des systèmes de mémoire fondamentalement distincts. En peu de mots, il semblerait qu'une inactivation du cortex frontal prélimbique retarde le processus de changement de stratégie [RAGOZZINO, 2007]. Les auteurs en concluent que la région prélimbique est cruciale pour les comportements flexibles.

Dans le contexte d'un conditionnement instrumental chez le rat, une transition entre stratégie par surentraînement liée à une discrimination bimodale a été étudiée dans HAD-

DON et KILLCROSS [2011]. Chaque groupe de rats a été entraîné à appuyer sur l'un des deux leviers disponibles en fonction de la présentation d'un stimulus. Une modalité (visuelle ou auditive) a ensuite été associée à un entraînement modéré ou extensif. Chaque rat apprend les deux modalités sur deux sessions successives pendant plusieurs jours. En phase d'extinction, les rats sont testés avec un stimulus auditif et un stimulus visuel en même temps. Dans un cas, les deux stimuli ont été associés au même levier pendant la phase d'entraînement (congruent). Dans l'autre cas (incongruent), les deux stimuli ont été associés à deux leviers différents. Les auteurs se concentrent sur ce choix de levier. Le rat choisit-il l'association stimulus-levier renforcée en entraînement modéré ou en entraînement extensif ? Si le rat appuie sur le levier renforcé en entraînement modéré, ce choix est considéré comme une réponse correcte. Le groupe contrôle montre ainsi un nombre d'erreurs plus élevé lorsque les deux stimuli sont non congruents, c'est-à-dire une tendance significative à appuyer sur le levier renforcé en entraînement extensif sans porter attention au stimulus associé à l'autre levier en entraînement modéré. Cet effet est renversé pour le groupe avec inactivation de la région infralimbique du cortex préfrontal. Les auteurs en concluent donc que la région infralimbique est nécessaire pour atténuer l'influence du comportement lié à un but.

Au niveau de l'apprentissage spatial, le rôle du cortex préfrontal a été étudié dans RICH et SHAPIRO [2007]. Les auteurs ont ainsi fait alterner l'apprentissage d'une stratégie de lieux et l'apprentissage d'une stratégie de réponse sur le labyrinthe en croix. Au moment d'une transition entre les stratégies, les régions prélimbique et infralimbique ont été conjointement inactivées. Si les rats sont effectivement capables de changer de stratégie, l'effet d'une inactivation n'apparaît que 24 heures après le changement de stratégie. Si les rats sont testés 20 min après inactivation, aucune différence avec le groupe contrôle n'apparaît. Les rats testés 24h après utilisent la stratégie qui était en place avant l'inactivation des régions prélimbique et infralimbique. Néanmoins, cet effet disparaît au cours du temps. Après plusieurs transitions, l'inactivation n'a plus d'effet et le groupe infusé change tout aussi bien de stratégie que le groupe contrôle.

Dans la continuité de cette étude, RICH et SHAPIRO [2009] ont enregistré l'activité neurale dans les régions infralimbique et prélimbique pour la même expérience de changement de stratégie. Ils ont ainsi révélé une dynamique complexe des neurones en fonction des transitions. Certains neurones vont s'éteindre progressivement après une transition de stratégie de lieux vers une stratégie de réponse. D'autres neurones vont décharger à l'endroit de la récompense après un changement vers une stratégie de réponse. En moyenne, la variation d'activité des neurones des régions prélimbique et infralimbique est la plus élevée durant des changements de stratégie comparé à une absence de changement. Pour les neurones dont l'activité décroît après un changement de stratégie, il n'existe pas de différence entre les deux régions. Pour les neurones dont l'activité croît, la région prélimbique est significativement en avance sur la région infralimbique. De plus, la région prélimbique est aussi en avance sur une augmentation des performances. Pendant cette période initiale, le rat persiste dans les trajectoires de l'ancienne stratégie accumulant les erreurs. Les auteurs en concluent que les changements d'activité dans la région prélimbique prédisent la capacité de l'animal à adopter une nouvelle stratégie.

2.4.4 Progrès récents

La dernière étude de ce chapitre est FERBINTEANU [2016]. Jusqu'à présent, les animaux apprennent les stratégies séparément et séquentiellement. Dans certains cas, les transitions s'effectuent par surentraînement [PACKARD et MCGAUGH, 1996] pour passer

du comportement lié à un but à un comportement habituel. Dans d'autres cas, les transitions se succèdent en fonction des performances [RICH et SHAPIRO, 2007] tout en exigeant qu'une seule stratégie soit apprise. Dans FERBINTEANU [2016], l'auteure a voulu mesurer une triple dissociation entre l'hippocampe, le striatum dorso-latéral et le striatum dorso-médian sur l'apprentissage d'une statégie de lieux ou de réponse dans le labyrinthe en croix. Cependant, les stratégies peuvent être apprises concurrentiellement ou séquentiellement.

Pour la tâche nécessitant une stratégie de lieux, les animaux sont récompensés pour se souvenir de la position de la récompense dans un bras. Si l'animal entre dans le bras correct pendant 9 essais sur 10, la position de la récompense change et un nouveau bloc commence. Pour la tâche nécessitant une stratégie de réponse, le rat doit associer la présence d'un indice positionné en dehors du labyrinthe avec une réponse droite ou gauche. Les rats sont entraînés sur 45 essais.

Dans la version séquentielle, un premier groupe de rats apprend la stratégie de lieux et un deuxième groupe apprend la stratégie de réponse. Les rats sont ensuite divisés en sous-groupes pour recevoir une lésion d'une des structures. Après chirurgie, la rétention de la stratégie apprise avant chirurgie est mesurée sur 5 jours. Puis les rats apprennent l'autre stratégie sur 5 autres jours.

Dans la version concurrentielle, tous les rats sont d'abord entraînés pour la stratégie de réponse. Lorsque les rats sont efficients dans cette première tâche, la seconde tâche leur est présentée. Ils sont ensuite entraînés sur chaque tâche *indépendemment* chaque jour dans un ordre aléatoire. Lorsque les rats atteignent 80 % de performances sur les deux tâches, ils sont assignés à un groupe de lésion. En phase de test, la performance est mesurée sur les deux tâches comme en phase d'entraînement.

Les résultats sont présentés dans la figure 2.11. On observe ainsi que

1. le groupe contrôle ne montre pas de variation de performances dans les deux versions et dans chaque tâche
2. une lésion du striatum dorso-médian diminue la performance pour les deux tâches et dans les deux versions
3. une lésion du striatum dorso-latéral affecte la stratégie de réponse dans la version séquentielle et les deux stratégies dans la version concurrentielle
4. une lésion de l'hippocampe affecte la stratégie de lieux dans la version séquentielle et les deux stratégies dans la version concurrentielle

Les résultats dans la version séquentielle sont consistants avec la littérature telle que présentée tout au long de ce chapitre. L'hippocampe sous-tend la stratégie de lieux et le striatum dorso-latéral sous-tend la stratégie de réponse [MCDONALD et WHITE, 1993; PACKARD et collab., 1989; PACKARD et MCGAUGH, 1996; YIN et collab., 2004]. La version concurrentielle offre un tableau de résultats différents par rapport à la théorie des systèmes de mémoire établie. De plus, ces résultats sont formulés avec l'outil principal pour forger la théorie, c'est-à-dire la lésion d'une structure. Cependant, une littérature sur l'interaction entre l'hippocampe et le striatum s'est développée récemment [DECOTEAU et collab., 2007; VAN DER MEER et collab., 2010; RAGOZZINO et collab., 2001; VOERMANS et collab., 2004]. Il est ainsi fort probable que certains processus d'interaction entre systèmes de mémoire attendent encore d'être révélés.

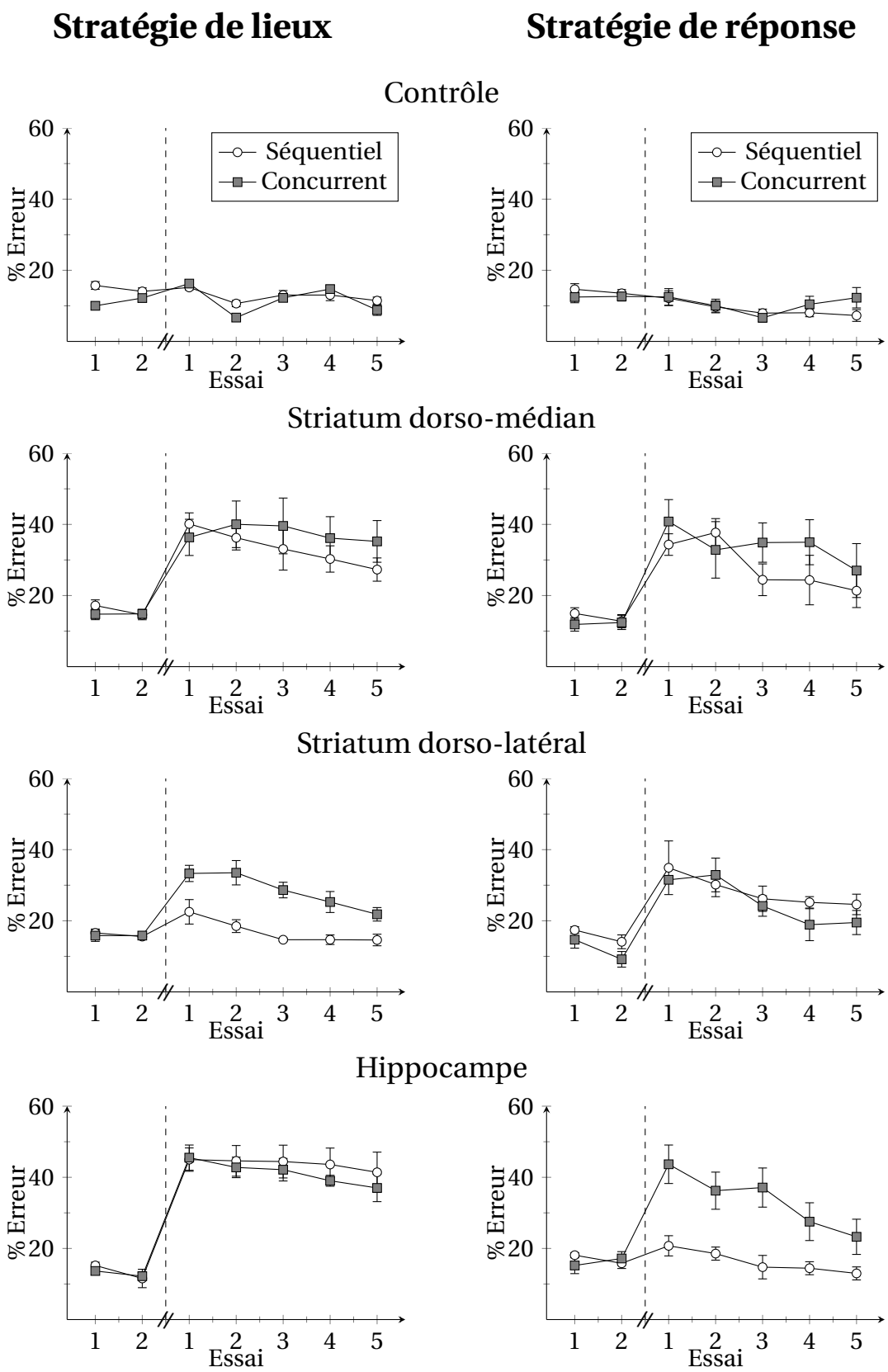

FIGURE 2.11 – Rétention d'information dans deux conditions. Chaque graphique est le pourcentage d'erreurs avant et après la chirurgie (barre verticale). Quand les animaux sont entraînés concurrentiellement dans une tâche de réponse à un stimulus et dans une tâche de navigation spatiale, les lésions hippocampiques diminuent aussi les performances dans la tâche de réponse à un stimulus. De même, pour les lésions du striatum dorso-latéral (DSL), la tâche de navigation spatiale est aussi affectée. Les lésions du striatum dorso-médian affectent dans les deux conditions. Adapté de FERBINTEANU [2016]

2.5 Conclusion

Le nombre de travaux faisant explicitement référence à la question des systèmes de mémoire parallèles diminue grandement au milieu des années 2000. Comme nous l'avons vu au cours de ce chapitre, la quantité de données empiriques a suffisamment convergé pour asseoir l'existence de systèmes neuronaux aux processus d'encodage, de rétention et de restitution distincts. Le fruit était donc mûr pour tomber sur la tête du modélisateur, ce qui sera le sujet des chapitres suivants.

Chapitre 3

Modèles de l'apprentissage par renforcement

Sommaire

3.1 Introduction

La théorie de l'apprentissage par renforcement propose un cadre formel permettant de définir l'interaction entre un agent apprenant et son environnement en termes d'états, d'actions et de récompenses. Issue principalement de l'intelligence artificielle, elle se distingue néanmoins de l'apprentissage supervisé qui apprend à partir d'exemples fournis par un superviseur externe. Dans notre cas, l'accent est mis sur l'interaction d'un agent avec son environnement à travers un signal de récompense obtenue après mise en oeuvre d'une action. Celui-ci apprend donc de sa propre expérience. De plus, il est capable de s'adapter pour toutes les situations à travers une interaction en continu avec l'environnement, ce que ne permet pas l'apprentissage supervisé puisque toute connaissance dépendra des exemples choisis par le superviseur. A travers cette brève introduction se dessinent les formes du problème. D'une part, l'agent doit «sentir» le monde et doit «agir» dessus. D'autre part, l'agent doit posséder un but à atteindre en lien avec les états du monde qu'il perçoit.

Historiquement, le domaine de l'apprentissage par renforcement s'est développé en intelligence artificielle à partir du milieu du XX^{e} siècle pour finalement arriver à maturité au début des années 2000. Depuis cette époque, le coeur de la théorie n'a pas évolué. Pour présenter ces concepts-clés, ce chapitre a été construit en suivant SUTTON et BARTO [1998] considéré comme la référence du domaine et couvrant tous les éléments nécessaires pour la compréhension des modèles présentés ultérieurement.

Dans une première partie, le formalisme de l'apprentissage par renforcement sera décrit avec notamment l'interface agent-environnement et la formalisation du but d'un agent. Dans une deuxième partie, les algorithmes utilisés par un agent pour apprendre à résoudre un problème d'apprentissage par renforcement seront présentés.

3.2 Formalisme de l'apprentissage par renforcement

3.2.1 L'interface agent-environnement

L'agent et l'environnement interagissent dans une succession de temps discrets t comme représentés dans la figure 3.1. A chaque instant t, l'agent reçoit l'état du monde sous la forme d'un symbole $s \in \mathcal{S}$ puis choisit une action $a \in \mathcal{A}$. En conséquence de son action, l'agent reçoit à l'instant $t+1$ une récompense numérique $r_{t+1} \in \mathbb{R}$.

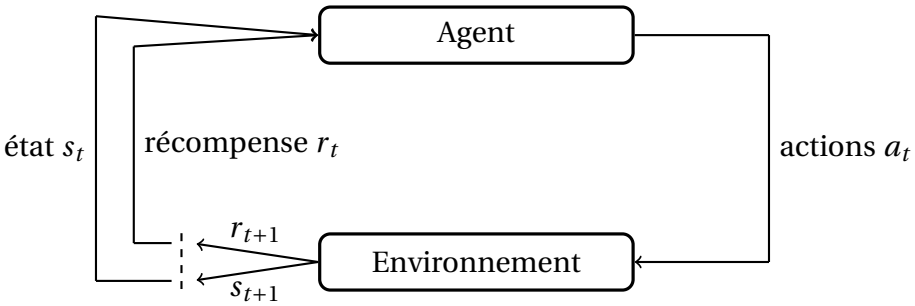

FIGURE 3.1 – L'interface agent-environnement dans l'apprentissage par renforcement

Connue sous le nom de processus de décision markovien [BELLMAN, 1957], cette représentation d'un problème d'apprentissage par renforcement se formalise ainsi :

— un ensemble d'états $\mathcal{S} = \{s_1, s_2, \ldots, s_n, \ldots\}$

— un ensemble d'actions $\mathscr{A} = \{a_1, a_2, \ldots, a_i, \ldots\}$

— une fonction de transition $\mathscr{T} : \mathscr{S} \times \mathscr{A} \times \mathscr{S} \rightarrow [0;1]$ définissant la probabilité de se retrouver dans un état s_n à partir d'un état s_m en choisissant une action a_i.

— une fonction de récompense $\mathscr{R} : \mathscr{S} \times \mathscr{A} \rightarrow \mathbb{R}$ associant une valeur de récompense r à une action a_i prise dans un état s_n.

De plus, le comportement de l'agent sera déterminé par une politique π définie comme :

$$\pi : \mathscr{S} \rightarrow \mathscr{A} \tag{3.1}$$

si la politique est déterministe ou

$$\pi : \mathscr{S} \times \mathscr{A} \rightarrow [0,1] \tag{3.2}$$

si la politique est probabiliste.

Dans le cas idéal, un état concentre toute l'information nécessaire pour prendre une décision. Cet état possède donc la propriété de Markov puisque sa signification est indépendante du chemin ou histoire qui l'a amené. Un exemple simple est celui d'un jeu d'échecs. La configuration courante de toutes les pièces du jeu servirait d'état markovien car cela résume tout ce qu'il y a d'important sur la situation du jeu pour pouvoir décider du prochain coup.

Sans propriété de Markov, la dynamique de l'environnement à l'instant $t+1$ en réponse aux décisions de l'agent à l'instant t peut être définie en termes probabilistes comme :

$$\mathrm{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0), \tag{3.3}$$

pour tous états et récompenses s' et r, sachant toutes valeurs possibles d'évènements passés : $s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0$. Avec la propriété de Markov, la dynamique de l'environnement à l'instant $t+1$ dépend seulement de l'état et de l'action à l'instant t et s'écrit en termes probabilistes :

$$\mathrm{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t). \tag{3.4}$$

En d'autres termes, un état possède la propriété de Markov si :

$$\mathrm{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0) = \mathrm{P}(s_{t+1} = s', r_{t+1} = r | s_t, a_t) \tag{3.5}$$

pour tout s', r, et toutes histoires $s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0$. Une tâche d'apprentissage par renforcement qui satisfait la propriété de Markov est ainsi appelée *Processus de décision markovien.*

3.2.2 Objectif de l'agent

Du point de vue de l'agent, son objectif est de maximiser la quantité totale de récompense qu'il reçoit à chaque instant t sous la forme d'une grandeur numérique $r_t \in \mathbb{R}$. Toutes les équations et méthodes suivantes s'écrivent donc en fonction de cette finalité. Dans le cas d'un horizon temporel fini , la somme des récompenses obtenues (ou retour-attendu) à partir d'un instant t s'écrit :

$$\mathrm{R}_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_{\mathrm{T}} \tag{3.6}$$

avec T la dernière étape.

Cette finitude dans le nombre d'étapes est adaptée à l'environnement possédant un état terminal (par exemple la sortie d'un labyrinthe) ou pouvant être décomposé en sous-séquences. Néanmoins, le cas le plus général est celui où l'interaction entre l'agent et son environnement est possiblement à horizon infini. La dernière étape se situe donc à $T = \infty$ et rend problématique la définition du retour attendu puisque celui-ci pourrait aussi bien être infini.

Pour gagner le droit à la récursion d'un modèle maximisant une récompense dans une temporalité infinie, le premier concept nécessaire est celui d'*atténuation*[1]. Plus précisément, un agent choisit une action a_t qui maximise le *retour attendu atténué* :

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{3.7}$$

avec γ ($0 \leq \gamma < 1$) le taux d'atténuation

Ce taux d'atténuation permet donc d'évaluer la valeur présente des récompenses futures. Si $\gamma < 1$ et les valeurs de récompenses $\{r_k\}$ limitées, la somme infinie R_t a donc une valeur finie. Dans le cas ou γ approche 1, l'agent prend en compte les récompenses futures. Si $\gamma = 0$, l'agent est «myope» et seule r_{t+1} est maximisée.

3.2.3 La fonction de valeur

Tous les algorithmes d'apprentissage par renforcement sont basés sur l'estimation d'une fonction de valeur qui associe à chaque état (ou couple état-action) l'intérêt de l'agent d'être dans cet état (ou de choisir une action dans un état). Cette fonction de valeur est définie en fonction de R_t et s'écrit :

$$V_\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}, \tag{3.8}$$

La fonction V_π donne ainsi la valeur attendue atténuée si l'agent suit la politique π. Dans le cas d'un couple état-action, la fonction $V_\pi(s)$ devient $Q_\pi(s,a)$ et s'écrit :

$$Q_\pi(s,a) = E_\pi\{R_t | s_t = s, a_t = a\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}. \tag{3.9}$$

Le point important ici est que ces fonctions V_π et Q_π peuvent être estimées à partir de l'expérience. Si un agent maintient en mémoire la récompense qui a suivi un état s (ou une action a sachant un état s) en suivant une politique π, la valeur moyenne de ces récompenses convergera vers $V_\pi(s)$ (ou $Q_\pi(s,a)$). Les algorithmes de ce type pour résoudre un problème d'apprentissage par renforcement sont appelés *méthode de Monte-Carlo* et nécessitent de moyenner sur un échantillonage aléatoire de la politique π.

Pour finir, la récursion promise apparaît au terme du développement suivant (l'astuce consiste à écrire l'équation 3.8 sur deux pas de temps consécutifs) :

1. Le deuxième concept est présenté dans 3.2.3

$$
\begin{aligned}
V_\pi(s) &= E_\pi\Big\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s\Big\} && (3.10)\\
&= E_\pi\Big\{r_{t+1}+\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t=s\Big\} && (3.11)\\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{T}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma E_\pi\Big\{\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t=s\Big\}\Big] && (3.12)\\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{T}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma V_\pi(s')\Big] && (3.13)
\end{aligned}
$$

L'équation 3.13 s'appelle *équation de Bellman pour* V_π et exprime la récursivité du modèle. Plus précisément, si la fonction de transition $\mathcal{T}$ et la fonction de récompense $\mathcal{R}$ sont connues, c'est-à-dire si l'agent possède ce que l'on appelle conventionnellement un «modèle du monde», il est possible d'évaluer entièrement la valeur $V_\pi(s)$ à l'état s grâce aux valeurs $V_\pi(s')$ des états suivants. Ces équations sont équivalentes pour la fonction $Q_\pi(s,a)$ à l'exception près qu'il faut considérer un couple état-action (s',a') suivant le choix de l'action a dans l'état s.

Comme tout bon modèle se voulant récursif, nous avons ainsi obtenu le droit de calculer la valeur d'un état (ou d'une action dans un état) au prix d'une limitation de l'horizon temporel : seul l'état suivant est important pour avancer dans le processus récursif et celui-ci contient lui-même la valeur de tous ses états suivants. Bien évidemment, une telle quantité d'informations est rarement disponible et entreront ainsi en scène les algorithmes nécessaires à la résolution (ou plus exactement l'approximation) de l'équation de Bellman (voir section 3.3).

3.2.4 Décision et optimalité

Jusqu'à présent, la politique π était considérée constante. Néanmoins, il est évident que le choix d'une politique va influencer le calcul de la fonction de valeur. En retour, cette fonction de valeur va induire un ordre partiel sur chaque politique. En d'autres termes et sachant que le but de l'agent est de maximiser le taux de récompense, une politique π est supérieure à une politique π' si $V_\pi(s) \geq V_{\pi'}(s) \forall s \in \mathcal{S}$. En supposant qu'il existe toujours au moins une politique supérieure aux autres, les politiques optimales π^* se déterminent en maximisant la fonction de valeur :

$$V^*(s) = \max_\pi V_\pi(s)\ \forall s \in \mathcal{S} \tag{3.14}$$

Pour la fonction Q considérant le couple état-action, l'équation s'écrit :

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)\ \forall s,a \in \mathcal{S}\times\mathcal{A} \tag{3.15}$$

S'il peut exister plusieurs politiques optimales, elles partagent toutes la même fonction de valeur optimale V^* qui est unique pour un problème donné. Il est donc possible de réécrire l'équation 3.13 sans référence à une politique donnée :

$$
\begin{aligned}
V^*(s) &= \max_a E\Big\{r_{t+1}+\gamma V^*(s_{t+1})|s_t=s,a_t=a\Big\} && (3.16)\\
&= \max_a \sum_{s'}\mathcal{T}^a_{ss'}\Big[\mathcal{R}^a_{ss'}+\gamma V^*(s')\Big] && (3.17)
\end{aligned}
$$

Les équations 3.16 et 3.17 sont appelées *équations d'optimalité de Bellman*. Intuitivement, elles expriment le fait que la valeur d'un état sous une politique optimale doit être égale à la valeur retournée en choisissant la meilleure action dans cet état.

L'équation d'optimalité de Bellman pour Q s'écrit :

$$\begin{aligned} Q^*(s,a) &= E\left\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') | s_t = s, a_t = a\right\} && (3.18) \\ &= \sum_{s'} \mathcal{T}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q^*(s', a')\right] && (3.19) \end{aligned}$$

L'équation d'optimalité de Bellman peut être écrite pour chaque état donnant ainsi un système d'équations non linéaires. Si les dynamiques de l'environnement sont connues ($\mathcal{R}^a_{ss'} et \mathcal{T}^a_{ss'}$), il est théoriquement possible de résoudre les équations analytiquement pour trouver une solution optimale V^*. Dans la condition de pouvoir les résoudre, le pouvoir récursif des équations de Bellman apparaît ainsi au plus fort. En utilisant V^* pour avancer dans le processus de décision, le retour attendu à long-terme est équivalent à la quantité locale $V^*(s')$ immédiatement disponible pour chaque état. En d'autres termes, une prise de décision à un pas de temps dans le futur est équivalente à la décision optimale sur un temps infini.

3.3 Les méthodes de résolution par différence temporelle

Le pouvoir récursif coûte généralement très cher en mathématiques [2]. Nous le payons une première fois avec l'équation 3.7 en limitant volontairement l'horizon des récompenses avec la condition $\gamma < 1$. En effet, nous nous assurons que la somme des récompenses à un certain pas de temps dans le futur devienne quantité négligeable. Nous le payons une deuxième fois avec l'équation 3.13 de Bellman. Il est ainsi possible d'augmenter la complexité du problème en développant l'équation 3.13 sur plusieurs pas de temps formant ainsi un *arbre* de possibilités et permettant un meilleur calcul de la fonction de valeur. Le choix de s'arrêter à l'état suivant offre donc un compromis. Pour finir, le coût computationnel de résolution analytique d'un système d'équations d'optimalité de Bellman augmente proportionnellement à la taille du problème. Dans SUTTON et BARTO [1998], les auteurs donnent ainsi l'exemple du Backgammon avec ces 10^{20} états rendant le calcul de la fonction de valeur V^* incalculable.

De manière globale, la solution des équations de Bellman suppose :

1. une connaissance précise des fonctions $\mathcal{T}$ et $\mathcal{R}$ liées à la dynamique de l'environnement (en d'autres termes, la connaissance d'un «modèle de l'environnement»)
2. un système d'équation *a priori* tractable ou suffisamment de ressources computationnelles
3. la propriété de Markov pour l'environnement

Ces trois conditions sont rarement réunies pour des problèmes d'apprentissage par renforcement. Pour cette raison, des méthodes ont ainsi été développées pour approximer les solutions d'un système d'équations de type V^* ou Q^*.

Une première classe de ces méthodes regroupées sous le terme *Programmation Dynamique* consiste principalement à améliorer itérativement une politique à partir de la

2. Du moins pour des modèles appliqués généralement à des problèmes physiques concrets comme par exemple le problème à n-corps.

fonction de valeur pour atteindre la politique optimale. Néanmoins, ces méthodes nécessitent une connaissance parfaite du modèle de l'environnement et sont souvent très coûteuses en termes de calcul.

Une deuxième classe de méthodes, précédemment citées, est appelée méthode de *Monte-Carlo*. Au contraire de la programmation dynamique, le but est d'apprendre de l'expérience sans modèle du monde. Les méthodes de Monte-Carlo estiment la fonction de valeur optimale à partir d'échantillonnages aléatoires de l'environnement. Le point important ici est l'utilisation de l'*expérience* de l'agent explorant son environnement.

A l'intersection de ces deux classes de méthodes apparaît ainsi le sujet de cette section : les méthodes d'*apprentissage par différence temporelle*. Très brièvement, l'apprentissage par différence temporelle repose sur l'expérience de l'agent en modifiant «au vol» la valeur d'un état en fonction de l'expérience de l'agent et de la valeur des actions suivantes.

Par ailleurs, c'est peut-être là où se cache l'appétance des neurosciences *comportementales théoriques* pour ces techniques d'apprentissage par renforcement : une mémorisation (puisque ce sont des valeurs qui sont modifiées en fonction *et* au cours de l'expérience) offre la récursivité (l'anticipation de l'instant futur à partir de l'instant présent). Des trois grandes familles d'apprentissage en intelligence artificielle (apprentissage non supervisé, supervisé et par renforcement), seul l'apprentissage par renforcement amène la combinaison de ces objets d'études spécifiques de l'étude du cerveau (mémoire, décision et anticipation).

3.3.1 La différence temporelle

Pour rappel, le développement des équations de Bellman est :

$$\begin{aligned} V_\pi(s) &= E_\pi\Big\{R_t | s_t = s\Big\} && (3.20)\\ &= E_\pi\Big\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\Big\} && (3.21)\\ &= E_\pi\Big\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s\Big\} && (3.22)\\ &= E_\pi\Big\{r_{t+1} + \gamma V_\pi(s_{t+1}) | s_t = s\Big\} && (3.23) \end{aligned}$$

Dans les méthodes de Monte-Carlo, la fonction de valeur est mise à jour en fonction de R_t (donc au niveau de l'équation 3.20) selon l'équation suivante :

$$V(s_t) \leftarrow V(s_t) + \alpha\Big[R_t - V(s_t)\Big] \tag{3.24}$$

L'hypothèse ici est qu'il existe des états terminaux dans l'environnement permettant d'obtenir des *épisodes* finis d'échantillonnage de manière à évaluer R_t.

Dans les méthodes de différence temporelle, l'équation 3.24 devient :

$$V(s_t) \leftarrow V(s_t) + \alpha\Big[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\Big] \tag{3.25}$$

et consiste à évaluer le retour au niveau de l'équation 3.23. La modification de la fonction de valeur $V(s_t)$ se fait donc au pas de temps $t+1$ pour le pas de temps t. La mise à jour globale devient locale et l'hypothèse d'épisodes d'échantillonnage finis n'est plus nécessaire.

3.3.2 SARSA

La version de l'équation 3.25 en terme de couple état-action Q(s,a) est appelée *SARSA* [RUMMERY, 1995; RUMMERY et NIRANJAN, 1994] et s'écrit :

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\Big[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\Big] \tag{3.26}$$

Dans SARSA, on considère les transitions entre couple état-action et l'algorithme nécessite donc l'ensemble : $< s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1} >$. L'algorithme 3.1 décrit SARSA.

Algorithme 3.1 : SARSA

Initialiser Q *arbitrairement*

pour tous les *épisodes* **faire**
- Initialiser s
- Choisir a à partir de s en utilisant la police dérivée de Q
- **répéter**
  - Observer r, s'
  - Choisir a' à partir de s' en utilisant la politique dérivée de Q
  - $Q(s,a) \leftarrow Q(s,a) + \alpha\Big[r + \gamma Q(s',a') - Q(s,a)\Big]$
  - $s \leftarrow s'$
  - $a \leftarrow a'$
- **jusqu'à** *fin de l'épisode*

fin

L'algorithme SARSA retarde ainsi la mise à jour puisque la modification de la valeur Q(s,a) est effectuée après avoir choisi l'action suivante a'. Dans SUTTON et BARTO [1998], les auteurs parlent d'apprentissage sur politique.

3.3.3 Q-Learning

L'algorithme qui a permis de rendre la mise à jour indépendante de la politique s'appelle *q-learning* [WATKINS, 1989] et est présenté dans 3.2.

Algorithme 3.2 : Q-learning

Initialiser Q *arbitrairement*

pour tous les *épisodes* **faire**
- Initialiser s
- **répéter**
  - Choisir a à partir de s en utilisant la politique dérivée de Q
  - Observer r, s'
  - $Q(s,a) \leftarrow Q(s,a) + \alpha\Big[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\Big]$
  - $s \leftarrow s'$
- **jusqu'à** *fin de l'épisode*

fin

L'introduction de l'opérateur *max* permet ainsi de rendre la mise à jour indépendante de la politique. Dans SUTTON et BARTO [1998], les auteurs parlent d'apprentissage hors politique.

3.3.4 Acteur-critique

La dernière classe de méthodes basées sur l'apprentissage par différence temporelle est appelée *acteur-critique* [BARTO et collab., 1983; KONDA et TSITSIKLIS, 2003; WITTEN, 1977]. L'acteur correspond à la politique qui sélectionnne l'action et le critique est la fonction de valeur qui «critique» l'action choisie par l'acteur. L'apprentissage est donc sur politique et s'effectue à la fois dans l'acteur et dans le critique sur la base de l'erreur de prédiction calculée par le critique. Cette architecture des méthodes acteur-critique est représentée dans la figure suivante :

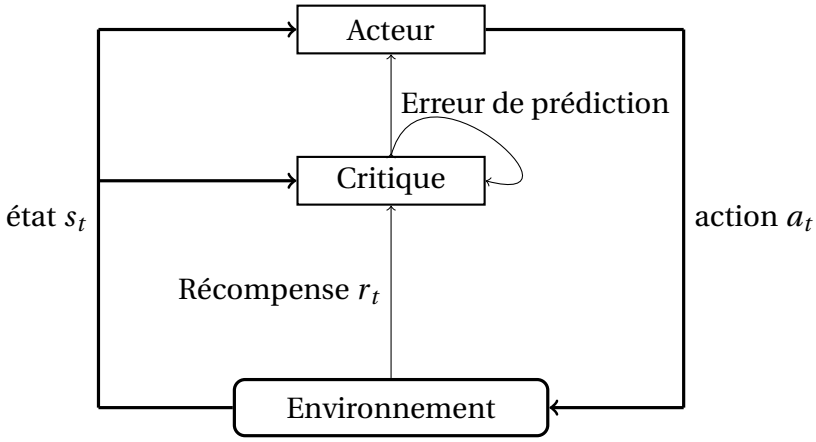

FIGURE 3.2 – Architecture acteur-critique

L'erreur de prédiction calculée par le critique est :

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3.27}$$

avec V la fonction de valeur. Cette erreur est ensuite utilisée pour évaluer l'action choisie par l'acteur. Si la politique π est sous la forme d'une fonction de valeur $Q(s,a)$, la mise à jour est possible selon l'équation suivante :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_2 \delta_t \tag{3.28}$$

avec α_2 un autre paramètre de mise à jour.

3.3.5 Le compromis exploration / exploitation

Pour toutes ces méthodes d'approximation de la fonction de valeur par différence temporelle, il est apparu très tôt que la façon dont la politique sélectionne l'action influence la convergence de l'algorithme. En effet, la mise à jour s'effectue après une sélection d'action et influence les sélections d'actions futures. La question qui se pose ici est celle de l'exploration de l'environnement par l'agent. Est-il nécessaire pour un agent d'explorer tout son environnement plusieurs fois pour approximer la fonction de valeur ? La réponse dépend bien entendu du problème donné et aussi de l'algorithme choisi. Dans certains cas, l'agent a intérêt à ne pas choisir une action qui semble optimale pour, au contraire, choisir une action *sous-optimale* mais qui peut, potentiellement, augmenter la

récompense à long-terme ou apporter une information qui sera cruciale à cette réussite future.

Pour répondre à cette problématique, des méthodes ont été développées dans le but de contrôler le compromis entre exploitation de la fonction de valeur et exploration de l'environnement. Une première manière de s'éloigner d'une politique en choisissant aléatoirement une action est la règle de sélection ϵ-*greedy*. L'action a est choisie selon la règle suivante :

$$a = \begin{cases} \text{action aléatoire si X} \leq \epsilon \text{ avec X}\sim\mathscr{U}(0,1) \\ \underset{a}{\text{argmax}}\, \text{Q}(s_t, a) \text{ sinon} \end{cases} \tag{3.29}$$

La principale faiblesse de cette approche est la sélection uniforme de l'action lors d'une phase d'exploration. Il peut exister des cas où l'information contenue dans la politique est importante, par exemple avec une action menant à une récompense très négative. La régle du *soft-max* permet ainsi de convertir la fonction de valeur en distribution de probabilité :

$$\text{P}(a|s) = \frac{e^{\beta \text{Q}(s,a)}}{\sum\limits_{b \in \mathscr{A}} e^{\beta \text{Q}(s,b)}} \tag{3.30}$$

En augmentant le paramètre β, les contrastes entre les valeurs d'action seront magnifiés. En diminuant β, la probabilité d'action s'approchera d'une distribution uniforme.

3.4 Modèle de planification

Un agent peut obtenir un niveau d'anticipation et de contrôle supérieur en construisant un modèle de son environnement. Dans les méthodes précédemment décrites, les fonctions de valeur sont évaluées incrémentalement au cours de l'expérience. On dit que l'agent est «model-free» dans le sens où il ne connaît ni n'essaie d'apprendre de modèle de l'environnement, car il ignore les fonctions de transition et de récompense. Il se contentera des signaux de récompense envoyés par l'environnement dans différents états. A l'inverse, si l'agent est capable d'approximer la fonction de transition entre les états et la fonction de récompense, il lui est alors théoriquement possible d'évaluer la fonction de valeur optimale en réalisant une «simulation mentale», c'est-à-dire un parcours exhaustif d'un graphe de transition entre l'état pondéré par la fonction de transition et la fonction de récompense. Le rôle de l'expérience est ainsi d'améliorer le modèle du monde dans la perspective d'améliorer la prédiction de la fonction de valeur.

La contrepartie de ces méthodes est évidemment un coût computationnel et une connaissance absolue de l'ensemble des états que l'agent rencontrera. Dans SUTTON et BARTO [1998], les auteurs parlent d'apprentissage sur modèle («model-based»).

Dyna-Q déterministe

La combinaison de l'algorithme du q-learning et d'un modèle de planification s'appelle *Dyna-Q* et fut développée dans SUTTON [1990]. Intuitivement, Dyna-Q est basé sur l'idée d'approximer la fonction de valeur en utilisant les algorithmes d'apprentissage par différence temporelle sur une expérience «virtuelle». Dans l'algorithme 3.3 du Dyna-Q, le terme *Modèle* désigne la structure permettant de mémoriser les transitions $< s, a, s' >$ ainsi que la récompense r associée. Pour finir, la convergence de la fonction de valeur augmente proportionnellement au nombre N d'échantillonnages du monde virtuel de l'agent.

Algorithme 3.3 : Dyna-Q

Initialiser $Q(s,a)$ *et* $Modèle(s,a)$

répéter

  $s \leftarrow$ état non terminal courant

  $a \leftarrow \sigma - greedy(s,Q)$

  Observer r, s'

  $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$

  $Modèle(s,a) \leftarrow s', r$

  répéter

    $s \leftarrow$ état précédent choisi aléatoirement

    $a \leftarrow$ action choisie à partir de s

    $s', r \leftarrow Modèle(s,a)$

    $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$

  jusqu'à N

jusqu'à *l'infini*

Dyna-Q stochastique

La version stochastique de Dyna-Q permet d'apprendre dans un environnement avec des transitions entre états incertaines. Le modèle prend ainsi la forme d'une table de transitions probabilistes entre états $\hat{\mathcal{T}}^a_{ss'}$ et d'une table de récompense attendue $\hat{\mathcal{R}}^a_{ss'}$. Pendant la phase de simulation interne, la mise à jour s'effectue selon :

$$Q(s,a) \leftarrow \sum_{s'} \hat{\mathcal{T}}^a_{ss'}\left[\hat{\mathcal{R}}^a_{ss'} + \gamma \max_{a'} Q(s',a')\right] \tag{3.31}$$

Heuristique de recherche

Pour finir, les méthodes de planification dans un espace d'états à partir d'un modèle et ne cherchant pas à modifier la fonction de valeur sont appelées méthodes de recherche par heuristique. Le but ici est de prendre la meilleure décision sachant une fonction de valeur donnée.

Dans la version la plus simple, l'idée est de parcourir un arbre de transition à partir de l'état courant vers tous les états futurs. Selon SUTTON et BARTO [1998], une recherche dans l'arbre des transitions à une profondeur k telle que γ^k soit très petit assure une politique optimale. Evidemment, le coût computationnel augmente proportionnellement à k. Néanmoins, ces méthodes sont potentiellement plus efficaces que des méthodes telles que Dyna-Q puisqu'elles ciblent les actions et les états qui suivent l'état courant réduisant ainsi grandement l'espace d'états.

3.5 Conclusion

Ce chapitre a donc permis de présenter les outils formels nécessaires à la compréhension des modèles présentés dans les chapitres suivants. En effet, comme nous allons le voir plus loin, ce formalisme permet d'algorithmiser la plupart des tâches issues de la psychologie et des neurosciences comportementales [BALLEINE et collab., 2008; DEZFOULI et BALLEINE, 2012; NIV, 2009].

Dans la suite de ce manuscrit, la distinction entre apprentissage sans modèle (SARSA, q-learning, acteur-critique) et apprentissage sur modèle (Dyna, heuristique) va se révéler être des plus pertinentes pour une série d'études sur les systèmes de mémoire en neurosciences. Très brièvement, certaines études ont ainsi capitalisé sur les différences de propriétés entre ces algorithmes notamment au niveau de la vitesse de convergence et de la flexibilité [DAW et collab., 2005; KERAMATI et collab., 2011] pour expliquer des mécanismes d'interactions entre systèmes de mémoire parallèles. Une revue exhaustive de ces études et de leurs implications sera donc le sujet du chapitre suivant.

Chapitre 4

Coordination de stratégies

Sommaire

Ce chapitre se décompose en trois sous-parties, chacune traitant d'un courant de recherche se développant à la fois parallèlement et en relation. Par souci de regroupement thématique, l'ordre chronologique de publication des études n'est pas respecté et les sections sont relativement indépendantes.

La première section regroupe des modèles de coordination de systèmes de mémoire pour la navigation intentionnellement développés pour reproduire les expériences de dissociation entre l'hippocampe et le striatum chez le rat [PACKARD, 1999; PACKARD et collab., 1989; PACKARD et MCGAUGH, 1996]. Dans la deuxième section, les modèles de coordination dans le cadre de l'étude du conditionnement instrumental sont présentés [BALLEINE et DICKINSON, 1998; DICKINSON, 1985; KILLCROSS et COUTUREAU, 2003]. La dernière section illustre brièvement une application possible de ce sujet : la robotique bio-inspirée.

Ces études sont-elles comparables ? Il existe des différences certaines entre une tâche de navigation et une tâche de conditionnement instrumental notamment du point de vue du type d'informations manipulées. Dans le premier cas, plusieurs stratégies peuvent être differenciées sur des critères comme l'entrée du système (sensoriel, proprioceptif, interne) ou le repère de référence (égocentrique ou allocentrique) [ARLEO et RONDI-REIG, 2007]. Dans le second cas, la tâche est éminemment stéréotypée.

Néanmoins, le point commun entre ces deux littératures est (dans la plupart des cas) la coordination entre un système d'apprentissage sur modèle (cf sec 3.4) et un système d'apprentissage par différence temporelle (cf sec 3.3) [KHAMASSI et HUMPHRIES, 2012]. Une évaluation de cette double littérature a ainsi pour objectif de transférer si possible des principes computationnels de coordination qui auraient été appliqués avec succès dans une littérature mais pas encore testés dans l'autre. Dans les contributions présentées plus loin dans ce manuscrit, nous comparerons même certains de ces critères de coordination issus des deux littératures sur des mêmes données chez l'homme.

4.1 En navigation

4.1.1 Développement précurseur

Au milieu des années 1990, des travaux d'enregistrement de neurones dopaminergiques dans l'aire tegmentale ventrale en conditionnement pavlovien [SCHULTZ et collab., 1993, 1997] ainsi que les intuitions de plusieurs théoriciens de l'apprentissage par renforcement [MONTAGUE et collab., 1996] offrirent une «révolution» en neuroscience. En effet, le taux de décharge de neurones dopaminergiques était identifiable proportionnellement à l'erreur de prédiction δ_t tel qu'énoncé à l'équation 3.27. Réciproquement, cette erreur de prédiction ou différence temporelle permettait de donner du sens à une activité neuronale autrement mystérieuse.

En d'autres termes, un système formel issu de la recherche en intelligence artificielle décodait un système naturel. Encore autrement, le système formel prédisait le fonctionnement du système naturel. Porté par un tel succès, il était donc normal de vouloir étendre le décodage d'un système d'apprentissage par renforcement vers d'autres observations de l'activité neuronale. Dans les modèles d'apprentissage par renforcement, l'ensemble des états constitue le cœur du système formel puisqu'ils permettent au pouvoir récursif d'apparaître (cf. Chapitre 3 ; A plus forte raison, le pouvoir récursif intrinsèque à tout modèle constitue en lui-même la seule justification de l'existence d'un ensemble d'états $\mathcal{S}$).

Bien que les enregistrements dopaminergiques de SCHULTZ et collab. [1997] soient issus de tâches de conditionnement pavlovien, les premiers modèles computationnels ont

très tôt appliqué l'idée d'apprentissage par différence temporelle basé sur la dopamine à des paradigmes de navigation. Dans FOSTER, MORRIS et DAYAN [2000], les auteurs proposent un décodage de l'ensemble des états $\mathscr{S}$ vers l'activité des cellules de lieux observée dans l'hippocampe [O'KEEFE et NADEL, 1978] à travers un système acteur-critique. Si le signal dopaminergique est la valeur de différence temporelle, les cellules de lieux sont les états du monde dans lequel l'agent/animal évolue.

Dans ce chapitre, cette étude constitue notre premier exemple d'un modèle computationnel de coordination de systèmes. Toutefois, c'est GUAZZELLI et collab. [1998] qui fut la première étude à proposer un système formel de la sorte. De même, ARLEO et GERSTNER [2000] propose à la même époque un modèle de navigation très proche d'une coordination de systèmes. Ces modèles spécifiques à la navigation seront discutés plus loin.

Au-delà de l'origine de la coordination de systèmes de mémoire, d'autres modèles de l'hippocampe simulant un agent apprenant la position d'une récompense dans un labyrinthe avaient déjà été proposés [BURGESS et collab., 1994; REDISH et TOURETZKY, 1997; TRULLIER et MEYER, 1997]. Ces études ne modélisaient qu'une seule stratégie (tout en faisant «coopérer» néanmoins une mémoire épisodique et l'apprentissage par renforcement). Deux limites apparaissaient clairement. Dans le premier cas, une unique cellule de lieux est associée au but de l'agent par un signal de récompense mais ne permet pas d'associer simplement les autres cellules de lieux aux actions menant à cette récompense [BURGESS et collab., 1994]. Dans un deuxième cas, les cellules de lieux sont associées à une coordonnée métrique permettant de calculer la direction vers le but [REDISH et TOURETZKY, 1997]. Évidemment, le défaut de cette approche est la position des coordonnées métriques de nature relative qui peuvent être rendues caduques par le positionnement aléatoire de l'animal en début d'essai (celui-ci perdant alors l'origine des coordonnées de l'essai précédent).

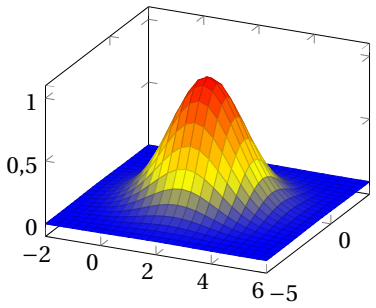

FIGURE 4.1 – Représentation idéalisée de l'activité d'une cellule de lieux dans FOSTER et collab. [2000].

La première étape de modélisation dans FOSTER et collab. [2000] est de représenter l'activité d'une cellule de lieux $i = 1 \ldots N$ placée au hasard dans l'espace du labyrinthe en fonction de la distance de l'agent par rapport au centre de la cellule de lieux s_i :

$$f_i(p) = exp(-\frac{||p - s_i||^2}{2\sigma^2}) \tag{4.1}$$

avec p la position de l'agent et σ l'étalement de la cellule de lieux. Pour illustration, cette simplification donne ainsi une forme d'activité telle que représentée dans la figure 4.1. Dans une deuxième étape, les auteurs ont choisi de modéliser l'acteur-critique sous la forme d'un réseau de neurones. De fait, une formalisation des états sous forme vectorielle au lieu d'une forme symbolique telle que présentée dans le chapitre 3 n'augmente ni ne diminue le cœur de la théorie de l'apprentissage par renforcement. Le critique est considéré comme une seule unité recevant la totalité de l'activité des cellules de lieux à travers un vecteur de poids $W \in \mathbb{R}^N$:

$$C(p) = \sum_i w_i f_i(p) \tag{4.2}$$

La fonction de valeur est approximée en changeant les poids w_i selon :

$$\Delta w_{i,t} \propto \delta_{i,t} f_{i,t}(p) \tag{4.3}$$

$$\delta_{i,t} = r_t + \gamma C(p_{t+1}) - C(p_t) \tag{4.4}$$

Étant dépendant de la position p, l'activité C du critique convergera vers un gradient dont la valeur maximale sera située au niveau de la récompense. Pour l'acteur, l'agent possède 8 directions différentes vers lesquelles se déplacer et sélectionne une direction avec la règle du soft-max 3.30. La valeur d'une action est calculée à chaque instant t en fonction de la position donnée par les cellules de lieux :

$$a_j(p) = \sum_i z_{ji} f_i(p) \text{ avec } j \in 1,\ldots,8 \tag{4.5}$$

La matrice de poids z_{ji} est modifiée selon :

$$\Delta z_{ji} \propto \delta_{i,t} f_{i,t}(p) g_{j,t} \tag{4.6}$$

avec $g_{j,t} = 1$ si l'action j a été choisie au temps t et $g_{j,t} = 0$ sinon. Ainsi, une cellule de lieux active en même temps qu'une cellule d'action à un instant où la valeur est augmentée verra sa connection augmentée de la même façon.

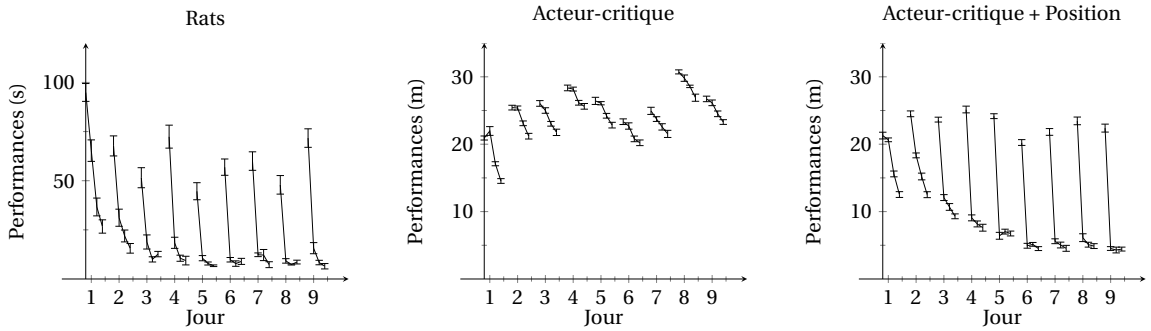

FIGURE 4.2 – Performances (temps de fuite en secondes pour les rats et distance avant la plateforme pour les modèles) pour trouver une plateforme immergée par un groupe de rats (*gauche*), le modèle d'acteur-critique simulé (*centre*) et le modèle combinant acteur-critique et position relative (*droite*). La plateforme est changée chaque jour. Adapté de FOSTER et collab. [2000]

Si un rat est entraîné à nager dans une piscine de Morris vers une plateforme immergée dont la localisation ne change pas, le modèle acteur-critique capture très bien le comportement du rat. Les difficultés apparaissent si la position de la plateforme est changée quotidiennement. Les résultats sont présentés dans la figure 4.2 et montrent ainsi la différence de comportement entre l'animal (gauche) et l'agent simulé (centre). Dans les premiers jours, l'apprentissage chez le rat est graduel mais devient quasiment instantané à partir du jour 5 et il est fort probable que le rat ait compris la structure de la tâche. A l'opposé, l'acteur-critique ne montre aucunement ce comportement et se contente d'apprendre une nouvelle fonction de valeur chaque jour. C'est un processus lent par essence et ralenti encore plus par la fonction de valeur apprise le jour précédent qui doit être oubliée.

Pour dépasser cette limitation, les auteurs proposent d'augmenter les capacités de leur agent en lui offrant la possibilité d'apprendre à calculer, à partir de l'activité de ses cellules de lieux, sa position exacte dans un repère cartésien en fonction de ses mouvements propres et de l'activité des cellules de lieux. Deux unités X et Y sont ainsi définies

et leur activité est calculée selon :

$$\mathrm{X}(p) = \sum_i w_i^{\mathrm{X}} f_i(p) \tag{4.7}$$

$$\mathrm{Y}(p) = \sum_i w_i^{\mathrm{Y}} f_i(p) \tag{4.8}$$

La mise à jour des poids s'effectue de manière incrémentale, par exemple pour X, selon :

$$\Delta w_i^{\mathrm{X}} \propto (\Delta x_t + \mathrm{X}(p_{t+1}) - \mathrm{X}(p_t)) \sum_{k=1}^{t} f_i(p_k) \tag{4.9}$$

avec Δx_t la distance parcourue dans la direction X. En convergeant, les unités X et Y offrent donc un couple de valeurs (x, y) différentes pour chaque position du labyrinthe.

Finalement, nous entrons dans le vif du sujet de ce chapitre avec l'adjonction d'une unité d'action a_{coord} aux 8 unités d'actions déjà présentes dans l'acteur-critique. L'unité a_{coord} se conforme aussi à l'équation 4.5 pour l'activité et à l'équation 4.6 pour la mise à jour de sa connection avec les cellules de lieux. Si le *soft-max* sélectionne a_{coord}, un vecteur de direction est calculé en fonction de la position estimée (X,Y) et de la position de récompense (X',Y') gardée en mémoire par deux autres unités. Si la récompense n'a pas encore été obtenue par l'agent, le vecteur de direction est calculé aléatoirement et la mise à jour 4.6 ne s'effectue pas.

Comme le montre la figure 4.2 de droite, la combinaison des deux systèmes permet donc de capturer le comportement des rats. L'amélioration graduelle des performances en début de tâche est essentiellement contrôlée par les actions du modèle acteur-critique qui est plus fiable que l'action a_{coord}. Lors de la convergence du système de coordonnées (X,Y) au jour 4, l'agent n'utilise plus que ce même système pour décider. La position (X,Y) étant apprise indépendamment du but, celle-ci permet de réaliser une navigation instantanée vers n'importe quelle récompense arbitraire. Cette capacité est bien entendu impossible pour l'acteur-critique.

La coordination de systèmes de mémoire telle qu'elle nous intéresse est ici sous une forme primitive. Néanmoins, elle sous-tend déjà les développements futurs par la façon de sélectionner des actions proposées par des systèmes formellement indépendants dans leur processus de prise de décision. De plus, ces systèmes possèdent des vitesses d'apprentissage différentes (lente pour l'acteur-critique et rapide pour le modèle de Position) qui une fois combinés permettent néanmoins de meilleures performances. Ils se différencient aussi par une flexibilité du modèle de Position qui n'existe pas pour le modèle acteur-critique. Si la position de la récompense change, le modèle acteur-critique doit réapprendre sa fonction de valeur. Le coût de calcul propre au modèle de Position n'est pas pris en compte ici dans la sélection de l'action. Tous ces détails ne sont pas exprimés explicitement dans l'étude. Toutefois, ils se retrouveront tous de façon formelle dans les études suivantes.

Au sujet du décodage de l'ensemble $\mathscr{S}$, FOSTER et collab. [2000] ne constitue qu'un exemple parmi la multitude d'autres modèles qui ont essayé d'assigner un rôle formel aux cellules de lieux [ARLEO et GERSTNER, 2000; GUAZZELLI et collab., 1998] tout en augmentant ces cellules de lieux avec de l'apprentissage par renforcement. De fait, le processus de sélection de l'action dans ARLEO et GERSTNER [2000] est similaire au modèle présenté dans cette section puisqu'il associe aussi les cellules de lieux à l'apprentissage par renforcement. De plus, un aspect multi-système existe aussi au niveau du processus de localisation des états/cellules de lieux de l'agent (deux systèmes de cellules de lieux interagissent, l'un utilisant l'information visuelle, l'autre l'information odométrique, pour permettre un

recalibrage et une localisation plus robuste). Dans GUAZZELLI et collab. [1998], les auteurs proposent un modèle complexe de plusieurs modules comprenant entre autres un modèle de construction d'un graphe (modélisant l'hippocampe et les cellules de lieux) et une stratégie de réponse utilisant un apprentissage par différence temporelle. La sélection de l'action est basée sur une somme de la récompense attendue par chaque module implémentant ainsi une fusion de stratégie. Les auteurs ont ensuite comparé les performances de leur modèle sur des expériences d'inversion de la position de la récompense dans un labyrinthe en T et un labyrinthe semi-radial chez le rat.

4.1.2 Coordination par différence temporelle

Inspiré par les résultats sur les dissociations entre hippocampe et striatum dorso-latéral (cf. Chapitre 2), un modèle de l'interaction de ces deux régions (et donc des deux stratégies qu'elles sous-tendent) a été proposé dans CHAVARRIAGA et collab. [2005]. Par souci de simplicité, tous les détails de ce modèle illustré dans la figure 4.3 ne seront pas décrits. Pour rappel, la stratégie de lieux (à droite dans 4.3) se base sur l'hippocampe et la stratégie de réponse (à gauche dans 4.3) se base sur le striatum dorso-latéral. Au contraire de FOSTER et collab. [2000], les cellules de lieux apprennent leur position à partir d'un vecteur mimant la vision de l'agent et d'une intégration de chemin. De manière similaire, les cellules de lieux sont aussi connectées à travers une matrice de poids W à 36 cellules d'actions, chacune proposant une direction différente. La stratégie de réponse possède aussi 36 cellules d'actions propres connectées directement à un vecteur d'entrée. Si un indice visuel s'aligne avec la direction préférée d'une partie i du vecteur, la position i de ce vecteur prend la valeur 1.

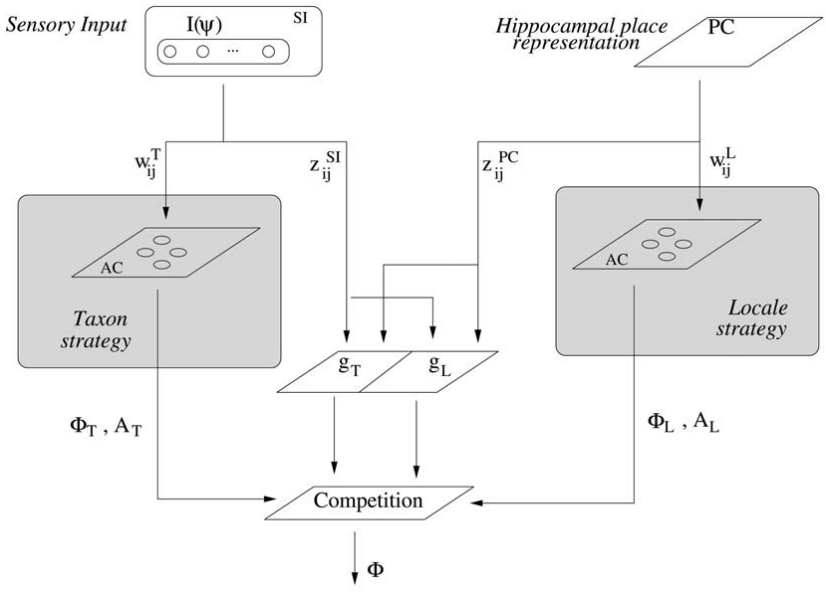

FIGURE 4.3 – Illustration du modèle de CHAVARRIAGA et collab. [2005]. La stratégie de réponse (*Taxon strategy*) est représentée à gauche et la stratégie de lieux (*Locale strategy*) est représentée à droite. Au milieu, le système d'arbitrage reçoit les entrées de chaque système pour décider et apprendre la meilleure politique de sélection de stratégie. Reproduit de CHAVARRIAGA et collab. [2005].

Chaque stratégie (ou expert) propose indépendamment une action. La sélection d'une stratégie est donc la question à résoudre à chaque instant pour l'agent. Le choix des modélisateurs a donc été de considérer chaque expert comme un «état» (dans le sens de la

théorie de l'apprentissage par renforcement) sur lequel peut se construire une fonction de valeur $g(k)$ avec $k \in$ {Stratégie de lieux (L), Stratégie de réponse (R)}. L'idée étant d'apprendre petit à petit quel expert a la plus grande valeur lorsqu'il contrôle le comportement à différents moments de la tâche. La fonction de valeur est calculée grâce à une matrice de poids z_{kj} connectant l'entrée $r_{k,j}$ de chaque expert avec la fonction de valeur selon :

$$g(k) = \sum_{j \in \mathrm{L}} (z_{\mathrm{L}j} r_j) + \sum_{j \in \mathrm{R}} (z_{\mathrm{R}j} r_j) \tag{4.10}$$

Cette fonction de valeur est ensuite utilisée pour choisir une stratégie selon :

$$\mathrm{P}(k) = \frac{g(k)\mathrm{A}(k)}{g(\mathrm{L})\mathrm{A}(\mathrm{L}) + g(\mathrm{R})\mathrm{A}(\mathrm{R})} \tag{4.11}$$

$\mathrm{A}(k)$ représente la valeur égale à la moyenne des deux actions $a_{k,i}$ les plus proches de la direction Φ_k choisie par l'expert k selon :

$$\Phi_k = \arctan \frac{\sum_i a_{i,k} sin(\phi_i)}{\sum_i a_{i,k} cos(\phi_i)} \tag{4.12}$$

avec ϕ_i la direction représentée par l'action $a_{i,k}$. Comme FOSTER et collab. [2000], la modification des poids se fait en utilisant les techniques d'apprentissage par renforcement et nécessite donc de calculer une erreur de prédiction δ_t. Néanmoins, la modification des poids des experts est pondérée par la probabilité $\mathrm{P}(k)$ d'avoir été sélectionné. Avant de passer aux résultats de test, il convient de remarquer que la force du modèle de CHAVARRIAGA et collab. [2005] est de pouvoir combiner *ad infinitum* des experts indépendants donc de mettre N experts différents avec le même principe si on le souhaite.

Ce modèle a été testé sur la tâche de navigation présentée dans PEARCE et collab. [1998] dont le résultat principal est représenté dans la figure 4.4.A. Très brièvement, les auteurs ont entraînés des rats à nager vers une plateforme située à une distance et une direction constantes d'un indice visuel. L'entraînement est effectué pendant 11 sessions de 4 essais. Au début de chaque session, l'indice et la plateforme sont placés aléatoirement sur 8 positions différentes pour rendre ineffective la stratégie de lieux. Néanmoins, l'indice donne une information sur la position proche de la plateforme. Dans la figure 4.4, le temps moyen nécessaire aux rats du groupe contrôle et du groupe avec lésion de l'hippocampe pour trouver la plateforme est représenté à l'essai 1 et l'essai 4 pour chaque session. On observe ainsi que :

1. tous les animaux apprennent une stratégie comme le montre la diminuation globale du temps moyen
2. le temps moyen diminue entre les sessions et à l'intérieur d'une session pour le groupe contrôle
3. la progression ne s'effectue qu'entre les sessions pour le groupe avec lésion de l'hippocampe
4. le temps moyen du groupe lésionné reste confiné entre les temps moyens du groupe sain
5. les performances se rejoignent dans la dernière session.

Pour les auteurs, l'interprétation est la suivante :

I. le groupe sain utilise une stratégie de lieux pour s'améliorer à l'intérieur d'une session

II. le groupe lésionné ne dispose pas de stratégie de lieux et ne peut donc pas s'améliorer à l'intérieur d'une session

III. l'utilisation d'une stratégie de lieux est pénalisant pour le groupe sain au premier essai puisqu'elle suggère au rat de chercher la position de l'essai précédent

IV. les 2 groupes lésionnés apprennent une stratégie de réponse utilisable pour toutes les sessions

V. la stratégie de lieux est plus efficace que la stratégie de réponse

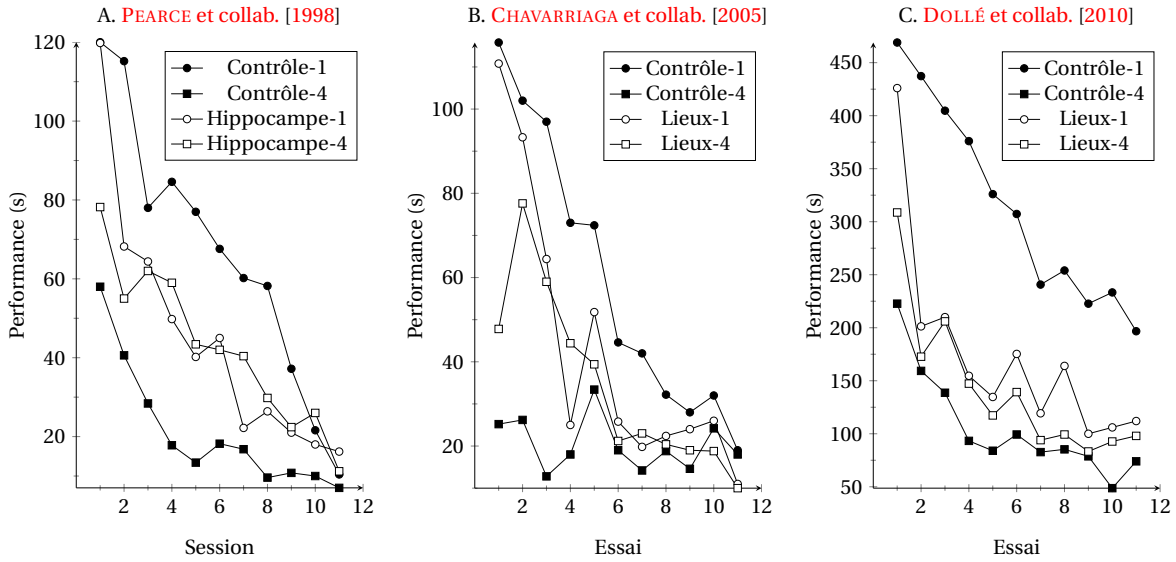

FIGURE 4.4 – A. Performances (temps de fuite en s) d'un groupe de rats sains et d'un groupe avec lésion de l'hippocampe dans une tâche de localisation de plateforme. Les performances sont représentées à l'essai 1 et l'essai 4 de chaque bloc. La position de la plateforme et de l'indice visuel change pour chaque bloc. Adapté de PEARCE et collab. [1998]. B. Performances du modèle de coordination des systèmes de mémoire de CHAVARRIAGA et collab. [2005] sur la tâche de PEARCE et collab. [1998]. C. Performances du modèle de coordination de systèmes de mémoire de DOLLÉ et collab. [2010] sur la tâche de PEARCE et collab. [1998].

Les résultats de la simulation d'un agent selon le modèle de CHAVARRIAGA et collab. [2005] sur cette tâche sont représentés dans la figure 4.4.B. Un agent Lieux (ne possédant plus le modèle de stratégie de lieux) correspond au groupe avec lésion de l'hippocampe. Globalement, le comportement de l'agent se conforme à celui des rats sur les différents points évoqués précédemment. La principale différence se situe au niveau de l'absence de progression de la performance pour le groupe sain entre les sessions. A l'essai 4 de toutes les sessions, le temps de fuite est minimal au début de l'expérience comme en fin d'expérience. Cela traduit une convergence de la stratégie de lieux plus rapide que celle des rats. Pour finir, les performances à partir de la session 5 sont indissociables pour Contrôle-4 et Lieux-4. L'adjonction d'une stratégie de lieux à la stratégie de réponses n'offre aucun avantage comportemental pour le modèle au contraire des rats.

Dans la continuité de la modélisation de l'hippocampe et du striatum dorso-latéral, DOLLÉ et collab. [2010] propose un autre modèle de coordination de systèmes de mémoire sur la tâche présentée dans PEARCE et collab. [1998]. Les résultats sont présentés dans la figure 4.4.C et le modèle est illustré dans la figure 4.5.

Sur le plan formel, le modèle de DOLLÉ et collab. [2010] est très similaire aux modèles présentés précédemment. La seule différence est qu'il combine un apprentissage sur modèle avec un apprentissage sans modèle au contraire de CHAVARRIAGA et collab. [2005]. La

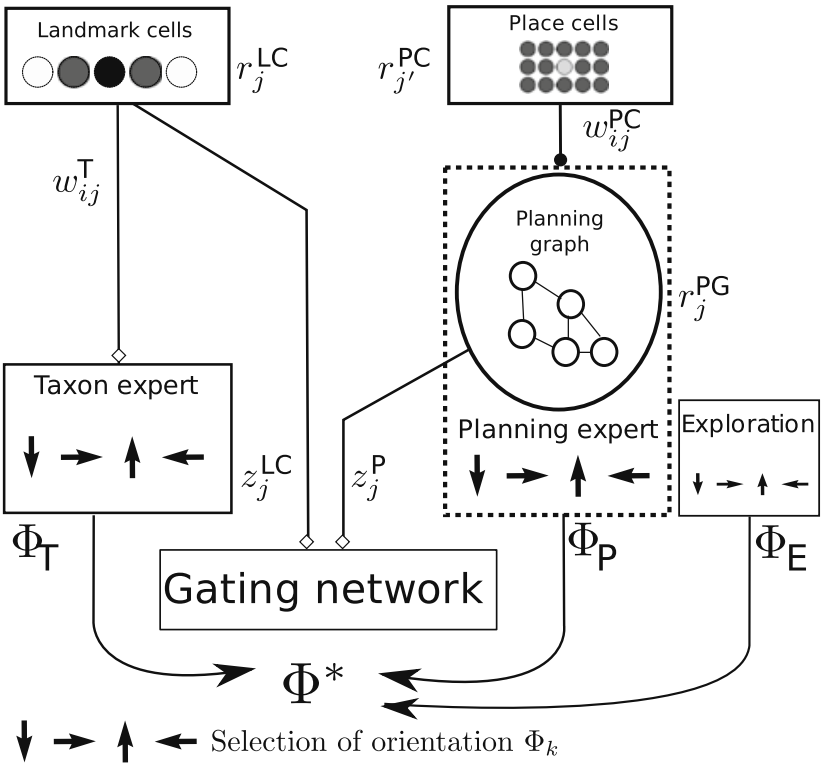

FIGURE 4.5 – Illustration du modèle de DOLLÉ et collab. [2010]. La stratégie de réponse (*Taxon expert*) est représentée à gauche et la stratégie de lieux (*Planning expert*) est représentée à droite. Au milieu, le système d'arbitrage reçoit les entrées de chaque système pour décider et apprendre la meilleure politique de sélection de stratégie grâce à un q-learning. Reproduit de DOLLÉ et collab. [2010].

stratégie de réponse (à gauche dans la figure 4.5) reçoit une entrée sous la forme d'un vecteur indiquant la présence dans toutes les directions d'un indice visuel à proximité. Cet état est ensuite converti en une valeur d'action par une matrice de poids permettant ainsi de dérouler les équations de l'apprentissage par renforcement. La stratégie de lieux est modélisée par un algorithme de recherche dans un graphe (selon les modèles présentés dans le chapitre précédent 3.4). Les noeuds du graphe sont associés au cours de l'expérience aux cellules de lieux positionnées aléatoirement au début de l'expérience et leur activité est proportionnelle à la distance avec l'agent (voir 4.1 pour illustration). Pendant le choix d'action, le lien entre un noeud N_i et un noeud N_j permet de guider l'agent de la position associée à N_i vers la position associée à N_j. Lors de l'obtention de la récompense par l'agent, le noeud le plus proche reçoit une valeur $G_i^* = 1$. Cette valeur est ensuite propagée pour tous les noeuds et réduite en fonction de la distance n (de transitions) avec le noeud G^* selon :

$$G_i = \alpha^n \text{ avec } \alpha < 1 \tag{4.13}$$

Cette fonction de valeur construite instantanément permet à l'agent de «remonter» son graphe vers la position de la récompense en choisissant toujours le noeud possédant la valeur maximale. Si le gain en vitesse de convergence du modèle dépasse largement celui d'un algorithme d'apprentissage par renforcement classique (et accessoirement augmente drastiquement la complexité du calcul), il convient de noter que nous ne sortons pas du cadre de la théorie de l'apprentissage par renforcement. La récompense, la fonction de valeur réduite et les états existent de la même manière.

Au niveau de la sélection de stratégie, les auteurs proposent d'associer une valeur g_k à

chaque expert. Un expert proposant un choix aléatoire permettant l'exploration de l'agent est adjoint aux deux autres experts. Tout comme CHAVARRIAGA et collab. [2005], la valeur g_k est calculée en fonction de l'entrée de chaque système à travers une matrice de poids $\mathbf{Z}$ (z_j^{LC} et z_j^{P} dans la figure 4.5). Pour simplifier les calculs suivants, on considère $\mathbf{R}$ le vecteur total d'entrée du système résultant de la concaténation des vecteurs d'entrée propres à chaque système et $\mathbf{G} = [\ldots, g_k, \ldots]$ le vecteur représentant la fonction de valeur associée à chaque expert (dans notre cas, $\mathbf{G} \in \mathbb{R}^3$). Le vecteur $\mathbf{G}$ est calculé selon :

$$\mathbf{G} = \mathbf{Z}.\mathbf{R} \tag{4.14}$$

A chaque instant t, l'action Φ^* choisie par l'agent provient de l'expert possédant la valeur g^* la plus grande. La mise à jour des poids $\mathbf{Z}$ s'effectue comme l'algorithme du q-learning selon :

$$\Delta\mathbf{Z} = \delta(t)\mathbf{E}(t) \tag{4.15}$$

$$\delta(t) = r(t) + \gamma \max_k \big(\mathbf{G}(t+1)\big) - g^* \tag{4.16}$$

La matrice $\mathbf{E}(t)$ représente les traces d'éligibilité permettant de renforcer uniquement les poids entre les entrées et les experts utilisés récemment :

$$\mathbf{E}(t+1) = \lambda\mathbf{E}(t) + \mathrm{R}.[exp(-(\Phi^*(t) - \mathbf{\Phi})) - exp(-\frac{\pi}{2})] \tag{4.17}$$

La deuxième partie de l'équation permet de renforcer négativement les experts qui auraient proposé récemment une action opposée à l'action $\Phi^*(t)$ choisie au temps t.

Confrontés à la tâche de PEARCE et collab. [1998], les résultats de la simulation du modèle (voir figure 4.4.C) montrent un apprentissage globalement similaire à celui des rats. Contrairement à CHAVARRIAGA et collab. [2005], ce modèle reproduit l'apprentissage progressif de l'essai 4 entre les sessions 1 et 5. Néanmoins, les temps de fuite pour Contrôle-1 sont largement supérieurs chez le modèle par rapport aux performances des rats. Concrètement, le modèle peine à inhiber sa stratégie de lieux en début de session. Pour finir, le gain d'une coordination entre stratégies existe toujours pour les derniers essais (différence significative entre Contrôle-4 et Lieux-4 selon les auteurs).

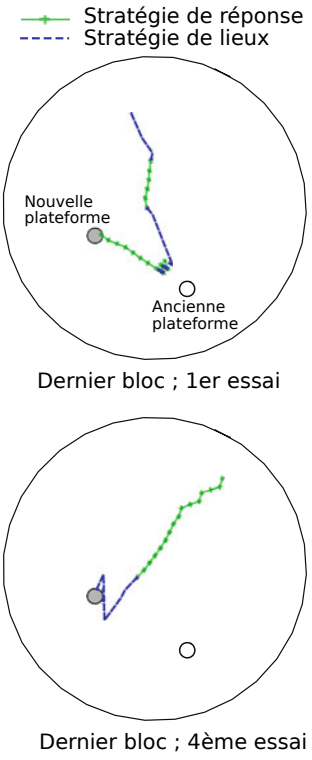

FIGURE 4.6 – Trajectoire typique du groupe contrôle en fin d'apprentissage pour l'essai 1 et l'essai 4 selon DOLLÉ et collab. [2010]

Une analyse spatiale et temporelle de la participation de chaque stratégie au cours du temps a été effectuée par les auteurs. Au niveau du degré d'utilisation, la stratégie de réponse se distingue de la stratégie de lieux puisque celle-ci est sélectionnée 60% du temps dans les derniers essais. Au niveau spatial, une trajectoire typique lors de sessions en fin d'apprentissage est illustrée dans la figure 4.6. Pour le premier essai, la trajectoire illustre bien la compétition entre les experts. La stratégie de lieux mène l'agent vers l'ancienne plateforme et la stratégie de réponse l'oriente vers l'indice visuel pour trouver la nouvelle plateforme. A l'essai 4, cette compétition disparaît et les deux stratégies sont complémentaires : l'approche permet de se localiser à proximité du but mais la stratégie de lieux,

plus précise, permet de trouver la localisation exacte de la plateforme. Pour conclure, il semblerait que le mécanisme d'arbitrage de ce modèle ne soit pas capable de résoudre efficacement ce problème de compétition initiale comme le montrent les mauvaises performances de Contrôle-1.

4.2 En apprentissage instrumental

Les résultats neurobiologiques dans une tâche de conditionnement instrumental ont été longuement discutés dans la section 2.4.2. Grâce à cette tâche, les rôles respectifs du striatum dorso-médian et du cortex préfrontal dans le comportement lié à un but et du striatum dorso-latéral dans le comportement habituel ont ainsi été exposés. Les études de modélisation de la transition entre ces systèmes de mémoire constituent donc le sujet de cette section.

Dans la plupart des cas, la coordination va s'effectuer entre un apprentissage sur modèle et un apprentissage par différence temporelle comme DOLLÉ et collab. [2010]. Toutefois, les modèles de coordination présentés ici sont radicalement différents puis qu'ils incorporent des notions d'incertitudes et de coût de calcul. L'hypothèse principale de cette section est la suivante : l'apprentissage sur modèle permet d'évaluer lentement et précisément une action et l'apprentissage par différence temporelle permet d'évaluer rapidement une action en contrepartie d'un apprentissage lent. Toute la finalité des modèles de coordination suivants va donc être de faire en sorte que le bon modèle soit utilisé au bon moment.

4.2.1 L'arbitrage selon l'incertitude

Pour réfléchir à la transition entre les systèmes de mémoire dans le conditionnement instrumental, il est nécessaire de formaliser le comportement lié à un but. Au cours de la section 2.4.2, nous avions évoqué brièvement les difficultés que posait théoriquement la finalité dans le comportement lié à un but, source des débats entre *cognitivistes* et *behavioristes*. Dans DAW et collab. [2005], les auteurs formalisent le comportement lié à un but grâce à une combinaison des techniques présentées à la section 3.4 sur les modèles de planification (que nous appellerons apprentissage sur modèle par la suite). De plus, cet apprentissage sur modèle est sous-tendu selon eux par une combinaison du cortex préfrontal et du striatum dorso-médian. Effectivement, le rôle du cortex préfrontal dans le comportement lié à un but a été observé par KILLCROSS et COUTUREAU [2003] et ce sont les résultats contenus dans cette étude que ces auteurs vont proposer de modéliser (cf 2.9). Plus précisément, ils vont reproduire la transition entre un comportement lié à un but vers un comportement habituel par surentraînement dans une tâche de conditionnement instrumental.

Comme illustré dans la figure 4.7, la tâche est formalisée comme un processus de décision markovien selon :

- $\mathscr{S} = \{S_0(\text{État initial}), S_1(\text{État transitoire}), S_2(\text{État sortant}), S_3(\text{État récompensant})\}$
- $\mathscr{A} = \{A_0(\text{Appuyer sur levier}), A_1(\text{Entrer dans l'auge })\}$
- $$\mathscr{R}(s_t, a_t) = \begin{cases} 1, & \text{if } (s_t, a_t) = (S_1, A_1). \\ 0, & \text{sinon.} \end{cases} \tag{4.18}$$
- La fonction de transition est rendue explicite dans la figure 4.7. Le processus de décision markovien est cyclique : les états S_2 et S_3 ramènent à l'état S_0.

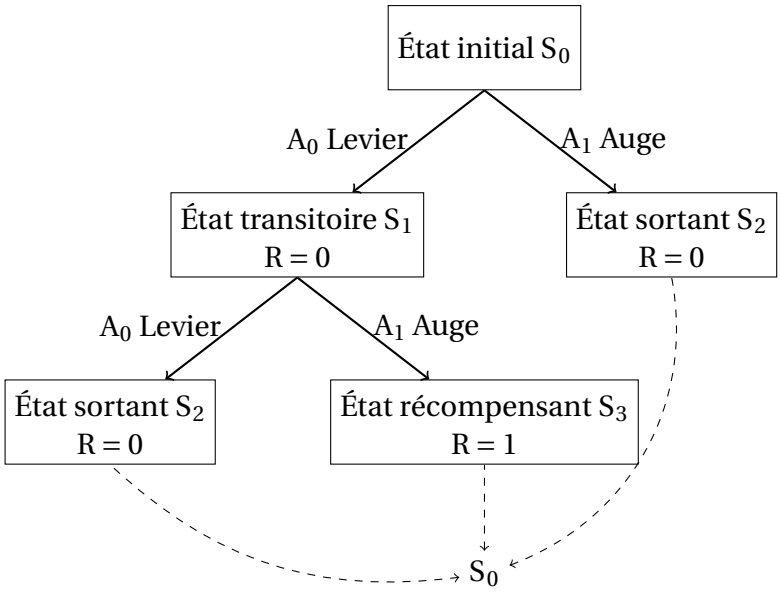

FIGURE 4.7 – Représentation sous la forme d'un processus de décision markovien d'une tâche standard de conditionnement instrumental. Adapté de DAW et collab. [2005]

.

Pour modéliser le système de mémoire sous-tendant l'apprentissage d'habitudes, les auteurs utilisent l'algorithme du q-learning version bayésienne [DEARDEN et collab., 1998]. En partie motivé par les progrès sur l'identification de l'erreur de prédiction au signal dopaminergique [MONTAGUE et collab., 1996; SCHULTZ et collab., 1993, 1997], ce choix de modèle se justifie par l'observation courante que les algorithmes d'apprentissage par différence temporelle sont lents à apprendre et inflexibles sur le choix de l'action une fois la fonction de valeur établie : deux observations imputables à un comportement habituel. Sans détailler le formalisme complet, la fonction de valeur $Q(s,a)$ dans la version bayésienne est remplacée par une distribution de probabilité. Les paramètres inhérents à toute loi de probabilité sont mis à jour grâce à la récompense et la valeur d'une action est calculée selon la moyenne de la distribution. De manière très simplifiée, le principal avantage de ce formalisme est de pouvoir garder en mémoire une mesure indirecte de la convergence de la fonction de valeur sous la forme de la variance associée à toute loi de probabilité. Si la fonction de valeur a convergé pour cette action, la variance associée à la valeur est faible.

Le système de mémoire modélisant le comportement lié à un but est conçu sous la forme de recherche dans un graphe en version bayésienne [DEARDEN et collab., 1999]. Les auteurs supposent que l'agent possède au début de la tâche la version complète de la fonction de transition. La justification de ce choix de modèle pour le comportement lié à un but est la suivante : la construction de la fonction de valeur «au vol» malgré le coût computationnel élevé permet à l'agent de s'adapter rapidement au changement de contingences dans l'environnement (c'est-à-dire les dévaluations dans le conditionnement instrumental). Tout comme le q-learning bayésien, l'apprentissage sur modèle bayésien permet aussi de calculer une incertitude sur la valeur d'une action de manière exacte à l'intérieur d'un essai.

Le cœur de la proposition de DAW et collab. [2005] est représenté dans la figure 4.8.A. La valeur moyenne des variables d'incertitudes pour l'action A_0 est calculée pour chaque système de mémoire en fonction des essais. Selon leur modèle de sélection de stratégie, le modèle ayant le minimum d'incertitudes contrôle le choix de l'action. Il apparaît ainsi

que l'incertitude du système d'apprentissage sur modèle est inférieure à l'incertitude du système d'apprentissage d'habitudes aux premiers essais. Vers l'essai 25, cette relation s'inverse. La fonction de valeur du q-learning a convergé et le peu d'incertitudes associées au processus computationnel simple de lire une valeur d'action dans une table de Q-valeurs permet à l'incertitude du q-learning d'être inférieure au modèle d'arbre de recherches. De plus, les auteurs modélisent la complexité et le coût computationnel induit par le calcul de la valeur selon l'arbre de transitions en augmentant de manière arbitraire (ajout d'un bruit) l'incertitude de l'apprentissage sur modèle.

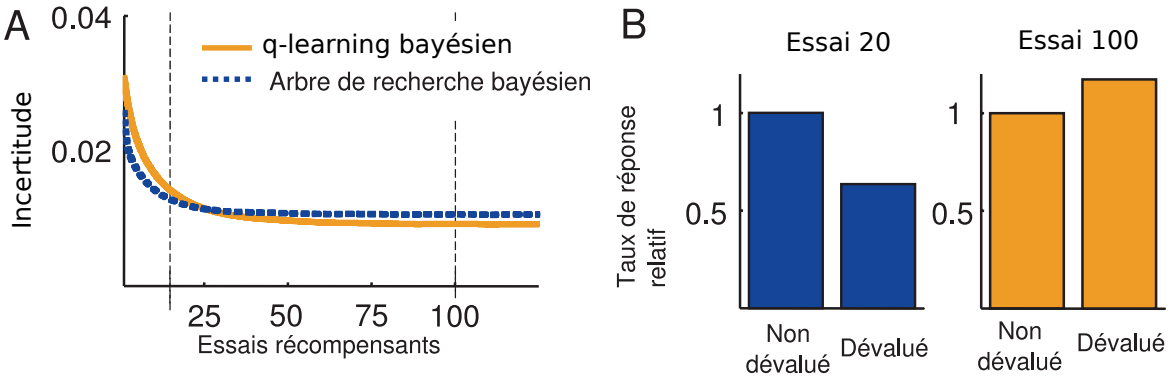

FIGURE 4.8 – A. Incertitude associée à chaque système permettant une transition dans le contrôle du comportement vers l'essai 25. B. Taux de réponse relatif dans l'appui de levier à l'essai 25 (donc contrôlé par le système d'apprentissage sur modèle) et à l'essai 100 (donc contrôlé par le système d'apprentissage d'habitudes). Adapté de DAW et collab. [2005].

Dans la figure 4.8, les résultats d'une dévaluation de la récompense ($\mathscr{R}(S_1, A_1) = 0$) après entraînement modéré (20 essais) et après entraînement intensif (100 essais) montrent la pertinence d'un tel choix de modélisation. Ces résultats reproduisent les observations faites chez le rat (voir figure 2.9 par KILLCROSS et COUTUREAU [2003]). Au début de l'entraînement, l'agent reste sensible à la valeur de la récompense. Une dévaluation de la récompense entraîne une diminution d'appui sur le levier. Si l'entraînement se poursuit, le système d'apprentissage d'habitudes prend le contrôle. Sachant la lenteur de ce modèle dans la mise à jour de sa politique, l'agent devient insensible à la dévaluation et continue d'appuyer sur le levier.

4.2.2 Le compromis entre vitesse et précision

La principale faiblesse de l'approche proposée dans DAW et collab. [2005] pour expliquer la transition entre comportement lié à un but et comportement habituel est la nécessité de lancer le calcul dans l'arbre des transitions (computationnellement coûteux) à chaque pas de temps pour évaluer l'incertitude. Ce calcul est néanmoins gratifiant puisqu'il permet d'obtenir une valeur précise de l'action. Le modèle d'habitudes ne souffre pas de cette limitation puisque l'incertitude se contente d'être lue en même temps que la valeur de l'action. Dans l'intention de dépasser ces limitations, KERAMATI et collab. [2011] propose un processus d'arbitrage construit sur un compromis entre la vitesse d'exploitation et la précision de l'estimation de la valeur.

Nous allons décrire le modèle de KERAMATI et collab. [2011] plus en détails ici, car il fait partie des modèles que nous avons comparés au nôtre, notamment dans la publication VIEJO et collab. [2015]. La réplication de ce modèle (après correction de certaines valeurs de paramètres) a d'ailleurs été validée et publiée dans VIEJO et collab. [2016].

Encore une fois, le système d'habitudes est modélisé par un algorithme d'apprentissage par différence temporelle : le Kalman q-learning [GEIST et collab., 2009]. Version simplifiée du q-learning bayésien [DEARDEN et collab., 1998], le Kalman q-learning permet de représenter la fonction de valeur sous la forme d'une loi normale. Au cours de l'apprentissage, une matrice de covariance Σ entre les couples état-action est mise à jour selon un gain de Kalman. Au moment du choix d'action, la moyenne de la loi normale sert directement de valeur de l'action.

Pour l'apprentissage sur modèle, l'agent maintient sa propre fonction de transition $p_T(s_t, a_t, s_{t+1}) \rightarrow [0,1]$ et sa propre fonction de récompense $R(s_t, a_t) \in \mathscr{R}$ qu'il évalue en fonction de son interaction avec l'environnement selon :

$$p_T(s_t, a_t, s_{t+1}) = (1-\phi)p_T(s_t, a_t, s_{t+1}) + \phi \tag{4.19}$$

$$R(s_t, a_t) = (1-\rho)R(s_, a_t) + \rho r_t \tag{4.20}$$

avec ϕ et ρ les paramètres du modèle. Lors de la prise de décision, le modèle mimant le comportement lié à un but entame un processus récursif selon :

$$Q^{mod}(s_t, a_i) = R(s_t, a_i) + \gamma \sum_{s'} p_T(\{s, a\} \rightarrow s') \max_{b \in A} Q(s', b)^{mod} \tag{4.21}$$

qui s'arrête au bout de 3 transitions dans le graphe 4.7. Pour simplifier le calcul, les auteurs supposent que l'agent a déjà une connaissance de la dimension du problème, c'est-à-dire de tous les états et les actions dans l'environnement. Pour mimer le processus de dévaluation de l'animal dans une tâche de conditionnement instrumental, la valeur $R(S_1, A_1)$ est arbitrairement changée à -1.

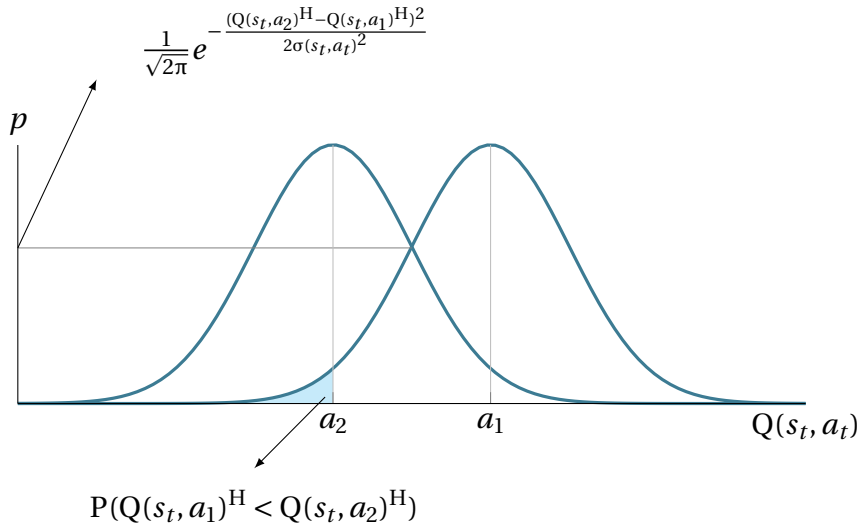

FIGURE 4.9 – Schéma du calcul de la valeur parfaite d'information de l'équation 4.23. Dans cette représentation idéalisée, l'action a_1 a une q-valeur supérieure à l'action a_2. Les flèches indiquent les deux parties de l'équation.

Pour arbitrer, l'agent va maintenir en mémoire une mesure du taux de récompense $\bar{R}$ évalué selon :

$$\bar{R} = (1-\sigma)\bar{R} + \sigma r_t \tag{4.22}$$

avec σ un paramètre du modèle. Cette valeur va ensuite être comparée à une valeur parfaite d'information $VPI(s_t, a)$ calculée pour chaque action possible dans l'état s_t. Pour

Algorithme 4.1 : KERAMATI et collab. [2011]

Initialisation
$Q(s,a)^{mod}, Q(s,a)^{hab}, \Sigma$
$\mathscr{R}(S_1, A_1) = 1$ *Fonction de récompense de l'environnement*
$\bar{R} = 0$ *Taux de récompense*
$R(s,a) = \{0, \ldots\}$ *Fonction de récompense de l'agent*

répéter
  $s_t \leftarrow S_0$
  si *Dévaluation* **alors**
    $\mathscr{R}(S_1, A_1) = 0$
    $R(S_1, A_1) = -1$
  fin
  tant que $s_t \neq S_3$ **faire**
    Sélection de l'action
    $\{a_1, \ldots, ai, \ldots\} \leftarrow sort(Q^{hab}(s_t, a_i))$
    Calculer :
    pour $a_i \neq a_1$ **faire**
      $VPI(s_t, a_i) \leftarrow$ 4.23
    fin
    $VPI(s_t, a_1) \leftarrow$ 4.24
    Comparer :
    pour $i \in \{a_1, a_2, \ldots, a_i, \ldots\}$ **faire**
      si $VPI(s_t, a_i) \geq \tau\bar{R}(t)$ **alors**
        $Q(s_t, a_i) \leftarrow Q^{model}(s_t, a_i)$
      fin
      sinon
        $Q(s_t, a_i) \leftarrow Q^{hab}(s_t, a_i)$
      fin
    fin
    $a_t \leftarrow soft - max[Q(s_t, a)]$
    $r_t = \mathscr{R}(s_t, a_t)$
    $s_{t+1} = \mathscr{T}(s_t, a_t)$
    Mise à jour (voir A.1)
  fin
jusqu'à *fin de l'apprentissage*

l'action possédant la plus grande q-valeur (et donc susceptible d'être sélectionnée), la mesure d'incertitude VPI se calcule selon :

$$\mathrm{VPI}(s_t,a_1) = (\mathrm{Q}(s_t,a_2)^{\mathrm{H}} - \mathrm{Q}(s_t,a_1)^{\mathrm{H}})\mathrm{P}(\mathrm{Q}(s_t,a_1)^{\mathrm{H}} < \mathrm{Q}(s_t,a_2)^{\mathrm{H}}) + \frac{\sigma(s_t,a_1)}{\sqrt{2\pi}} e^{-\frac{(\mathrm{Q}(s_t,a_2)^{\mathrm{H}} - \mathrm{Q}(s_t,a_1)^{\mathrm{H}})^2}{2\sigma(s_t,a_1)^2}} \tag{4.23}$$

Pour les autres actions avec une q-valeur inférieure, la VPI se calcule selon :

$$\mathrm{VPI}(s_t,a_i) = (\mathrm{Q}(s_t,a_i)^{\mathrm{H}} - \mathrm{Q}(s_t,a_1)^{\mathrm{H}})\mathrm{P}(\mathrm{Q}(s_t,a_i)^{\mathrm{H}} > \mathrm{Q}(s_t,a_1)^{\mathrm{H}}) + \frac{\sigma(s_t,a_t)}{\sqrt{2\pi}} e^{-\frac{(\mathrm{Q}(s_t,a_1)^{\mathrm{H}} - \mathrm{Q}(s_t,a_i)^{\mathrm{H}})^2}{2\sigma(s_t,a_t)^2}} \tag{4.24}$$

Pour avoir une intuition de ces deux longues équations, la première a été schématisée dans la figure 4.9 avec une séparation en 2 parties. Ces deux parties sont «orthogonales» comme on peut le voir dans la figure et permettent de rendre compte de toutes les relations possibles entre a_1 et a_2 (c'est-à-dire leur écartement/rapprochement ou leur étalement/concentration). Les mêmes principes s'appliquent pour la deuxième équation 4.24.

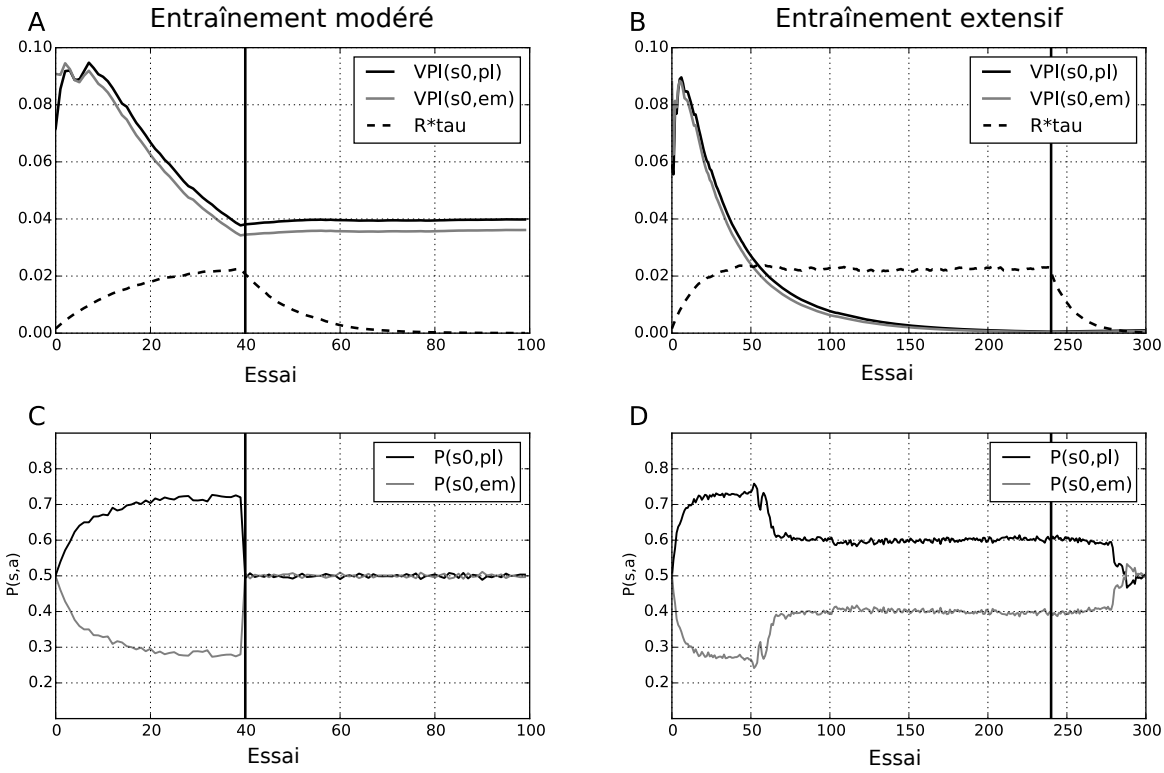

FIGURE 4.10 – **A.** Valeur d'information précise (VPI) pour l'action d'appuyer sur le levier (A0) et l'action d'entrer dans l'auge (A1) à l'état initial (S0) contre le taux récompense en apprentissage modéré. Les barres verticales représentent la dévaluation de la récompense. **B.** En apprentissage extensif. **C.** Probabilité d'action pour A0 et A1 dans l'état S0 en apprentissage modéré. **D.** En apprentissage extensif. Reproduit de VIEJO et collab. [2016] répliquant le travail de KERAMATI et collab. [2011] *(Code open-source disponible en téléchargement avec l'article sur le site web de ReScience).*

Du point de vue de l'agent, l'arbitrage entre le système d'apprentissage sur modèle et le système d'apprentissage d'habitudes se décrit ainsi :

— $\mathrm{VPI}(s_t,a_i) > \bar{\mathrm{R}}(t)$: l'action a_i est incertaine et le taux de récompense est bas. Si les récompenses arrivent lentement, il est préférable de perdre du temps dans le raffinage de la valeur de l'action en utilisant le modèle du monde pour être sûr d'obtenir une récompense

— $\text{VPI}(s_t, a_i) < \bar{\text{R}}(t)$: la valeur de l'action a_i est correctement encodée dans le système d'habitudes et les récompenses arrivent rapidement. Il est préférable de choisir rapidement une action en faisant confiance au comportement habituel.

Pour clarifier le processus, l'algorithme 4.1 résume les principales étapes de la coordination des systèmes de mémoire selon le compromis vitesse-précision. La version complète de l'algorithme du Kalman q-learning est détaillée dans l'annexe A.1.

Tout comme DAW et collab. [2005], les auteurs ont testé leur modèle sur une tâche classique de conditionnement instrumental que nous avons reproduit dans VIEJO et collab. [2016]. Dans cette tâche, un groupe de rats est entraîné à appuyer sur un levier pour ensuite consommer sa récompense dans une auge. Les résultats sont présentés dans la figure 4.10 pour un entraînement modéré (4.10.A et C) et un entraînement extensif (4.10.B et D). Le transfert de contrôle entre systèmes de mémoire s'illustre bien à l'essai 50 de la figure 4.10.B. En entraînement modéré, la dévaluation s'effectue avant le transfert de contrôle (avant que le modèle ait développé une habitude comportementale contrôlée par le q-learning) et le système d'apprentissage sur modèle continue de dominer la sélection de l'action. Cela se traduit par une probabilité d'action juste au regard de l'absence de récompense après la dévaluation (c'est-à-dire probabilité issue d'un mécanisme flexible qui s'est rapidement adapté à la dévaluation). Si le transfert de contrôle s'effectue (après un apprentissage plus long), le système d'apprentissage d'habitudes ne modifie pas immédiatement ces probabilités d'action. Comme observé dans la littérature expérimentale, l'habitude persiste après une dévaluation de la récompense. Néanmoins, nous observons que le Kalman q-learning continue d'apprendre et finalement retourne à des probabilités d'action exactes 50 essais après la dévaluation, ce qui n'est pas discuté par les auteurs.

C'est une prédiction intéressante des modèles de renforcement qui à notre connaissance est rarement discutée/évaluée dans la littérature : la prédiction est que même si l'apprentissage par différence temporelle est inflexible relativement à l'apprentissage sur modèle (et donc prédit une persistence du comportement même après dévaluation), si on laisse suffisamment de temps au modèle après la dévaluation (test en extinction suffisamment long (ici > 50 essais)), alors on devrait finir par observer un désapprentissage de l'habitude et donc une adaptation comportementale à la dévaluation. Or, la plupart des tests en extinction sont très courts et ne permettent donc pas d'observer ce phénomène.

Pour finir, le compromis entre vitesse et précision a aussi été exploré dans PEZZULO et collab. [2013] dans une proposition de modèle très similaire à celle de KERAMATI et collab. [2011]. La variable VPI est remplacée par une valeur d'information qui est cette fois-ci comparée à un seuil fixe (et non à un taux de récompense) pour arbitrer entre l'exploitation d'une table de q-valeurs ou l'utilisation d'un graphe de planification. Contrairement à KERAMATI et collab. [2011], les auteurs proposent de faire varier la profondeur de la planification dans le graphe de transition en fonction de l'incertitude associée à la q-valeur d'une action. Les auteurs ont ensuite testé leur modèle sur différentes versions d'un labyrinthe en double T en faisant varier l'incertitude à chaque point de décision. Si la capacité adaptative du modèle est ainsi démontrée, les auteurs n'ont pas effectué de comparaison avec des données réelles au contraire des études précédentes.

4.2.3 Interaction avec la mémoire de travail

Dans l'introduction, le succès de l'identification de la différence temporelle au signal dopaminergique a été discuté ainsi que certaines de ses conséquences [MONTAGUE et collab., 1996; SCHULTZ et collab., 1997]. Parmi ces conséquences figure le développement d'une littérature dédiée au décodage à différents niveaux de l'activité cérébrale par des

variables issues du formalisme de l'apprentissage par renforcement. A ce niveau, l'importance majeure du striatum a été démontrée. On peut ainsi citer O'DOHERTY et collab. [2004] qui montre une identification possible du striatum dorsal et ventral à un système acteur-critique. Dans SAMEJIMA et collab. [2005], les auteurs enregistrent l'activité des neurones du striatum pendant une tâche de conditionnement instrumental et montrent que l'activité de certains neurones dans le striatum prédit la valeur de l'action future. Dans PESSIGLIONE et collab. [2006], les auteurs décodent la valeur de la différence temporelle d'un modèle d'apprentissage par renforcement à l'activité hémodynamique du striatum chez des sujets humains. En électrophysiologie chez le rat, les auteurs de KHAMASSI et collab. [2008] trouvent que des neurones du striatum ventral ont bien un profil d'activité anticipatrice de récompense compatible avec la partie critique du modèle acteur-critique.

Néanmoins, le décodage de variables constitutives d'un modèle d'apprentissage par renforcement ne s'est pas arrêté au striatum. On peut citer notamment JOCHAM et collab. [2011] qui décode la valeur d'une action dans le cortex préfrontal ventro-médian en neuroimagerie. Dans TANAKA et collab. [2004], c'est une constellation de régions dans le cortex préfrontal et le cortex pariétal qui s'activent durant l'apprentissage d'une tâche modélisable par un algorithme d'apprentissage par renforcement. Dans BROVELLI et collab. [2008], c'est un réseau fronto-pariétal et striato-dorsal qui s'active dans les premiers essais d'une tâche induisant le transfert d'un comportement lié à un but à un comportement habituel. Dans CAVANAGH et collab. [2010], l'erreur de prédiction est décodée dans le cortex préfrontal en électro-encéphalographie.

Dans le but d'aider à identifier dans l'activité corticale ce qui pouvait provenir du système d'apprentissage par renforcement du striatum, les auteurs de FRANK et collab. [2001] ont cherché à modéliser les interactions entre apprentissage par renforcement et mémoire de travail. Pour aller plus loin, COLLINS et FRANK [2012] ont suggéré que cette interaction pouvait passer inaperçue lors d'une analyse du comportement et de l'activité cérébrale uniquement dans les termes de l'apprentissage par renforcement.

En effet, ils suggèrent au contraire qu'il peut y avoir interaction entre plusieurs systèmes différents, la mémoire de travail pouvant être l'un des ingrédients importants des systèmes d'apprentissage sur modèle qui manipulent et trient un grand nombre d'informations lors de la recherche dans leur graphe.

Dans COLLINS et FRANK [2012], les auteurs ont donc étudié la contribution respective d'un système d'apprentissage par différence temporelle et d'une mémoire de travail formalisés dans des tâches impliquant différents degrés de mémoire de travail. Nous allons décrire leur modèle plus en détail car il fait partie de ceux que nous avons comparés au nôtre dans VIEJO et collab. [2015] et qui sera présenté dans le chapitre suivant.

Pour mesurer l'impact de la mémoire de travail durant une tâche de conditionnement instrumental chez l'humain, les auteurs de COLLINS et FRANK [2012] proposent d'augmenter progressivement le nombre des états (stimuli visuels) n_s (2 à 6) avec un nombre d'actions constant ($n_a = 3$). Les sujets doivent donc associer un stimulus présenté dans un ordre aléatoire avec une réponse unique. Les courbes d'apprentissage des sujets en fonction du nombre de stimuli sont présentés dans la figure 4.11.D. Typiquement, l'augmentation du nombre de stimuli pénalise les sujets et leurs performances diminuent. Néanmoins, tous les sujets apprennent puisque les performances moyennes augmentent dans tous les cas.

Pour le processus de modélisation, le postulat de l'horizon temporel infini dans le formalisme de l'apprentissage par renforcement n'est pas nécessaire ici. Seul le passé de l'agent est important pour apprendre la fonction de valeur. Le modèle d'apprentissage

par renforcement apprend donc des q-valeurs Q^{AR} sans facteur γ selon :

$$Q^{AR}(s_t, a_t) \leftarrow Q^{AR}(s_t, a_t) + \alpha(r_t - Q^{AR}(s_t, a_t)) \tag{4.25}$$

avec α le taux d'apprentissage. La mémoire de travail est modélisée comme une liste finie représentant les derniers évènements Q^{MT}. La qualité de l'information retenue décroît progressivement selon :

$$Q^{MT}(s_t, a_t) \leftarrow Q^{MT}(s_t, a_t) + \epsilon(\frac{1}{n_A} - Q^{MT}(s_t, a_t)) \tag{4.26}$$

avec ϵ un facteur de décroissance. Un soft-max convertit les q-valeurs en probabilités d'actions qui sont ensuite combinées selon un poids $w(t)$:

$$p(a|s_t) = (1 - w(t))p^{AR}(a|s_t) + w(t)p^{MT}(a|s_t) \tag{4.27}$$

Ce poids w est mis à jour à chaque essai par une moyenne pondérée par la probabilité que chaque modèle ait obtenu la récompense selon :

$$w(t+1) = \frac{p^{MT}(r_t|s_t, a_t)w(t)}{p^{MT}(r_t|s_t, a_t)w(t) + p^{AR}(r_t|s_t, a_t)(1 - w(t))} \tag{4.28}$$

Un poids proche de 0 indique une préférence pour les q-valeurs issues de l'apprentissage par renforcement et un poids proche de 1 indique une préférence pour la mémoire de travail. Pour finir, la probabilité $p^{MT}(r_t|s_t, a_t)$ que la mémoire de travail ait obtenu la récompense est calculée selon :

$$p^{MT}(r_t|s_t, a_t) = min(1, \frac{C}{n_S}) \times \left\{ \begin{array}{c} Q^{MT}(s_t, a_t) \text{ si } r_t = 1. \\ 1 - Q^{MT}(s_t, a_t) \text{ si } r_t = 0. \end{array} \right\} + (1 - min(1, \frac{C}{n_S}))\frac{1}{n_A} \tag{4.29}$$

Dans ce modèle, les auteurs considèrent qu'il existe un seuil C de capacité de la mémoire de travail au-delà duquel les choix proposés par ce système sont de nature probabiliste. Si la capacité est dépassée ($C/n_S < 1$), la mémoire de travail a moins de chance de prédire la récompense et $p^{MT}(r_t|s_t, a_t)$ est diminué en conséquence (sauf dans le cas inverse ou $r_t = 0$). La probabilité d'obtention de la récompense du modèle d'apprentissage par renforcement est proportionnelle aux q-valeurs :

$$p^{AR}(r_t|s_t, a_t) = \left\{ \begin{array}{c} Q^{AR}(s_t, a_t) \text{ si } r_t = 1. \\ 1 - Q^{aR}(s_t, a_t) \text{ si } r_t = 0. \end{array} \right. \tag{4.30}$$

Pour capturer le comportement des sujets, les auteurs ont simulé le système d'apprentissage par renforcement seul (fig 4.11.A), le système de mémoire de travail seul (fig 4.11.B) et les deux systèmes combinés (fig 4.11.C). En optimisant les paramètres pour approximer au mieux les choix des sujets, l'avantage d'une combinaison apparaît ainsi clairement pour approximer les variations des performances des sujets en fonction du nombre d'états.

Pour conclure, ces auteurs ont montré que des fonctions cognitives de haut niveau sont impliquées dans des tâches facilement modélisables par des algorithmes traditionnels d'apprentissage par renforcement. Néanmoins, il est possible de recenser plusieurs aspects de la mémoire de travail qui n'ont pas été abordés dans COLLINS et FRANK [2012]. En effet, le système formel de mémoire de travail propose à chaque instant une quantité d'information constante ce qui semble peu en accord avec la flexibilité associée à ce type de mémoire. De plus, les auteurs ne se basent que sur les choix des sujets pour optimiser les paramètres. Si les sujets utilisent une mémoire de travail, les temps de réaction doivent potentiellement augmenter, ce qui devrait être explicable par le modèle. Cette double observation (choix et temps de réaction) fournit une contrainte de modélisation supplémentaire que nous utiliserons au chapitre suivant.

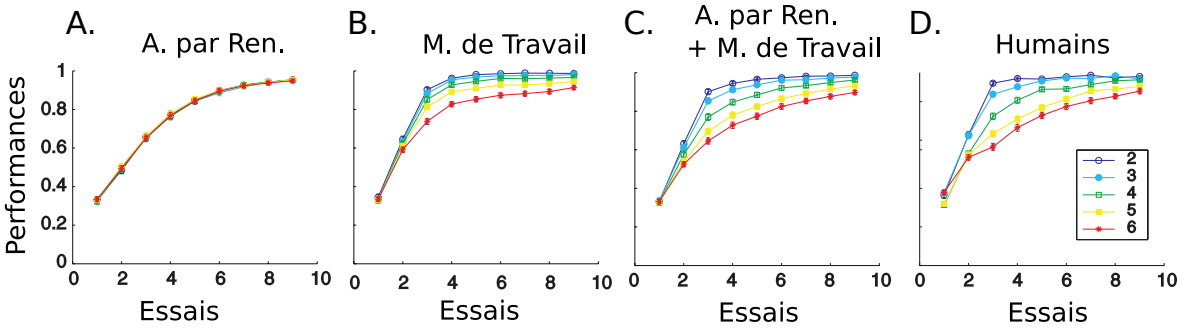

FIGURE 4.11 – A. Evolution de la performance du modèle d'apprentissage par renforcement en fonction du nombre de stimuli. B. Pour le modèle de mémoire de travail. C. Pour la combinaison du modèle d'apprentissage par renforcement et du modèle de mémoire de travail. D. Chez les sujets humains. Adapté de COLLINS et FRANK [2012].

4.3 En robotique

Peu discutée jusqu'à présent, la flexibilité comportementale qu'offre la parallélisation des systèmes de mémoire s'est révélée une source d'inspiration pour le développement de la robotique autonome. En effet, la capacité de changer de stratégie comportementale offre un degré de contrôle et un potentiel adaptatif non négligeable à un robot.

Nous discuterons brièvement ces travaux ici, non pas pour leur intérêt pour la robotique, mais parce que le test de modèles neuro-inspirés de coordination de systèmes de mémoire sur des robots permet en retour de donner une meilleure compréhension de quels principes de coordination fonctionnent ou ne fonctionnent pas dans telle ou telle situation expérimentale.

Dans GIRARD et collab. [2005], les auteurs ont implémenté une stratégie de lieux et une stratégie de réponse dans une simulation du robot rat Psikharpax (figure 4.12) qu'ils ont ensuite testé dans différentes configurations du labyrinthe en T. Dans ce modèle, la combinaison des systèmes ne passe pas par une sélection (un système est sélectionné au détriment de l'autre à chaque instant) mais par une fusion : les q-valeurs de la stratégie de réponse et de la stratégie de lieux sont sommées dans un modèle du striatum ventral selon une certaine pondération fixée par le modélisateur (et qui donc n'évolue pas au cours de l'apprentissage). Cette idée de fusion est similaire à la somme pondérée de q-valeurs de COLLINS et FRANK [2012] que nous avons exposée dans la section précédente. Dans une première partie, les auteurs montrent que les deux stratégies ensemble sont supérieures à une seule stratégie. Dans une deuxième partie, les deux stratégies sont testées en situation de conflit et la résolution du conflit se fait en changeant manuellement la pondération de la somme des q-valeurs. Ce modèle de coordination est très similaire à GUAZZELLI et collab. [1998] à l'exception près que les auteurs ont effectué une fusion pondérée des stratégies.

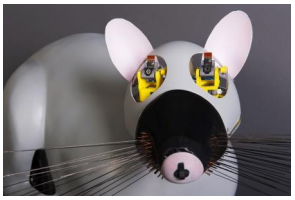

FIGURE 4.12 – Le robot-rat Psikharpax utilisé dans CALUWAERTS et collab. [2012]

Le modèle de DOLLÉ et collab. [2010] a été appliqué dans CALUWAERTS et collab. [2012] avec la seconde version du robot rat Psikharpax. Le modèle a été largement augmenté au niveau de la construction des états par les capteurs du robot mais le processus de sélection de stratégie reste le même. Toutefois, les auteurs ont proposé de connecter uniquement les cellules de lieux au système de sélection de stratégie (une cellule de lieux indique

quelle stratégie est préférée) et non l'entrée visuelle directe. En effet, certaines hypothèses propres aux modèles de navigation ne se retrouvent pas forcément en robotique notamment :

— le modèle a un accès parfait à sa position et son orientation

— la perception visuelle est parfaite, permettant à l'agent de distinguer sans erreurs les différents indices et leurs propagations dans le module approprié.

— le modèle est un point virtuel sans corps entier et sans difficultés possibles de mouvement.

Pour offrir un niveau de contrôle supérieur au robot, les auteurs ont ajouté un modèle simple de changement de contexte qui consiste à mémoriser les valeurs du système d'arbitrage à chaque fois que la position de la récompense est changée. Quand la récompense est remise à la même position, les valeurs d'arbitrage sont retrouvées par reconnaissance du contexte précédemment rencontré et par récupération directe des q-valeurs associées, évitant ainsi au modèle de réapprendre la coordination de stratégie. Les contextes dans ce modèle robotique sont très proches de la notion de «task-set» utilisée en neurosciences cognitives chez l'homme [COLLINS et KOECHLIN, 2012] avec l'idée qu'on doit parfois apprendre de nouveaux contextes, parfois reconnaître des contextes précédents, parfois explorer par une sorte de contexte par défaut (le système d'exploration du modèle). Il serait à ce titre intéressant de tester ce modèle robotique sur des données primates et de le comparer avec des modèles existants. Au final, le robot est capable d'associer chaque stratégie en fonction de la position dans le labyrinthe. Les auteurs montrent ainsi les avantages que procure la coordination de stratégie dans une tâche simple de navigation en robotique réelle.

Cette idée de contexte a aussi été reprise dans KHAMASSI et collab. [2011] pour une tâche d'apprentissage par renforcement spécialement adaptée à un robot humanoïde. Formalisé sous la forme d'un réseau de neurones, le cœur du modèle est constitué uniquement d'apprentissage par différence temporelle. Néanmoins, les auteurs font une proposition intéressante pour ce chapitre puisqu'ils proposent d'ajouter un système de méta-apprentissage qui permet d'ajuster dynamiquement le paramètre β réglant le compromis exploitation-exploration. Un ensemble de neurones «catégorisateurs» se modifient en fonction de la valeur de différence temporelle δ. Ces neurones influencent ensuite le paramètre β et peuvent être perçu comme une mémoire contextuelle.

Pour finir, l'approche de DAW et collab. [2005] et leur proposition de formalisation du comportement lié à un but et du comportement habituel a été reprise dans RENAUDO et collab. [2015] sur une tâche robotique de poussage de blocs sur un tapis roulant. Posé devant ce tapis, le robot reçoit une récompense en poussant un bloc du tapis. A un certain pas de temps, la vitesse du tapis change demandant un effort d'adaptation de la part du robot. Très brièvement, les auteurs ont testé des stratégies de sélection similaires à DAW et collab. [2005] et KERAMATI et collab. [2011] (basées sur de l'incertitude ou un taux de récompense) et des stratégies de combinaison des probabilités d'action de chaque expert (système de vote ou normalisation à l'intérieur d'un soft-max). Ces méthodes de coordination ont été comparées à une sélection de stratégie aléatoire. Néanmoins, les résultats ne sont pas en faveur des méthodes de coordination ou de combinaison. Les auteurs montrent que les critères de coordination issus des neurosciences, et considérant que l'apprentissage sur modèle est toujours le plus performant, bien que coûteux, ne sont pas toujours vrais en robotique et qu'il existe des tâches sur lesquelles un système d'apprentissage par différence temporelle ou une coordination aléatoire peuvent être plus avantageux.

Une observation importante à faire sur ce «passage au réel» est la difficulté que rencontre le modèle du comportement lié à un but. Formalisé sous la forme d'un graphe qui se construit au cours du temps, celui-ci devient beaucoup trop coûteux, lent et constamment incertain dans un monde continu où les états s'enchaînent. Dans les modèles de sélection de stratégie en conditionnement instrumental que nous avons présentés, l'hypothèse d'un monde synchrone, séquentiel et de petite dimension était toujours posée, même si elle l'était souvent de manière implicite. Pour contrer cette limitation, les auteurs proposent, entre autres, de ne pas replanifier à partir de zéro le modèle du monde pour chaque essai. Ceci constitue à nouveau une proposition qui mériterait d'être testée sur des données issues des neurosciences. Néanmoins, nous ne testerons pas cette prédiction dans ce manuscrit et nous contenterons de résumer les deux idées importantes que nous retenons de cette littérature robotique et qui participeront aux inspirations pour la conception du modèle présenté au chapitre suivant :

— le système d'apprentissage sur modèle nécessite beaucoup de temps de calcul avant de pouvoir prendre une décision. Il faut pouvoir évaluer dynamiquement dans les calculs de ce système si cela est toujours fructueux, ou alternativement quand (et dans quels états de la tâche) on peut s'en passer pour ne reposer que sur les décisions du système sans modèle.

— la décision finale de l'agent peut gagner en efficacité en reposant sur une fusion des propositions d'action faites par les deux systèmes plutôt que sur la sélection d'un seul système qui guide l'action à un moment donné.

4.4 Conclusion

Dans ce chapitre, nous avons décrit plusieurs modèles computationnels coordonnant des systèmes de mémoire. Dans la première partie du chapitre, des modèles de navigation sont présentés et ceux-ci distinguent la stratégie de lieux de la stratégie de réponse. Dans la seconde partie, des modèles du conditionnement instrumental distinguent le comportement lié à un but du comportement habituel (à l'exception de COLLINS et FRANK [2012] proposant une mémoire de travail différente de l'apprentissage sur modèle).

Dans KHAMASSI et HUMPHRIES [2012], les auteurs discutent des rapprochements possibles entre les deux littératures en comparant les études de lésion en navigation. Leur proposition de classification de modèles est synthétisée dans le tableau suivant :

		Sélection de l'action	
		Inflexible Lent à apprendre	Flexible Rapide à apprendre
Stratégie de navigation	Position	Association Lieux-action	Carte
	Indice	Habitudes	Planification

TABLEAU 4.1 – Taxonomie des stratégies de navigation basée sur la dichotomie entre l'apprentissage sur modèle et l'apprentissage par différence temporelle. Adapté de KHAMASSI et HUMPHRIES [2012].

La distinction entre une sélection flexible ou inflexible de l'action (ce qui correspond à la proposition de DAW et collab. [2005]) permet de clarifier les données des études de

lésion en navigation. Un modèle de décision en navigation ne sera pas défini par le type d'information qu'il utilise mais par la façon dont il l'utilise. Par exemple, cette distinction permet de comprendre que CHAVARRIAGA et collab. [2005] a utilisé un modèle d'habitudes et un modèle d'association lieux-actions ce qui, dans la proposition de DAW et collab. [2005], revient au même modèle de comportement habituel.

Nous pouvons distinguer trois méthodes de coordination en incluant les études de robotique :

1. la méthode de «porte» [CALUWAERTS et collab., 2012; CHAVARRIAGA et collab., 2005; DOLLÉ et collab., 2010]. La coordination est contrôlée par un modèle d'apprentissage par renforcement.
2. la méthode de sélection [DAW et collab., 2005; FOSTER et collab., 2000; KERAMATI et collab., 2011; RENAUDO et collab., 2015] qui est en soi un modèle avec ses propres hypothèses (par exemple la finalité du compromis vitesse-précision chez KERAMATI et collab. [2011])
3. la méthode de fusion [COLLINS et FRANK, 2012; GIRARD et collab., 2005; GUAZZELLI et collab., 1998].

Pour résumer, l'intérêt de ce chapitre est de pouvoir rapprocher les différentes méthodes de coordination de systèmes de mémoire issues de littératures diverses. Ces méthodes (plus précisément les méthodes 2 et 3) seront ainsi réutilisées au cours du chapitre suivant traitant de la coordination de la mémoire de travail et de l'apprentissage par renforcement. De plus, certaines des applications robotiques de ces modèles suggèrent des orientations pour combiner les critères de coordination de certains modèles tout en tirant profit de propriétés d'autres modèles (comme la fusion des sélections d'action de différents systèmes ou encore l'économie du calcul lors de la planification de l'action).

Chapitre 5

Mémoire de travail et apprentissage par renforcement

Sommaire

5.1 Introduction

Le contraste entre apprentissage par différence temporelle et apprentissage sur modèle est clairement devenu incontournable dans la littérature sur les systèmes de mémoire à partir de la fin des années 2000. Beaucoup de revues de cette littérature ont ainsi discuté et parfois assigné des rôles formels à la plupart des régions neuronales citées dans le chapitre 2 en fonction de la théorie de l'apprentissage par renforcement [Ashby et collab., 2010; Balleine et collab., 2007; Balleine et O'Doherty, 2010; Daw et O'Doherty, 2013; Doll et collab., 2015; Graybiel, 2008; Khamassi et Humphries, 2012; van der Meer et collab., 2012; Packard, 2009; Redish et collab., 2008; Yin et collab., 2008; Yin et Knowlton, 2006].

Une grande importance est ainsi accordée à la planification, c'est-à-dire regarder dans le futur de l'agent à travers son modèle du monde. Des raffinements ont été proposés dans plusieurs directions. Par exemple, dans Botvinick et Weinstein [2014], les auteurs discutent de modèles hiérarchiques qui consistent en fait en une concaténation des actions permettant une récursion «accélérée». Le but de l'apprentissage sur modèle est donc de garder en mémoire des informations sur cette récursion accélérée de manière à choisir la bonne concaténation d'actions lorsque le temps «ralentit» (c'est-à-dire les pas de temps où l'agent a la possibilité de décider). De même, des propositions de décodage du modèle permettant la planification de l'action ont été étudiées. Dans Gläscher et collab. [2010] et Lee et collab. [2014], les auteurs ont proposé un décodage de l'erreur de prédiction d'état (en écho à l'erreur de prédiction de récompense associée à la dopamine) dans le cortex préfrontal latéral, antérieur et le sillon intrapariétal en imagerie cérébrale.

Néanmoins, l'évaluation d'une action peut aussi nécessiter de regarder dans le passé de l'agent et ceci, grâce à un formalisme plus élaboré que le simple fait d'engranger des valeurs dans un tableau comme le font les algorithmes d'apprentissage par différence temporelle. Deux formes de mémoire sont ainsi concernées : la mémoire de travail et la mémoire épisodique. Ce chapitre concerne la mémoire de travail et la mémoire épisodique sera traitée dans le chapitre suivant.

Comme énoncé dans l'introduction du chapitre 2, la mémoire de travail est une théorie [Baddeley et Hitch, 1974]. Dans ses plus grandes lignes, la mémoire de travail se définit comme la rétention et la manipulation d'un nombre limité d'informations discrètes pendant un intervalle de temps fini. Cette finitude dans ses capacités de rétention, d'abord évaluées à 7 éléments [Miller, 1956] puis plus récemment à 4 [Conway et collab., 2001], est d'ailleurs ce qui caractérise le mieux la mémoire de travail. Toutefois, la littérature en psychologie distingue la limitation des capacités en termes d'éléments retenus de la limitation de capacités en termes de ressources consommées. Dans le premier cas, chaque élément occupe le même espace quelles que soient les caractériques qu'il possède. Dans le deuxième cas, les caractéristiques d'un élément ont des conséquences sur les capacités de rétention (un élément complexe consomme plus d'espace qu'un élément simple) [Bays et Husain, 2008]. Dans ce chapitre, nous modéliserons la mémoire de travail selon la première vision et chaque élément occupera le même espace formel.

La mémoire de travail est-elle un système de mémoire selon les critères du chapitre 2 ? Si le deuxième critère d'un ensemble de propriétés spécifiques est assez respecté (capacités limitées, information discrète, etc), le premier critère sur la dissociation et la spécificité à un substrat neuronal n'offre aucune consistence au regard des données empiriques. Jusqu'à très récemment, il était admis que le cortex préfrontal sous-tendait la mémoire de travail [Goldman-Rakic, 1995; Miller et Cohen, 2001]. Cette proposition a été rendue possible grâce à l'observation en imagerie d'une augmentation d'activité dans le cortex

préfrontal durant l'intervalle de rétention d'une tâche mesurant la mémoire de travail [LEVY et GOLDMAN-RAKIC, 2000]. Cette observation a aussi été faite chez les singes par l'enregistrement des neurones du cortex préfrontal dorso-latéral [PROCYK et GOLDMAN-RAKIC, 2006] dans une expérience que nous modéliserons en deuxième partie de chapitre.

Néanmoins, une autre vision a progressivement émergé pour tenter de réconcilier de nouvelles études empiriques (voir LARA et WALLIS [2015] pour une revue récente). Très brièvement, le contenu d'un élément en mémoire de travail est maintenu par les mêmes régions sensorielles (ou neurones) qui sont activées lorsque ce contenu est présent dans l'environnement. Ainsi, le rôle du cortex préfrontal n'est pas de garder l'information en mémoire, mais plutôt de garder l'attention et le contrôle sur ces éléments actifs dans les régions sensorielles [POSTLE, 2006]. De plus, le cortex préfrontal a été associé à beaucoup d'autres fonctions exécutives de haut niveau notamment le suivi et le changement de contexte [KOECHLIN et HYAFIL, 2007]. En définitive, il est évidemment plus facile d'allonger une liste de critères proposée dans les années 1990 que de réfuter point par point une telle quantité de données empiriques. Ce quatrième critère sera donc une exception : certains systèmes de mémoire sont sous-tendus par une interaction entre des structures neuronales et la mémoire de travail en constitue le principal exemple.

Un autre aspect de la mémoire de travail est sa relation avec la dopamine. Le cortex préfrontal contient une large quantité de récepteurs dopaminergiques et est sous influence du système dopaminergique contenu dans l'aire tegmentale ventrale [GOLDMAN-RAKIC, 1995]. L'influence de la dopamine sur la mémoire de travail a été démontrée pour la première fois dans BROZOSKI et collab. [1979]. Une diminution de la dopamine dans le cortex préfrontal chez le singe induisait une baisse de performances dans une tâche mesurant la mémoire de travail. Chez l'humain, une diminution de dopamine chez des patients parkinsoniens induit des effets d'hésitation et de ralentissement lors des processus de décision hautement incertains d'une tâche impliquant la mémoire de travail [PESSIGLIONE et collab., 2005]. Cet handicap est corrigé par l'administration d'un agoniste dopaminergique.

Toutefois, il semblerait que la relation entre dopamine et mémoire de travail soit infiniment plus complexe [COOLS et D'ESPOSITO, 2011]. Une injection d'agoniste dopaminergique pour un faible niveau naturel de dopamine améliorerait les performances et une injection d'agoniste pour un fort niveau naturel de dopamine diminuerait les performances chez le rat [FLORESCO et PHILLIPS, 2001]. Le même effet a été demontré chez le singe et l'humain (COOLS et D'ESPOSITO [2011] pour une revue récente).

Pour revenir à un sujet plus mathématique, les effets physiologiques de la dopamine sur la circuiterie du cortex préfrontal ont inspiré des modèles connectionnistes. Dans FRANK et collab. [2001], les auteurs ont proposé un modèle intégré du cortex préfrontal et du striatum pour expliquer le rôle de la dopamine dans des tâches de contrôle cognitif de haut niveau. Le rôle du striatum est celui d'une «porte» qui permet de contrôler grâce à la dopamine le maintien et la mise à jour des informations contenues dans le cortex préfrontal. Les auteurs soulignent l'importance de l'apprentissage par renforcement mais ne proposent pas d'implémentation. Dans ROUGIER et collab. [2005], le même modèle du striatum est remplacé par un modèle acteur-critique et est utilisé pour implémenter la variation du poids des «portes» grâce à un apprentissage par différence temporelle. Les auteurs montrent ainsi qu'il est possible de contrôler la mise à jour des éléments en mémoire de travail grâce à la différence temporelle qui est associée au signal dopaminergique phasique [SCHULTZ et collab., 1997]. Dans TODD et collab. [2009], ce modèle de porte a aussi été formalisé à un niveau plus symbolique dans un acteur-critique standard. La mémoire de travail est modélisée discrètement et sert à augmenter l'espace des

états en fonction d'«actions-portes». Un état est une concaténation de l'état courant et des états contenus dans la mémoire de travail. A chaque instant, une action est choisie ainsi qu'une «action-porte» qui va mettre à jour la case en mémoire de travail correspondante. Le critique du modèle va ensuite évaluer l'ensemble de ces actions. En omettant certains détails mathématiques et en généralisant, ces modèles utilisent la mémoire de travail comme une augmentation de la dimension de l'espace état-actions qui est évalué ensuite par l'apprentissage par différence temporelle. Ces modèles proposent aussi un modèle d'interaction synergétique entre le cortex préfrontal et le striatum. De fait, il semble difficile de pouvoir considérer ces modèles comme une interaction entre systèmes de mémoire. Chaque module occupe une place spécifique et certains modules n'ont pas pour rôle de proposer une action comme les modèles présentés au chapitre 4.

Le but de ce chapitre va donc être de proposer un modèle de mémoire de travail, c'est-à-dire incluant à la fois la rétention d'informations et la sélection d'actions et ceci, séparé d'un autre système de mémoire mimant l'apprentissage d'habitudes (q-learning). Ce modèle de mémoire de travail (ou délibératif) occupe le même rôle que l'apprentissage sur modèle décrit dans les chapitres 3 et 4 et identifié comme le comportement lié à un but dans Daw et collab. [2005]. Cette similarité va donc nous permettre d'étudier la transition entre comportement lié à un but et comportement habituel en utilisant certains modèles d'arbitrage présentés dans le chapitre 4 [Collins et Frank, 2012; Keramati et collab., 2011] ou en proposant un nouveau modèle d'interaction.

Dans la première partie de ce chapitre, nous allons donc décrire un nouveau modèle de mémoire de travail bayésienne que nous avons confronté spécifiquement avec les résultats de Brovelli et collab. [2008, 2011]. Dans Brovelli et collab. [2011], l'analyse en imagerie cérébrale dans une tâche d'association visuo-motrice montre une activation en premier du putamen dans les premiers essais de la tâche suivie d'une activation du noyau caudé lorsque les sujets engagent de grandes ressources cognitives, suivie d'un transfert progressif vers le putamen à nouveau lorsque l'habitude se consolide. Ces résultats permettent aux auteurs de discuter d'une possible implication lors de la phase d'exploration du même substrat neuronal impliqué dans le comportement habituel. Etant donné ces résultats, nous allons donc supposer que la tâche d'apprentissage de Brovelli et collab. [2011] représente un exemple canonique du conditionnement instrumental. Cette tâche peut ainsi être utilisée pour étudier l'acquisition et la consolidation initiale d'une tâche instrumentale durant laquelle le comportement lié à un but et le comportement habituel interagissent. Une fois consolidées, les relations visuo-motrices forment ainsi un ensemble d'habitudes.

Une observation cruciale pour le propos de ce chapitre concerne les temps de réaction de Brovelli et collab. [2011]. L'interaction dynamique entre les deux stratégies, plutôt que leur recrutement séquentiel, se reflète dans l'évolution des temps de réaction durant l'apprentissage. Étant donné cet ensemble de résultats, le travail formel s'est concentré sur l'explication de l'ensemble des observations comportementales (choix et temps de réaction) et pas seulement les choix des sujets.

Dans la deuxième partie de ce chapitre, nous avons testé la coordination de mémoire de travail et d'apprentissage par renforcement sur une tâche d'association visuo-motrice chez le singe [Khamassi et collab., 2015; Procyk et collab., 2000; Quilodran et collab., 2008]. Cette tâche a été conçue pour enregister l'activité neuronale dans le cortex cingulaire antérieur et le cortex préfrontal latéral dans une tâche qui alterne exploration et exploitation. De fait, il semblerait que ces structures soient impliquées dans des processus de contrôle de haut niveau impliquant la régulation du compromis exploitation et exploration, la valeur de l'action et la différence temporelle. Pour être efficace, la phase

d'exploration nécessite l'usage de la mémoire de travail pour retenir les options déjà essayées sans succès.

Dans KHAMASSI et collab. [2015], les auteurs ont proposé différentes variations du modèle d'apprentissage par différence temporelle pour expliquer le comportement des sujets. En optimisant les choix des singes uniquement, c'est un q-learning additionné de plusieurs heuristiques qui capture le mieux le comportement. Parmi ces heuristiques figurent «l'oubli» des q-valeurs, l'improbabilité de tester une action récompensée dans la session précédente ou un doublement du paramètre β (contrôlant le compromis exploitation et exploration) entre les essais de recherche et les essais de répétition.

Regroupées sous le terme de méta-apprentissage, ces idées ont déjà été explorées dans KHAMASSI et collab. [2011] dans une tâche robotique. Néanmoins, les auteurs, à l'instar de COLLINS et FRANK [2012], discutent de la possibilité d'une implication de la mémoire de travail comme stratégie séparée permettant une flexibilité comportementale que ne permettent pas les modèles lents d'apprentissage par renforcement. C'est cette hypothèse que nous allons tester dans la deuxième partie de ce chapitre.

Les temps de réaction des singes montrent une évolution dans certains cas similaire à ceux des sujets de BROVELLI et collab. [2011] et dans d'autres cas une évolution inverse. En capturant les choix et les temps de réaction dans une tâche d'association visuo-motrice chez le singe, cela nous permet aussi de tester la transférabilité de notre modèle de mémoire de travail ainsi que des modèles de coordination de stratégies.

5.2 Chez l'homme

5.2.1 Apprentissage visuo-moteur

Le but de la tâche décrite dans BROVELLI et collab. [2008, 2011] est de faire apprendre par essai-erreur une association fixe entre un stimulus visuel (3 cercles de couleurs différentes) et une action d'un doigt de la main droite à un groupe de sujets humains. A chaque essai, un stimulus est présenté au sujet et celui-ci doit répondre en moins de 1.5 seconde par un appui sur une touche d'un clavier avec un seul doigt de la main droite. Après un délai variant entre 4 et 12 secondes, une image est présentée lui indiquant si la réponse était correcte ou incorrecte. L'ordre de présentation des stimuli est mélangé et le sujet connaît l'indépendance entre les stimuli (la réponse correcte pour un stimulus ne prédit aucunement la réponse correcte pour un autre stimulus). Chaque participant réalise la tâche 4 fois et une session dure environ 42 essais. Pour résoudre la tâche, les sujets doivent mémoriser les réponses précédentes pour éviter de répéter des erreurs. Une fois la réponse positive obtenue, l'association correcte correspondante est la seule information utile pour le reste de la session.

Pour obtenir des performances stables, reproductibles et équivalentes entre les sujets, la tâche est manipulée et les associations stimulus-action ne sont pas définies *a priori*. Au contraire, elle sont assignées pendant chaque session en fonction des actions du sujet.

Les trois stimuli seront appelés par la suite S1, S3 et S4 pour les raisons expliquées maintenant. La première présentation de chaque stimulus est toujours suivie par une réponse négative et ce, quel que soit le choix du sujet. A la seconde présentation du stimulus S1, toute action nouvelle (donc non-choisie à la première présentation) sera considérée comme correcte. Pour le second stimulus S3, l'action correcte est assignée une fois que le sujet a réalisé trois essais incorrects (avec trois actions différentes). Pour le stimulus S4, la bonne réponse est donnée avec 4 essais incorrects (avec 4 actions différentes). En d'autres termes, la bonne réponse est donnée lors de la seconde action possible pour S1,

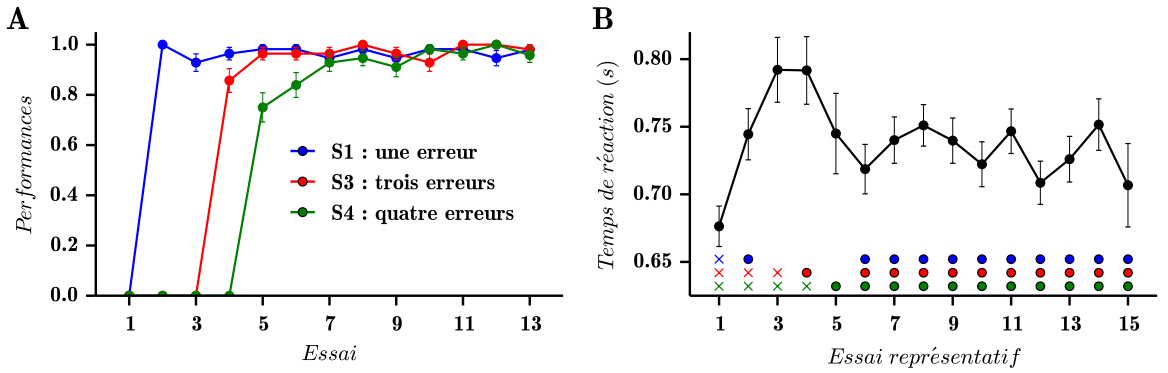

FIGURE 5.1 – A. Les trois courbes pour chaque stimulus représentent la probabilité de réponse correcte (Performances) en fonction du nombre de présentations du stimulus (et non en fonction de l'ordre de présentation durant la session) calculée selon le résultat (1 pour correct, 0 pour incorrect). B. Le temps de réaction moyen est calculé après une réorganisation des temps de réaction individuels selon des essais représentatifs [BROVELLI et collab., 2011]. Les rangées de symboles indiquent la position du temps de réaction en fonction du type de stimulus avant la réponse correcte (croix) et après la première réponse correcte (cercle).

la quatrième action possible pour S3 et la cinquième action possible pour S4. Cette manipulation permet ainsi de forcer les sujets à faire un nombre minimum constant d'erreurs durant la phase d'acquisition (1 pour S1, 3 pour S3, 4 pour S4). Cela permet aussi d'obtenir des performances stables entre les sujets comme le montre la variance faible de la figure 5.1.A de la performance moyenne.

Les temps de réaction (RT) sont mesurés dans l'intervalle entre la présentation du stimulus et l'appui sur une touche du clavier. Dans le but de visualiser l'évolution des temps de réaction (RTs) durant l'expérience, un processus de mise en ordre a été appliqué pour obtenir une valeur moyenne sur chaque essai représentatif. Cette méthode est reproduite de BROVELLI et collab. [2011]. Les cinq premiers essais représentatifs forment la phase d'acquisition pendant laquelle le sujet fait les erreurs requises. Les dix essais représentatifs suivants forment la phase de consolidation. Les valeurs utilisées pendant la phase d'acquisition pour chaque essai représentatif sont décrites comme suit :

1. moyenne des RTs de la première action incorrecte pour chaque stimulus
2. moyenne des RTs de la première action correcte pour S1 et de la seconde action incorrecte pour S3 et S4
3. moyenne des RTs de la troisième action incorrecte pour S3 et S4
4. moyenne des RTs de la première action correcte pour S3 et de la quatrième action incorrecte pour S4
5. moyenne des RTs de la première action correcte pour S4.

De l'essai 6 à l'essai 15, la moyenne des RTs est effectuée à partir de la seconde présentation du stimulus après que la bonne réponse a été trouvée. Pour aider à la compréhension du processus, la figure 5.1.B de l'évolution moyenne des RTs durant les essais représentatifs a été assortie d'indices du type d'essai utilisé pour chaque temps de réaction moyen. Comme le montre la figure, la phase d'acquisition se distingue par des temps de réaction qui augmentent de l'essai représentatif 1 à 3 puis diminuent de l'essai 4 à 5. Cette évolution non-linéaire illustre ainsi une augmentation de la charge cognitive lorsque les erreurs s'accumulent en début de session faisant ainsi ralentir les sujets dans leur choix de l'action. Si les actions correctes sont identifiées par le sujet, cette charge cognitive diminue et les sujets accélèrent leur prise de décision.

5.2.2 Modèles computationnels

Pour identifier les processus computationnels pouvant sous-tendre le comportement des sujets pendant la tâche, nous avons simulé plusieurs modèles tout en comparant leur capacité à approximer au mieux les observations comportementales faites à chaque essai. Tous les modèles sont construits sur l'hypothèse que le comportement des sujets repose sur un processus d'apprentissage par renforcement (AR), une mémoire de travail (MT) ou une combinaison des deux. Dans ce dernier cas, nous avons testé plusieurs principes de combinaison de stratégies dont certains ont été exposés précédemment.

Dans le but de reproduire le comportement des sujets, nous avons utilisé le formalisme du processus de décision Markovien à temps discrets selon :

— l'ensemble des états $\mathscr{S}$ = {Bleu, Rouge, Vert}

— l'ensemble des actions $\mathscr{A}$ = {Pouce, Index, Majeur, Annulaire, Auriculaire}

— la fonction de récompense $\mathscr{R}(s_t, a_t) \in \{0, 1\}$

— la fonction de transition est déterministe et consiste à présenter les états dans un ordre aléatoire par blocs de 3.

A chaque instant t, l'agent observe un état s_t et calcule la probabilité d'action $p(a_t|s_t)$ à partir de laquelle une action est échantillonnée. Ensuite, le modèle génératif est mis à jour selon la récompense r_t.

Stratégie habituelle

Nous avons choisi de modéliser le comportement habituel avec un q-learning [WATKINS et DAYAN, 1992], l'un des algorithmes standards de l'apprentissage par renforcement. L'algorithme d'apprentissage est considéré *model-free*. En effet, il apprend une table de valeurs étant capable, de manière très réactive, de sélectionner plusieurs actions dans plusieurs états du monde. Cela ne nécessite pas d'acquérir un modèle interne du monde qui aurait permis de calculer précisément la valeur d'une action dans un état donné. Les équations ont déjà été détaillées dans 3.3.3.

Pour utiliser le modèle de compromis vitesse-précision proposé dans KERAMATI et collab. [2011], nous avons utilisés le Kalman q-learning que nous avions déjà décrit dans la section 4.2.2 (voir A.1 pour une description plus détaillée).

Stratégie délibérative

Nous proposons un modèle de mémoire de travail bayésienne (MTB) pour expliquer le comportement lié à un but. Pour résumer, le principal attribut du modèle MTB est de stocker un nombre limité de descriptions d'essais passés. Au début de chaque session, le modèle est initialisé comme une liste vide et chaque élément est ajouté dans un ordre chronologique. Un élément t_i en «mémoire» (l'indice i indique la position dans la liste d'éléments) contient l'information d'un essai précédent particulier sous la forme d'un triptyque de fonctions de masses :

1. $p(a|t_i)$ est la probabilité d'avoir observé un certain état dans un essai passé
2. $p(a|s, t_i)$ est la probabilité d'avoir observé une action étant donné un état dans un essai passé
3. $p(r|a, s, t_i)$ est la probabilité d'avoir observé une récompense étant donné un état, une action dans un essai passé.

Un paramètre N contrôle le nombre maximal d'éléments pouvant être maintenus en mémoire. Les éléments en mémoire les plus anciens sont retirés de la liste quand le nombre d'éléments en mémoire dépasse la capacité N. Un élément en mémoire est ajouté après chaque essai selon les règles suivantes :

$$p(s|t_1) = \left\{ \begin{array}{ll} 1 & \text{si s est le dernier état} \\ 0 & \text{sinon} \end{array} \right\} \tag{5.1}$$

$$p(a|s,t_1) = \left\{ \begin{array}{ll} 1 & \text{si a est la dernière action} \\ 0 & \text{sinon} \end{array} \right\} \tag{5.2}$$

$$p(r=1|a,s,t_1) = \left\{ \begin{array}{ll} 1 & \text{si } r_t > 0 \\ 0 & \text{sinon} \end{array} \right\} \tag{5.3}$$

Après chaque essai, la liste est mise à jour en convoluant avec une distribution uniforme chaque élément $p(s|t_i), p(a|s,t_i), p(r|a,s,t_i)$ pour mimer l'affaiblissement mnésique selon :

$$p(..|..,t_i) = (1-\epsilon)p(..|..,t_i) + \epsilon\mathscr{U} \tag{5.4}$$

Plus l'élément en mémoire est ancien, plus ces fonctions de masses probabilistes approcheront d'une distribution uniforme engendrant une perte d'information. Le niveau d'uniformisation du contenu de la mémoire de travail est contrôlé par ϵ.

Lors d'une étape de décision, la probabilité d'action $p(a|s,t_{0\to i})$ est calculée itérativement en utilisant la règle de Bayes. Le terme $t_{0\to i}$ représente le nombre d'éléments en mémoire traités et peut changer d'essai en essai. Pour calculer la probabilité d'action, la première étape est de calculer la distribution de probabilité jointe :

$$p(s,a,r|t_{0\to i}) = p(s,a,r|t_{0\to i\text{-}1}) + p(s|t_i)p(a|s,t_i)p(r|a,s,t_i) \tag{5.5}$$

Nous faisons l'hypothèse d'indépendance entre des éléments de mémoire successifs permettant de les sommer. Le processus de décision au regard de la tâche commence avec :

$$p(a,r|s_t,t_{0\to i}) = \frac{p(s,a,r|t_{0\to i})}{p(s_t)} \tag{5.6}$$

Un état s_t est présenté à l'agent avec certitude. Ainsi, $p(s_t) = 1$ ce qui permet de calculer $p(a,r|s,t_{0\to i})$. En utilisant la règle bayésienne, cette probabilité est réduite à

$$p(a|r,s_t,t_{0\to i}) = \frac{p(a,r|s_t,t_{0\to i})}{\sum_a p(a,r|s_t,t_{0\to i})} \tag{5.7}$$

Dans cette tâche, il existe seulement deux résultats $r \in \{0,1\}$. Selon les règles de la tâche, seulement une action mène à une récompense positive et les actions associées avec une récompense négative doivent être évitées. Au début de la tâche, seules les récompenses négatives $r = 0$ ont été reçues et les actions non essayées doivent être favorisées. Si la seule action associée à la récompense positive a été observée, la probabilité de cette action aux essais suivants doit être maximale. Ce raisonnement est résumé dans l'équation suivante :

$$\mathrm{Q}(s_t,a) = \frac{p(a|r=1,s_t,t_{0\to i})}{p(a|r=0,s_t,t_{0\to i})} \tag{5.8}$$

Contrairement au système d'habitudes, nous n'avons pas utilisé de soft-max pour calculer $p(a|t_{0\to i})$ mais un simple calcul de normalisation qui permet d'éviter un paramètre β additionnel.

Processus de décision de la mémoire de travail bayésienne

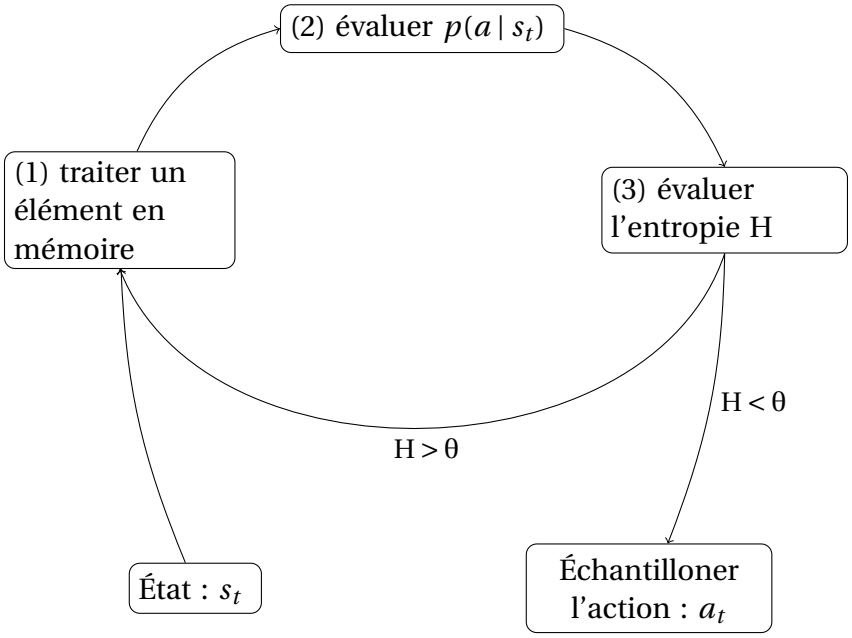

FIGURE 5.2 – Le processus de décision du modèle de mémoire de travail est décomposé en 3 étapes. (1) La première étape infère la loi jointe $p(s,a,r|t_{0\to i})$. (2) La deuxième étape évalue à partir de cette loi jointe les probabilités d'action. (3) La troisième étape calcule l'entropie H à partir de ces probabilités d'action. Le but de ce cycle est de réduire itérativement l'entropie H.

L'index $0 \to i$ représente le nombre d'éléments utilisés pour calculer $p(s,a,r)$ (si t_0, aucun élément de mémoire n'est traité et l'action est échantillonnée à partir d'une distribution uniforme). Dans le processus de décision, seul un sous-ensemble d'information disponible peut être pertinent pour le choix de l'action. Les états sont indépendants et tous les éléments ne doivent pas nécessairement être traités. Si la bonne action a été choisie au dernier essai, la décision peut se baser sur le premier élément dans la mémoire de travail (encodant le dernier essai) et les éléments sur les actions fausses n'ont pas besoin d'être utilisés. Au contraire, tous les éléments en mémoire à propos d'un certain stimulus doivent être traités quand l'agent est toujours en train de chercher la réponse correcte. Nous avons résolu ce problème en mesurant l'entropie de Shannon sur les probabilités d'action :

$$\mathrm{H} = -\sum_a \left(p(a|t_{0\to i}) \times \log_2 p(a|t_{0\to i}) \right) \tag{5.9}$$

Ainsi, la sélection de l'action est faite quand le niveau d'entropie H est inférieur à un certain seuil θ.

Pour comprendre la pertinence de l'entropie dans cette tâche, nous pouvons considérer $p(a|r=1,s_t)$ et $p(a|r=0,s_t)$ deux possibles éléments en mémoire encodant respectivement un essai réussi et un essai non réussi après avoir choisi une action j effectuée dans un essai précédent. Ils sont symétriques puisque $p(a=j|r=1,s_t) = p(a=j|r=0,s_t)$ (dans deux mondes parallèles, l'agent se rappelle que l'action j est bonne/mauvaise avec la même probabilité). Mais la division dans l'équation 5.8 n'est pas commutative. Il apparaît ainsi que $\mathrm{Q}(s_t,a=j)^{r=1} > \mathrm{Q}(s_t,a=j)^{r=0}$.

En normalisant les q-valeurs, $p(a)^{r=0}$ est proche d'une distribution uniforme moins une action ce qui implique une faible baisse d'entropie. Au contraire, $p(a)^{r=1}$ est proche d'une fonction de Dirac ce qui réduit drastiquement l'entropie quand l'information est

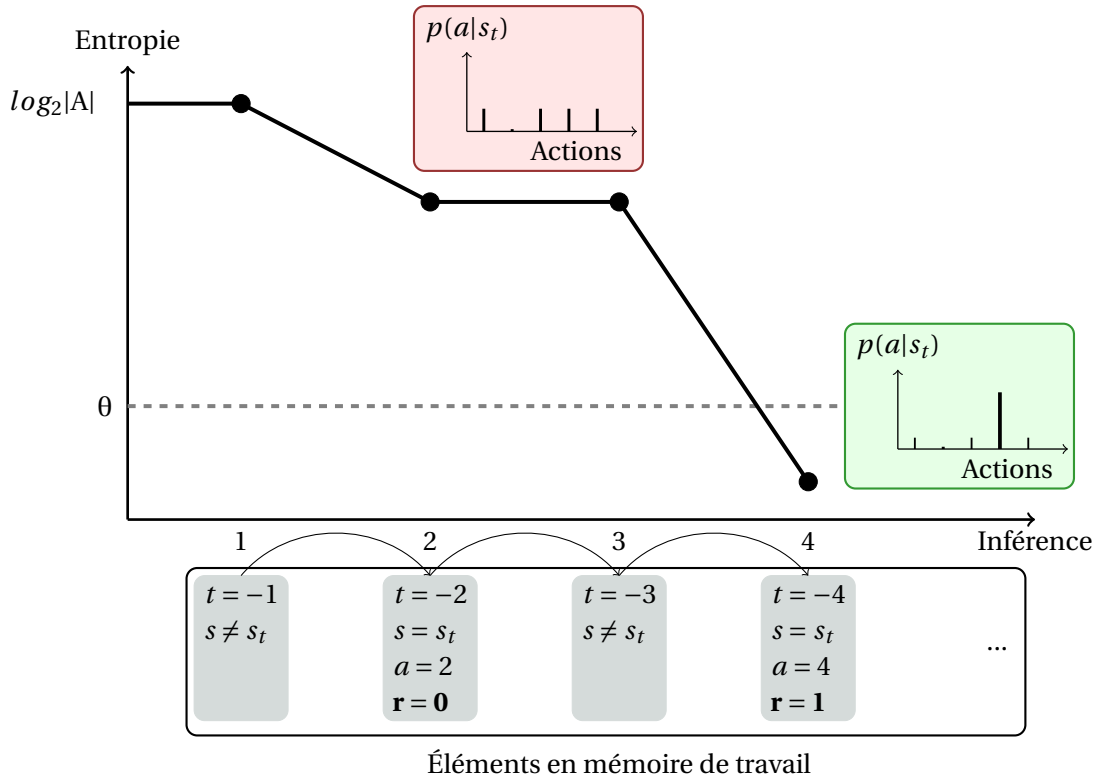

FIGURE 5.3 – Exemple théorique de l'évolution de l'entropie du modèle de mémoire de travail bayésienne durant un processus de décision. Dans cet exemple, l'agent observe un stimulus s_t et doit décider. Pour garder l'exemple simple, nous n'avons pas considéré la convolution avec une loi uniforme permettant de dégrader les éléments en mémoire. La rangée inférieure représente les éléments en mémoire rangés dans l'ordre chronologique de leur observation passée. Le premier élément (c'est-à-dire l'essai précédent) représente un stimulus différent. Lors de son traitement, aucune information n'est gagnée, la probabilité d'action reste uniforme et l'entropie $\mathrm{H}[p(a|s_t)]$ reste égale à l'entropie maximale $log_2|\mathrm{A}|$. Le second élément est un rappel négatif ($r = 0$) et son traitement modifie légèrement $p(a|s_t)$ puisque la probabilité de l'action 2 est largement diminuée. L'entropie H diminue en conséquence sans toutefois franchir le seuil requis θ. Le traitement du quatrième élément en mémoire permet d'inférer l'action correcte ($r = 1$). Les probabilités d'action sont modifiées pour favoriser uniquement cette action et l'entropie H chute. Le seuil θ est franchi ce qui permet à l'agent de décider.

rappelée par la mémoire de travail. Ainsi, le gain d'information est crucial puisque :

$$\mathrm{H}\big(p(a)^{r=1}\big) \ll \mathrm{H}\big(p(a)^{r=0}\big) \tag{5.10}$$

Une illustration du gain d'information dépendant des éléments en mémoire traités est donnée dans la figure 5.3.

Dans cette figure, l'entropie H est représentée selon le nombre d'inférences d'un ensemble spécifique d'éléments. La propriété de division des probabilités de notre modèle est spécifique à la tâche. Une seule action amène une récompense positive étant donné un état. Plus précisément, cette règle de minimisation de l'entropie déclenchant la décision est en miroir des règles de la tâche. Toutefois, cela n'empêche pas notre modèle de se généraliser à d'autres tâches possédant des temporalités d'obtention de récompenses différentes. Pour finir, le paramètre θ a besoin d'être estimé avec précaution puisqu'il peut forcer une décision à être prise sans une quantité d'information suffisante. Dans certains cas, aucun gain d'information ne peut se produire après l'évaluation des éléments en mémoire de travail et la décision est prise pour cause d'absence d'autres éléments.

Temps de réaction simulés

Dans l'introduction, nous avions mis en avant l'importance de l'évolution des temps de réaction et leur relation probable avec l'hypothèse de l'utilisation de systèmes de mémoire distincts. En construisant notre modèle de mémoire de travail, nous avons naturellement mis en parallèle les temps de réaction observés (supposés refléter la charge cognitive des sujets) et le nombre d'éléments en mémoire traités qui peut changer d'essai en essai en fonction du gain d'information.

En effet, le concept d'accumulation d'information a déjà été exploré dans différents modèles de course (*race models*) et ceux-ci ont été utilisés pour expliciter une grande quantité de données empiriques sur les temps de réaction [CARPENTER et collab., 2009; REDDI et CARPENTER, 2000]. Un travail particulièrement intéressant est la tentative de Norwich d'unifier les lois de perception qui permettent de prédire les temps de réaction [NORWICH, 2003]. Il a ainsi proposé que « as adaptation proceeds, entropy (potential information) falls, and information is gained ». A partir de cette hypothèse, un modèle descriptif très général des temps de réaction (RT) pour la détection de stimuli est construit. De manière simplifiée, la quantité d'information minimale nécessaire pour réagir peut être quantifiée comme une différence d'entropie $\triangle \mathrm{H} = \mathrm{H}(\mathrm{I}, t_0) - \mathrm{H}(\mathrm{I}, t_r)$ avec I l'intensité du stimulus et $t_r - t_0$ le temps de réaction. Pour réduire l'entropie, les récepteurs sensoriels doivent être échantillonnés n fois dans le but de gagner de l'information et ce taux d'échantillonnage détermine t_r.

Néanmoins, ces modèles se contentent d'une transformation monotone de l'entropie dans la distribution d'évidence en faveur de différentes options (au début l'entropie est maximale, puis elle décroît à mesure qu'on accumule de l'évidence pour telle ou telle option). Or dans les résultats comportementaux présentés dans la figure 5.1.B, les temps de réaction augmentent lorsque le sujet accumule de l'information, puis décroissent (évolution non monotone).

Nous allons donc ici restreindre au q-learning une relation monotone entre entropie et temps de réaction. Nous proposons que le temps de réaction simulé (sRT) du modèle est

$$s\mathrm{RT}(trial) = \mathrm{H}(p(a|s_t)) \tag{5.11}$$

De fait, $\mathrm{H}(p(a|s_t))$ décroît lentement avec les progrès de l'habituation. Nous supposons que cette variable peut être utilisée pour décoder le processus général d'habituation pendant une session.

Pour le système de mémoire de travail, nous allons supposer que le temps de réaction dépend (1) du nombre d'éléments encodés en mémoire de travail et (2) de la décision que fait le modèle de chercher (inférer) ces éléments en mémoire de travail ou pas. Dans ce cas, le temps de réaction simulé (sRT) du modèle est dépendant du logarithme du nombre d'éléments en mémoire traités i additionné à la valeur d'entropie calculée sur la probabilité d'action finale selon :

$$s\mathrm{RT}(trial) = (log_2(i+1))^{\sigma} + \mathrm{H}(p(a|s_t)) \tag{5.12}$$

σ est un paramètre libre contrôlant la proportion de la première partie de l'équation dans sRT. Dans le cas du modèle d'habitudes, $log_2(i+1)$ est égal à 0 et sRT est égal à l'entropie calculée sur les probabilités d'action.

5.2.3 Modèles de coordination

Jusqu'à présent, nous avons décrit un q-learning classique comme modèle d'apprentissage d'habitudes et un nouveau modèle de mémoire de travail bayésienne comme modèle d'apprentissage de comportement lié à un but. Les deux modèles décrits précédemment ont été proposés dans le but de tester l'hypothèse que ni un QL, ni un MTB seuls ne peuvent expliquer entièrement le comportement des sujets dans la tâche de BROVELLI et collab. [2008]. Nous allons tester l'hypothèse alternative que les choix et les temps de réaction des sujets s'expliquent mieux par une coordination des deux systèmes plutôt que par un système unique. Pour ce faire, nous allons comparer 3 critères différents de coordination de ces systèmes, deux existants [COLLINS et FRANK, 2012; KERAMATI et collab., 2011] et une nouvelle proposition de coordination.

Sélection sur VPI

Le premier modèle d'interaction est un processus de sélection directement adapté de [KERAMATI et collab., 2011] (cf 4.2.2). Pour rappel, la dépendance à la durée d'entraînement sur la sensibilité à la dévaluation d'une action est modélisée par un compromis entre la vitesse et la précision. La Valeur d'Information Précise (VPI) est proportionnelle à la mesure d'incertitude calculée sur les q-valeurs dans le système habituel qui décroît avec l'entraînement. Si une incertitude élevée est contenue dans les q-valeurs du système habituel, l'agent sera incité à exploiter au maximum sa mémoire de travail. Tout comme KERAMATI et collab. [2011], la VPI est évaluée pour chaque $Q(s_t, a)$ et comparée avec une moyenne glissante représentant le taux de récompense qui, pour rappel, est calculé selon :

$$\bar{R}(s_{t+1}) \leftarrow (1-\sigma_r)\bar{R}(s_t) + \sigma_r r_t, \tag{5.13}$$

Cette variable spécifie ainsi un coût pour les niveaux de cognition élevés. La règle qui détermine l'utilisation d'une stratégie est la même que dans KERAMATI et collab. [2011] à ceci près qu'une seule condition $VPI(s_t, a) > R(s_t)$ suffit à déclencher l'utilisation de la mémoire de travail utilisable pour toutes les actions.

Mélange pondéré

Le second modèle d'interaction que nous avons testé est le mélange pondéré adapté de COLLINS et FRANK [2012]. Décrite dans 4.2.3, cette interaction a été proposée spécifiquement pour une mémoire de travail et un q-learning. Malgré les différences entre leur modèle de mémoire de travail et MTB (COLLINS et FRANK [2012] n'utilisent pas de recherche adaptative dans la mémoire de travail), nous avons intégré la principale idée. La décision résulte d'une somme pondérée de chaque système :

$$p(a|s_t) = (1 - w(t, s_t))\, p(a|s_t)^{\mathrm{QL}} + w(t, s_t)\, p(a|s_t)^{\mathrm{MTB}} \tag{5.14}$$

Le processus de mise à jour des poids a déjà été détaillé dans la section 4.2.3 et ne change pas ici. Tout comme le modèle de sélection sur VPI, les systèmes sont séparés et produisent une probabilité d'action séparément.

Coordination par entropie

En complément de ces deux modèles adaptés d'études précédentes, nous proposons une troisième interaction appelée coordination par entropie. Ce modèle explore les possibilités d'une interaction rapprochée entre les modèles. Le premier point est de différencier les deux mesures d'entropie associées avec chaque stratégie individuelle.

H^{QL} est l'entropie d'information calculée selon les probabilités d'action du q-learning. Cette valeur décroît après chaque essai au fur et à mesure que l'apprentissage accroît progressivement les différences entre la valeur de la meilleure action et les autres valeurs. Cela fournit une information sur la progression dans la tâche d'apprentissage. A l'opposé, H^{MTB} est évaluée à l'intérieur du processus de décision de la mémoire de travail. Au début de chaque essai, H^{MTB} est égale à la valeur maximale d'entropie $H^{max} = \log_2(|Action|)$. Au fur et à mesure de la progression de la mémoire de travail, H^{MTB} va décroître (comme illustré dans la figure 5.3). Pour résumer, H^{QL} évolue entre les essais et H^{MTB} évolue à l'intérieur d'un essai.

Le second aspect est l'interaction dynamique entre les stratégies. A l'intérieur de la mémoire de travail, nous proposons de remplacer le choix déterministe entre décider et inférer des éléments en mémoire avec un choix probabiliste binaire. Au lieu de comparer H^{MTB} avec θ, une sous-action de l'ensemble $\{décider, inférer\}$ est échantillonnée avec les probabilités $p(décider|t_{0\to i}, H^{MTB}, H^{QL})$ et :

$$p(inférer|t_{0\to i}, H^{MTB}, H^{QL}) = 1 - p(décider|t_{0\to i}, H^{MTB}, H^{QL}) \tag{5.15}$$

Pour échantillonner l'une de ces deux sous-actions après chaque élément de mémoire traité, ces probabilités sont calculées avec l'équation logistique suivante :

$$p(décider|t_{0\to i}, H^{MTB}, H^{QL}) = \frac{1}{1 + \lambda_1(n-i)\exp^{-\lambda_2(2H^{max} - H^{MTB}_{0\to i} - H^{QL})}} \tag{5.16}$$

avec $n \leq N$ le nombre d'éléments contenus dans la mémoire de travail pendant un essai, i le nombre d'éléments déjà traités et λ_1, λ_2 des paramètres de gain. Si l'information sur les essais passés est contenue dans la mémoire de travail (n augmente), ces éléments doivent être traités jusqu'à ce que H^{MTB} soit suffisament basse. Ainsi, une augmentation de n favorise la sous-action *inférer* et non la sous-action *décider*. La variable n est indispensable dans un comportement lié à un but car nous voulons que l'agent exploite au maximum toutes les informations disponibles pour choisir la meilleure action. Néanmoins, la différence $n - i$ agit comme un coût dynamique à l'intérieur d'un essai qui augmente $p(décider)$ au fur et à mesure des tours dans l'inférence des éléments. De fait, la décision doit être prise dans un laps de temps fini. La probabilité $p(décider)$ est calculée à chaque fois qu'un élément est traité et permet de choisir entre chercher plus d'information avec la probabilité $1 - p(décider)$ ou s'engager dans le processus décisionnel.

Si la décision est engagée, les q-valeurs de chaque stratégie (QL et MTB) sont simplement sommées selon l'équation suivante :

$$Q(s_t, a) = Q(s_t, a)^{MTB}_{0\to i} + Q(s_t, a)^{QL} \tag{5.17}$$

Pour finir, les probabilités d'action sont calculées grâce à un soft-max avec un paramètre β_{final} différent du paramètre β utilisé pour normaliser les q-valeurs du q-learning. Au-delà de l'habituel compromis entre exploitation et exploration, le soft-max est important à cause de sa propriété de symétrie translationnelle que nous pouvons utiliser. De fait, les q-valeurs finales sont toujours la somme de deux systèmes. Au début de la tâche, la mémoire de travail peut avoir extrait beaucoup d'informations qui, grâce au soft-max, ne seront pas perturbées par l'ajout de q-valeurs uniformes d'un q-learning ignorant.

Pour comprendre le modèle de manière intuitive, il convient de discuter trois phases de la tâche instrumentale :

1. au premier essai, MTB est une liste vide ; QL et MTB fournissent des q-valeurs uniformes et $H^{QL} = H^{MTB} = H^{max}$. Ainsi, $p(décider) = 1$ et la décision est nécessairement prise. De fait, l'agent commence la tâche avec aucun indice.

2. durant la phase d'acquisition, H^{QL} est proche de H^{max} puisque le q-learning est un algorithme d'apprentissage lent. Après plusieurs essais entraînant des récompenses négatives, le nombre n d'éléments en mémoire augmente dans MTB. En combinant ces facteurs, $p(décider)$ est faible au début d'un essai et le processus d'inférence sera favorisé dans MTB au détriment d'une réponse automatique issue du QL.

3. la phase de consolidation démarre quand une action correcte a été trouvée. Dans ce cas, H^{QL} décroît graduellement tandis que H^{MTB} décroît en peu d'inférences. De fait, MTB ne contient quasiment plus que les essais corrects. L'asymétrie dans $Q(s_t, a)^{MTB}$ issue de la division entre les probabilités de se souvenir d'actions correctes et les probabilités de se souvenir d'actions incorrectes influence directement $p(décider)$. De fait, l'entropie H^{MTB} chute instantanément quand une action correcte est inférée. Néanmoins, les éléments encodant des réponses correctes ne sont pas la seule influence sur MTB. Le processus d'apprentissage de q-valeurs dans le QL fait décroître H^{QL}. En conséquence, $p(décider)$ est plus élevée au début de l'essai. Ainsi, la décision est prise plus rapidement durant les derniers essais de la tâche.

Pour résumer, la coordination par entropie équilibre le processus de mémoire de travail à partir de l'incertitude du système lié à un but et du système d'habitudes.

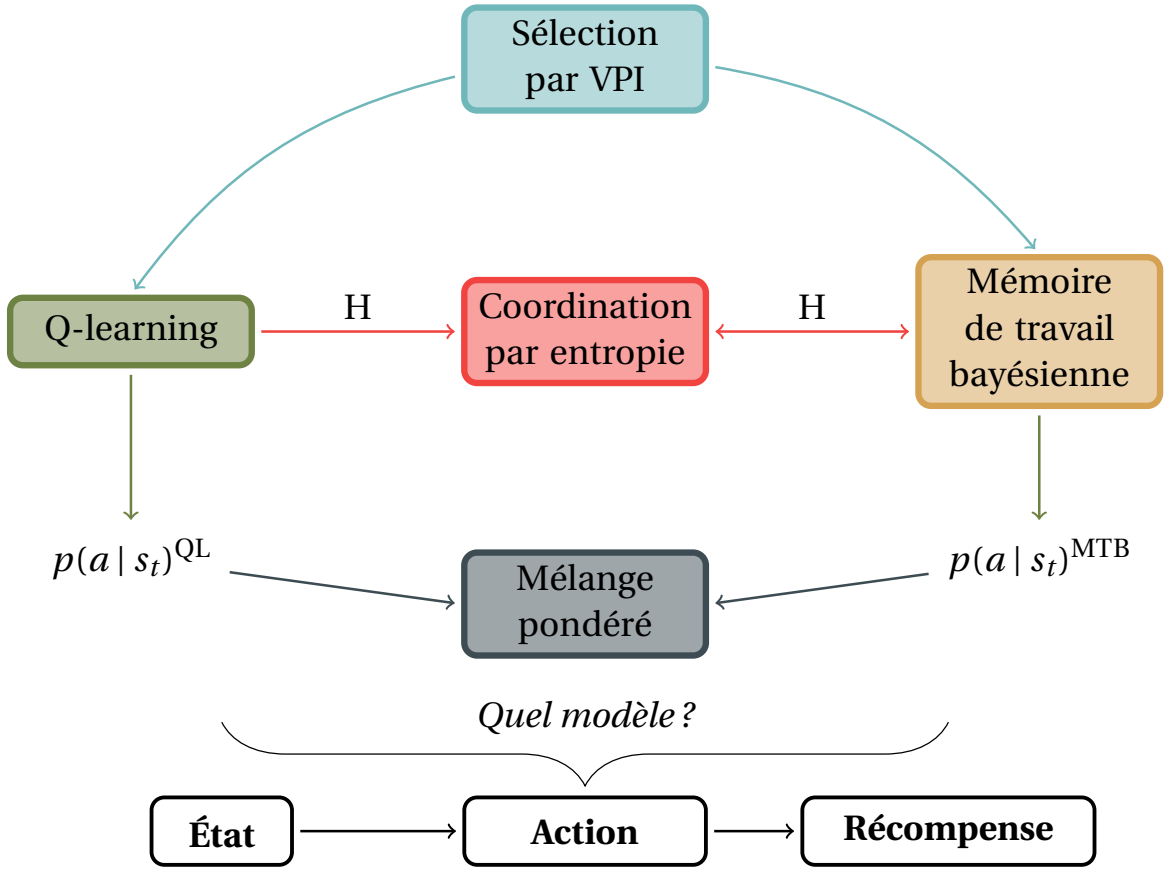

FIGURE 5.4 – Pour tous les essais, l'agent choisit une action grâce à un modèle possible (sur les cinq modèles possibles). Le comportement lié à un but est représenté par le modèle de mémoire de travail bayésienne (MTB) et le comportement habituel est représenté par le q-learning (QL). Entre les deux, les différents modèles pour l'interaction sont la sélection par VPI [KERAMATI et collab., 2011], la coordination par entropie et le mélange pondéré [COLLINS et FRANK, 2012].

5.2.4 Méthodes pour comparaison de modèles

Jusqu'à présent, nous avons présenté cinq modèles :

1. QL : seulement q-learning
2. MTB : seulement mémoire de travail bayésienne
3. Coordination entropique entre QL et MTB
4. Mélange pondéré entre QL et MTB
5. Sélection entre QL et MTB par VPI

Chaque modèle peut ainsi choisir une action et prédire un temps de réaction vRT suivant l'équation 5.12. Le meilleur modèle génératif est défini par sa capacité à répliquer les choix et les temps de réaction des sujets essai par essai. Cela correspond ainsi à deux objectifs à remplir à travers une optimisation des paramètres. A cause de ce doublement des objectifs, le problème de paramétrisation des modèles a été transposé dans un cadre d'optimisation multiobjective que nous avons choisi de résoudre en utilisant le programme SFERES [Mouret et Doncieux, 2010]. En utilisant l'algorithme d'évolution standard NSGA-2, un individu est défini comme un vecteur de paramètres θ_{model} pour chaque modèle. Le tableau 5.1 résume les paramètres libres de chaque modèle.

L'algorithme d'évolution consiste à démarrer la recherche du meilleur individu (vecteur de paramètres) à partir d'un groupe d'individus initialisé aléatoirement constituant ainsi la première génération. A chaque itération, le meilleur groupe d'individus est sélectionné de manière à converger vers le paramétrage optimal pour approximer les choix et les temps de réaction des sujets. L'algorithme NSGA-2 inclut aussi parmi ces objectifs une fonction mesurant la diversité parmi la population. Mouret et Doncieux [2010] ont ainsi montré que maximiser une fonction de diversité améliore la convergence de l'algorithme.

A chaque génération, les meilleurs individus sont sélectionnés en générant le modèle paramétré correspondant qui est ensuite évalué sur ces trois objectifs. Très brièvement, le premier objectif est de mesurer la vraisemblance que le modèle choisisse la même action que le sujet [Daw et collab., 2011]. Le second objectif est une erreur des moindres carrés entre la moyenne des RTs des sujets et la moyenne des vRTs du modèle. Ces moyennes sont calculées en fonction des essais représentatifs comme montré à la figure 5.1.C. Le troisième objectif est la mesure de diversité (distance euclidienne dans l'espace des paramètres) entre l'ensemble des paramètres évalués et les autres ensembles de paramètres dans la population.

L'optimisation des paramètres est faite indépendamment pour chaque sujet et chaque modèle. Après l'optimisation, l'algorithme d'évolution propose un ensemble de solutions P (c'est-à-dire des paramètres) qui maximise les objectifs f_i selon :

$$f_i : \mathrm{P} \rightarrow \mathbb{R}, i = 1, \ldots, n \tag{5.18}$$

On note $f = (f_1, \ldots, f_n)$ le vecteur des valeurs prises par chaque objectif. Étant donné deux solutions $a, b \in \mathrm{P}$, la solution a domine b si $f_i(a) \geq f_i(b), i = 1, \ldots, n$ et qu'il existe i tel que $f_i(a) > f_i(b)$. En d'autres termes, nous gardons les solutions qui sont strictement meilleures sur au moins un objectif et cet ensemble de solutions constitue le front de Pareto [Deb et collab., 2000].

En appliquant les règles de constitution d'un front de Pareto, plusieurs solutions peuvent exister et certains compromis doivent être effectués. Pour déterminer le meilleur modèle pour chaque sujet, la première étape est de combiner les fronts de Pareto des différents

modèles. La population des meilleurs individus de chaque modèle est mélangée pour recréer un nouveau front de Pareto des meilleurs individus pour chaque sujet.

A partir de ce nouveau front de Pareto mélangeant les modèles, la prochaine étape consiste en une agrégation des coordonnées numériques $\{x_1, x_2\}^{pareto}$ représentant la position d'une solution dans l'espace cartésien en une relation de préférences entre les solutions. Le but est bien évidemment de vérifier la qualité en choisissant une seule solution qui sera ensuite simulée. Dans la littérature correspondante sur les processus de décision sur front de Pareto, plusieurs fonctions d'agrégation ont été développées [EMROUZNEJAD et MARRA, 2014]. Néanmoins, le but de ce travail n'est pas de les comparer ni de les étudier. Le principal intérêt des fonctions d'agrégation est de prendre en compte mathématiquement le fait de perdre une certaine qualité d'optimisation sur un objectif tout en gagnant en qualité sur un autre objectif. Nous avons testé trois fonctions d'agrégation classiques : Tchebytchev, OWA et Distance. De fait, la plupart des solutions choisies par les fonctions d'agrégation se recouvrent. Les détails mathématiques sont donnés dans l'annexe A.2.

Par souci de vérification de notre méthode, nous avons réalisé deux versions de cette optimisation : une pour les choix seulement (en désactivant l'objectif des temps de réaction), l'autre pour les choix et temps de réaction.

Modèle	**Symbole**	**Limites**	**Description**
Q-L seul	α	$0 < \alpha < 1$	Taux d'apprentissage
	β	$0 < \beta < 100$	Température du soft-max
MTB seul	N	$1 < N < 10$	Taille de la mémoire
	θ	$0 < \theta < log\|A\|$	Seuil d'entropie fixe
	ϵ	$0 < \epsilon < 0.1$	Bruit
Sélection sur VPI	η	$0.00001 < \eta < 0.001$	Initialisation de la covariance
	σ_r	$0 < \sigma_r < 1$	Convergence du taux de récompense
Mélange pondéré	w_0	$0 < w_0 < 1$	Poids initial
Coordination sur Entropie	λ_1, λ_2	$0.00001 < \lambda_i < 1000$	Paramètres sigmoïdes
	β_{final}	$0 < \beta_{final} < 100$	Température du soft-max
	σ	$0 < \sigma < 20$	vRT

TABLEAU 5.1 – Tous les paramètres pour chaque stratégie sont aussi présents dans les modèles de coordination de stratégie à l'exception de θ qui disparaît dans le modèle de coordination sur Entropie et α qui disparaît dans le modèle de sélection par VPI.

5.2.5 Résultats

Optimisation des choix

Dans une première partie, nous allons considérer les résultats découlant de l'optimisation des choix seulement. Comme explicité précédemment, le processus d'optimisation est fait pour chaque sujet et chaque modèle. Cela retourne un ensemble de solu-

tions et parmi elles, la solution $\theta^{max}_{modèle}$ qui maximise la fonction de vraisemblance $\hat{L} = \sum_a P(a|modèle, \theta_{modèle})$.

Dans cette première approche directe, nous assignons à chaque sujet le meilleur modèle en comparant la vraisemblance brute. Le but ici est de vérifier si nous pouvons reproduire avec nos données l'observation de COLLINS et FRANK [2012] que le critère de sélection BIC sur-pénalise la complexité inhérente à tout modèle dual bien que celui-ci soit meilleur en termes de vraisemblance brute. Dans une deuxième étape, nous montrons la comparaison des modèles en fonction du critère de sélection BIC [1] qui permet de diminuer la vraisemblance d'un modèle en fonction du nombre de paramètres libres qu'il contient. Au final, nous obtenons le résultat suivant : le modèle de coordination sur Entropie est le meilleur pour 8 sujets et le modèle de mélange pondéré est le meilleur pour 6 sujets (voir 5.5.A).

Pour vérifier la capacité de chaque modèle gagnant à répliquer les résultats comportementaux, nous avons testé chaque jeu de paramètres. Chaque modèle paramétré fait ainsi ses propres choix (possiblement différents de ceux du sujet). Tout comme dans la figure 5.1, nous avons calculé la probabilité de réponses correctes (Perfomances) pour chaque stimulus moyenné sur les quatorze modèles paramétrés. La performance des modèles est montrée dans la figure 5.5.B au-dessus des performances humaines correspondantes.

Si la probabilité de réponse correcte est pratiquement indistinguable pour S1 et S3 (les courbes bleues et rouges), nous observons une grande différence pour S4 (la courbe verte). Pour comparer les courbes d'apprentissage entre les sujets et les modèles, nous avons calculé un test de Pearson χ_2 pour chaque stimulus et chaque essai sur les pourcentages de réponses correctes. Nous avons trouvé 9 essais significativement différents entre les sujets et les modèles. Comme cela peut être observé dans la figure 5.5.C, six essais discordants sont pour S4 : essai 5, 6, 7, 8, 10, 12. La plus grande différence concerne la cinquième présentation du stimulus S4 (test de χ_2, 1 dl, t = 15.52, P<0.001). Le modèle simulé fait légèrement plus d'erreurs de répétition que le sujet quand il cherche la réponse correcte. Ainsi, il n'atteint pas la performance moyenne des sujets durant la phase de consolidation.

Traditionnellement, le processus de sélection de modèle est réalisé en incluant un terme de pénalité pour la complexité du modèle [DAW, 2011; KHAMASSI et collab., 2015]. Comme énoncé précédemment, nous avons utilisé BIC [SCHWARZ, 1978] qui correspond à une approximation asymptotique de la probabilité postérieure. Le terme de pénalité est calculé selon le nombre de paramètres libres de chaque modèle. Dans notre étude, le modèle le plus simple est le q-learning (3 paramètres libres) et le plus complexe est le modèle de coordination par entropie (7 paramètres libres) (voir Table 5.1). En appliquant le critère BIC, nous obtenons ainsi des résultats drastiquement différents puisque les modèles les plus simples sont favorisés. Ainsi, le q-learning est assigné à 9 sujets et le modèle de mémoire de travail est assigné à 4 sujets. Seul le modèle de coordination par entropie survit au processus de pénalisation en étant assigné à un sujet comme l'indique la figure 5.5.B. Ainsi, comme dans l'étude de COLLINS et FRANK [2012] étudiant la contribution relative de la mémoire de travail et de l'apprentissage par renforcement, l'utilisation d'un terme de pénalité pour la complexité favorise les modèles à stratégie unique sur les modèles à double stratégie.

Le résultat d'une simulation des modèles sélectionnés selon le critère BIC est montré dans la figure 5.5.D. Au niveau statistique, seul l'essai 5 est significativement différent (test du χ_2, 1 dl, t=8.44, p<0.01).

1. BIC : *Bayesian Information Criterion*

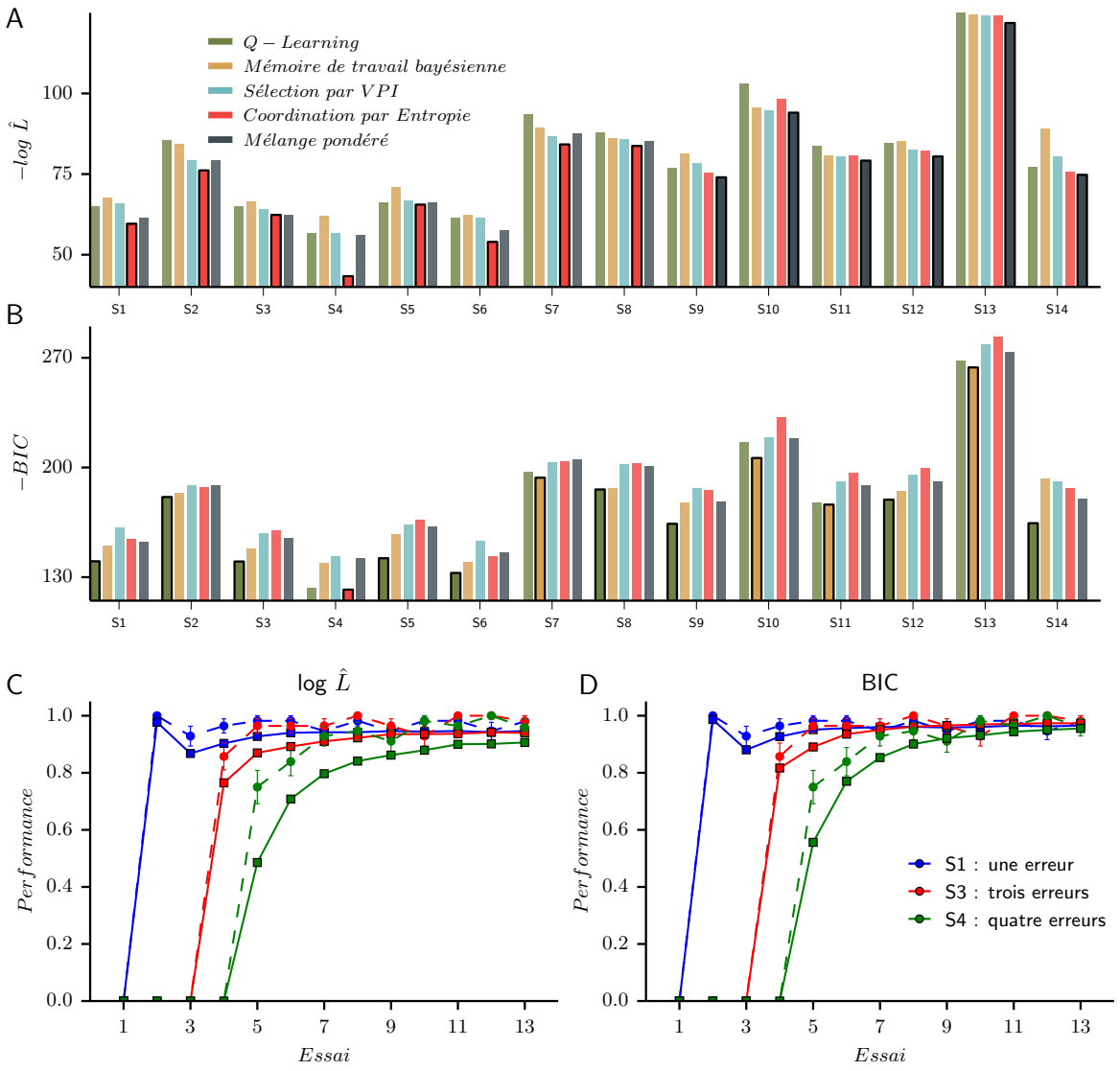

FIGURE 5.5 – A. Sans correction de la fonction de vraisemblance, l'ensemble des meilleurs modèles est composé de 8 modèles de coordination par entropie et de 6 modèles de mélange pondéré. L'histogramme de chaque meilleur modèle est cerclé d'une ligne noire pour chaque sujet. B. En appliquant la correction BIC, 9 modèles de q-learning, 4 modèles de mémoire de travail bayésienne et 1 modèle de coordination par entropie sont sélectionnés comme étant les meilleurs modèles. C. Simulation du groupe des meilleurs modèles selon la vraisemblance brute. D. Selon la correction BIC.

Dans la dernière phase d'analyse, nous allons voir que l'optimisation des choix et des temps de réaction va considérablement réduire la capacité des modèles uniques à correspondre aux résultats comportementaux des sujets.

Les fronts de Pareto pour les choix et les temps de réaction

Une nouvelle optimisation a été réalisée pour chaque modèle en incluant les deux objectifs : choix et temps de réaction. Les fronts de Pareto pour chaque modèle pour un sujet sont représentés dans la figure 5.6. Chaque point sur le front de Pareto représente une possible paramétrisation d'un modèle et d'un sujet. Cet ensemble de solutions «domine» au sens de Pareto toutes les solutions sous-optimales et n'est dominé par aucun autre ensemble. Les solutions pour le même modèle sont représentées dans la même couleur et connectées par une ligne. Le fait de maximiser à la fois les choix et les temps de réaction peut graphiquement être interprété comme une population de solutions qui forme un front de Pareto se déplaçant vers la solution optimale située dans le coin droit supérieur de la figure 5.6.A.

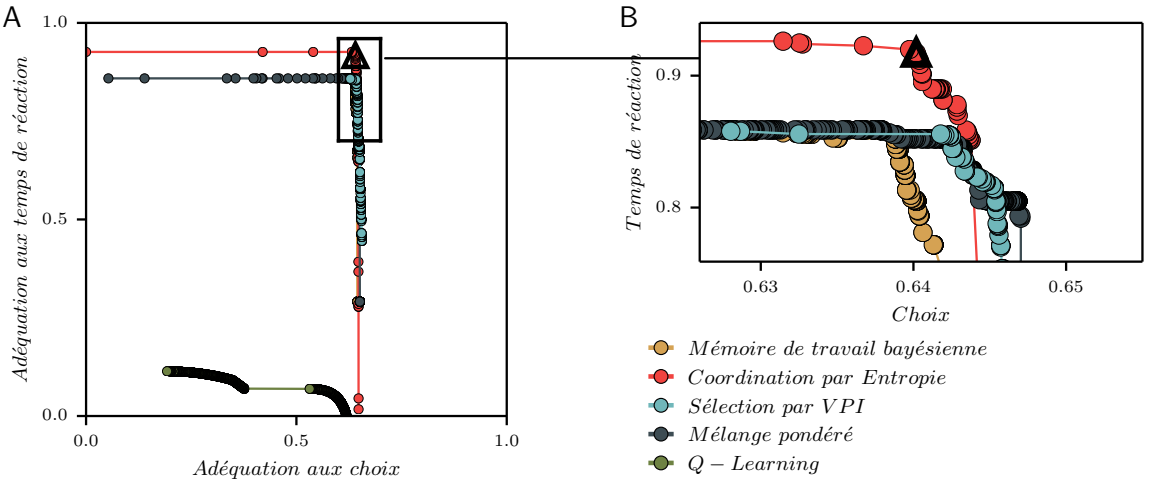

FIGURE 5.6 – A. A la fin du processus d'optimisation, le front de Pareto pour un sujet est construit en conservant des solutions qui ne peuvent être améliorées dans une dimension sans se dégrader dans une autre dimension et ce, quel que soit le modèle. Les meilleures solutions sont situées dans le coin supérieur droit de chaque sous-figure. B. Le coin supérieur droit des fronts de Pareto est élargi pour montrer la position relative de chaque modèle. Le triangle noir est superposé à la solution sélectionnée par une fonction de compromis entre la qualité d'adéquation aux choix et la qualité de l'adéquation aux temps de réaction. Pour ce sujet, la meilleure solution est donnée par le modèle de coordination par entropie. Celui-ci augmente largement la qualité d'adéquation au temps de réaction comparé aux autres modèles. Ce processus de sélection est réalisé pour chaque sujet.

A partir de la taille et de la position des fronts de Pareto, nous pouvons déjà observer une diversité de solutions. Par exemple, le q-learning montre une très mauvaise adéquation des temps de réaction comparé aux autres modèles. Dans la figure 5.6.B, un zoom est fait sur le coin du front de Pareto. Démarrant à partir de la meilleure solution située sur l'axe x du repère cartésien, les trois modèles duaux sont pratiquement équivalents. En escaladant le front de Pareto, le niveau d'adéquation aux choix des sujets diminue tout en améliorant l'adéquation aux temps de réaction. Passé un certain seuil, une dissociation entre les modèles apparaît. Pour ce sujet, le modèle de coordination sur Entropie donne un niveau d'adéquation au RT supérieur aux autres modèles. Ainsi, le point crucial d'une

sélection de solution à l'intérieur d'un front de Pareto est le niveau d'adéquation qu'il est acceptable de perdre sur les choix pour en gagner sur les RTs.

Dans l'exemple donné dans la figure 5.6.A, l'ensemble des solutions possibles ne contient que des paramétrages de modèles duaux et exclut les modèles à stratégie unique. En d'autres termes, les fronts des modèles uniques (MTB et QL) sont complètement dominés par les fronts de modèles duaux.

Le processus de sélection de la meilleure solution à l'intérieur du front de Pareto est un problème complexe dans le domaine de la décision multi-critère [ZITZLER et THIELE, 1999]. Dans la figure 5.6, la position de la solution sélectionnée avec la fonction d'agrégation de Tchebychev pour le sujet 6 est illustrée avec un triangle noir.

Sujet	- Bloc 1	- Bloc 2	- Bloc 3	- Bloc 4	Tous les blocs
1	**Coord-E**	W-Mix	**Coord-E**	**Coord-E**	W-Mix
2	**Select-VPI**	Coord-E	Coord-E	Coord-E	Coord-E
3	Coord-E	Coord-E	Coord-E	Coord-E	Coord-E
4	Coord-E	Coord-E	Coord-E	Coord-E	Coord-E
5	**W-Mix**	Coord-E	Coord-E	Coord-E	Coord-E
6	Coord-E	Coord-E	Coord-E	**W-Mix**	Coord-E
7	**Coord-E**	**Select-VPI**	**Select-VPI**	W-Mix	W-Mix
8	**W-Mix**	Select-VPI	Select-VPI	**Coord-E**	Select-VPI
9	**Select-VPI**	**Select-VPI**	**Select-VPI**	**Select-VPI**	W-Mix
10	Select-VPI	Select-VPI	Select-VPI	Select-VPI	Select-VPI
11	Coord-E	Coord-E	Coord-E	Coord-E	Coord-E
12	Coord-E	**W-Mix**	Coord-E	**W-Mix**	Coord-E
13	Coord-E	Coord-E	Coord-E	Coord-E	Coord-E
14	Coord-E	Coord-E	Coord-E	Coord-E	Coord-E

TABLEAU 5.2 – Résultat de la validation croisée. Chaque bloc est retiré systématiquement pour l'optimisation. Les modèles étant discordants avec les modèles résultant de l'optimisation entière sont représentés en gras. (Q-L : q-learning, MTB : Mémoire de travail bayésienne, Select-VPI : Sélection sur VPI, W-Mix : Mélange pondéré, Coord-E : Coordination par ntropie)

L'ensemble des meilleurs modèles pour chaque sujet en appliquant la fonction d'agrégation de Tchebychev est constitué de 9 modèles de coordination par entropie, 3 modèles de mélange pondéré et 2 modèles de sélection sur VPI. Bien que différents du premier ensemble de modèles sélectionnés à la sous-section 5.2.5, les modèles duaux sont toujours présents.

Pour nous assurer de la robustesse de notre méthode, nous avons fait une validation croisée dans le but de vérifier l'assignation du meilleur modèle à chaque sujet. Un bloc sur quatre est systématiquement retiré des sessions d'entraînement par sujet. L'optimisation est faite sur les choix et les temps de réaction de tous les sujets avec seulement trois blocs. La même fonction d'agrégation est appliquée dans le but de comparer les modèles. Les résultats sont montrés dans le tableau 5.2.

Les colonnes -Bloc *i* regroupent les résultats de l'optimisation faite sans le bloc correspondant. La dernière colonne donne les résultats originaux. Les cases en surbrillance indiquent les modèles discordants. Seuls trois sujets (1,7,9) donnent, dans une majorité de cas, un résultat différent. Pour 6 sujets sur 14, le résultat est le même (3,4,10,11,13,14). De plus, le meilleur modèle est le même dans 3 tests pour 3 sujets (2,5,6). En tout, le pourcentage d'erreur est de 30% (17/56). De plus, l'observation la plus pertinente est la suprématie encore une fois des modèles duaux qui apparaissent dans tous les cas. La validation

croisée confirme donc l'hypothèse de départ : une combinaison de stratégies est nécessaire pour expliquer les observations comportementales.

Simulation des modèles

Une dernière façon de tester l'adéquation de nos modèles est de les simuler sur le même ordre temporel de stimulus que les sujets. De manière à apprécier la qualité d'adéquation offerte par les modèles duaux, la fonction d'agrégation de Tchebytchev est aussi utilisée sur les fronts de Pareto de chaque modèle individuel (MTB et QL) pour sélectionner les jeux de paramètres optimaux correspondants. Concrètement, nous avons simulé un comportement moyen de choix et de temps de réaction pour l'ensemble des meilleurs paramètres de QL, l'ensemble des meilleurs paramètres MTB et l'ensemble des meilleurs paramètres pour tous les meilleurs modèles.

Les courbes d'apprentissage pour la simulation des quatorze modèles paramétrés sont représentées dans la figure 5.7 pour les modèles uniques (voir figure 5.7.A pour QL et 5.7.B pour MTB) et pour les meilleurs modèles (voir figure 5.7.C). Chaque comportement généré est superposé avec les courbes d'apprentissage des sujets. Pour chaque fonction d'agrégation et chaque modèle, la probabilité de réponse correcte est calculée sur la séquence de résultats binaires ($r_t \in 0,1$) pour chaque stimulus comme dans la figure 5.1.B et 5.5.

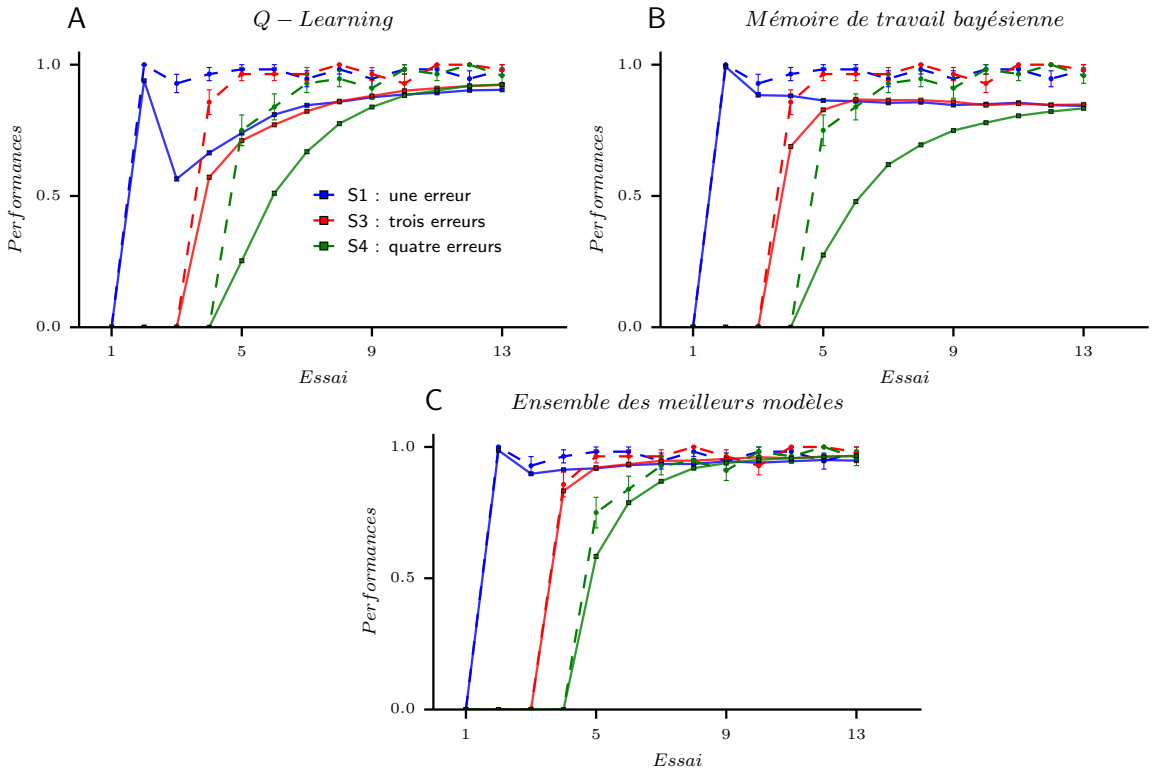

FIGURE 5.7 – A. Simulation sur les choix des meilleurs paramétrages de q-learning pour chaque sujet selon l'optimisation des choix et des temps de réaction. B. Pour le modèle de mémoire de travail bayésienne. C. Pour l'ensemble des meilleurs modèles ce qui correspond à 9 coordinations par entropie, 3 mélanges pondérés, et 2 sélections sur VPI, c'est-à-dire uniquement des modèles duaux.

Pour les meilleurs modèles et comme les meilleurs modèles sélectionnés précédemment avec BIC, seul le cinquième essai de S4 montre une différence significative entre les performances des sujets et du modèle (test du χ_2, T=5.57, p<0.05). Dans les 44 autres cas,

nous n'avons pas trouvé de différences significatives entre les modèles et les sujets (test du χ_2, T<2.16, p>0.14).

Concernant l'ensemble des meilleurs paramétrages du q-learning, 18 essais sur 36 se sont révélés significativement différents (test du χ_2, T>4.04, p<0.05). La plupart de ces essais discordants se retrouvent en début de tâche comme le montre la figure 5.7.A.

Pour l'ensemble des meilleurs paramétrages de MTB , 22 essais étaient significativement différents des performances des sujets (test du χ_2, T>3.87, p<0.05) comme le montre la figure 5.7.B.

Performance pour le stimulus S1. Seul le modèle MTB et les meilleurs modèles duaux ont réussi à reproduire les performances des sujets pour la première réponse correcte de S1 au deuxième essai. En moyenne, la performance du QL est inférieure ($PCR^{Q-L}_{S1|t=2}$ = 98.6%). La diminution de la valeur associée à l'action choisie au premier essai ne suffit pas à prévenir cette action d'être re-choisie au deuxième essai car le q-learning est un algorithme d'apprentissage lent. Dans les essais suivants, la performance du q-learning reste autour de 60% et augmente doucement ensuite.

Pour le modèle MTB, les performances aux essais suivants chutent à 80% et restent ensuite constantes contrairement aux performances des sujets qui continuent d'augmenter. A l'exception d'une légère chute de performance au premier essai, l'ensemble des meilleurs modèles duaux fournit des performances qui ne sont pas significativement différentes pour le stimulus S1.

Performances pour les stimuli S3 et S4. Pour les deux autres stimuli, les performances des sujets ont été entièrement reproduites par l'ensemble des meilleurs modèles duaux. Néanmoins, nous observons que la probabilité de réponse correcte des modèles est inférieure à celle des sujets pour le cinquième essai du stimulus S3. A cet essai, une seule action reste possible (les 4 autres actions étant déjà associées à une récompense négative) et les modèles duaux font plus d'erreurs répétitives que les sujets (les modèles répètent les actions déjà associées à une récompense négative). Pour le q-learning, les performances des stimuli S3 et S4 sont en-dessous des performances des sujets à l'exception des derniers essais. Une telle observation illustre la propriété de convergence lente de la stratégie habituelle. Comme énoncé précédemment, cette observation contraste avec la stratégie liée à un but : les performances pour tous les stimuli convergent vers une probabilité de réponse correcte constante qui n'évolue pas entre les essais.

Les temps de réaction. La simulation de la seconde observation comportementale est représentée dans la figure 5.8 à partir du même comportement généré dans la figure 5.7 des performances. Dans cette figure, nous avons appliqué deux traitements consécutifs : une mise à l'échelle et une mise en ordre pour pouvoir comparer et discuter l'évolution des temps de réaction moyens des sujets et des modèles. Une mise à l'échelle est nécessaire puisque nous comparons une distribution de temps de réaction en secondes et une distribution de temps de réaction simulés en unité arbitraire. Nous avons choisi de normaliser chaque distribution selon sa médiane et son écart interquartile. Le processus de mise en ordre est le même que dans la figure 5.1.C et appliqué aux temps de réaction simulés générés par un modèle à chaque essai. Une fois de plus, les observations issues des sujets et des modèles sont superposées l'une à l'autre dans la figure 5.8.

Au niveau du comportement du q-learning, nous observons que les vRTs ne font que décroître. De fait, les vRTs de QL sont calculés selon l'entropie des probabilités d'action finales. En d'autres termes, la stratégie habituelle devient de plus en plus rapide pour

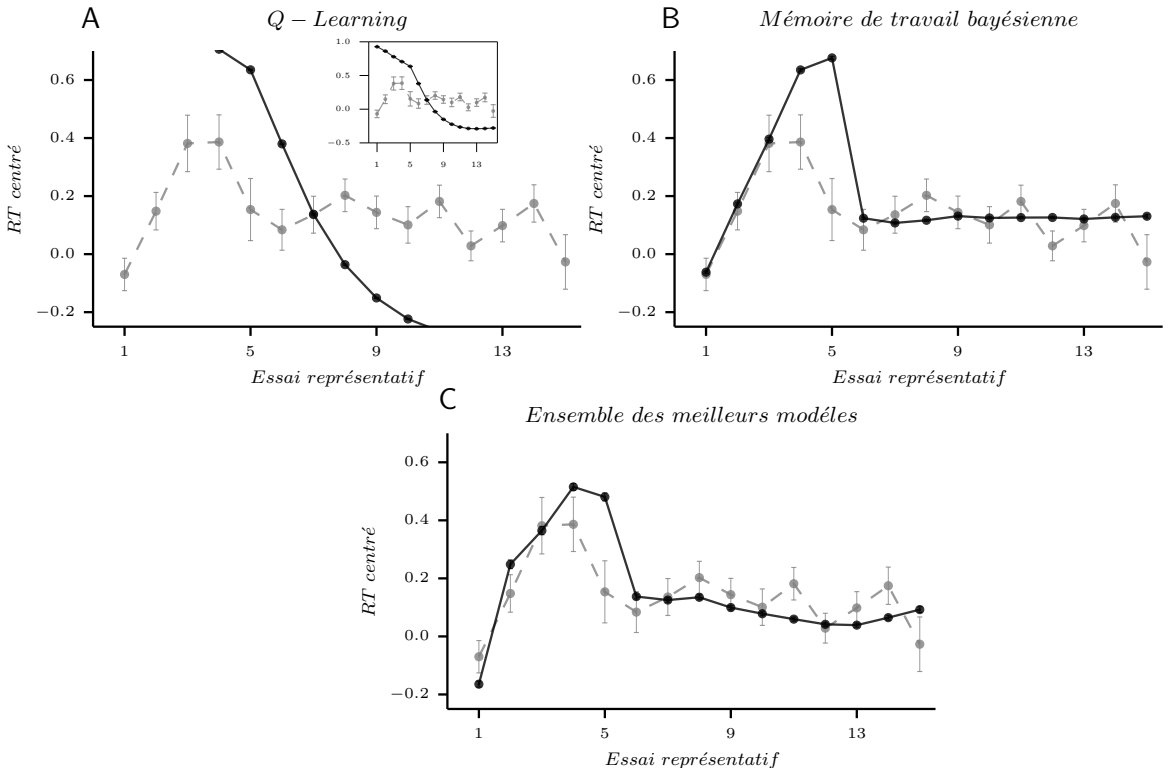

FIGURE 5.8 – A. Simulation des temps de réaction moyens pour les meilleurs paramétrages de q-learning pour chaque sujet selon l'optimisation des choix et des temps de réaction. B. Pour le modèle de mémoire de travail bayésienne. C. Pour l'ensemble des meilleurs modèles. Ces temps de réaction ont été obtenus en même temps que la simulation des choix présentée dans la figure 5.7

répondre durant l'apprentissage, ce qui diffère du comportement des sujets. Ce résultat rejoint l'observation faite sur les fronts de Pareto du sujet 6 (voir figure 5.6) au sujet de l'adéquation des RTs du QL qui était largement inférieure aux autres modèles.

Le modèle MTB produit un comportement beaucoup plus riche. Nous observons ainsi une augmentation des temps de réaction des essais représentatifs 1 à 5 suivie par une diminution vers un temps de réaction constant dans les essais suivants. Malgré le fait que cette évolution soit aussi présente chez les sujets, nous observons une discordance entre les deux distributions. En utilisant un test de Mann-Whitney, nous avons trouvé 7 essais représentatifs significativement différents (essais 1, 4, 5, 6, 10, 12, 15 ; test U de Mann-Whitney, $p<0.05$).

Pour finir, l'évolution des temps de réaction des sujets est mieux reproduite pour l'ensemble des meilleurs modèles duaux. Seuls 6 essais représentatifs sont significativement différents (essais 2, 4, 5, 6, 10, 15 ; test U de Mann-Whitney $p<0.05$). Pendant la phase de consolidation, les temps de réaction simulés décroissent graduellement grâce à la contribution du q-learning. Néanmoins, la différence la plus grande est toujours observée pour le cinquième essai.

Les temps de réaction par sujet. La qualité de l'adéquation des temps de réaction pour chaque sujet en simulant l'ensemble des meilleurs modèles est donnée dans la figure 5.9. Pour chaque sujet, l'évolution des temps de réaction moyens pour chaque essai représentatif est superposée avec les temps de réaction moyens simulés. Autrement dit, les temps de réaction moyens de la figure 5.8 peuvent être séparés en temps de réaction moyens individuels comme dans la figure 5.9.

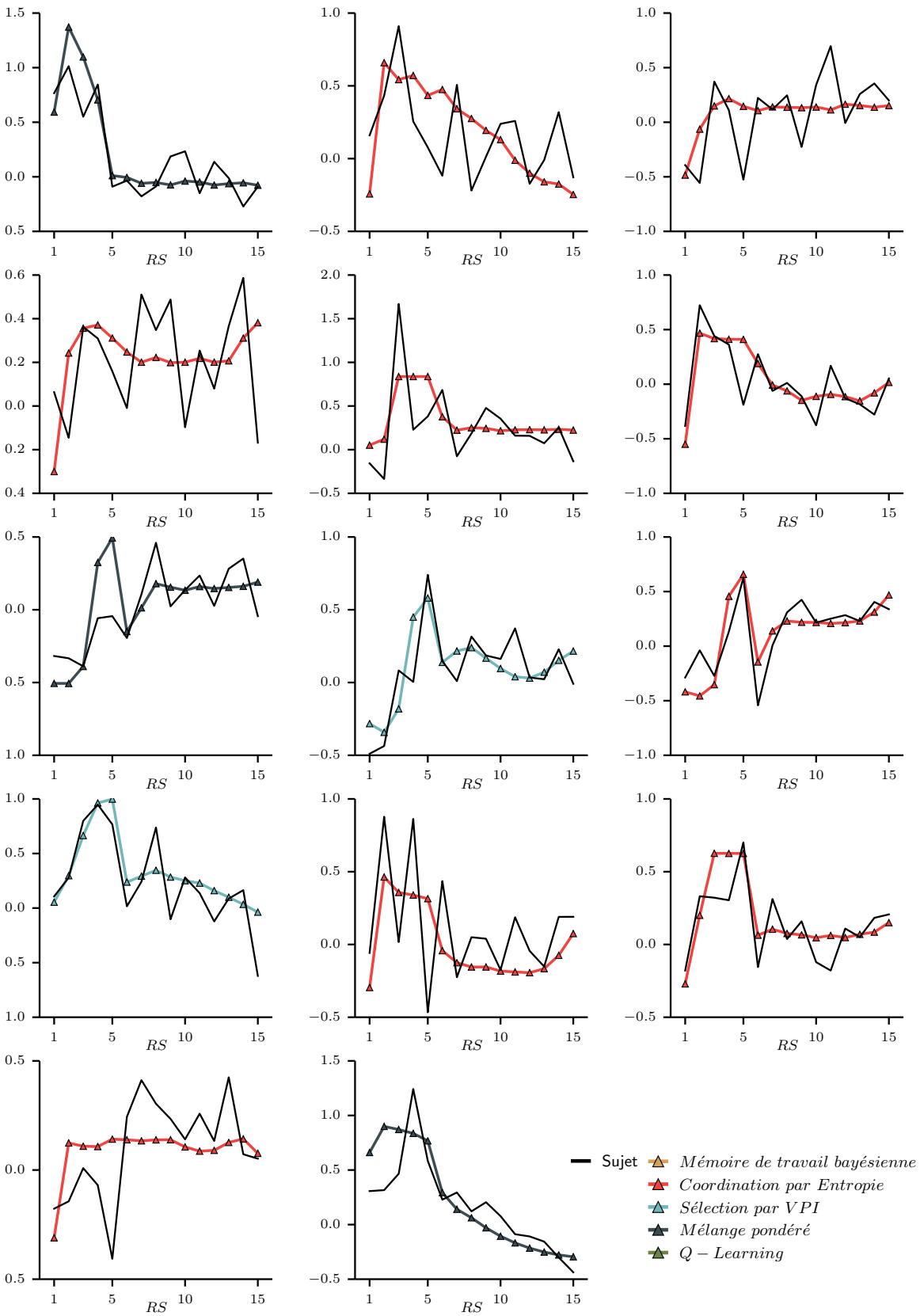

FIGURE 5.9 – Pour chaque figure, le temps de réaction moyenné en simulation (triangle coloré) pour chaque essai représentatif est représenté au-dessus des temps de réaction moyens des sujets (trait noir). La couleur des temps de réaction simulés indique le modèle sous-jacent.

La première observation que l'on peut faire sur les temps de réaction individuels concerne leur instabilité et leur stochasticité, ce qui génère de grandes différences inter-individuelles. Néanmoins, les temps de réaction simulés générés par l'ensemble des meilleurs modèles ne constituent pas des variables stéréotypées copiant uniquement la valeur moyenne de tous les sujets. Les modèles duaux capturent les différences inter-individuelles au niveau des temps de réaction.

Contribution relative de chaque système de mémoire

L'ensemble des meilleurs modèles est uniquement composé d'une combinaison de stratégies ce qui permet d'étudier la contribution relative du comportement lié à un but et du comportement habituel. Ces observations sont montrées dans la figure 5.10 et la contribution relative est moyennée pour les sujets ayant les mêmes meilleurs modèles.

Pour le mélange pondéré, la contribution relative est contenue dans le poids $w(t, s_t)$ qui évolue à chaque essai. Tout aussi direct est le modèle de sélection sur VPI basé sur le compromis vitesse et précision. Mais la nature du modèle de coordination par entropie ne permet pas d'évaluer directement la contribution de chaque système de mémoire directement. Il n'existe pas de variable dans ce modèle qui représente directement le jeu de coordination. Pour étudier la contribution relative, nous avons procédé à une «lésion» d'un système en observant le résultat à travers l'entropie. A chaque étape de la simulation, l'entropie H est évaluée à partir de la probabilité d'action complète ou de la probabilité d'action sans la contribution du MTB ou du QL. Le résultat de cette étude de «lésion» est montré dans la figure 5.10.A à C.

Nous observons ainsi que l'entropie du modèle de coordination sur entropie est inférieure à l'entropie du même modèle sans MTB ou sans QL. En d'autres termes, la quantité d'information contenue dans les probabilités d'action pendant une combinaison de systèmes de mémoire est plus grande que pour une stratégie seule. Ceci est différent du modèle de mélange pondéré et du modèle de sélection de stratégie pour lesquels l'entropie est encadrée par l'entropie des modèles «lésionnés». Ainsi, le modèle de coordination par entropie est le seul modèle qui montre une augmentation du gain d'information lorsque les deux systèmes de mémoire sont combinés.

Le mélange pondéré montre une préférence claire pour la stratégie habituelle avec un poids faible qui ne fait que décroître comme le montre la figure 5.10.D. Le poids décroît de façon monotone signifiant ainsi que la contribution du MTB décroît au cours d'une session au contraire du QL qui voit sa contribution augmenter. Néanmoins, le poids w_t faible tout au long de l'expérience indique une contribution toujours plus importante de QL sur MTB. Ceci est aussi observé dans la figure 5.10.B en comparant les entropies du modèle entier et du modèle «lésionné». Toutefois, la contribution de MTB est nécessaire de l'essai 1 à l'essai 5 dans le but de résoudre la tâche.

Pour finir, le modèle de sélection montre un comportement cohérent au regard de l'hypothèse qu'il incarne, c'est-à-dire le compromis vitesse et précision permettant l'expression du comportement lié à un but au début de la tâche et l'expression du comportement habituel à la fin de la tâche. La VPI est supérieure au taux de récompense durant la phase d'acquisition favorisant ainsi le modèle MTB. Aux cinquième et sixième essais, la VPI diminue grandement, ce qui correspond à la fin de la phase d'acquisition. L'agent a reçu toutes les associations positives et la VPI décroît massivement. Néanmoins, la VPI est supérieure au taux de récompense. Le changement de stratégie s'effectue vers l'essai 8. Dans la figure 5.10.C, on observe ainsi que l'entropie du modèle de sélection entier approche doucement de l'entropie du modèle de sélection avec QL seulement.

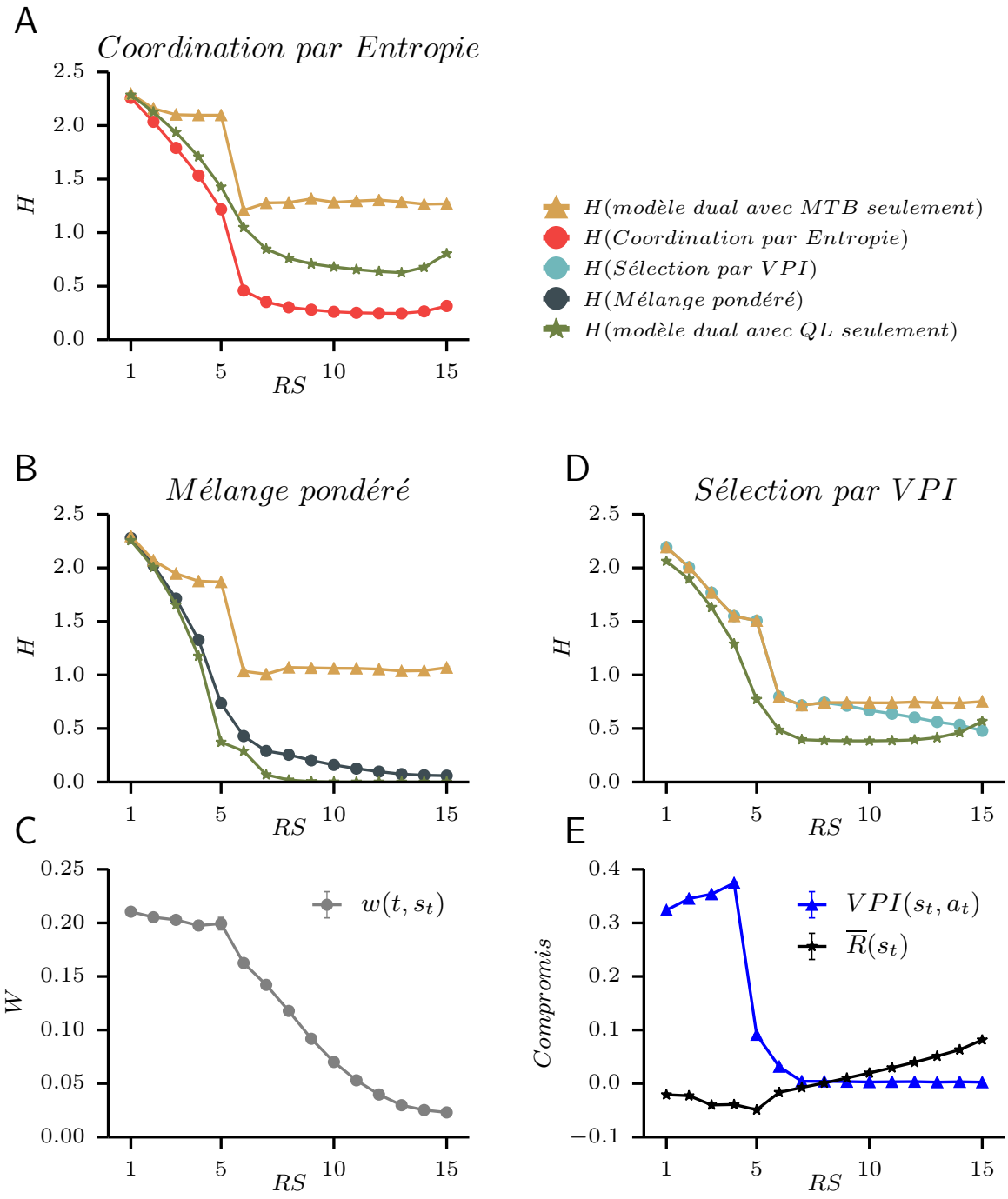

FIGURE 5.10 – Contribution moyenne à chaque essai représentatif du modèle de mémoire de travail et du q-learning à partir de l'ensemble des meilleurs modèles. A, B, D. Pour chaque modèle, la contribution est évaluée en retirant la q-valeur du modèle de mémoire de travail ou du q-learning séparément et en évaluant l'entropie des probabilités d'action en résultant. L'entropie finale pour le modèle dual entier (MTB+QL) est aussi représentée dans chaque cas. La moyenne est calculée sur les essais représentatifs pour l'ensemble des sujets pour lesquels ce modèle a été assigné selon l'optimisation multi-objective. C. Le poids w_t du modèle de mélange pondéré est représenté montrant la préférence du QL sur MTB. D. Le compromis vitesse et précision du modèle de sélection par VPI montre un changement de stratégie vers l'essai 8.

5.2.6 Conclusion

Dans cette première partie de chapitre, nous avons capturé les résultats comportementaux dans une tâche d'apprentissage instrumental conçue par A. Brovelli [BROVELLI et collab., 2008, 2011] et qui permet d'étudier l'interaction entre le comportement lié à un but et le comportement habituel. Nous avons ainsi proposé un nouveau modèle de mémoire de travail bayésienne comme comportement lié à un but. Nous avons aussi proposé un modèle de coordination par Entropie permettant de coupler le comportement lié à un but et le comportement habituel (algorithme du q-learning). Pour comparer avec la littérature correspondante, nous avons adapté un modèle de mélange pondéré [COLLINS et FRANK, 2012] et un modèle de sélection par VPI [KERAMATI et collab., 2011].

Pour optimiser les paramètres libres de chaque modèle possible (stratégie d'apprentissage seule ou combinée), nous avons utilisé l'algorithme d'évolution multi-objectif NSGA-2 [MOURET et DONCIEUX, 2010] appliqué au comportement de chaque sujet (choix et temps de réaction). De plus, nous avons utilisé une fonction de diversité pour garantir la convergence du processus d'optimisation. Nous avons contraint l'algorithme d'évolution à maximiser la vraisemblance sur les choix des sujets et à minimiser l'erreur des moindres carrés entre les temps de réaction des sujets moyennés et les temps de réaction simulés. Au final, la sélection d'une solution sur le front de Pareto résultant fournit uniquement des paramétrages correspondant à des modèles duaux.

L'ensemble des meilleurs modèles duaux est composé de 9 modèles de coordination par Entropie, 3 modèles de mélange pondéré et 2 modèles de sélection par VPI. L'intérêt des modèles duaux pour un apprentissage instrumental est clairement établi dans cette tâche à condition de s'intéresser à la fois au choix et au temps de réaction. Cette conclusion est-elle valable dans d'autres tâches ? Cette approche computationnelle peut-elle nous aider à quantifier la relative contribution de l'apprentissage par renforcement et de la mémoire de travail dans des tâches du même type chez le singe ? C'est ce que nous avons vérifié dans la section suivante en testant ces mêmes modèles chez le singe dans un apprentissage instrumental similaire.

5.3 Chez le singe

5.3.1 Tâche de résolution de problèmes

Dans QUILODRAN et collab. [2008] et KHAMASSI et collab. [2015], des singes ont été entraînés à découvrir par essai-erreur une cible récompensante parmi 4 cibles possibles. Un problème typique commence avec la phase de *recherche* durant laquelle l'animal réalise des essais incorrects (INC) jusqu'à la découverte de la bonne cible (CO1). Ensuite, une période de *répétition* commence durant laquelle l'animal peut répéter le même choix récompensant pendant un nombre d'essais variant de 3 à 11 (ceci permet de réduire l'anticipation de la fin du problème). A la fin de la phase de répétition, un signal est fourni indiquant le commencement d'un nouveau problème. La nouvelle cible récompensante est dans 90% des cas différente de la précédente. Les évènements successifs à l'intérieur d'un essai puis d'un problème (donc une suite d'essais) sont représentés dans la figure 5.11.

5.3.2 Modèles computationnels

Nous avons testé le q-learning, le modèle de mémoire de travail (5.2.2), le modèle de mélange pondéré (5.2.3) et le modèle de coordination par entropie (5.2.3). N'ayant pas

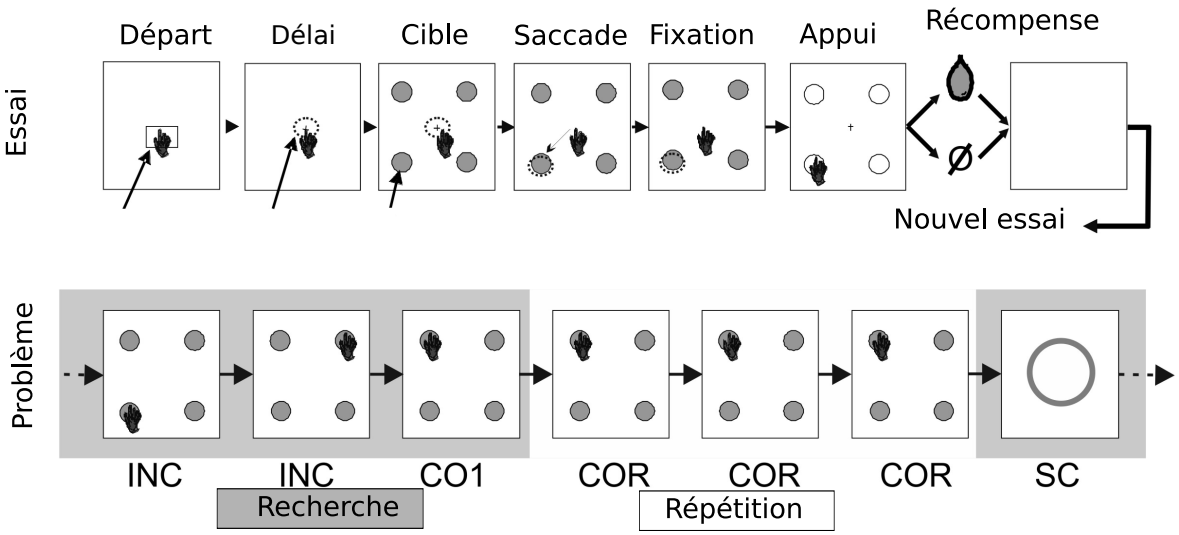

FIGURE 5.11 – Tâche de résolution de problèmes chez le singe. Les singes doivent trouver par essai-erreur la cible récompensante. Si l'essai est correct, un jus est délivré à l'animal. La phase de répétition peut durer de 3 à 11 essais. Adapté de KHAMASSI et collab. [2015].

capturé beaucoup de comportements dans la section précédente, le modèle de sélection sur VPI n'a pas été retenu pour cette étude.

En plus des 4 modèles testés identiquement à la section précédente (Variation 1), nous avons testé plusieurs variations de ces modèles dans le but de mieux capturer le comportement des singes. Les variations sont :

2. une optimisation du γ du q-learning. Dans la section précédente, le γ du q-learning est nul puisqu'il n'existait pas de relation entre les transitions entre les états. Dans le cas présent, l'état est unique et bouclant ce qui rend l'utilisation du γ possible. Le γ est aussi optimisé pour toutes les variations suivantes.
3. Etant donné que chaque singe est entraîné sur de nombreux essais, il est possible que l'utilisation de la mémoire habituelle ait une influence d'un problème à l'autre. Pour modéliser cet effet, le q-learning n'est pas réinitialisé au début d'un essai.
4. Néanmoins, il est probable que cette stratégie de non-réinitialisation soit plus efficace si les q-valeurs du q-learning s'effacent au fur et à mesure de la tâche. Tout comme KHAMASSI et collab. [2015], nous avons ainsi testé une version du q-learning avec oubli et sans réinitialisation. Les couples état-action qui n'ont pas été sélectionnés dans un essai sont mis à jour selon :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + (1-\kappa)(Q_0 - Q(s_t, a_t)) \tag{5.19}$$

avec $Q_0 = 0$ et $0 \leq \kappa < 1$ un paramètre de mise à jour.

5. Dans certains cas (singe m et singe p ; figure A.1), le profil de temps de réaction est opposé aux observations de temps de réaction de la section précédente que nous avons capturés par une combinaison de stratégies. Les temps de réaction en phase d'exploration sont inférieurs aux temps de réaction en phase d'exploitation ce qui impliquerait que la stratégie délibérative et donc la mémoire de travail soit utilisée seulement en phase de répétition des essais corrects. Etant donné que les singes sont entraînés sur des milliers d'essais, il est fort probable que le processus de décision durant la phase d'exploration soit en partie automatisé. Pour autant, les singes doivent quand même opérer un processus délibératif en évitant de répéter les actions négatives. Une possibilité que nous avons explorée dans cette dernière version

est celle d'une anticipation de l'action par la mémoire de travail. Durant la phase de mise à jour des systèmes de mémoire par la récompense, nous avons testé une heuristique simple qui consiste, pour les essais incorrects uniquement, à réinférer l'ensemble des éléments en mémoire de travail de manière à préparer la distribution de probabilité de chaque action utilisée au début de l'essai suivant. Ainsi, l'entropie des probabilités d'actions $H(p(a|s_t))$ des q-valeurs combinées (quel que soit le modèle de coordination utilisé) diminuera pendant les essais explorateurs sans que celle-ci ne s'accompagne d'une augmentation de la charge cognitive de la stratégie délibérative (ce qui ferait augmenter les temps de réaction). Pour rappel, nous avons modélisé le temps de réaction selon $sRT(trial) = (log_2(i+1))^\sigma + H(p(a|s_t))$ avec i le nombre d'éléments en mémoire de travail inférés. Cette heuristique n'est pas appliquée pendant la phase de répétition de l'action correcte.

Tous les modèles testés pour chaque singe sont résumés dans le tableau suivant :

	MTB	Q-L	Mélange	Coordination
Variation 1	Voir 5.2.2	Q-L($\gamma = 0$)	Voir 5.2.3	Voir 5.2.3
Variation 2	∅	$\gamma \in [0,1[$	$\gamma \in [0,1[$	$\gamma \in [0,1[$
Variation 3	∅	$\gamma \in [0,1[$ + ¬ INIT(Q-L)	$\gamma \in [0,1[$ + ¬ INIT(Q-L)	$\gamma \in [0,1[$ + ¬ INIT(Q-L)
Variation 4	∅	$\gamma \in [0,1[$ + ¬ INIT(Q-L) OUBLI(Q-L)	$\gamma \in [0,1[$ + ¬ INIT(Q-L) OUBLI(Q-L)	$\gamma \in [0,1[$ + ¬ INIT(Q-L) OUBLI(Q-L)
Variation 5	HEURISTIQUE (anticipation par MTB)	∅	$\gamma \in [0,1[$ + ¬ INIT(Q-L) OUBLI(Q-L) HEURISTIQUE	$\gamma \in [0,1[$ + ¬ INIT(Q-L) OUBLI(Q-L) HEURISTIQUE

TABLEAU 5.3 – Tableau des 5 variations du modèle de mémoire de travail (MTB), du q-learning (Q-L), du modèle de mélange pondéré (Mélange) et du modèle de coordination par entropie (Coordination). Le symbole ∅ désigne les modèles non concernés par la variation testée. Le symbole ¬ est utilisé ici pour désigner l'absence de réinitialisation du q-learning au début d'un nouveau problème.

5.3.3 Résultats

Les résultats de l'optimisation sont présentés dans la figure 5.12 pour les cinq singes. Tout comme la section précédente, nous avons maximisé la vraisemblance que le modèle fasse les mêmes choix que les sujets et nous avons minimisé l'erreur des moindres carrés entre les moyennes des temps de réaction aux essais représentatifs.

Les essais représentatifs sont différents de la section précédente étant donné qu'il n'y a pas un nombre minimal d'erreurs à effectuer pour chaque singe. Pour obtenir des essais représentatifs, nous avons divisé chaque problème en fonction du nombre d'essais de recherche (1 à 5). Pour la phase de répétition, seuls les trois premiers essais sont conservés. Les temps de réaction sont moyennés permettant de donner une courbe d'évolution

par sujet et par problème. L'optimisation de chaque modèle s'attelle donc à minimiser la différence avec chacune de ces courbes moyennes d'évolution des temps de réaction pour un sujet. Les fronts de Pareto obtenus après l'optimisation sont présentés dans la figure 5.12 pour chaque singe. Pour chaque modèle, un front de Pareto est créé en mélangeant les variations dudit modèle.

La première observation que nous pouvons faire est, encore une fois, la nécessité de combiner des systèmes de mémoire pour expliquer dans cette tâche les choix et les temps de réaction. Les modèles de coordination par entropie et de mélange pondéré surpassent les modèles simples de mémoire de travail bayésienne et de q-learning. A l'exception du singe r pour laquelle la variation 5 semble la plus adaptée, les autres fronts de Pareto de chaque modèle mélangent plusieurs variations. Par exemple, pour le modèle de coordination par entropie appliqué au singe s, la meilleure adéquation aux choix commence par la variation 2 (en bas du front de Pareto) puis, vers le milieu du front, c'est la variation 5 qui se rapproche le plus des temps de réaction (tout en perdant de l'adéquation aux choix).

Comme il n'existe de mesure directe de la performance des singes comme dans la section précédente et que ce travail est principalement préliminaire, nous avons seulement testé le paramétrage de chaque modèle sur la reproduction des temps de réaction en suivant la séquence de choix du singe (et non en roue libre). Le paramétrage testé correspond donc à la solution qui maximise l'adéquation aux temps de réaction sur les fronts de Pareto. Implicitement, cette solution correspond à la première solution qui dépasse le modèle uniforme dans l'adéquation aus choix du singe (l'adéquation aux choix est normalisée entre un «modèle» ne proposant que des probabilités uniformes et la réplication parfaite des choix).

Les résultats sont présentés dans la figure 5.13 pour chaque singe avec la version associée à chaque modèle. Les phases d'exploration et d'exploitation sont séparées par une barre verticale. La moyenne des temps de réaction pour chaque type de problème et pour chaque singe est représentée dans la figure 5.13 par une courbe noire pointée.

Comme le montrent les fronts de Pareto, des solutions améliorant l'adéquation aux choix existent (bien qu'elles diminuent l'adéquation aux temps de réaction). Pour finir, nous avons aussi appliqué le processus de sélection de solution de Tchebytchev (A.2) pour extraire un seul paramétrage des fronts de Pareto. Représentée par une étoile noire sur chaque front de Pareto, cette solution est un compromis entre l'adéquation aux choix et l'adéquation aux temps de réaction. Les résultats de simulation contraints par les choix des singes sont présentés dans la figure annexe A.1.

Pour la figure 5.13, le modèle capturant le mieux les temps de réaction est le modèle de coordination par entropie version 5 appliqué au singe g. Pour rappel, la version 5 correspond à l'utilisation du gamma dans le q-learning, l'oubli des q-valeurs, la non-réinitialisation et une anticipation par le modèle de mémoire de travail pendant la phase de recherche.

Le deuxième modèle capturant le mieux l'évolution des temps de réaction est le modèle de coordination par entropie version 3 appliqué au singe r. L'évolution des temps de réaction du singe r est très similaire à l'évolution moyenne des temps de réaction dans Brovelli et collab. [2011] de la section précédente et ne semble pas nécessiter d'anticipation de la mémoire de travail.

Au contraire, les singes s et p présentent des profils largement inverse de Brovelli et collab. [2011] et de fait sont mieux approximés par une anticipation de la mémoire de travail (avec toutefois des différences certaines pour le singe p pour 0INC,3REP et 1INC,3REP). Pour finir, le singe m est faiblement capturé par le modèle de mélange pondéré version 4.

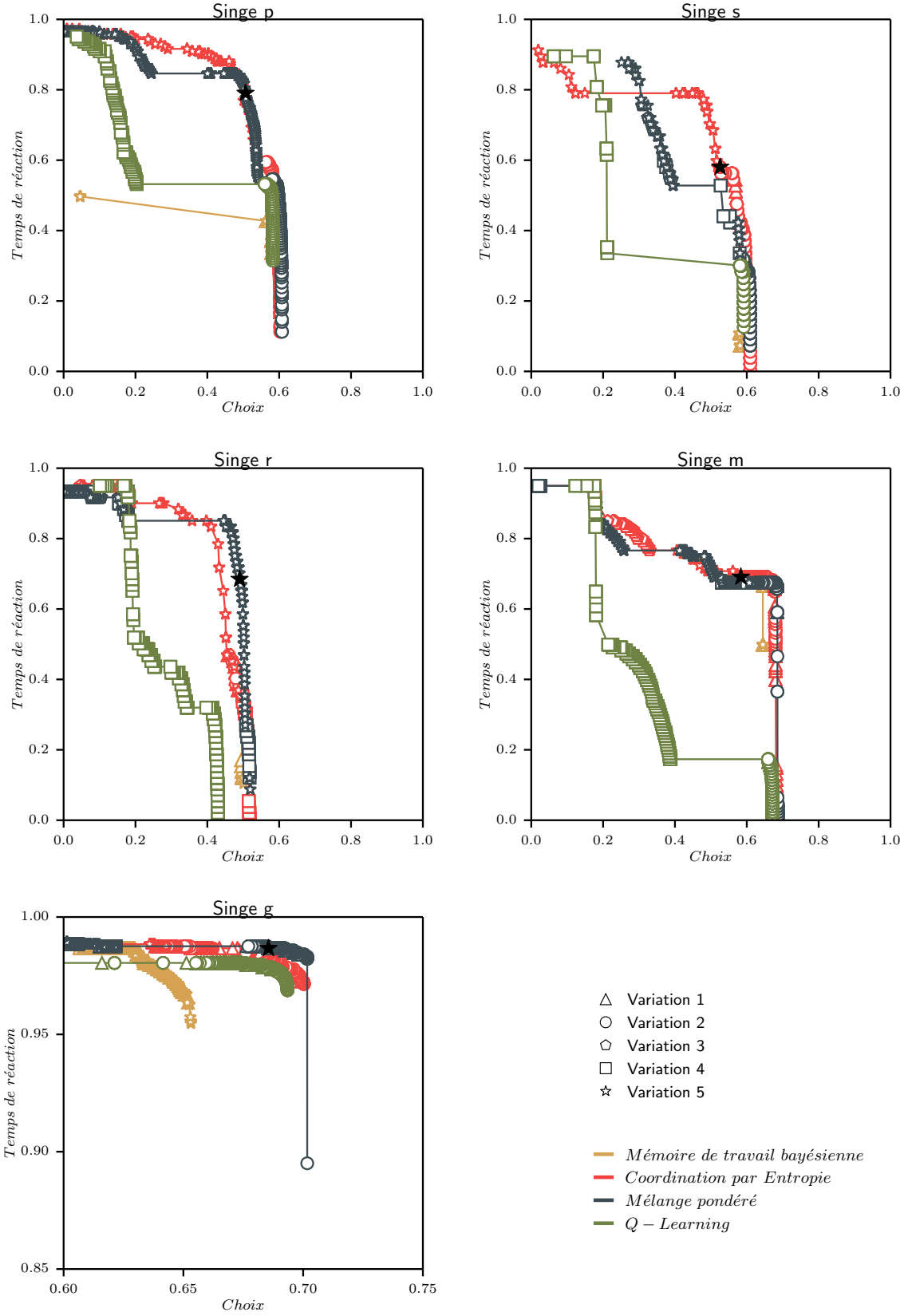

FIGURE 5.12 – Pour chaque singe, le front de Pareto pour chaque modèle est construit en mélangeant les variations possibles. Nous avons optimisé le q-learning, le modèle de mémoire de travail bayésienne, le modèle de mélange pondéré et le modèle de coordination par entropie. Les variations sont décrites dans la section 5.3.2. L'étoile noire représente la solution sélectionnée par l'opérateur de Tchebytchev.

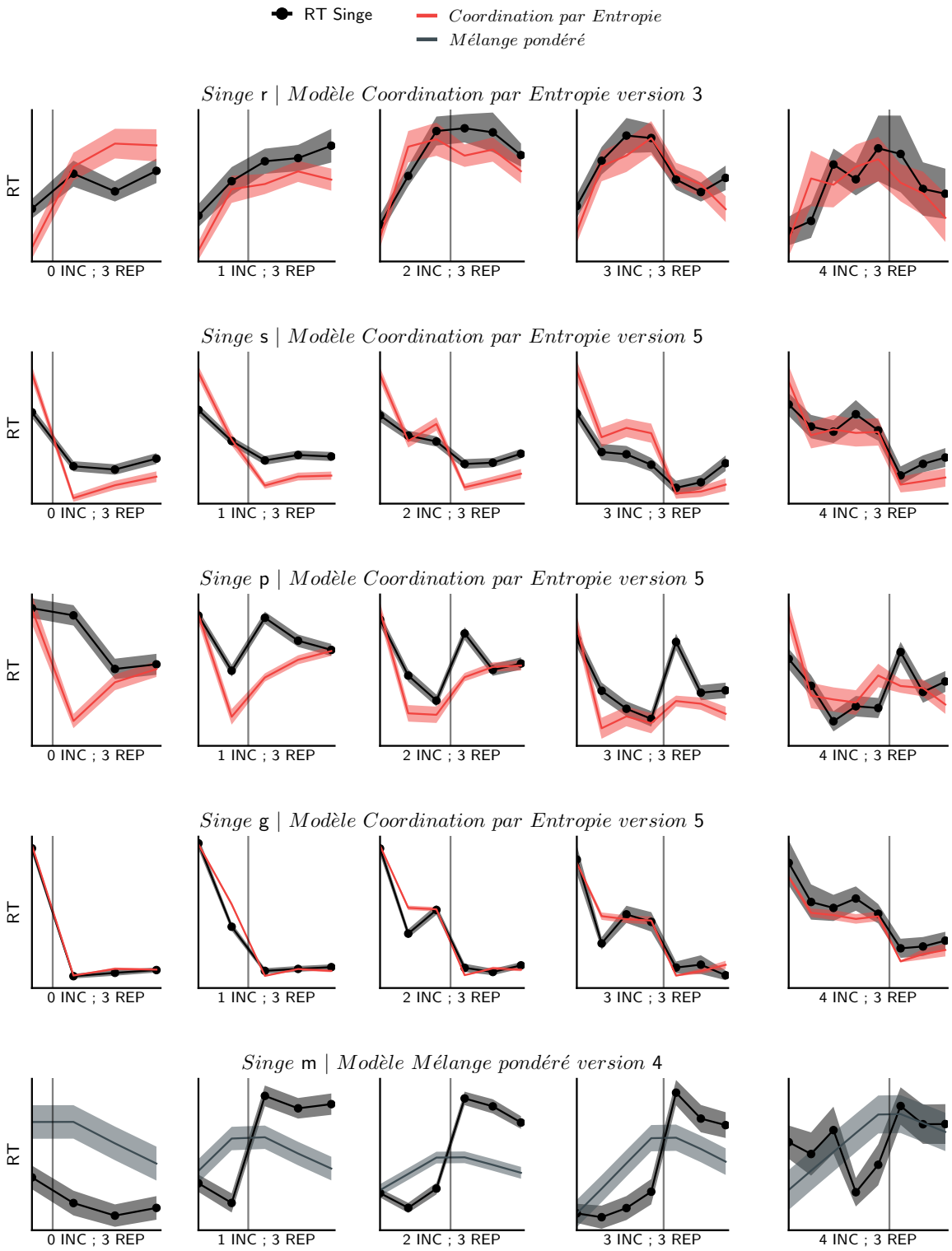

FIGURE 5.13 – Simulation (moyenne ± erreur type) des temps de réaction contraints par la séquence de choix du singe. Pour chaque singe, la moyenne des temps de réaction est effectuée en séparant les problèmes en fonction du nombre d'essais dans la phase d'exploration.

5.3.4 Discussion

Dans cette section, nous avons exploré très brièvement les possibilités de transfert de notre modèle de mémoire de travail ainsi que des modèles de coordination vers une tâche d'association visuo-motrice chez le singe. Selon ce travail préliminaire, Le modèle de coordination par entropie a réussi à capturer l'évolution des temps de réaction pour les singes r et g et partiellement les singes s et p. Pour m, le modèle de mélange pondéré est celui qui minimise la différence des moindres carrés sans toutefois capturer l'évolution des temps de réaction.

Cette tâche a été étudiée dans une série d'articles chez le singe [KHAMASSI et collab., 2015; PROCYK et collab., 2000; QUILODRAN et collab., 2008; ROTHÉ et collab., 2011] mais aussi chez l'humain en imagerie cérébrale (IRMf) [AMIEZ et collab., 2012] et en électroencéphalographie (EEG ; avec 5 actions possibles au lieu de 4) [SALLET et collab., 2013]. Dans les deux cas, les auteurs ont cherché à corréler une erreur de prédiction de récompense RPE = $r_{obtenue} - p_{correct}.r_{attendu} (= 1)$ avec l'activité cérébrale enregistrée pendant la phase exploratrice. En IRMf, le cortex cingulaire antérieur, le cortex fronto-insulaire, le striatum, le cortex rétrosplénial et le cortex préfrontal dorsolatéral moyen corrèlent positivement avec une RPE positive (en d'autres termes : plus la RPE est élevée, moins la récompense était attendue et plus l'activité cérébrale est élevée). Néanmoins, cette corrélation disparaît pour une RPE négative. En EEG, les auteurs s'intéressent aux potentiels évoqués lors de la réception de la récompense. Contrairement aux résultats en IRMf, le potentiel évoqué pour les régions frontales corrèle à la fois avec la RPE positive et la RPE négative. De plus, les auteurs montrent que le même potentiel évoqué apparaît au moment du signal indiquant le début d'un nouveau problème impliquant ainsi un suivi de la structure de la tâche par les sujets et non seulement les récompenses positives et négatives.

En outre, une similitude existe avec la tâche étudiée dans cette section au niveau des résultats comportementaux. Tout comme les singes g, p et s, les sujets humains en IRMf et en EEG montrent la même tendance d'accélération des temps de réaction au fur et à mesure de l'élimination des cibles et une légère augmentation du temps de réaction en phase de répétition.

Ces études chez le singe [KHAMASSI et collab., 2015; PROCYK et collab., 2000; QUILODRAN et collab., 2008; ROTHÉ et collab., 2011] et chez l'humain [AMIEZ et collab., 2012; SALLET et collab., 2013] montrent clairement une activité cérébrale de haut niveau reliée à l'évaluation, l'encodage mais aussi le suivi de l'incertitude associée à la décision pendant l'obtention de la récompense. Dans sa version originale (Version 1), ces caractéristiques ne sont pas modélisées par notre modèle de mémoire de travail que ce soit comme modèle unique ou comme modèle combiné à un q-learning. Celui-ci ne fait qu'encoder la description de l'essai. La seule version qui se rapproche d'une mesure concernant l'incertitude courante concerne la version 5 avec une réévaluation du contenu de la mémoire de travail induisant une diminution d'entropie pendant l'obtention de la récompense. Néanmoins, les auteurs se sont principalement intéressés à une erreur de prédiction de récompense qui est un signal plus proche de celui de l'erreur de différence temporelle utilisé par notre q-learning. Les modèles de coordination entre un q-learning et un modèle de mémoire de travail que nous proposons sont ainsi appropriés pour chercher de nouvelles variables computationnelles pouvant être utilisées pour chercher des corrélats dans l'activité neurophysiologique, en plus de la RPE, pour mieux comprendre le rôle des différentes structures étudiées.

De plus, une différence concerne le nombre d'essais effectués. Chez les sujets humains, ce sont une dizaine de problèmes en IRMf et une centaine de problèmes en EEG

qui sont résolus. Au contraire, les singes sont confrontés à un millier d'essais de cette tâche et il est ainsi fort probable qu'une stratégie habituelle se développe. Néanmoins, cette stratégie d'automatisation de la recherche de bonne réponse semble assez éloignée d'un q-learning tel qu'utilisé dans ce travail de modélisation qui sera toujours lié à l'association correcte qui change d'un essai à l'autre pénalisant la réussite de la tâche. Le q-learning classique semble peu adapté comme modèle d'une mémoire inflexible et tournée vers la résolution de cette tâche.

Cette stratégie habituelle peut s'incarner dans des méta-paramètres comme le propose KHAMASSI et collab. [2011]. Dans cette étude, le q-learning comme modèle central de décision est augmenté par des méta-paramètres β qui sont différents en fonction de la phase d'un problème. Une variation simple de nos modèles pourrait être de considérer l'ajout d'un module (en plus de la mémoire de travail et du q-learning) apprenant l'entropie moyenne de chaque type d'essais (multipliant ainsi le nombre d'états pour ce module). Cette entropie moyenne pourrait être utilisée dans le processus de décision de la mémoire de travail en modulant par exemple le seuil d'entropie qui déclenche la décision.

Ce type de méta-apprentissage, où le modèle apprend progressivement les entropies moyennes dans différents états de la tâche pour ainsi biaiser les décisions du modèle, semble une piste prometteuse qui pourrait nous permettre de mieux rendre compte de l'apprentissage progressif de la structure de la tâche au cours des milliers d'essais réalisés par les singes. Etant donné que les humains testés sur la même tâche [SALLET et collab., 2013] ont un profil de temps de réaction proche de celui des singes étudiés ici, il se pourrait que les instructions données aux sujets humains et la facilité de la tâche aient pu permettre à ces sujets de comprendre très vite la structure de la tâche, sans avoir besoin d'effectuer des milliers d'essais. C'est une des perspectives que nous pourrions étudier par la suite en faisant une comparaison homme-singe avec nos modèles sur cette tâche.

Pour conclure, le travail de modélisation présenté dans cette section constitue principalement un test de transférabilité des modèles et de la méthode d'optimisation développés dans la section précédente. Dans un cas (singe g), le modèle de coordination par entropie avec heuristique d'anticipation est le meilleur modèle pour les choix et capture très bien les temps de réaction. Pour les autres singes, d'autres variations des modèles seront ainsi nécessaires pour capturer entièrement l'évolution des temps de réaction et des choix.

5.4 Conclusion

Dans ce chapitre, nous avons présenté un modèle de mémoire de travail bayésienne qui, à partir d'une liste d'éléments en mémoire, minimise l'entropie d'information en évaluant itérativement chaque élément. Quand l'entropie est inférieure à un certain seuil, l'agent considère qu'il a assez d'information pour décider et une action est donc choisie. Ce modèle a été conçu pour modéliser précisément la tâche d'association visuo-motrice étudiée dans BROVELLI et collab. [2008, 2011]. Dans ces études, les auteurs émettent l'hypothèse d'une interaction entre une stratégie délibérative et une stratégie habituelle. Cette hypothèse nous a conduit à adjoindre un q-learning comme stratégie habituelle (selon les propositions de la littérature correspondante [DAW et collab., 2005; KERAMATI et collab., 2011]). Comme processus d'interaction, nous avons proposé un modèle de coordination par entropie qui permet de régler l'instant de la décision selon l'entropie inférée des éléments en mémoire de travail mais aussi de l'entropie contenue dans les probabilités d'action du q-learning. Nous avons comparé ce modèle à un modèle de mélange pondéré

adapté de COLLINS et FRANK [2012] et à un modèle de sélection par VPI adapté de KERAMATI et collab. [2011]. En optimisant les paramètres sujet par sujet selon un algorithme d'évolution multi-objectif, nous montrons ainsi qu'il est possible de capturer l'évolution de la performance et des temps de réaction par l'un des trois modèles d'interaction proposés.

Dans une seconde partie de chapitre, nous avons testé brièvement nos modèles et notre approche d'optimisation sur une tâche similaire chez le singe. Si un processus d'interaction semble très bien s'appliquer pour expliquer les temps de réaction pour un singe, les observations comportementales pour les autres singes ne sont pour l'instant modélisées que partiellement. De fait, cette tâche est différente de celle étudiée chez BROVELLI et collab. [2011]. Il n'existe qu'un seul état, il n'y a pas de manipulation de la tâche et les singes sont entraînés sur des milliers d'essais. Si le modèle de mémoire de travail bayésienne ainsi que les processus de coordination proposés semblent constituer une base possible pour la modélisation de la tâche, d'autres développements et ajustements semblent nécessaires pour capturer entièrement les observations de choix et de temps de réaction.

Pour conclure, ce chapitre constitue notre proposition de modélisation de deux systèmes de mémoire séparés : la mémoire de travail et la mémoire procédurale. Dans le chapitre suivant, nous allons nous intéresser à la mémoire épisodique qui est une autre forme de mémoire pouvant être utilisée dans les processus de décision impliquant l'apprentissage par renforcement. Confrontés à la modélisation du comportement de souris dans un labyrinthe, cette utilisation de la mémoire épisodique, selon les modèles computationnels proposés dans ce chapitre, correspond ainsi à une intégration/coopération (plutôt qu'une séparation) d'un processus de mémoire de travail avec un processus d'apprentissage par renforcement.

Chapitre 6

Apprentissage de séquences avec mémoire rétrograde

Sommaire

6.1 Introduction

Les processus de décision markoviens tels que présentés dans le chapitre 3 souffrent du défaut propre à tout système récursif : l'état suivant ne dépend que de l'état courant. La détermination de la valeur d'une transition entre états se construit uniquement sur un signal de récompense et cela détermine entièrement la trajectoire de l'agent. L'historique des états passés n'influence pas une transition immédiate.

Néanmoins, il ne fait aucun doute que le souvenir de ces états passés apporte un avantage dans un processus de décision, ou pour le dire dans les termes de l'apprentissage par renforcement, cela permet une meilleure évaluation de l'action. Par exemple, une tâche de reconnaissance de forme avec *delay* (section 2.2.3) ou une tâche d'alternance spatiale dans un labyrinthe radial (section 2.2.3) ne nécessitent pas le formalisme classique de l'apprentissage par renforcement. Pour résoudre ces tâches, un agent doit prendre une décision en fonction des états passés. Par ailleurs, c'est ce que nous avons par exemple observé au chapitre précédent avec le résultat montrant qu'un modèle qui utilise, en plus de l'apprentissage par renforcement, une mémoire de travail sur les états, actions, récompenses passés, permet de mieux rendre compte de données comportementales chez des sujets humains et primates.

Dans la littérature en neuroscience et en psychologie, le terme consacré au souvenir des expériences passées est celui de mémoire épisodique [TULVING, 1972, 1985, 2002]. Nous l'avions déjà cité dans l'introduction sur les différentes formes de mémoire comme appartenant à une théorie (tableau 2.1). Le terme neutre est mémoire autobiographique. Selon la définition classique, la mémoire épisodique est la capacité d'encoder et de se rappeler d'évènements passés personnels en répondant à trois questions sur le souvenir : «où, quand et quoi». A l'origine, la mémoire épisodique est postulée par Tulving pour la distinguer de la mémoire sémantique. Il était ainsi possible de différencier dans une tâche en psychologie expérimentale le rappel de faits généraux du rappel des évènements personnellement vécus par le sujet. Tulving postule aussi une dimension temporelle imaginaire (associée à une dimension spatiale) que le sujet peut parcourir à volonté. Les souvenirs sont ainsi indexés sur cette dimension en fonction de leur successivité temporelle. Une fois placés sur cet axe temporel, il est possible de remonter le fil des évènements successifs.

Ce manuscrit étant un travail de modélisation et donc par extension de simulation, nous pouvons déjà remarquer la facilité avec laquelle nous allons pouvoir modéliser la mémoire épisodique. La capacité d'indexer des cases, de leur assigner un contenu et de pouvoir lire ce contenu dans un certain ordre fixe est la définition même du fonctionnement d'un ordinateur. Par ailleurs, il n'est pas anodin de remarquer que la question d'amener le passé dans le présent d'un point de vue formel est aussi trivial que la question inverse est difficile (c'est-à-dire amener le futur dans le présent). Pour preuve, le lecteur est invité à consulter les références incluses dans les sections 4.1.2 et 4.2 sur la modélisation du comportement lié à un but et d'apprécier le niveau de complexité des systèmes formels.

Néanmoins, cette modélisation a été aussi contrainte par les observations en neurobiologie sur la mémoire épisodique et notamment sur le fonctionnement du substrat neuronal concerné. A ce sujet, des éléments de réponse ont déjà été donnés dans le chapitre 2 sur l'implication de l'hippocampe. Dans la classification des mémoires (figure 2.3) selon SQUIRE [2004], la mémoire épisodique était considérée comme étant une particularité de la mémoire déclarative et associée au lobe temporal. Si l'on considère les caractéristiques d'un système de mémoire telles qu'exposées dans l'introduction du chapitre 2,

la mémoire épisodique est une forme de mémoire déclarative. Au contraire, Tulving parle de système pour discuter de la mémoire épisodique tout en reconnaissant que ce choix dépend de la problématique posée [TULVING, 2002]. Pour trouver un juste milieu à ce débat, nous allons donc considérer la mémoire épisodique comme une forme de mémoire que peut exprimer en fonction de la demande cognitive le système de mémoire contenu dans l'hippocampe (une autre forme est par exemple la mémoire spatiale). De plus, ce compromis s'accorde bien au système formel hybride que nous allons proposer par la suite.

Cela nous amène donc à l'hippocampe comme siège possible de la mémoire épisodique et de l'apprentissage de séquences [FOSTER et KNIERIM, 2012; RONDI-REIG et collab., 2006]. Néanmoins, nous n'allons pas quitter la sphère de l'apprentissage par renforcement pour les raisons qui suivent. Pour rappel, le premier modèle de coordination de stratégies exposé dans le chapitre 4 correspondait, entre autres, à une proposition de décodage de l'ensemble des états d'un acteur-critique vers les cellules de lieux de l'hippocampe [FOSTER et collab., 2000]. Cette proposition a ensuite été consolidée par l'observation du phénomène de répétition hippocampique [FOSTER et WILSON, 2006]. Très brièvement, l'ordre d'activitation des cellules de lieux lorsque l'animal parcourt son environnement est rejoué en accéléré lorsque l'animal se repose ou consomme sa récompense. Pour les auteurs de FOSTER et WILSON [2006], ce phénomène est identifiable au processus de simulation des algorithmes d'apprentissage sur modèle comme le dyna-q présenté dans la section 3.4. La fonction de valeur converge plus vite si l'agent rejoue «en simulation» les transitions entre états effectuées pendant l'exploration. De fait, le lien entre ces répétitions hippocampiques et la notion de récompense a été démontré dans PFEIFFER et FOSTER [2013]. Pour toutes ces raisons, le formalisme de l'apprentissage par renforcement sera aussi utilisé dans ce chapitre.

Pour achever cette introduction, la dernière étape est de définir la tâche permettant d'étudier cet apprentissage de séquences. Dans FOUQUET et collab. [2010], les auteures discutent, entre autres, de l'impossibilité du rappel déclaratif chez des rongeurs en opposition au rappel conscient d'un souvenir autobiographique humain. Pour étudier de manière générique la mémoire épisodique chez le rongeur, ces auteures proposent un nouveau labyrinthe : le *star-maze* (voir figure 6.1). Dans sa version complète, le *star-maze* permet d'étudier l'expression de plusieurs stratégies (stratégie de lieux, stratégie de réponses, etc) [RONDI-REIG et collab., 2006] chez le rongeur. La capacité de se souvenir de ses choix successifs (et donc de leur ordre temporel) aux différents croisements du labyrinthe et d'agir en conséquence est reconnue comme la stratégie séquentielle égocentrée et constitue de fait l'expression d'une mémoire épisodique.

A la faveur d'une collaboration avec B. Babayan et L. Rondi-Reig , nous avons ainsi pu étudier dans sa forme la plus pure la relation qui existe entre la mémoire épisodique et l'apprentissage par renforcement en modélisant le comportement de souris apprenant une séquence d'actions dans une version simplifiée du *star-maze*. De fait, la version simplifiée correspond à une absence d'indices visuels empêchant l'utilisation d'une stratégie de lieux ou d'une stratégie de réponse. Seule la mémorisation des choix successifs est utile pour localiser la plateforme.

L'acquisition des données comportementales a été effectuée par B. Babayan ainsi que l'analyse de l'imagerie c-Fos. Les détails de cette analyse ainsi que les réseaux neuronaux révélés sont donnés dans BABAYAN [2014]. Ce chapitre concerne principalement les résultats de modélisation, c'est-à-dire les choix de formalisme effectués et les résultats de simulation du comportement d'un agent dans un labyrinthe. Toutefois, certains résultats de l'imagerie c-Fos seront décrits à la fin de ce chapitre. En effet, B. Babayan a effectué un

travail de corrélation de l'activité c-Fos [1] avec les paramètres libres d'un des modèles d'apprentissage par renforcement que nous avons optimisés. Par souci de clarté, ce décodage des modèles vers des observations biologiques sera aussi abordé à la fin de ce chapitre et nous nous tournons maintenant vers le travail de modélisation en commençant par décrire la tâche.

6.2 Tâche de navigation chez la souris

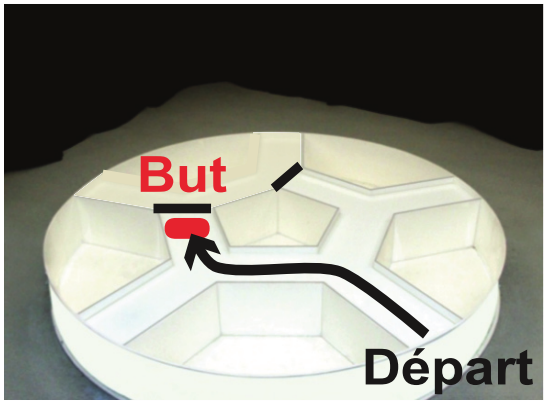

FIGURE 6.1 – Les souris doivent apprendre un chemin composé de deux intersections pour trouver une plateforme immergée. La position de la plateforme et la position de départ ne changent pas au cours de l'expérience. Reproduit de BABAYAN [2014]

Dans cette tâche, des souris doivent nager dans un labyrinthe en double Y pour trouver une plateforme immergée. Une image du labyrinthe est donnée dans la figure 6.1 avec la position de la plateforme et la position de départ qui restent constantes pour tous les essais.

Durant la première phase d'exploration, les souris sont naïves vis-à-vis du but de la tâche. Pour atteindre la phase d'exploitation, les souris doivent être capables de suivre le chemin optimal composé de deux points d'intersection menant à la plateforme. Au total, ce sont 30 souris qui ont été entraînées jusqu'à 6 jours avec 16 essais par jour. Pour être considérée comme ayant appris la tâche, une souris doit être capable d'avoir 75% des essais corrects pendant un jour d'entraînement et une performance parfaite sur 4 sessions le lendemain (voir 6.2.A). Ce critère permet d'assurer l'homogénéité dans le niveau de maîtrise de la séquence apprise. Au final, ce sont 15 souris qui ont atteint ce critère. Ces souris sont divisées en 3 groupes selon le temps qu'il leur a fallu pour maîtriser la tâche (4 jours, 5 jours ou 6 jours). Néanmoins, ces groupes ont un niveau de performances comparable sur les deux derniers jours d'apprentissage (voir la distance parcourue dans le labyrinthe de la figure 6.2.B).

Comme le montre la figure 6.2.A (à gauche), un autre groupe de souris a été placé dans le labyrinthe sans toutefois avoir le temps d'apprendre à résoudre la tâche. Ce groupe «exploratoire» sert de contrôle aux analyses c-Fos effectuées par B. Babayan. Il permet de soustraire l'activité motrice et perceptuelle propre à la tâche pour ne conserver que le réseau de structures spécifiques à l'apprentissage. Dans la suite de ce chapitre, seul le groupe «exploitateur» est considéré pour le travail de modélisation.

1. C-Fos est une technique d'imagerie post-mortem qui sera explicité plus en détails dans la section 6.5.

Le dernier point important de la tâche est le guidage. Si la souris n'est pas capable de trouver la plateforme en moins de 60s, celle-ci est remise au point de départ et guidée vers la plateforme en suivant le chemin le plus court.

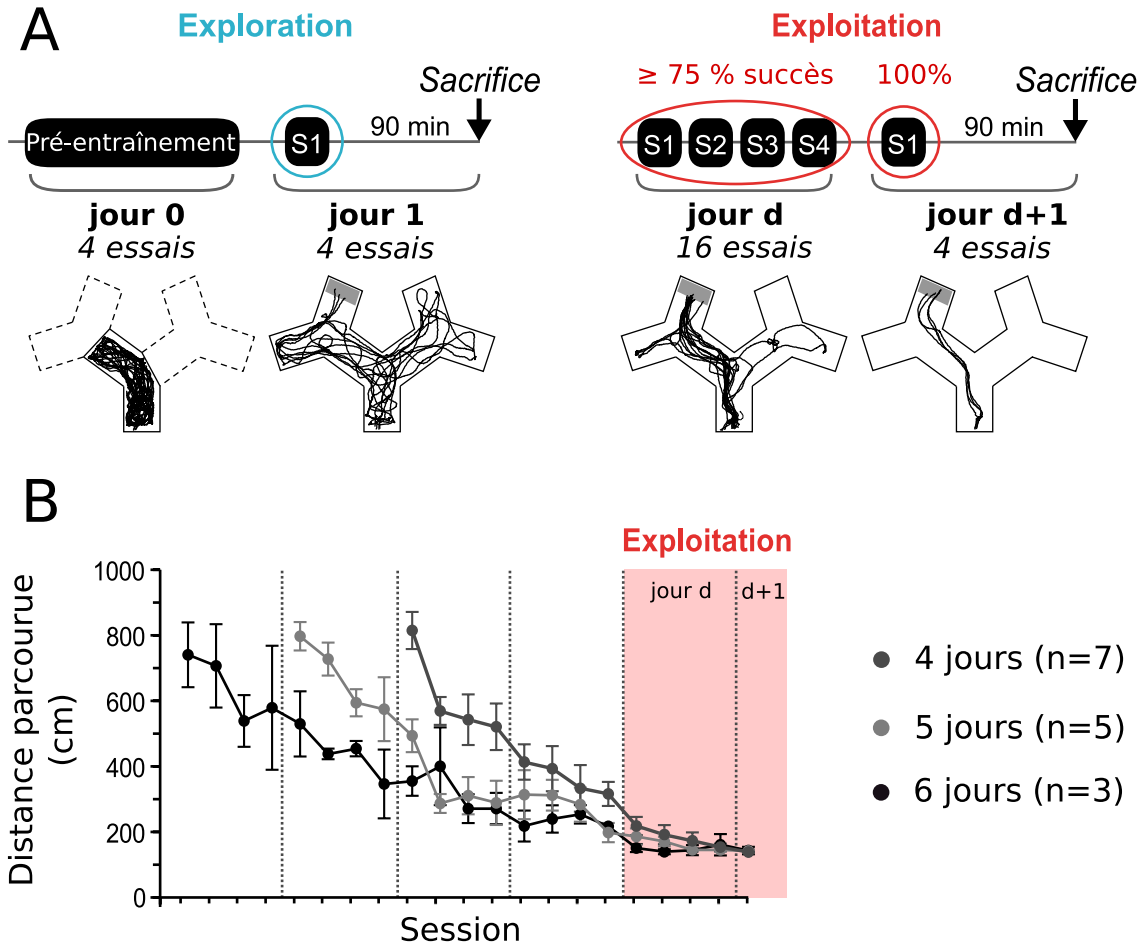

FIGURE 6.2 – A. A gauche, un groupe de souris «exploratoire» effectue une session de pré-entraînement ainsi qu'une unique session d'entraînement le jour suivant pour découvrir la tâche. La trajectoire d'une souris est montrée en exemple en-dessous. A droite, un groupe de souris «exploitateur» effectue 4 sessions par jour jusqu'à la performance requise de 75% de sessions correctes sur un jour d'entraînement et 100% de sessions correctes le jour suivant. La trajectoire d'une souris «exploitatrice» est aussi montrée en exemple en-dessous. B. La distance parcourue (cm) est représentée en fonction de la session pour 3 sous-groupes de souris ayant atteint le critère en 4 jours, 5 jours ou 6 jours. Reproduit de BABAYAN [2014]

6.3 Modèles théoriques

Le labyrinthe a été formalisé dans un processus de décision markovien à temps discrets selon :

— l'ensemble des états $\mathscr{S}$ = {Couloir (I), Cul-de-sac (u), Croisement (Y)} (c'est-à-dire un encodage du contour local du labyrinthe selon la position)

— l'ensemble des actions $\mathscr{A}$ = {Avant (F), Droite (R), Gauche (L), Demi-tour (U)}

— la fonction de récompense :

$$\mathscr{R} = \begin{cases} 1 & \text{si l'agent arrive sur la plateforme} \\ 0 & \text{sinon} \end{cases} \tag{6.1}$$

— la fonction de transition déterministe est construite selon la topologie du labyrinthe comme montré dans la figure 6.3.

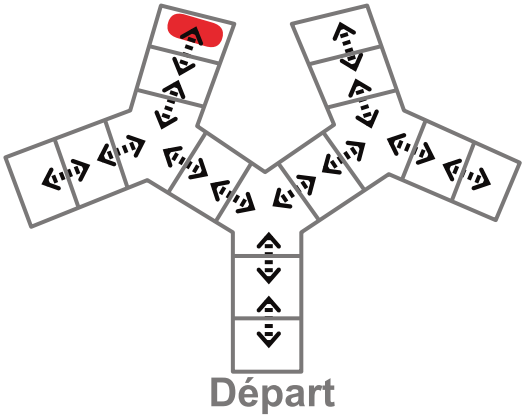

FIGURE 6.3 – Discrétisation du double labyrinthe en Y utilisé pour la modélisation du comportement des souris. Le départ et l'arrivée (en rouge) sont constants au cours de l'expérience. Chaque case représente un état identifiable par l'agent comme étant soit un couloir (I), une intersection (Y) ou un cul-de-sac (u). Les actions offertes à l'agent dans chaque case sont représentées par des flèches. Reproduit de BABAYAN [2014]

En fonction de la position de l'agent dans le labyrinthe, certaines actions ne sont pas proposées au processus de décision. L'ensemble des actions possibles est donc en fonction de l'état courant selon :

— $\mathscr{A}(\mathrm{I}) = \{\mathrm{F}, \mathrm{U}\}$

— $\mathscr{A}(u) = \{\mathrm{U}\}$

— $\mathscr{A}(\mathrm{Y}) = \{\mathrm{L}, \mathrm{R}, \mathrm{U}\}$

La position de la récompense est constante dans toute l'expérience.

Pour identifier le processus pouvant expliquer l'acquisition de séquences, nous avons testé la capacité de trois différents modèles d'apprentissage à capturer le comportement des souris. Nous avons ainsi comparé :

1. un algorithme d'apprentissage sur modèle
2. un algorithme d'apprentissage par différence temporelle
3. un algorithme d'intégration de chemin.

Le choix des deux premiers modèles est directement inspiré par la littérature sur la coordination des systèmes de mémoire comme présenté dans le chapitre 4. L'hypothèse ici est que l'apprentissage de séquences est sous-tendu dans le premier cas par la construction d'un modèle de l'environnement permettant d'inférer la valeur d'une action ou dans le deuxième cas par la mise à jour régulière d'une fonction de valeur. Pour finir, il est aussi possible que l'activité des cellules hippocampiques reflète une intégration de temps et de distance à partir du point de départ [MCNAUGHTON et collab., 2006]. Pour modéliser cet effet possible, nous avons aussi testé un algorithme d'intégration de chemin.

6.3.1 Apprentissage sur modèle

L'algorithme d'apprentissage sur modèle construit graduellement et par exploration un arbre de transitions internes des états rencontrés. Chaque transition reçoit ainsi une probabilité qui est ensuite utilisée pour planifier le chemin le plus court en utilisant un

algorithme de recherche dans un graphe [MARTINET et collab., 2008]. Les états peuvent apparaître dans plusieurs positions du labyrinthe. Néanmoins, les états se différencient entre eux par leurs relations de succession dans le graphe.

Les règles de la construction du graphe sont les suivantes :

1. Au début d'un essai, l'agent considère sa position initiale dans le graphe comme étant toujours le même noeud N_0.
2. Quand l'agent quitte un noeud N_k et essaye une nouvelle action a_i jamais essayée auparavant, un nouveau noeud N_{m+1} est ajouté aux m noeuds déjà existants.
3. Si l'action demi-tour U est choisie, l'agent ne retourne pas aux noeuds précédents mais ajoute un nouveau noeud au graphe.
4. Si l'agent arrive dans un nouveau noeud N_{m+1} contenant la plateforme, la valeur du noeud est changée selon $R(N_{m+1}) = 1$. La valeur par défaut est 0.

Un exemple de construction du graphe est donné dans la figure 6.4 pour deux essais consécutifs. Au troisième pas de temps du deuxième essai, l'agent choisit une action différente ce qui entraîne la construction d'une nouvelle branche dans le graphe.

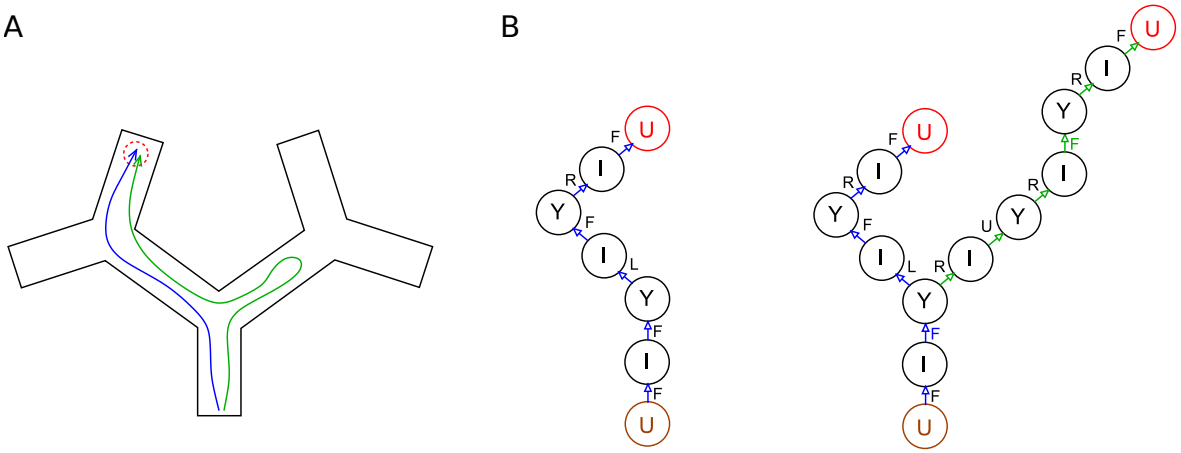

FIGURE 6.4 – A. Deux trajectoires successives de l'agent (bleue puis verte) dans le labyrinthe. B. Le graphe correspondant se construit progressivement. A la première intersection de la deuxième trajectoire, l'agent choisit l'action d'aller à droite ce qui entraîne la création d'une nouvelle branche dans l'arbre des transitions qui le mènera finalement à la récompense. Néanmoins, la propagation de la valeur à partir du noeud récompensant favorisera l'autre branche puisque son noeud récompensant est plus proche.

Lors de la création d'une transition dans le graphe, une probabilité de transition est assignée selon :

$$T(N_k, a_i, N_{m+1}) = \eta \tag{6.2}$$

Si la transition existe déjà, la probabilité de transition est mise à jour selon :

$$T(N_m, a_i, N_n) \leftarrow T(N_m, a_i, N_n) + \eta(1 - T(N_m, a_i, N_n)) \tag{6.3}$$

Le paramètre η représente la vitesse d'apprentissage du modèle.

En utilisant ce modèle de la topologie du labyrinthe, l'agent peut évaluer la valeur d'une action selon l'état courant en propageant l'information sur la position de la récompense. Une valeur V est ainsi assignée à chaque noeud en fonction de sa distance avec le noeud récompensant selon un facteur d'atténuation γ. Cette itération de la valeur est répétée jusqu'à convergence pour tous les noeuds selon :

$$V(N) \leftarrow max\big(R(N), V(N), max_i(\gamma T(N_m, a_i, N_n)V(N_n)\big) \tag{6.4}$$

Chaque action reçoit aussi une q-valeur $Q(N_m, a_i)$ selon le noeud courant N_m. La valeur de l'action est calculée selon :

$$Q(N_m, a_i) = \begin{cases} 0 & \text{si l'action n'a jamais été essayée} \\ \gamma T(N_m, a_i, N_n) V(N_n) & \text{sinon (avec } N_n \text{ le noeud suivant)} \end{cases} \quad (6.5)$$

La sélection finale est effectuée grâce à l'équation du soft-max. Ce modèle est caractérisé par trois paramètres :

1. η la vitesse d'apprentissage
2. γ le facteur d'atténuation de la valeur
3. β le compromis exploitation-exploration du soft-max.

6.3.2 Apprentissage par différence temporelle

Le modèle d'apprentissage par différence temporelle utilisé dans ce chapitre est un acteur-critique standard (cf chapitre 3.3.4). Néanmoins, le principal problème de ce processus d'apprentissage est l'absence d'indices externes dans l'environnement pour distinguer les intersections du labyrinthe. Dans la version classique, l'apprentissage associera toutes les intersections avec la même action parce qu'il n'y a aucune information sensorielle qui permette de distinguer les 3 intersections du labyrinthe comme ne correspondant pas à un seul et même état. Or, la première intersection nécessite de tourner à gauche et la seconde intersection nécessite de tourner à droite. Cette limitation a été confirmée dans une simulation du modèle classique d'acteur-critique qui s'est révélé incapable d'atteindre le niveau de performances des souris (voir figure 6.5). Comme solution, nous proposons de lever l'ambiguïté en augmentant la dimensionnalité de l'ensemble des états pour ce modèle. Cette augmentation de la dimensionnalité a déjà été proposée dans plusieurs approches différentes [Lin et Mitchell, 1992; McCallum, 1995; Zilli et Hasselmo, 2008]. Dans Zilli et Hasselmo [2008], l'étude concerne précisément une inclusion de mémoire épisodique dans le formalisme d'apprentissage par renforcement et propose une «factorisation» des états $S = S_1 \times S_2 \times \ldots \times S_n$.

Dans notre cas, l'augmentation de la dimensionnalité va se faire différemment. Un état s_t est construit par adjonction à l'entrée «sensorielle» $S \in \{I, u, Y\}$ d'une mémoire des n actions passées $a \in \{F, L, R, U\}$ selon :

$$s_t = (S_t, a_{t-1}, \ldots, a_{t-n}) \quad (6.6)$$

Par la suite, nous considérerons que $n = 3$ pour toutes les souris et nommerons TD-3 un tel modèle en opposition à TD-0 le modèle sans mémoire. Avec une telle mémoire, la première intersection peut être distinguée de la seconde puisque seule la seconde intersection est précédée par un tournant. Le fait de fixer la taille de la mémoire épisodique nous permet aussi de ne pas considérer la pénalisation des modèles en fonction de la complexité car tous auront ainsi le même nombre de paramètres.

A chaque pas de temps, l'agent construit son propre état s_t qui est ensuite évalué selon les équations d'apprentissage par renforcement standard :

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (6.7)$$

$$V(s_t) \leftarrow V(s_t) + \eta \delta_t \quad (6.8)$$

Comme dans tout acteur-critique, la différence temporelle δ_t est ensuite utilisée pour mettre à jour la distribution de probabilités d'action selon :

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \eta \delta_t \quad (6.9)$$

La sélection de l'action finale est aussi effectuée selon un soft-max. Les paramètres de ce modèle au nombre de 3 sont donc :

1. η la vitesse d'apprentissage de l'acteur et du critique
2. γ le facteur d'atténuation de la valeur
3. β le compromis exploitation-exploration du soft-max.

6.3.3 Intégration de chemin

Le modèle d'intégration de chemin maintient à chaque pas de temps une estimation probabiliste de la position $P_t(x,y)$ de l'agent durant ses déplacements en fonction de son point de départ. Nous sortons ici du cadre discret propre au processus de décision markovien que nous avons utilisé pour les deux autres modèles. La position probabiliste $P_t(x,y)$ est modélisée grâce à une distribution gaussienne en 2 dimensions $(x,y) \in \mathbb{R}^2$ centrée sur la position de l'agent. Pour modéliser l'accumulation d'erreurs intrinsèques à l'intégration de chemin, la déviation standard σ augmente progressivement à chaque pas de temps selon :

$$P_t(x,y) = \mathcal{N}\left((x,y), t \times \sigma_o\right) \tag{6.10}$$

La position de la plateforme est représentée initialement sous la forme d'une distribution uniforme de l'espace en deux dimensions. Cette distribution est ensuite mise à jour à chaque fois que la plateforme est atteinte en utilisant l'estimation courante de la position $P_t(x,y)$ selon :

$$P_{but} \leftarrow (1-\eta)P_{but} + \eta P_t(x,y) \tag{6.11}$$

Une estimation précise de la position courante permet ainsi de positionner finement la plateforme dans l'environnement. Au contraire, une estimation imprécise de la position ne modifiera pas beaucoup la distribution uniforme initiale de la position de la plateforme.

Le processus de décision est assez fastidieux d'un point de vue méthodologique. En effet, l'agent ne possède qu'une estimation de sa position courante. Le choix de l'action doit donc se faire en fonction de toutes les positions possibles dans l'environnement. Par souci de simplicité d'implémentation de l'algorithme, nous avons donc choisi de discrétiser les distributions de probabilités dans une grille de 30 par 30 cases, ce qui représente 900 positions à évaluer pour chaque direction possible. Ces directions possibles α sont calculées dans chaque position discrète du labyrinthe en fonction des actions offertes à l'agent. La valeur d'une direction α est calculée selon :

$$P_t(direction = \alpha) = \frac{\sum_{(P_{but},P_t)\in\alpha} P_t(x,y)P_{but}(x,y)}{\sum_{(P_{but},P_t)} P_t(x,y)P_{but}(x,y)} \tag{6.12}$$

Concrètement et dans les faits, le nombre d'actions possibles détermine une séparation de l'environnement en autant de cadrans d'angles égaux de directions fixes par rapport au repère cartésien. Par exemple, un état Y offre toujours trois actions possibles, ce qui correspond à 3 cadrans (à noter que les cadrans ne sont pas constants pour un état ; le premier Y du labyrinthe fournira une division en cadrans différente des deux autres Y). Pour calculer la valeur d'un cadran (et donc de l'action associée), l'algorithme va parcourir toutes les cases de l'environnement en multipliant à chaque fois la probabilité P_t d'être dans la case pour l'agent avec la somme des probabilités P_{but} des cases incluses dans le cadran. Une sommation de toutes ces sous-valeurs associées à une case permet de déterminer la valeur de l'action et la division dans l'équation 6.12 permet de normaliser en probabilités d'action. Pour finir, la sélection de l'action est effectuée avec un soft-max.

Ce modèle possède donc trois paramètres :

1. σ_0 permettant d'initialiser la déviation standard de la position
2. η la vitesse d'apprentissage de la distribution de probabilité de la récompense
3. β le compromis exploitation-exploration du soft-max.

6.3.4 Méthodes pour comparaison de modèles

Les techniques de comparaison de modèles sont les mêmes que dans la section 5.2.4 à l'exception près que nous avons considéré seulement 2 objectifs pour le processus d'optimisation des paramètres : les choix des souris et la diversité dans les solutions proposées par l'algorithme d'optimisation. Les choix des souris ont ainsi été obtenus en discrétisant la trajectoire réelle des animaux selon la division en cases de la figure 6.3. Les paramètres pour chaque modèle ainsi que les limites autorisées pour l'optimisation sont regroupés dans le tableau 6.1.

Modèle	**Symbole**	**Limites**	**Description**
Graphe de transition	η	$0 < \alpha < 1$	Taux d'apprentissage
	β	$0 < \beta < 200$	Température du soft-max
	γ	$0 < \gamma < 1$	Facteur d'atténuation
Acteur-critique	η	$0 < \eta < 1$	Taux d'apprentissage
	β	$0 < \beta < 200$	Température du soft-max
	γ	$0 < \gamma < 1$	Facteur d'atténuation
Intégration de chemin	η	$0.001 < \eta < 1.0$	Taux d'apprentissage
	β	$0 < \beta < 200$	Température du soft-max
	σ_0	$0 < \gamma < 1$	Initialisation de la variance

TABLEAU 6.1 – Tableau des paramètres et des limites associées pendant l'optimisation pour les modèles d'apprentissage par graphe, le modèle d'acteur-critique avec mémoire des 3 actions passées et le modèle d'intégration de chemin.

Nous n'avons pas utilisé de fonction d'agrégation pour un front de Pareto comme au chapitre 5 puisque seuls les choix nous intéressent. Le meilleur modèle est donc celui qui maximise la vraisemblance que le modèle fasse les mêmes choix que les souris.

6.4 Résultats

6.4.1 Simulation du modèle TD-0

Le premier résultat de la figure 6.5 est la vérification que l'acteur-critique sans mémoire soit incapable de résoudre la tâche. Nous avons donc optimisé TD-0 pour chaque souris et sélectionné les paramètres qui maximisaient la vraisemblance que le modèle fasse les mêmes choix que la souris. Pour comparer le modèle avec le temps nécessaire (en secondes) pour trouver la plateforme en fonction des sessions, nous avons normalisé le nombre de transitions dans notre labyrinthe à cases avec la vitesse moyenne de nage des souris. La durée d'une transition en simulation est ainsi équivalente à 1.05s. Le guidage de l'agent s'effectue après 57 transitions dans le labyrinthe.

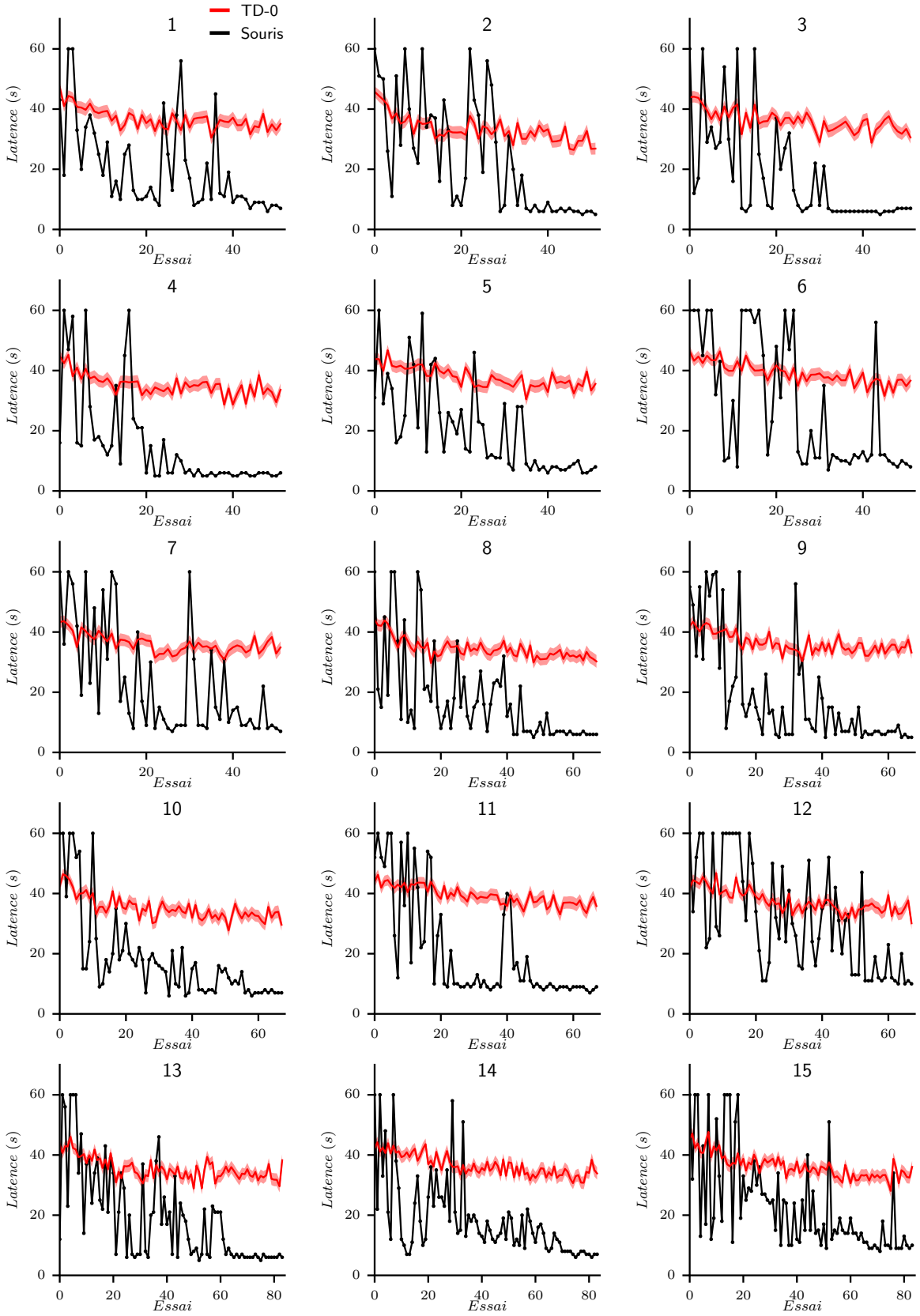

FIGURE 6.5 – Simulation du modèle TD-0 pour toutes les souris. La courbe d'apprentissage du modèle est moyennée sur 100 simulations.

Le résultat est sans appel et le modèle TD-0 ne progresse quasiment pas durant toute la tâche. Ce résultat conforte ainsi notre hypothèse sur la nécessité d'une mémoire des actions passées pour permettre aux modèles d'apprentissage par différence temporelle de résoudre la tâche.

6.4.2 Vraisemblance des modèles sur les choix

Pour comparer la capacité de chaque modèle à capturer le comportement des souris, nous avons donc optimisé la vraisemblance que les modèles fassent les mêmes choix que les souris. Les résultats sont présentés dans la figure 6.6 et montrent que le modèle TD-3 capture le comportement de 13 souris selon ce critère de vraisemblance. Pour les souris 7 et 14, le meilleur modèle est l'apprentissage sur modèle.

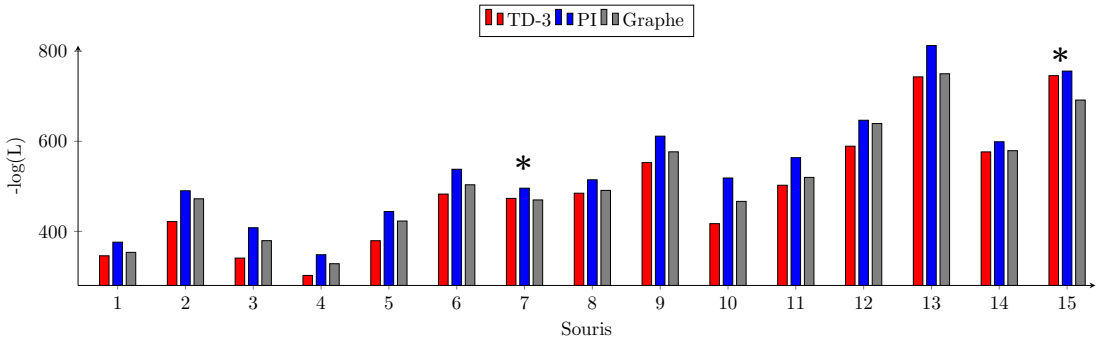

FIGURE 6.6 – Logarithme de la vraisemblance (-log(L)) que le modèle fasse les mêmes choix que la souris. L'optimisation de cette vraisemblance a été effectuée pour 15 souris ayant atteint le critère d'inclusion. TD-3 : modèle d'apprentissage par différence temporelle avec mémoire des 3 dernières actions passées. PI : modèle d'intégration de chemin. Graphe : modèle d'apprentissage par recherche dans un graphe. Le modèle TD-3 est le meilleur modèle pour toutes les souris à l'exception de 2 souris (marquées par une étoile) pour lesquelles le modèle Graphe est le meilleur.

6.4.3 Simulation des modèles optimisés

Pour chaque souris, nous avons ainsi identifié le meilleur paramétrage de chaque modèle. Tout comme la simulation du modèle TD-0, nous avons simulé chaque modèle optimisé pour chaque souris et les résultats sont présentés dans les figures 6.9, 6.7, 6.8.

La différence dans les principes d'apprentissages sous-tendant chaque modèle apparaît en comparant la progression des performances. Pour le modèle d'intégration de chemin (figure 6.7), il n'y a pas de progression dans l'apprentissage. Les performances restent constantes à partir du début de l'apprentissage. De fait, la raison est attribuée au guidage. A la fin du premier essai, il est peu probable que l'agent ait trouvé la récompense. Le guidage va ainsi placer dans le repère cartésien la distribution de probabilité associée au but ce qui permettra aux essais suivants de trouver la récompense directement. La seule solution pour l'optimisation des paramètres est de jouer sur le niveau de bruit du modèle. Cela permet d'obtenir des performances moyennes qui sont un compromis entre les performances disparates au début de l'apprentissage et les performances optimales en fin de tâche. Étant donné les résultats de simulation, il est peu probable que les souris utilisent une intégration de chemin (du moins telle que nous l'avons modélisée).

Le modèle de graphe semble montrer le même problème (figure 6.8). Son apprentissage est beaucoup trop rapide comparé à celui des souris. Une fois la récompense trouvée, la valeur est propagée dans tous les noeuds du graphe, ce qui assure à l'agent de trouver la plateforme en exploitant la fonction de valeur. Le même compromis semble effectué par l'optimisation des choix. Le modèle atteint très rapidement un temps de latence constant qui n'évolue pas et qui reste très loin des performances des souris à la fin de la tâche (à l'exception de la souris 2).

Pour finir, la simulation confirme les résultats de l'optimisation de la vraisemblance. L'acteur-critique avec mémoire des dernières actions est le meilleur modèle pour capturer l'apprentissage de séquence chez la souris (figure 6.9). Les performances évoluent pendant la simulation. A l'exception de certains cas (souris 6,7,15), le TD-3 atteint des performances optimales à la fin de la tâche. La fonction de valeur qui se met progressivement à jour en fonction de l'expérience de l'agent est ce qui semble capturer le mieux le comportement des souris.

Nous avons effectué une dernière vérification en calculant l'erreur des moindres carrés entre la courbe de performances moyennes simulées et la courbe de performances réelles pour chaque souris. Les résultats sont présentés dans la figure 6.10 et confirment que le modèle TD-3 est celui qui est le plus proche des performances des souris.

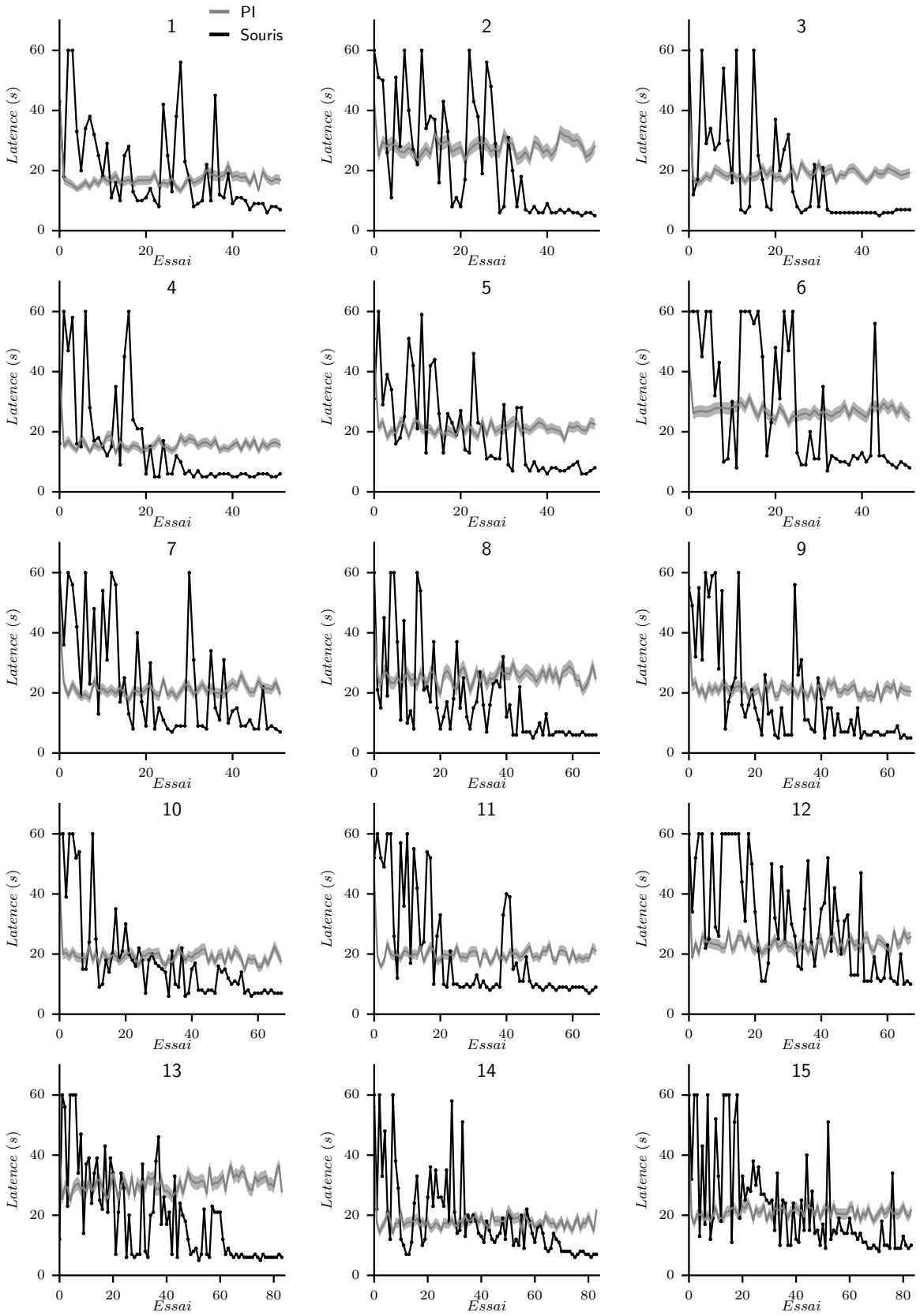

FIGURE 6.7 – En gris, simulation du modèle d'intégration de chemin (PI) pour toutes les souris. La courbe d'apprentissage du modèle est moyennée sur 100 simulations. En noir, la performance des souris pour chaque essai.

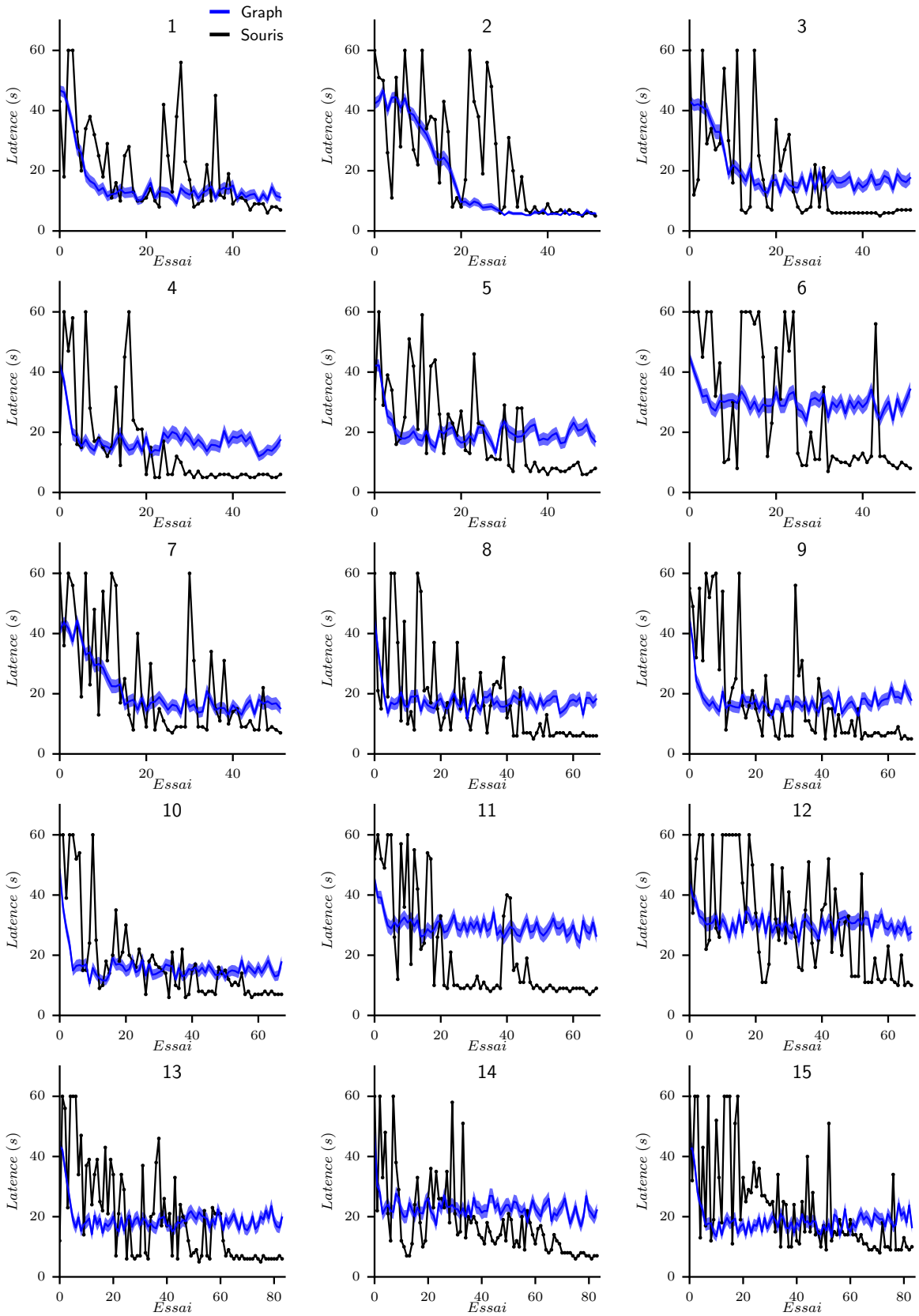

FIGURE 6.8 – En bleu, simulation du modèle d'apprentissage sur graphe pour toutes les souris. La courbe d'apprentissage du modèle est moyennée sur 100 simulations. En noir, la performance des souris pour chaque essai.

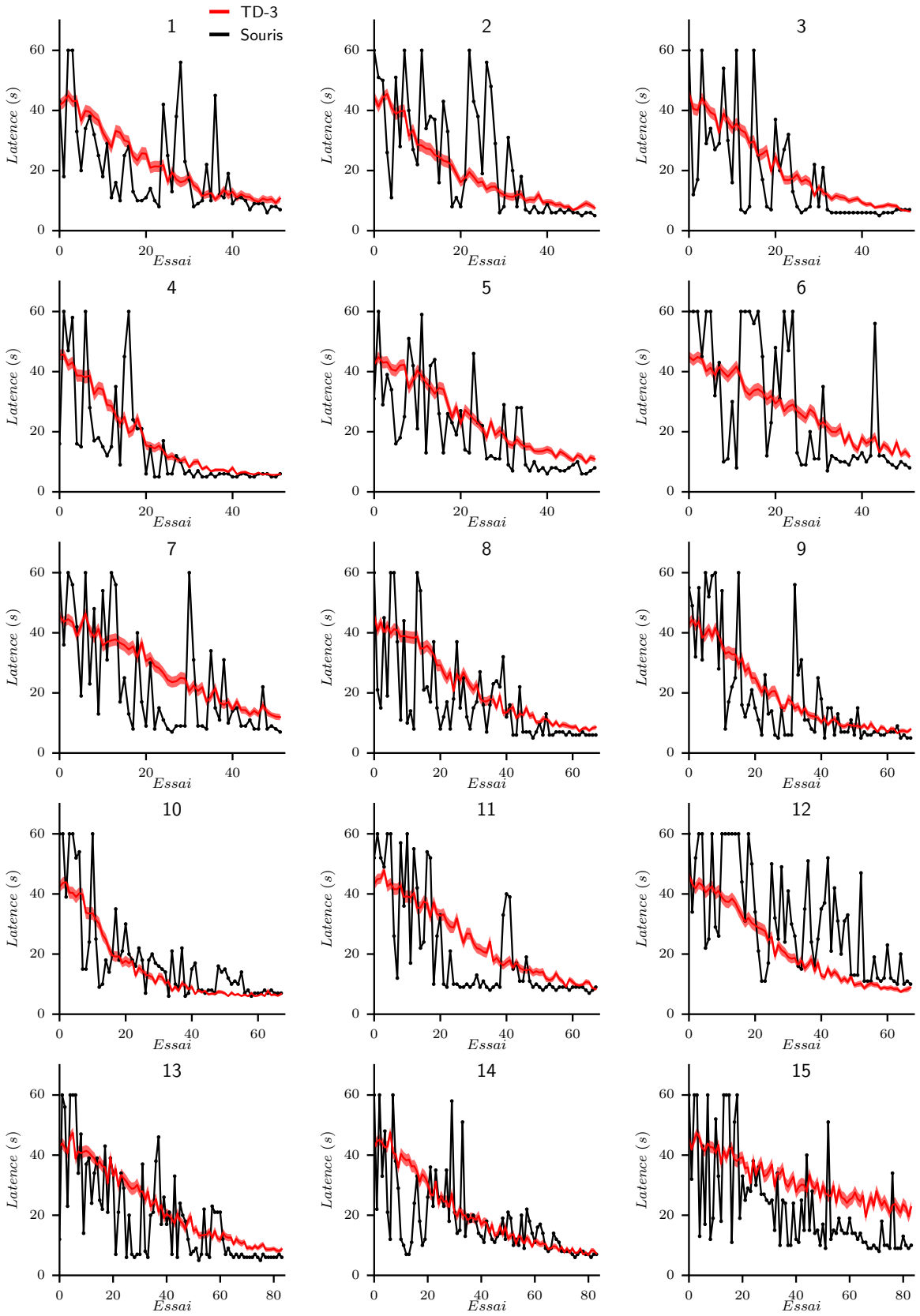

FIGURE 6.9 – En rouge, simulation du modèle d'apprentissage par différence temporelle avec mémoire des 3 dernières actions (TD-3) pour toutes les souris. La courbe d'apprentissage du modèle est moyennée sur 100 simulations. En noir, la performance des souris pour chaque essai.

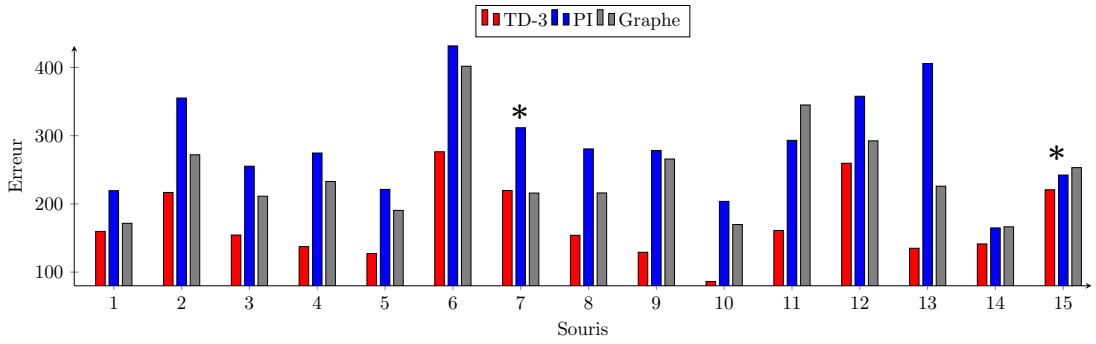

FIGURE 6.10 – Pour chaque modèle, l'erreur des moindres carrés a été calculée entre la performance simulée et la performance de la souris correspondante. Les souris 7 et 15 (marquées d'une étoile) ont reçu le modèle de graphe comme meilleur modèle selon la vraisemblance sur les choix. Néanmoins, on observe que pour la souris 15, le modèle TD-3 est plus proche selon l'erreur des moindres carrés.

6.5 Corrélation des paramètres avec l'activité c-Fos

Comme promis dans l'introduction, nous allons maintenant décrire les résultats de l'imagerie c-Fos effectuée par B. Babayan pour pouvoir ensuite présenter la corrélation des paramètres du modèle d'apprentissage par différence temporelle avec les structures activées.

La protéine c-Fos est un marqueur d'activité neuronale qui est exprimé dans le noyau des cellules. Dans BABAYAN [2014], ce marqueur a été utilisé pour identifier *a posteriori* les structures activées pendant une session. B. Babayan a ainsi comparé les structures activées dans 34 régions du cerveau des souris exploratrices et des souris exploitatrices (voir figure 6.2). Pour identifier l'activité c-Fos spécifique au contenu mnésique de la tâche et pour soustraire l'activité sensori-motrice, un groupe contrôle a aussi nagé à durée égale dans deux branches du labyrinthe sans plateforme. L'activité c-Fos a été normalisée en fonction de ce groupe de souris contrôle. Pour finir, un algorithme de regroupement markovien a été appliqué sur les matrices de corrélation de l'activité c-Fos normalisée. Cette procédure permet de générer un réseau des structures co-actives pendant la tâche. Les résultats sont présentés dans la figure 6.11.

Pour chaque groupe de souris, le réseau révélé est riche et complexe. Néanmoins, une évolution apparaît clairement et il semblerait que le réseau engagé en exploration subisse une réorganisation fonctionnelle majeure pendant l'apprentissage d'une séquence. L'observation la plus importante faite par l'analyse du réseau révèle l'émergence de l'hippocampe dorsal et des lobules IV/V du cervelet. Ces régions du cerveau sont identifiables aux noeuds centraux du réseau révélé. Dit autrement, ces deux régions sont des pivots du réseau, c'est-à-dire les noeuds par lesquels passent préférentiellement les chemins les plus courts permettant de relier un noeud du graphe à un autre. Il semblerait qu'il existe ainsi une communication importante avec le cervelet lors de l'exploitation d'une mémoire de séquence apprise dans une tâche de navigation.

Le travail de modélisation entre maintenant en scène. Nous avons cherché la présence d'éventuelles corrélations entre les variations de marquage c-Fos d'un individu à l'autre, pour chaque région mesurée, avec les variations individuelles des paramètres des modèles optimisés. Ainsi, B. Babayan a mesuré la covariance entre l'activité c-Fos des souris

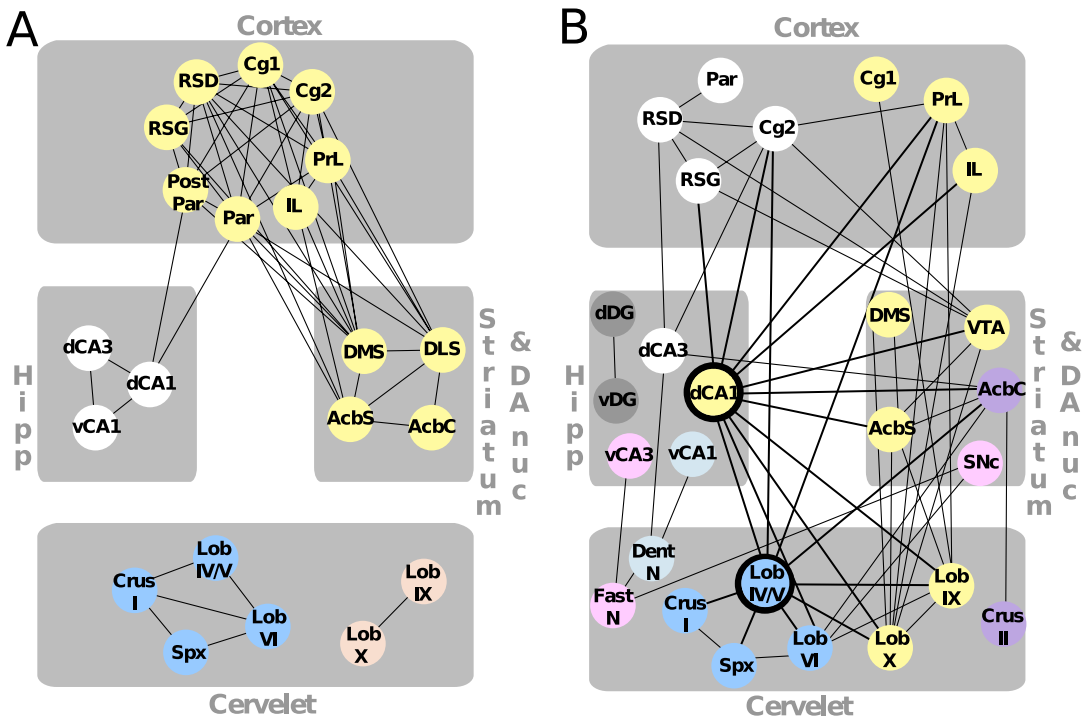

FIGURE 6.11 – A. Structures co-actives pour le groupe explorateur B. Pour le groupe exploitateur. Abréviations : cortex : auditif primaire (Au1), prélimbique (PrL), infralimbique (IL), cingulaire 1 et 2 (Cg1, Cg2), dysgranulaire et granulaire rétrosplénial (RSD, RSG), pariétal et pariétal postérieur (Par, PostPar), entorhinal médian (MEC) ; striatum et noyau dopaminergique (DA nuc) : striatum dorso-médian (DMS), striatum dorso-latéral (DLS), coeur du noyau accumbens (AcbC) et coquille (AcbS), aire tegmentale ventrale (VTA), substance noire pars compacta (SNc) ; hippocampe : CA1 dorsal (dCA1), CA3 dorsal (dCA3), CA1 ventral (vCA1), CA3 ventral (vCA3), CA2 dorsal (dCA2), gyrus denté dorsal et ventral(dDG, vDG) ; cervelet : lobules IV/V (Lob IV/V), VI (Lob VI), VII (Lob VII), IX (Lob IX), X (Lob X), Simplex (Spx), denté (Dent N), fastigial (Fast N) et noyau interpositus (IntP N).

exploitatrices et les paramètres des 13 souris auxquelles ce modèle a été assigné. Le facteur d'atténuation γ proche de 1 ne varie pas entre les individus et n'a pas été retenu pour la corrélation.

Au final, des corrélations ont été obtenues pour l'hippocampe et le cervelet uniquement. Plus spécifiquement, le CA1 dorsal, le CA3 ventral, les lobules IV/V, le Crus I et le noyau fastigial corrèlent positivement avec le taux d'apprentissage α comme montré dans la figure 6.12. En d'autres termes, un taux d'apprentissage élevé est corrélé avec une densité élevée d'activité c-Fos dans ces 5 structures. Pour le paramètre de compromis β, 4 structures sont corrélées négativement : le CA3 dorsal, le CA3 ventral, le noyau denté et le noyau fastigial comme le montre la figure 6.13. Ainsi, un β faible, donc un compromis orienté vers l'exploration, est corrélé à une densité c-Fos élevée dans ces 4 structures.

Pour résumer, seules les régions de l'hippocampe et du cervelet ont une activité c-Fos qui corrèle avec les paramètres du modèle d'apprentissage par différence temporelle. De plus, cette analyse de corrélation est indépendante de l'analyse de regroupement présentée au début de cette section. Cependant, les deux analyses révèlent que certaines régions de l'hippocampe et du cervelet, dont celles identifiées comme pivots du réseau, jouent un rôle non seulement dans l'apprentissage de séquences mais aussi dans la dynamique d'acquisition de ce comportement à travers le taux d'apprentissage et le compromis exploration-exploitation.

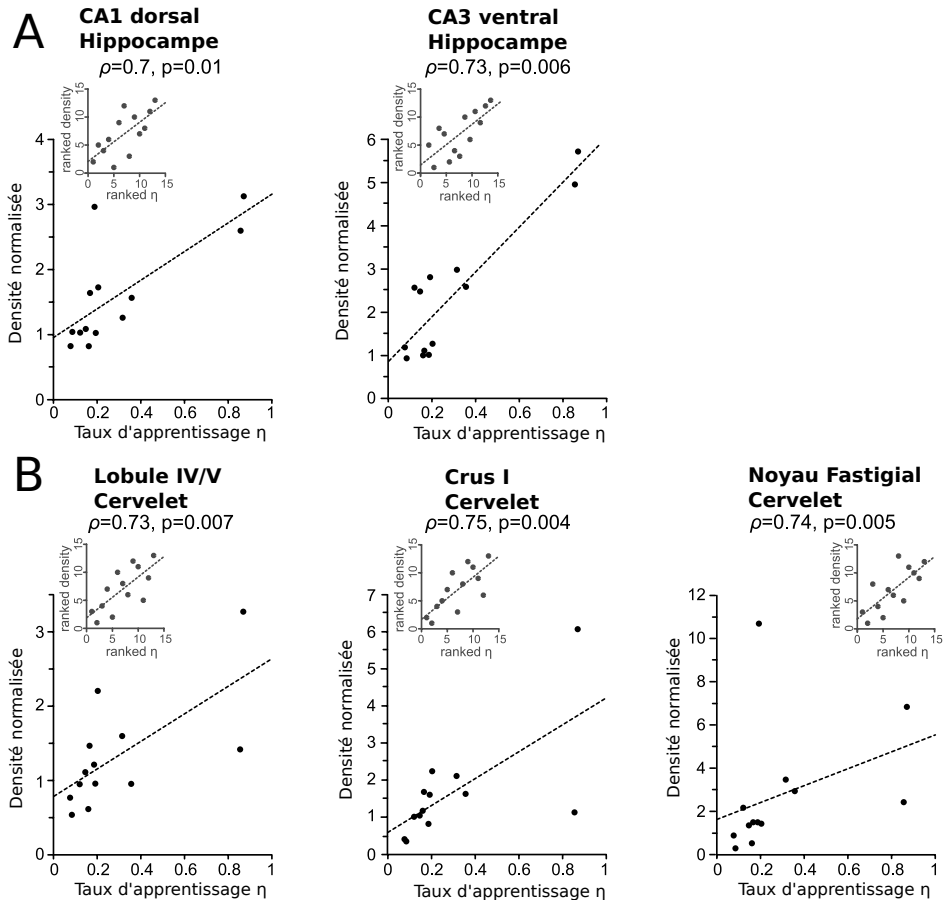

FIGURE 6.12 – Corrélation du taux d'apprentissage avec l'imagerie c-Fos. Seul l'hippocampe (A) et les structures du cervelet (B) ont montré des corrélations significatives entre la densité Fos et les taux d'apprentissage individuels estimés pour le modèle d'apprentissage par différence temporelle. Chaque figure montre la corrélation brute. La sous-figure montre la même corrélation avec les données ordonnées, qui permet de calculer la corrélation de Spearman.

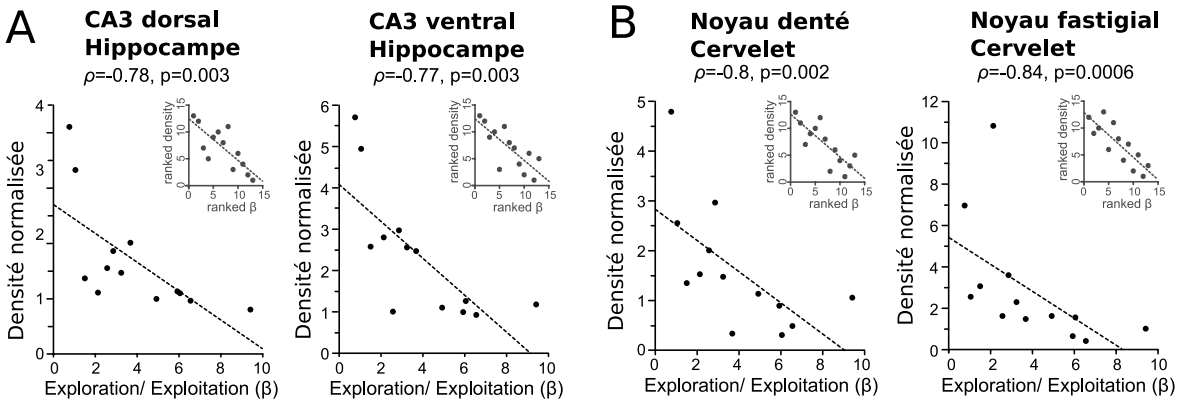

FIGURE 6.13 – Corrélation du compromis exploration/exploitation avec l'imagerie c-Fos. Seuls l'hippocampe (A) et les structures du cervelet (B) ont montré des corrélations significatives entre la densité c-Fos et les taux d'apprentissage individuels estimés pour le modèle d'apprentissage par différence temporelle. Chaque figure montre la corrélation brute. La sous-figure montre la même corrélation avec les données ordonnées, qui permet de calculer la corrélation de Spearman.

6.6 Conclusion

Nous avons proposé un formalisme combinant une mémoire épisodique avec un modèle d'apprentissage par différence temporelle que nous avons confronté avec succès à une tâche d'apprentissage de séquences chez la souris. Nous l'avons comparée avec un apprentissage sur modèle et une intégration de chemin.

La différence formelle entre les deux premiers modèles testés est issue directement de la proposition de DAW et collab. [2005] pour le conditionnement instrumental et reprise dans DOLLÉ et collab. [2010] pour la navigation. Dans une stratégie de lieux comme dans un comportement lié à un but, l'agent exploite un modèle de l'environnement pour calculer finement la valeur d'une action [DAW et collab., 2005; DOLLÉ et collab., 2010; KERAMATI et collab., 2011]. Il faut toutefois noter que les modèles de navigation cités sont utilisés dans des tâches nécessitant l'intégration d'indices visuels externes permettant des calculs complexes. Dans notre cas, l'apprentissage sur modèle se révèle trop rapide et trop efficace pour reproduire l'apprentissage lent des souris (tout comme l'est l'intégration de chemin qui n'évolue pas dans le temps).

Pour résumer, le besoin pour la souris d'effectuer plusieurs essais pour propager la valeur de récompense jusqu'à la position de départ s'adapte bien au principe de la mise à jour par différence temporelle. Néanmoins, notre modèle est différent des modèles d'habitudes équivalents au modèle stimulus-réponse en navigation présenté précédemment puisqu'il inclut une mémoire des actions passées. Exprimé autrement, la valuation de l'état dans les modèles classiques est remplacée par une valuation d'une séquence d'action. Les séquences d'action sont ensuites concaténées de telle sorte qu'une séquence d'action présente induit la transition vers une séquence d'action future selon le principe récursif propre aux modèles d'apprentissage par renforcement.

Ce résultat nous amène à la conclusion suivante : la planification d'une action future servie par un apprentissage sur modèle n'est pas nécessaire pour l'apprentissage de séquences dans un environnement sans indice externe. Cette absence de modèle du monde dans notre système formel remet en question l'implication de l'hippocampe comme énoncé dans l'introduction.

Pourtant, le décodage des paramètres libres du modèle TD-3 ainsi que la construction du réseau selon l'activité c-Fos présenté dans la section précédente convergent vers une interaction entre le cervelet et l'hippocampe comme siège possible de l'acquisition et de la restitution de séquences d'action.

L'apprentissage sur une séquence d'action que nous proposons est-il compatible avec les observations sur le fonctionnement de l'hippocampe comme système de mémoire ? Au niveau formel choisi (c'est-à-dire très loin d'une modélisation d'un réseau de neurones), nous ne pouvons que spéculer sur un possible lien. Toutefois, le CA3 dorsal qui corrèle avec le paramètre β d'exploration est communément perçu comme le lieu de stockage d'associations de toutes sortes qui permet ensuite de nourrir le CA1 d'où sont issues les séquences d'activation des cellules de lieux observées en électrophysiologie [LISMAN et collab., 2005]. Dans notre cas, il est possible que le dCA3 soit la structure qui contient la mémoire des actions passées confirmant ainsi le rôle de l'hippocampe comme structure organisatrice des évènements passés [SHAPIRO et collab., 2006].

Néanmoins, il existe aussi des différences majeures qui apparaissent si l'on considère les développements récents de la littérature sur les «simulations» de séquences hippocampiques discutés dans l'introduction de ce chapitre. Ainsi, certaines études ont montré des propriétés de ces répétitions de séquences hippocampiques qui sont hors de portée du formalisme classique de l'apprentissage par renforcement [DIBA et BUZSÁKI, 2007;

DRAGOI et TONEGAWA, 2013; GROSMARK et BUZSÁKI, 2016; SILVA et collab., 2015; WANG et collab., 2015]. Un exemple simple est donné dans VILLETTE et collab. [2015] avec l'observation que les séquences apparaissent de manière ordonnée dans l'obscurité et sans récompense fournie à l'animal. Si les cellules de lieux hippocampiques sont effectivement des états de l'apprentissage par renforcement, quel est l'avantage computationnel pour les états de former une chaîne par défaut ? Même en étant considérée comme séquence «par défaut» dans le système hippocampique, c'est une propriété qui n'apparaît pas naturellement dans le système formel. Les états dans un système acteur-critique (ou dans un algorithme d'apprentissage par renforcement classique) ne tendent pas à s'organiser spontanément.

Pour finir, le cervelet est généralement associé au contrôle moteur et à l'apprentissage supervisé, c'est-à-dire un système qui apprend d'un signal d'erreur [DOYA, 2000]. Une revue exhaustive de toutes les fonctions et observations liées au cervelet dépasse largement le sujet ici. Néanmoins, l'interaction entre l'hippocampe gauche et le Crus I du cervelet droit chez des sujets humains confrontés à une tâche de navigation en réalité virtuelle nécessitant l'apprentissage d'une séquence a déjà été observée dans IGLÓI et collab. [2014]. De plus, les auteurs de OHMAE et MEDINA [2015] ont montré récemment que le signal de l'olive inférieure envoyé au cervelet dans une simple tâche de stimulus-réponse était similaire à l'erreur de différence temporelle que nous avons présentée au chapitre 3 et qui est maintenant associée classiquement au signal dopaminergique dans le striatum [SCHULTZ et collab., 1997]. Pour conclure, la superposition du réseau révélé par l'imagerie c-Fos et des corrélations des paramètres libres du modèle d'apprentissage par différence temporelle suggèrent une implication majeure du cervelet dans la mise à jour de la fonction de valeur servant à l'apprentissage de séquences.

Chapitre 7

Discussions & Perspectives

Sommaire

7.1 Contributions

Après s'être frayé un chemin dans ce dédale de modèles, le lecteur peut se satisfaire du nombre de mémoires qu'il a fait travailler : mémoire sémantique, mémoire de travail, mémoire iconique... Pour autant, une rapide introspection sur ce jeu des mémoires n'offre aucun indice sur le processus de coordination. Est-ce de la fusion, de la sélection ou cette supposée coordination n'est-elle qu'une vue de l'esprit ? La difficulté de cette question justifie en soi la problématique développée ici. Néanmoins, l'entropie de la mémoire des premiers chapitres est sans doute forte. Alors soyons méthodiques et synthétisons les contributions de ce manuscrit.

Mémoire de travail et apprentissage par différence temporelle

La première contribution en réalité n'est pas liée spécifiquement à la problématique de ce manuscrit mais concerne un enjeu majeur pour les sciences modernes : la réplication des modèles computationnels. Dans VIEJO et collab. [2016], nous avons répliqué le modèle de sélection de stratégie en fonction du compromis vitesse-précision présenté dans KERAMATI et collab. [2011]. Nous avons ainsi montré la validité des résultats du modèle appliqué à une tâche de conditionnement instrumental chez le rongeur.

Dans la continuité de KERAMATI et collab. [2011] et en général des modèles suivant la proposition de DAW et collab. [2005] sur une coordination de l'apprentissage sur modèle et de l'apprentissage par différence temporelle, nous avons proposé une modélisation du comportement observé dans la tâche d'association visuo-motrice de BROVELLI et collab. [2008, 2011]. Etant donné la nature de la tâche (nécessitant une décision en fonction des actions passées), cela nous a conduit à proposer un modèle de mémoire de travail. Nous avons choisi une approche bayésienne pour formaliser ce modèle. Un élément en mémoire, décrivant un essai passé, est encodé sous la forme de fonctions probabilistes discrètes permettant des calculs inférentiels rapides au regard de la faible dimension de la tâche.

L'hypothèse centrale de ce modèle de mémoire de travail est l'utilisation de l'entropie d'information. A travers une inférence séquentielle et chronologiquement inverse des éléments en mémoire, la finalité de ce modèle est de réduire itérativement l'entropie contenue dans les probabilités d'action, donc de gagner de l'information, pour pouvoir décider. Les capacités limitées de la mémoire de travail se retrouvent dans un bruit uniforme qui est ajouté à chaque mise à jour et par une limitation du nombre d'éléments possiblement encodables. En simulant le modèle, l'évolution du nombre d'éléments en mémoire utilisés s'est rapprochée de l'évolution des temps de réaction des sujets, ce qui nous a conduit à proposer un temps de réaction proportionnel à la charge computationnelle de la mémoire de travail.

Les résultats précédant ce travail en neuro-imagerie de la tâche ont indiqué une activation du noyau caudé en phase d'acquisition (augmentation de la charge cognitive et ralentissement du temps de réaction chez les sujets) et une activation du putamen en phase de consolidation (diminution de la charge cognitive et accélération du temps de réaction) BROVELLI et collab. [2011]. Cette séparation des substrats neuronaux nous a ainsi conduit à opposer la mémoire de travail à une mémoire procédurale que nous avons modélisée sous la forme d'un q-learning. Ce choix rejoint celui de plusieurs études sur la coordination de stratégies qui toutes ont choisi l'apprentissage par différence temporelle comme mémoire procédurale lente à apprendre et inflexible qui pourrait correspondre à un apprentissage d'habitudes comportementales [CHAVARRIAGA et collab., 2005; COLLINS et FRANK, 2012; DAW et collab., 2005; DOLLÉ et collab., 2010; KERAMATI et collab., 2011].

Bien que chaque modèle puisse apprendre la tâche séparément, nous avons fait l'hypothèse d'une interaction entre la mémoire de travail et le q-learning. Cela nous a conduit à proposer un nouveau modèle de coordination par entropie. En fonction de la quantité d'information contenue dans les probabilités d'action du q-learning et les probabilités d'action de la mémoire de travail, le processus inférentiel propre à la mémoire de travail s'arrête pour échantillonner l'action ou continue pour gagner plus d'information. C'est ce jeu d'entropie qui permet la coordination des systèmes induisant une augmentation de la charge de calcul en début d'apprentissage et une inhibition de la charge de calcul en fin d'apprentissage du seul fait que le q-learning est un algorithme d'apprentissage lent.

Pour comparer avec la littérature, nous avons ensuite adapté le modèle de KERAMATI et collab. [2011] sur la sélection en fonction du compromis vitesse-précision et le modèle de COLLINS et FRANK [2012] proposant un mélange pondéré en fonction de l'historique de récompense de l'agent. Puis deux processus d'optimisation distincts (choix puis choix et temps de réactions) ont été réalisés en utilisant un algorithme d'évolution standard NSGA-2 [MOURET et DONCIEUX, 2010].

En optimisant les choix seulement, les modèles duaux se sont révélés les meilleurs (presque autant de mélange pondéré que de coordination par entropie). Toutefois, une correction de la vraisemblance en fonction de la complexité du modèle a permis aux modèles simples n'utilisant que le q-learning ou que la mémoire de travail de dominer l'adéquation aux choix. Ce résultat nous a conforté dans notre décision de capturer les choix et les temps de réaction. Après avoir construit des fronts de Pareto sur ces 2 objectifs, nous avons appliqué des fonctions de décision multi-critères permettant de sélectionner le meilleur modèle pour chaque sujet. Le résultat de l'optimisation double permet d'expliquer 9 sujets par le modèle de coordination, 3 sujets par le modèle de mélange pondéré et 2 sujets par le modèle de sélection. Après une vérification des paramètres par une simulation, la séparation des temps de réaction moyens en temps de réaction individuels nous a indiqué que nos modèles étaient capables de refléter l'évolution des temps de réaction des sujets.

Ce travail confirme que le comportement de sujets humains dans cette tâche peut s'expliquer le mieux et le plus parcimonieusement par une coordination dynamique entre deux systèmes de mémoire. Un travail préliminaire appliquant la même méthode à des données primates suggère qu'une coordination de systèmes de mémoire semble également en jeu dans ce cas.

Apprentissage de séquences avec mémoire épisodique

La troisième contribution concerne la caractérisation formelle de l'apprentissage de séquences chez la souris. Confrontées à un labyrinthe en double Y, les souris doivent apprendre une trajectoire stéréotypée sans aucun indice externe. De manière à capturer individuellement le comportement de 15 souris, nous avons optimisé en utilisant NSGA-2 les choix avec un apprentissage par différence temporelle (acteur-critique), un apprentissage sur modèle et une intégration de chemin. Pour donner sa chance au premier modèle d'apprendre dans un environnement sans indice externe, nous avons augmenté la dimensionnalité des états par une mémoire épisodique des 3 dernières actions. Ce choix s'est révélé payant puisque ce modèle capture au mieux l'apprentissage pour 13 souris (le comportement des deux autres souris est expliqué par l'apprentissage sur modèle).

Pour finir, ce travail de modélisation a bouclé sur le travail expérimental présenté dans BABAYAN [2014]. En analysant l'imagerie c-Fos des souris naïves et des souris ayant appris la tâche, B. Babayan a ainsi montré une réorganisation fonctionnelle majeure des struc-

tures actives après apprentissage. D'une activation striato-corticale au début de la tâche, le réseau se réorganise autour d'une activation cérébello-hippocampique à la fin de la tâche. De manière indépendante, nous avons testé les éventuelles corrélations entre les densités c-Fos de chaque individu et les paramètres de leur meilleur modèle. Nous avons également obtenu des effets significatifs pour diverses subdivisions de l'hippocampe et du cervelet, en particulier pour les noeuds centraux du réseau identifiés par la première méthode, renforçant ainsi ce résultat. Nous avons donc conclu que l'apprentissage de séquences (et son expression) est sous-tendu par une interaction entre l'hippocampe et le cervelet.

7.2 Critiques

Au regard des contributions de ce manuscrit, nous pouvons affirmer la nécessité d'une modélisation de systèmes de mémoire séparés ou fusionnés pour expliquer, dans certains cas, le comportement chez l'homme, le singe et la souris. Toutefois, des limites inhérentes à tout processus de modélisation apparaissent et seront maintenant discutées.

En premier, il semblerait qu'il existe une dissonance latente lorsque les deux chapitres de résultat sont mis en perspective et cela concerne le rôle de l'apprentissage par différence temporelle. Dans les deux cas, il ne semble pas pointer vers la même structure. En deuxième, nous discuterons succinctement des choix faits pour modéliser la mémoire de travail. En dernier, nous parlerons des progrès récents sur l'apprentissage de séquences dans l'hippocampe.

Du rôle de l'apprentissage par différence temporelle

Dans le chapitre 5, nous avons modélisé un comportement d'apprentissage lent procédural chez l'humain par un q-learning. Dans le chapitre 6, nous avons modélisé l'apprentissage de séquences chez la souris par un acteur-critique. Dans les deux cas, le processus en jeu est un apprentissage par différence temporelle. Pour autant, l'équivalence formelle implique-t-elle une équivalence naturelle ?

En revenant d'abord aux résultats de dissociation du striatum dans le chapitre 2, la conclusion était donc que le striatum dorso-médian est responsable du comportement lié à un but (en interaction avec l'hippocampe) et le striatum dorso-latéral est responsable du comportement habituel. Le putamen est l'équivalent humain du striatum dorso-latéral et a aussi été associé au comportement habituel en neuro-imagerie [TRICOMI et collab., 2009]. Dans BROVELLI et collab. [2011], le putamen s'active aussi pendant la phase de consolidation de l'association visuo-motrice. Si aucune dévaluation de la récompense n'a été effectuée pour mesurer la force de l'association stimulus-réponse, il est fort probable que cette activation ne reflète que le début de l'acquisition d'une habitude comportementale. Notre choix du q-learning se situe donc dans la ligne proposée par DAW et collab. [2005] et les études dopaminergiques de SCHULTZ et collab. [1997].

En étudiant l'acquisition de séquences chez la souris, le meilleur modèle s'est révélé aussi être l'apprentissage par différence temporelle sous la forme d'un acteur-critique. De même, aucune dévaluation n'a été effectuée pour mesurer le niveau d'habituation de la tâche. Néanmoins, les souris réalisent une suite d'actions très stéréotypées pendant plusieurs jours (entre 4 et 6 jours) ce qui laisse supposer une forme d'habitude dans l'enchaînement des actions motrices.

Le travail de corrélation des paramètres du modèle avec l'activation c-Fos et la construction du réseau de structures co-actives selon c-Fos convergent vers un réseau cérébello-

hippocampique. Le réseau striato-dorsal n'est impliqué qu'au début de l'apprentissage contrairement à ce qu'aurait prédit la théorie proposée par DAW et collab. [2005]. Récemment, les auteurs de OHMAE et MEDINA [2015] ont présenté un signal de différence temporelle dans le cervelet pendant une tâche de conditionnement pavlovien chez le rongeur. On peut ainsi supposer une localisation des deux parties de l'acteur-critique. Le cervelet serait responsable de l'évaluation de l'état et l'hippocampe soutiendrait la mémoire des actions passées. Ce rôle de l'hippocampe se conforme bien avec les études récentes sur la mémoire épisodique et l'apprentissage de séquences [LISMAN et collab., 2005; TULVING et DONALDSON, 1972]. La question est donc de savoir s'il existe des formes d'habitudes réparties dans plusieurs structures neuronales et, par extension, si le modèle d'habitudes est trop général.

Actuellement, les termes habitudes et apprentissage par différence temporelle sont utilisés de manière interchangeable dans certaines revues [DOLAN et DAYAN, 2013; DOLL et collab., 2015]. Considérant comme acquise cette équivalence, beaucoup d'études se concentrent ainsi sur le décodage du modèle de planification dans des régions variées du cortex [GLÄSCHER et collab., 2010; LEE et collab., 2014].

En opposition, une littérature se développe remettant cette équivalence en question. Dans DEZFOULI et BALLEINE [2012], la motivation des auteurs pour proposer un autre modèle d'habitudes est liée, en partie, à ce que nous avons observé dans la réplication de KERAMATI et collab. [2011] (voir figure 4.10). Après dévaluation et en laissant l'algorithme avancer, le modèle d'habitudes finit par désapprendre au contraire des observations comportementales en conditionnement instrumental où l'habitude persiste ce qui n'est pas en contradiction avec les données expérimentales qui suggèrent que le test en extinction est trop court pour observer un désapprentissage mais que si le test est plus long, on commence effectivement à observer un désapprentissage [DEZFOULI et BALLEINE, 2012]. Comme second argument, les auteurs discutent aussi de la propriété de certaines séquences d'actions habituelles A-B observées expérimentalement : B se déclenche automatiquement après A sans avoir besoin d'identifier le nouvel état après A, et souvent même en anticipation du changement d'état induit par A. Les méthodes de différence temporelle n'expliquent pas ce phénomène car elles ont besoin d'identifier un nouvel état avant de déclencher une nouvelle action. Pour se rapprocher au mieux des observations comportementales en dévaluation, les auteurs proposent un modèle de construction de séquences d'actions habituelles indépendantes des états et supervisées par la planification, mimant ainsi une hiérarchie de systèmes de mémoire. La séquence d'action permettant l'expression d'une habitude devient ainsi insensible à l'état courant et à la récompense.

Un deuxième exemple est donné dans MILLER et collab. [2016] avec la proposition de remplacer l'apprentissage par différence temporelle comme modèle d'habitudes par un apprentissage hebbien. Ainsi, le but de leur modèle est de «libérer» les habitudes du processus de maximisation de récompense propre à l'apprentissage par renforcement [1]. Pour ces auteurs, la distinction entre apprentissage sur modèle et apprentissage par différence temporelle appliqué au comportement animal ne constitue ainsi qu'une argutie.

Le point commun entre ces deux études est l'utilisation de la planification (reflétant supposément le comportement lié à un but) pour entraîner le modèle d'habitudes. Sur ce point précis, une proposition différente a été faite dans TOPALIDOU et collab. [2016] dans le but d'expliquer les résultats d'inactivation chez le singe de PIRON et collab. [2016]. Très brièvement, les auteurs ont montré qu'une inactivation du *globus pallidus* interne (GPi :

1. Ce qui les conduit à proposer le comportement habituel comme *value-free* en opposition au comportement lié à un but comme *value-based*.

la sortie du striatum vers le cortex) empêche l'apprentissage de nouvelles habitudes mais ne perturbe pas une habitude apprise auparavant (et pendant plusieurs mois). Pour capturer cette observation, les auteurs de TOPALIDOU et collab. [2016] proposent un modèle connectionniste du cortex et du striatum permettant une simulation de l'inactivation du GPi. Au niveau formel, l'habitude se situe dans le cortex associatif sous la forme d'un apprentissage hebbien qui ne dépend pas de la récompense mais juste du choix de l'action effectuée par le striatum qui lui implémente un algorithme d'apprentissage par différence temporelle. Comme les deux précédents modèles, c'est le comportement lié à un but qui conduit à la construction de l'habitude. Dans ce dernier cas, le comportement lié à un but est implicitement formalisé sous la forme d'un apprentissage par différence temporelle.

Pour résumer, il existe actuellement un faisceau d'indices permettant de questionner la proposition d'équivalence entre le comportement habituel et l'apprentissage par différence temporelle héritée de DAW et collab. [2005] et que nous avons appliquée dans ce manuscrit. Il semblerait que le système de mémoire sous-tendant le comportement habituel soit gradué et réparti dans plusieurs structures neuronales (striatum, cortex associatif ou cervelet). D'autres études seront ainsi nécessaires pour décrire formellement le comportement habituel dans toutes ses variations possibles [WOOD et RÜNGER, 2016].

Pour conclure, cette dernière proposition n'empêche pas le questionnement sur la capacité des algorithmes d'apprentissage par renforcement à capturer une «trop» grande variété de comportements et donc, *in fine*, à caractériser des systèmes de mémoire intrinsèquement distincts. Une position radicale déciderait ainsi de balayer ces modèles car trop généraux. Pour appuyer cette position, on peut citer SADACCA et collab. [2016] qui montre récemment que le signal dopaminergique pourrait se comporter selon un apprentissage par différence temporelle mais aussi selon un apprentissage sur modèle. A l'opposé, une position modérée rétorquerait que ces modèles de décision (planification, apprentissage par différence temporelle, apprentissage hebbien) coexistent dans le cerveau et sont identifiables à des systèmes de mémoire différents. Enfin, une position relative suggérerait que la généralité est inscrite dans la nature même des modèles d'apprentissage par renforcement en rappelant leur origine. La conceptualisation d'un modèle en système (agent) constitué d'états (perceptions) qui est sous influence d'un environnement (tâche) imprimant une force (récompense) déterminant une trajectoire (état suivant) à travers le pouvoir récursif (cf Chapitre 3) est une proposition issue directement de la physique classique. La finalité des modèles de la physique est leur «universalité» qui est un terme qui pourrait tout aussi bien s'appliquer à un acteur-critique ou à un q-learning (mais rarement à un système vivant). L'implication d'un modèle issu de la physique classique pour l'explication des systèmes biologiques présente des avantages et des inconvénients qu'il faut savoir différencier.

De la formalisation de la mémoire de travail

Notre modèle de mémoire de travail est différent des modèles d'apprentissage par renforcement puisqu'il ne maximise pas une récompense à long-terme mais minimise une incertitude à court-terme. Cette notion de minimisation d'incertitudes sous la forme d'entropie se retrouve dans des modèles de perception [NORWICH, 2003]. Elle se retrouve aussi dans des modèles théoriques d'explication du comportement avec minimisation de *free-energy* qui permet de lever l'ambiguïté sur les états cachés [FRISTON et collab., 2016]. A notre connaissance, nous n'avons pas trouvé d'utilisation d'entropie dans les modèles précédents de mémoire de travail. De plus, les modèles en psychologie se situent généralement à un niveau conceptuel qui permet rarement une simulation [BADDELEY, 2012].

Toutefois, des modèles de mémoire de travail existent comme cité dans l'introduction du chapitre 5 [FRANK et collab., 2001; ROUGIER et collab., 2005; TODD et collab., 2009]. Dans ces modèles, la mise à jour de la mémoire de travail est contrôlée par le signal de différence temporelle. Si la tâche utilisée dans notre cas ne s'y prête pas directement, une extension évidente et simple de notre modèle pourrait être faite dans cette direction. L'ajout de chaque élément à la liste serait conditionné par le signal d'erreur provenant du q-learning.

Pour finir, une dernière proposition d'extension de notre modèle serait de combiner la mémoire de travail à l'apprentissage sur modèle. On peut ainsi imaginer une tâche qui nécessite à la fois d'inférer à partir d'éléments en mémoire de travail et de planifier dans un modèle du monde. Une entropie commune permettrait ainsi de commencer dans la mémoire de travail puis de continuer dans le modèle du monde si l'incertitude est toujours forte. Des propositions similaires concernant la mémoire épisodique ont été faites récemment en plus de la dichotomie de l'apprentissage sur modèle et de l'apprentissage par différence temporelle [GERSHMAN et DAW, 2016; LENGYEL et DAYAN, 2007].

Des limites biologiques de l'acteur-critique pour l'apprentissage de séquences

Dans le chapitre 6, l'acteur-critique fusionné avec une mémoire des dernières actions s'est révélé être le meilleur modèle pour expliquer l'apprentissage de séquences chez la souris. Si cela a été vérifié par le processus d'optimisation, il est toutefois possible de remarquer quelques défauts en comparant les courbes de simulation avec les courbes d'apprentissage réel. Au début de l'expérience, les performances des souris sont clairement disparates et chaotiques. De manière à décharger l'acteur-critique de la responsabilité de capturer ce comportement initial, une proposition simple serait de le combiner avec un sélecteur d'action aléatoire tout comme DOLLÉ et collab. [2010], ce qui modéliserait très grossièrement le réseau fronto-striatal observé par B. Babayan en début de tâche. Une variable mesurant le taux de récompense permettrait ensuite de donner la responsabilité du choix de l'action à l'acteur-critique.

Néanmoins, ce dédoublement ne permettrait sans doute pas de réduire le paramètre d'atténuation γ qui a été exclu du travail de corrélation car identique pour toutes les souris et proche de 1. Cette imperfection dans le processus de modélisation s'explique toutefois par la nature de la tâche. Un γ élevé indique que l'agent a appris une structure dans la résolution de la tâche qui permet d'assurer des transitions fortes entre états [SUTTON et BARTO, 1998]. De fait, le labyrinthe en double Y offre une trajectoire très stéréotypée dans l'espace des états qui a de plus été augmenté par la mémoire des dernières actions.

Pour conclure, il existe des découvertes récentes sur l'apprentissage de séquences par l'hippocampe qui restent, pour l'instant, hors de portée de notre modèle. Ces découvertes sont résumables en une seule question : l'hippocampe prédit-il le futur [BENDOR et SPIERS, 2016; BUHRY et collab., 2011] ? Si cette question ne semble pas directement liée à notre sujet, le substrat neuronal impliqué est commun (du moins pour l'hippocampe, le cervelet ne semble pas avoir été impliqué). Une série d'études récentes [DRAGOI et TONEGAWA, 2013; GROSMARK et BUZSÁKI, 2016; OLAFSDOTTIR et collab., 2015; PFEIFFER et FOSTER, 2013] montrent ainsi que les cellules de lieux s'assemblent pour former des chaînes d'activation d'états passés, d'états futurs déjà explorés ou d'états futurs jamais explorés. Ces propriétés de l'hippocampe pour la navigation n'ont pas encore été adressées par des modèles courants d'apprentissage par renforcement. De plus, des limitations ont aussi été soulevées concernant les seuils statistiques de détection de ces événements [SILVA et collab., 2015]. D'autres développements seront ainsi nécessaires pour élucider

les fonctionnements de l'hippocampe dans l'apprentissage de séquences.

7.3 Conclusions

Ce manuscrit présente un travail de modélisation du comportement humain et animal. Notre hypothèse de départ est l'existence de systèmes de mémoire aux processus de rétention et de restitution de l'information distincts. Cette hypothèse a été confortée par plusieurs décennies d'observations de dissociation du substrat neuronal chez l'humain, le singe et le rongeur.

La théorisation, et plus précisément la simulation du comportement en fonction de systèmes de mémoire parallèles, constitue de fait un projet plus récent. Le vecteur de référence de ce projet est l'apprentissage par renforcement. Issue de l'intelligence artificielle, cette théorie nous a fourni des modèles initiaux à partir desquels nous avons pu proposer d'autres modèles de mémoire ainsi que des modèles possibles de coordination. Nous avons ensuite démontré leur capacité à capturer le comportement d'un groupe de sujets humains, d'un groupe de singes et d'un groupe de souris.

L'implication de la théorie de l'apprentissage par renforcement dans les neurosciences est aussi confortée par ses capacités de prédiction de l'activité neuronale. Nous souhaitons ainsi que les modèles présentés dans ce manuscrit puissent fournir une matrice de variables suffisante pour décoder à la fois le comportement mais aussi les processus mis en jeu lorsqu'un ou plusieurs systèmes de mémoire s'expriment.

Annexe A

Annexes

A.1 Kalman Q-Learning

Dans le Kalman q-learning, la fonction de valeur est modélisée comme un vecteur θ suivant une marche aléatoire et peut s'exprimer comme :

$$\begin{cases} \theta_i = \theta_{i-1} + \mu_i & \text{équation d'évolution} \\ r_i = g_{t_i}(\theta_i) + n_i & \text{équation d'observation} \end{cases} \tag{A.1}$$

avec mu_i et n_i les bruits d'observation de variance P_{v_i} et P_{n_i}. L'algorithme du Kalman q-learning s'écrit :

Algorithme A.1 : Kalman Q-Learning

Initialisation $\hat{\theta}_{0|0}$ $P_{0|0}$

répéter

Observer la transition :

$t_i = (s_i, a_i, s_{i+1})$

phase de prédiction

$\hat{\theta}_{i|i-1} = \hat{\theta}_{i-1|i-1}$

$P_{i|i-1} = P_{i-1|i-1} + P_{v_{i-1}}$

Calcul des sigma-points

$\Theta_{i|i-1} = \{\hat{\theta}^j_{i|i-1},\ 0 \geq j \geq 2p\}$

$W = \{w_j,\ 0 \geq j \geq 2p\}$

$R_{i|i-1} = \{\hat{r}^j_{i|i-1} = \hat{Q}_{\hat{\theta}^j_{i|i-1}}(s_i, a_i) - \gamma max_{b \in A} \hat{Q}_{\hat{\theta}^j_{i|i-1}}(s_{i+1}, b),\ 0 \geq j \geq 2p\}$

Calcul des statistiques d'intérêt

$\hat{r}_{i|i-1} = \sum_{j=0}^{2p} w_j \hat{r}^j_{i|i-1}$

$P_{\theta r_i} = \sum_{j=0}^{2p} w_j (\hat{\theta}^j_{i|i-1} - \hat{\theta}_{i|i-1})(\hat{r}^j_{i|i-1} - \hat{r}_{i|i-1})$

$P_{r_i} = \sum_{j=0}^{2p} w_j (\hat{r}^j_{i|i-1} - \hat{r}_{i|i-1})^2 + P_{n_i}$

Phase de correction

$K_i = P_{\theta r_i} P_{r_i}^{-1}$

$\hat{\theta}_{i|i} = \hat{\theta}_{i|i-1} + K_i(r_i - \hat{r}_{i|i-1})$

$P_{i|i} = P_{i|i-1} - K_i P_{r_i} K_i^T$

jusqu'à T

A.2 Fonctions d'agrégation

Une fonction d'agrégation permet de combiner des valeurs numériques $x_1, \ldots, x_m$ en une seule valeur $M(x_1, \ldots, x_m)$ de telle sorte que le résultat final de l'agrégation prenne en compte chaque valeur individuelle. Dans le problème du chapitre 5, une optimisation des paramètres propose un ensemble de solutions en 2 dimensions : une valeur mesurant l'adéquation aux choix et une valeur mesurant l'adéquation aux temps de réaction. Le but de la fonction d'agrégation est donc d'ordonner ces solutions selon une valeur unique.

De manière à pouvoir comparer les solutions, la première étape consiste à normaliser chaque valeur d'adéquation dans l'intervalle unitaire $[0,1]$. Pour normaliser, la borne supérieure et la borne inférieure sont choisies comme la meilleure et la pire valeur de la mesure d'adéquation.

Agrégation selon la distance euclidienne. Une première fonction d'agrégation consiste à mesurer la distance euclidienne de chaque solution à partir d'un point de référence $p \in \mathbb{R}^m$. Dans notre cas (le plus simple), la valeur de chaque solution est donc égale à $p = \mathbf{1}$. Ce point correspond à la limite supérieure de chaque dimension du problème à optimiser. La solution avec la distance euclidienne minimale est ainsi sélectionnée.

Agrégation selon la distance de Tchebychev. La première méthode peut être raffinée en utilisant la distance de Tchebychev. La valeur d'une solution est définie comme la distance à la cible en utilisant une norme infinie. Un vecteur de poids $\lambda \in \mathbb{R}^m_+$ est introduit pour biaiser le classement si certaines mesures d'adéquation sont plus importantes que d'autres. Le point de référence p est le point idéal $\alpha \in \mathbb{R}^m$ défini comme $\alpha_i = sup_{x \in \mathbb{X}} x_i$. Le point idéal est différent pour chaque front de Pareto. A l'inverse, on définit le Nadir comme étant la pire combinaison de scores comme $\beta_i = inf_{x \in \mathbb{X}} x_i$. Pour finir, la fonction d'agrégation est définie comme :

$$t(x) = \max_{i \in M} \lambda_i \frac{\alpha_i - x_i}{\alpha_i - \beta_i} + \epsilon \sum_{i=1}^{m} \lambda_i \frac{\alpha_i - x_i}{\alpha_i - \beta_i} \tag{A.2}$$

avec ϵ suffisament petit. Pour plus de détails sur cette méthode, le lecteur est invité à consulter WIERZBICKI [1986].

Agrégation selon une moyenne pondérée ordonnée. La dernière fonction d'agrégation testée est la moyenne pondérée ordonnée (*OWA*). A partir d'une permutation σ telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(m)}$ et un vecteur de poids $w = (w_1, \ldots, w_m)$, $w \in [0,1]$, la fonction d'agrégation est définie comme :

$$owa(x) = \sum_{i=1}^{m} w_i x_{\sigma(i)} \tag{A.3}$$

Les détails de cette fonction sont donnés dans YAGER [2004].

A.3 Figure annexe

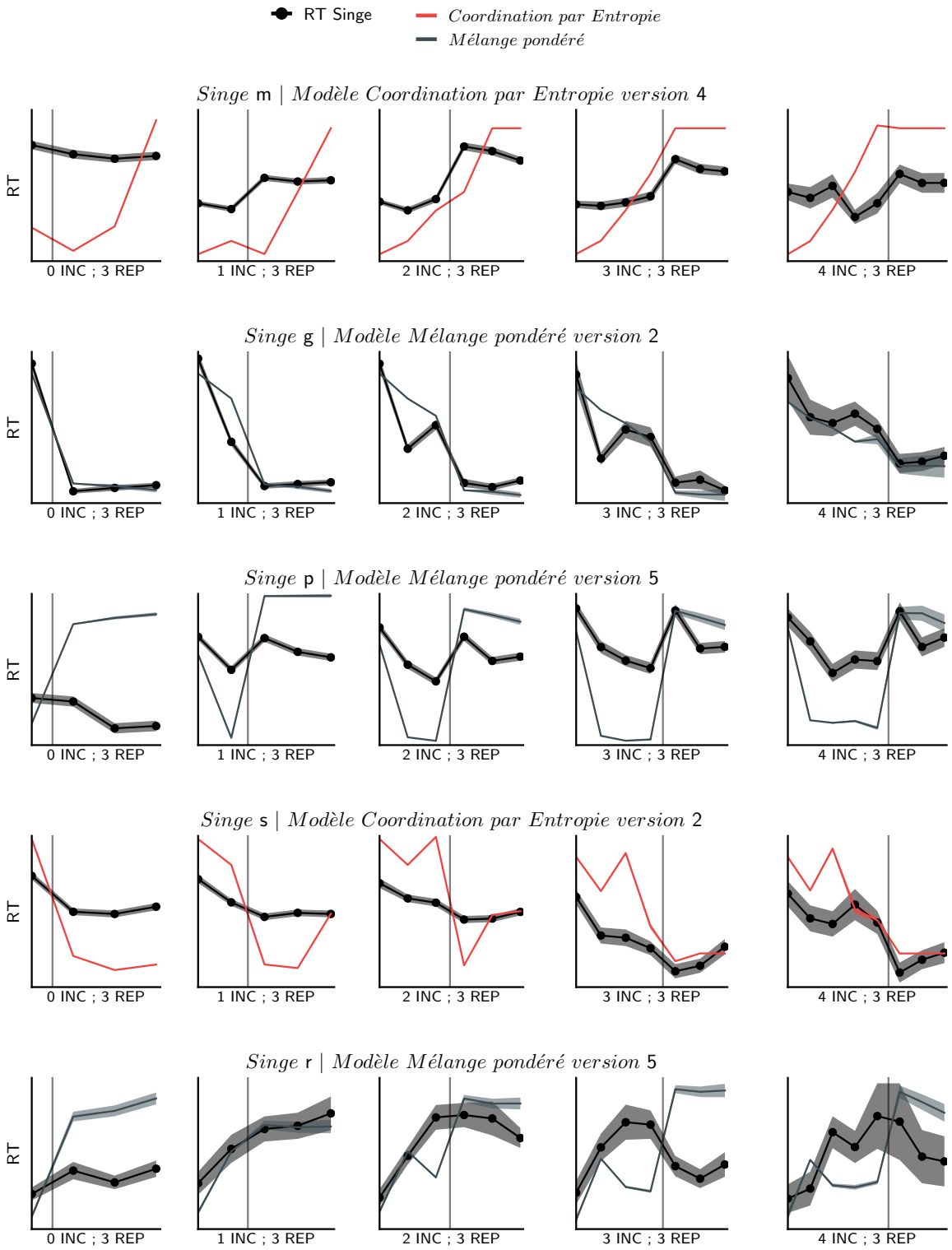

FIGURE A.1 – Simulation (moyenne ± erreur type) des temps de réaction contraints par la séquence de choix du singe. Pour chaque singe, la moyenne des temps de réaction est effectuée en séparant les problèmes en fonction du nombre d'essais dans la phase d'exploration. Les solutions sont issues d'une sélection selon l'opérateur de Tchebytchev appliquée aux fronts de Pareto représentés dans la figure 5.12.

Bibliographie

AMIEZ, C., J. SALLET, E. PROCYK et M. PETRIDES. 2012, «Modulation of feedback related activity in the rostral anterior cingulate cortex during trial and error exploration», *Neuroimage*, vol. 63, n° 3, p. 1078–1090. 97

ARLEO, A. et W. GERSTNER. 2000, «Spatial cognition and neuro-mimetic navigation : a model of hippocampal place cell activity», *Biological cybernetics*, vol. 83, n° 3, p. 287–299. 44, 46

ARLEO, A. et L. RONDI-REIG. 2007, «Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms», *Journal of integrative neuroscience*, vol. 6, n° 03, p. 327–366. 43

ASHBY, F., B. TURNER et J. HORVITZ. 2010, «Cortical and basal ganglia contributions to habit learning and automaticity», *Trends Cogn Sci*, vol. 14, n° 5, p. 208–215. 66

ATKINSON, R. C. et R. M. SHIFFRIN. 1968, «Human memory : A proposed system and its control processes», *Psychology of learning and motivation*, vol. 2, p. 89–195. 4, 10

BABAYAN, B. 2014, *Unraveling the neural circuitry of sequence-based navigation using a combined fos imaging and computational approach*, thèse de doctorat, Paris 5. iv, 102, 103, 104, 105, 116, 123

BADDELEY, A. 2012, «Working memory : theories, models, and controversies», *Annual review of psychology*, vol. 63, p. 1–29. 126

BADDELEY, A. D. et G. HITCH. 1974, «Working memory», *Psychology of learning and motivation*, vol. 8, p. 47–89. 9, 66

BALLEINE, B., M. DELGADO et O. HIKOSAKA. 2007, «The role of the dorsal striatum in reward and decision-making», *J Neurosci*, vol. 27, n° 31, p. 8161–8165. 66

BALLEINE, B. et J. O'DOHERTY. 2010, «Human and rodent homologies in action control : corticostriatal determinants of goal-directed and habitual action», *Neuropsychopharmacology*, vol. 35, n° 1, p. 48–69. 66

BALLEINE, B. W., N. D. DAW et J. P. O'DOHERTY. 2008, «Multiple forms of value learning and the function of dopamine», *Neuroeconomics : decision making and the brain*, vol. 36, p. 7–385. 40

BALLEINE, B. W. et A. DICKINSON. 1998, «Goal-directed instrumental action : contingency and incentive learning and their cortical substrates», *Neuropharmacology*, vol. 37, n° 4, p. 407–419. 21, 24, 43

BARTO, A. G., R. S. SUTTON et C. W. ANDERSON. 1983, «Neuronlike adaptive elements that can solve difficult learning control problems», *IEEE transactions on systems, man, and cybernetics*, , nº 5, p. 834–846. 38

BAYS, P. M. et M. HUSAIN. 2008, «Dynamic shifts of limited working memory resources in human vision», *Science*, vol. 321, nº 5890, p. 851–854. 66

BELLMAN, R. 1957, «Dynamic programming», *Princeton University Press*, p. 151. 31

BENDOR, D. et H. J. SPIERS. 2016, «Does the hippocampus map out the future ?», *Trends in cognitive sciences*, vol. 20, nº 3, p. 167–169. 127

BOTVINICK, M. et A. WEINSTEIN. 2014, «Model-based hierarchical reinforcement learning and human action control», *Philos Trans R Soc Lond B Biol Sci*, vol. 369, nº 1655. 66

BROVELLI, A., N. LAKSIRI, B. NAZARIAN, M. MEUNIER et D. BOUSSAOUD. 2008, «Understanding the neural computations of arbitrary visuomotor learning through fmri and associative learning theory», *Cerebral Cortex*, vol. 18, nº 7, p. 1485–1495. 59, 68, 69, 76, 91, 98, 122

BROVELLI, A., B. NAZARIAN, M. MEUNIER et D. BOUSSAOUD. 2011, «Differential roles of caudate nucleus and putamen during instrumental learning», *NeuroImage*, vol. 57, nº 4, p. 1580–1590. iii, 68, 69, 70, 91, 94, 98, 99, 122, 124

BROZOSKI, T. J., R. M. BROWN, H. ROSVOLD et P. S. GOLDMAN. 1979, «Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey», *Science*, vol. 205, nº 4409, p. 929–932. 67

BUCKNER, R. L., S. E. PETERSEN, J. G. OJEMANN, F. M. MIEZIN, L. R. SQUIRE et M. RAICHLE. 1995, «Functional anatomical studies of explicit and implicit memory retrieval tasks», *The Journal of Neuroscience*, vol. 15, nº 1, p. 12–29. 11

BUHRY, L., A. H. AZIZI et S. CHENG. 2011, «Reactivation, replay, and preplay : how it might all fit together», *Neural plasticity*, vol. 2011. 127

BURGESS, N., M. RECCE et J. O'KEEFE. 1994, «A model of hippocampal function», *Neural networks*, vol. 7, nº 6-7, p. 1065–1081. 44

CALUWAERTS, K., M. STAFFA, S. N'GUYEN, C. GRAND, L. DOLLÉ, A. FAVRE-FÉLIX, B. GIRARD et M. KHAMASSI. 2012, «A biologically inspired meta-control navigation system for the psikharpax rat robot», *Bioinspiration & biomimetics*, vol. 7, nº 2, p. 025 009. iii, 61, 64

CARPENTER, R., B. REDDI et A. ANDERSON. 2009, «A simple two-stage model predicts response time distributions», *J Physiol*, vol. 587, nº 16, p. 4051–4062. 75

CAVANAGH, J. F., M. J. FRANK, T. J. KLEIN et J. J. ALLEN. 2010, «Frontal theta links prediction errors to behavioral adaptation in reinforcement learning», *Neuroimage*, vol. 49, nº 4, p. 3198–3209. 59

CERMAK, L. S., M. VERFAELLIE et K. A. CHASE. 1995, «Implicit and explicit memory in amnesia : An analysis of data-driven and conceptually driven processes.», *Neuropsychology*, vol. 9, nº 3, p. 281. 10

CHANG, Q. et P. E. GOLD. 2003, «Intra-hippocampal lidocaine injections impair acquisition of a place task and facilitate acquisition of a response task in rats», *Behavioural brain research*, vol. 144, n° 1, p. 19–24. 20

CHAVARRIAGA, R., T. STRÖSSLIN, D. SHEYNIKHOVICH et W. GERSTNER. 2005, «A computational model of parallel navigation systems in rodents», *Neuroinformatics*, vol. 3, n° 3, p. 223–241. iii, 47, 48, 49, 51, 64, 122

CHOI, C.-Y., S.-R. HAN, G.-T. YEE et C.-H. LEE. 2010, «A understanding of the temporal stem», *Journal of Korean Neurosurgical Society*, vol. 47, n° 5, p. 365–369. 12

COHEN, N. J., H. EICHENBAUM, B. S. DEACEDO et S. CORKIN. 1985, «Different memory systems underlying acquisition of procedural and declarative knowledgea», *Annals of the New York Academy of Sciences*, vol. 444, n° 1, p. 54–71. 10

COLLINS, A. et E. KOECHLIN. 2012, «Reasoning, learning, and creativity : frontal lobe function and human decision-making», *PLoS Biol*, vol. 10, n° 3, p. e1001 293. 62

COLLINS, A. G. et M. J. FRANK. 2012, «How much of reinforcement learning is working memory, not reinforcement learning ? a behavioral, computational, and neurogenetic analysis», *European Journal of Neuroscience*, vol. 35, n° 7, p. 1024–1035. iii, 59, 60, 61, 63, 64, 68, 69, 76, 78, 81, 91, 99, 122, 123

CONWAY, A. R., N. COWAN et M. F. BUNTING. 2001, «The cocktail party phenomenon revisited : The importance of working memory capacity», *Psychonomic bulletin & review*, vol. 8, n° 2, p. 331–335. 66

COOLS, R. et M. D'ESPOSITO. 2011, «Inverted-u–shaped dopamine actions on human working memory and cognitive control», *Biological psychiatry*, vol. 69, n° 12, p. e113–e125. 67

CORKIN, S. 1968, «Acquisition of motor skill after bilateral medial temporal-lobe excision», *Neuropsychologia*, vol. 6, n° 3, p. 255–265. 10

CORKIN, S. 1984, «Lasting consequences of bilateral medial temporal lobectomy : Clinical course and experimental findings in hm», dans *Seminars in Neurology*, vol. 4, © 1984 by Thieme Medical Publishers, Inc., p. 249–259. 10

COUTUREAU, E. et S. KILLCROSS. 2003, «Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats», *Behavioural brain research*, vol. 146, n° 1, p. 167–174. iii, 21, 25

DAW, N. 2011, *Decision Making, Affect, and Learning : Attention and Performance XXIII*, chap. Trial-by-trial data analysis using computational models, Oxford University Press, p. 1–26. 81

DAW, N., S. GERSHMAN, B. SEYMOUR, P. DAYAN et R. DOLAN. 2011, «Model-based influences on humans' choices and striatal prediction errors», *Neuron*, vol. 69, n° 6, p. 1204–1215. 79

DAW, N. D., Y. NIV et P. DAYAN. 2005, «Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control», *Nature neuroscience*, vol. 8, n° 12, p. 1704–1711. iii, 41, 52, 53, 54, 58, 62, 63, 64, 68, 98, 119, 122, 124, 125, 126

DAW, N. D. et J. P. O'DOHERTY. 2013, «Multiple systems for value learning», *Neuroeconomics : Decision Making, and the Brain,*. 66

DEARDEN, R., N. FRIEDMAN et D. ANDRE. 1999, «Model based bayesian exploration», dans *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., p. 150–159. 53

DEARDEN, R., N. FRIEDMAN et S. RUSSELL. 1998, «Bayesian q-learning», dans *AAAI/IAAI*, p. 761–768. 53, 55

DEB, K., S. AGRAWAL, A. PRATAP et T. MEYARIVAN. 2000, «A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization : Nsga-ii», *Lecture Notes in Comput. Sci.*, vol. 1917, p. 849–858. 79

DECOTEAU, W. E., C. THORN, D. J. GIBSON, R. COURTEMANCHE, P. MITRA, Y. KUBOTA et A. M. GRAYBIEL. 2007, «Learning-related coordination of striatal and hippocampal theta rhythms during acquisition of a procedural maze task», *Proceedings of the National Academy of Sciences*, vol. 104, n° 13, p. 5644–5649. 27

DEVAN, B., R. MCDONALD et N. WHITE. 1999, «Effects of medial and lateral caudate-putamen lesions on place-and cue-guided behaviors in the water maze : relation to thigmotaxis», *Behavioural brain research*, vol. 100, n° 1, p. 5–14. 22

DEVAN, B. D. et N. M. WHITE. 1999, «Parallel information processing in the dorsal striatum : relation to hippocampal function», *The Journal of neuroscience*, vol. 19, n° 7, p. 2789–2798. 22, 23

DEZFOULI, A. et B. W. BALLEINE. 2012, «Habits, action sequences and reinforcement learning», *European Journal of Neuroscience*, vol. 35, n° 7, p. 1036–1051. 40, 125

DIBA, K. et G. BUZSÁKI. 2007, «Forward and reverse hippocampal place-cell sequences during ripples», *Nature neuroscience*, vol. 10, n° 10, p. 1241–1242. 119

DICKINSON, A. 1985, «Actions and habits : the development of behavioural autonomy», *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 308, n° 1135, p. 67–78. 21, 22, 43

DICKINSON, A. et B. BALLEINE. 1994, «Motivational control of goal-directed action», *Animal Learning & Behavior*, vol. 22, n° 1, p. 1–18. 21, 22

DICKINSON, A., B. BALLEINE, A. WATT, F. GONZALEZ et R. A. BOAKES. 1995, «Motivational control after extended instrumental training», *Animal Learning & Behavior*, vol. 23, n° 2, p. 197–206. 21

DOLAN, R. et P. DAYAN. 2013, «Goals and habits in the brain», *Neuron*, vol. 80, n° 2, p. 312–325. 125

DOLL, B. B., D. SHOHAMY et N. D. DAW. 2015, «Multiple memory systems as substrates for multiple decision systems», *Neurobiology of learning and memory*, vol. 117, p. 4–13. 66, 125

DOLLÉ, L., D. SHEYNIKHOVICH, B. GIRARD, R. CHAVARRIAGA et A. GUILLOT. 2010, «Path planning versus cue responding : a bio-inspired model of switching between navigation strategies», *Biological cybernetics*, vol. 103, n° 4, p. 299–317. iii, 49, 50, 51, 52, 61, 64, 119, 122, 127

DOYA, K. 2000, «Complementary roles of basal ganglia and cerebellum in learning and motor control», *Current opinion in neurobiology*, vol. 10, nº 6, p. 732–739. 120

DRAGOI, G. et S. TONEGAWA. 2013, «Distinct preplay of multiple novel spatial experiences in the rat», *Proceedings of the National Academy of Sciences*, vol. 110, nº 22, p. 9100–9105. 120, 127

EMROUZNEJAD, A. et M. MARRA. 2014, «Ordered weighted averaging operators 1988 - 2014 : A citation-based literature survey», *Int. J. Intell. Syst.*, vol. 29, nº 11, p. 994–1014. 80

FERBINTEANU, J. 2016, «Contributions of hippocampus and striatum to memory-guided behavior depend on past experience», *The Journal of Neuroscience*, vol. 36, nº 24, doi : 10.1523/JNEUROSCI.0840-16.2016, p. 6459–6470. URL http://www.jneurosci.org/content/36/24/6459.abstract. iii, 26, 27, 28

FLORESCO, S. B., O. MAGYAR, S. GHODS-SHARIFI, C. VEXELMAN et T. MARIC. 2006, «Multiple dopamine receptor subtypes in the medial prefrontal cortex of the rat regulate set-shifting», *Neuropsychopharmacology*, vol. 31, nº 2, p. 297–309. 25

FLORESCO, S. B. et A. G. PHILLIPS. 2001, «Delay-dependent modulation of memory retrieval by infusion of a dopamine d1 agonist into the rat medial prefrontal cortex.», *Behavioral neuroscience*, vol. 115, nº 4, p. 934. 67

FOSTER, D., R. MORRIS et P. DAYAN. 2000, «A model of hippocampally dependent navigation, using the temporal difference learning rule», *Hippocampus*, vol. 10, nº 1, p. 1–16. iii, 44, 45, 46, 47, 48, 64, 102

FOSTER, D. J. et J. J. KNIERIM. 2012, «Sequence learning and the role of the hippocampus in rodent navigation», *Current opinion in neurobiology*, vol. 22, nº 2, p. 294–300. 102

FOSTER, D. J. et M. A. WILSON. 2006, «Reverse replay of behavioural sequences in hippocampal place cells during the awake state», *Nature*, vol. 440, nº 7084, p. 680–683. 102

FOUQUET, C., C. TOBIN et L. RONDI-REIG. 2010, «A new approach for modeling episodic memory from rodents to humans : the temporal order memory», *Behavioural brain research*, vol. 215, nº 2, p. 172–179. 102

FRANK, M. J., B. LOUGHRY et R. C. O'REILLY. 2001, «Interactions between frontal cortex and basal ganglia in working memory : a computational model», *Cognitive, Affective, & Behavioral Neuroscience*, vol. 1, nº 2, p. 137–160. 59, 67, 127

FRISTON, K., T. FITZGERALD, F. RIGOLI, P. SCHWARTENBECK, J. O'DOHERTY et G. PEZZULO. 2016, «Active inference and learning», *Neuroscience & Biobehavioral Reviews*, vol. 68, p. 862–879. 126

GABRIELI, J. D., D. A. FLEISCHMAN, M. M. KEANE, S. L. REMINGER et F. MORRELL. 1995, «Double dissociation between memory systems underlying explicit and implicit memory in the human brain», *Psychological Science*, vol. 6, nº 2, p. 76–82. 11

GEHRING, W. J. et R. T. KNIGHT. 2000, «Prefrontal–cingulate interactions in action monitoring», *Nature neuroscience*, vol. 3, nº 5, p. 516–520. 24

GEIST, M., O. PIETQUIN et G. FRICOUT. 2009, «Kalman temporal differences : the deterministic case», dans *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, IEEE, p. 185–192. 55

GERSHMAN, S. J. et N. D. DAW. 2016, «Reinforcement learning and episodic memory in humans and animals : An integrative framework», *Annual Review of Psychology*, vol. 68, nº 1. 127

GIRARD, B., D. FILLIAT, J.-A. MEYER, A. BERTHOZ et A. GUILLOT. 2005, «Integration of navigation and action selection functionalities in a computational model of cortico-basal-ganglia–thalamo-cortical loops», *Adaptive Behavior*, vol. 13, nº 2, p. 115–130. 61, 64

GLÄSCHER, J., N. DAW, P. DAYAN et J. O'DOHERTY. 2010, «States versus rewards : dissociable neural prediction error signals underlying model-based and model-free reinforcement learning», *Neuron*, vol. 66, nº 4, p. 585–595. 66, 125

GOLD, P. E. 2004, «Coordination of multiple memory systems», *Neurobiology of learning and memory*, vol. 82, nº 3, p. 230–242. 9

GOLDMAN-RAKIC, P. S. 1995, «Cellular basis of working memory», *Neuron*, vol. 14, nº 3, p. 477–485. 66, 67

GRAYBIEL, A. 2008, «Habits, rituals, and the evaluative brain», *Annu Rev Neurosci*, vol. 31, p. 359–387. 66

GROSMARK, A. D. et G. BUZSÁKI. 2016, «Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences», *Science*, vol. 351, nº 6280, p. 1440–1443. 120, 127

GUAZZELLI, A., M. BOTA, F. J. CORBACHO et M. A. ARBIB. 1998, «Affordances. motivations, and the world graph theory», *Adaptive Behavior*, vol. 6, nº 3-4, p. 435–471. 44, 46, 47, 61, 64

HADDON, J. E. et S. KILLCROSS. 2011, «Inactivation of the infralimbic prefrontal cortex in rats reduces the influence of inappropriate habitual responding in a response-conflict task», *Neuroscience*, vol. 199, p. 205–212. 25

HAMMOND, L. J. 1980, «The effect of contingency upon the appetitive conditioning of free-operant behavior», *Journal of the experimental analysis of behavior*, vol. 34, nº 3, p. 297–304. 21

HARLEY, C. W. 1972, «Hippocampal lesions and two cue discrimination in the rat», *Physiology & behavior*, vol. 9, nº 3, p. 343IN5347–346 348. 15

HARTLEY, T. et N. BURGESS. 2005, «Complementary memory systems : competition, cooperation and compensation», *Trends in Neurosciences*, vol. 28, nº 4, p. 169–170. 9

HICKS, L. H. 1964, «Effects of overtraining on acquisition and reversal of place and response learning.», *Psychological Reports*. 20

HIRSH, R. 1974, «The hippocampus and contextual retrieval of information from memory : A theory», *Behavioral biology*, vol. 12, nº 4, p. 421–444. 15

HULL, C. 1943, «Principles of behavior», . 9

IGLÓI, K., C. F. DOELLER, A.-L. PARADIS, K. BENCHENANE, A. BERTHOZ, N. BURGESS et L. RONDI-REIG. 2014, «Interaction between hippocampus and cerebellum crus i in sequence-based but not place-based navigation», *Cerebral Cortex*, p. bhu132. 120

JOCHAM, G., T. A. KLEIN et M. ULLSPERGER. 2011, «Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices», *The Journal of neuroscience*, vol. 31, nº 5, p. 1606–1613. 59

KERAMATI, M., A. DEZFOULI et P. PIRAY. 2011, «Speed/accuracy trade-off between the habitual and the goal-directed processes», *PLoS Comput Biol*, vol. 7, nº 5, p. e1002 055. iii, 41, 54, 56, 57, 58, 62, 64, 68, 71, 76, 78, 91, 98, 99, 119, 122, 123, 125

KHAMASSI, M. et M. D. HUMPHRIES. 2012, «Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies», *Frontiers in behavioral neuroscience*, vol. 6, p. 79. 43, 63, 66

KHAMASSI, M., S. LALLÉE, P. ENEL, E. PROCYK et P. F. DOMINEY. 2011, «Robot cognitive control with a neurophysiologically inspired reinforcement learning model», *Frontiers in neurorobotics*, vol. 5, p. 1. 62, 69, 98

KHAMASSI, M., A. B. MULDER, E. TABUCHI, V. DOUCHAMPS et S. I. WIENER. 2008, «Anticipatory reward signals in ventral striatal neurons of behaving rats», *European journal of neuroscience*, vol. 28, nº 9, p. 1849–1866. 59

KHAMASSI, M., R. QUILODRAN, P. ENEL, P. DOMINEY et E. PROCYK. 2015, «Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex», *Cereb Cortex*. 68, 69, 81, 91, 92, 97

KILLCROSS, S. et E. COUTUREAU. 2003, «Coordination of actions and habits in the medial prefrontal cortex of rats», *Cerebral Cortex*, vol. 13, nº 4, p. 400–408. iii, 21, 24, 43, 52, 54

KIM, J. J. et M. G. BAXTER. 2001, «Multiple brain-memory systems : the whole does not equal the sum of its parts», *Trends in neurosciences*, vol. 24, nº 6, p. 324–330. 9

KNOWLTON, B. J., J. A. MANGELS et L. R. SQUIRE. 1996, «A neostriatal habit learning system in humans», *Science*, vol. 273, nº 5280, p. 1399. iii, 11, 12

KOECHLIN, E. et A. HYAFIL. 2007, «Anterior prefrontal function and the limits of human decision-making», *Science*, vol. 318, p. 594–598. 67

KONDA, V. R. et J. N. TSITSIKLIS. 2003, «On actor-critic algorithms», *SIAM journal on Control and Optimization*, vol. 42, nº 4, p. 1143–1166. 38

KORIAT, A. et M. GOLDSMITH. 1996, «Memory metaphors and the real-life/laboratory controversy : Correspondence versus storehouse conceptions of memory», *Behavioral and Brain Sciences*, vol. 19, nº 02, p. 167–188. 3

KRIECKHAUS, E. et G. WOLF. 1968, «Acquisition of sodium by rats : interaction of innate mechanisms and latent learning.», *Journal of comparative and Physiological Psychology*, vol. 65, nº 2, p. 197. 21

LARA, A. H. et J. D. WALLIS. 2015, «The role of prefrontal cortex in working memory : a mini review», *Frontiers in systems neuroscience*, vol. 9. 67

LEE, S., S. SHIMOJO et J. O'DOHERTY. 2014, «Neural computations underlying arbitration between model-based and model-free learning», *Neuron*, vol. 81, nº 3, p. 687–699. 66, 125

LENGYEL, M. et P. DAYAN. 2007, «Hippocampal contributions to control : The third way.», dans *NIPS*, vol. 20, p. 889–896. 127

LEVY, R. et P. S. GOLDMAN-RAKIC. 2000, «Segregation of working memory functions within the dorsolateral prefrontal cortex», dans *Executive control and the frontal lobe : Current issues*, Springer, p. 23–32. 67

LIN, L.-J. et T. M. MITCHELL. 1992, «Memory approaches to reinforcement learning in non-markovian domains», cahier de recherche. 107

LISMAN, J. E., L. M. TALAMINI et A. RAFFONE. 2005, «Recall of memory sequences by interaction of the dentate and ca3 : a revised model of the phase precession», *Neural Networks*, vol. 18, nº 9, p. 1191–1201. 119, 125

MARR, D. 1982, «Vision : A computational approach», . 4

MARTINET, L.-E., J.-B. PASSOT, B. FOUQUE, J.-A. MEYER et A. ARLEO. 2008, «Map-based spatial navigation : A cortical column model for action planning», dans *International Conference on Spatial Cognition*, Springer, p. 39–55. 106

MCCALLUM, R. A. 1995, «Instance-based utile distinctions for reinforcement learning with hidden state», dans *ICML*, p. 387–395. 107

MCDONALD, R. J. et N. M. WHITE. 1993, «A triple dissociation of memory systems : hippocampus, amygdala, and dorsal striatum.», *Behavioral neuroscience*, vol. 107, nº 1, p. 3. iii, 17, 18, 19, 22, 23, 27

MCNAUGHTON, B. L., F. P. BATTAGLIA, O. JENSEN, E. I. MOSER et M.-B. MOSER. 2006, «Path integration and the neural basis of the'cognitive map'», *Nature Reviews Neuroscience*, vol. 7, nº 8, p. 663–678. 105

VAN DER MEER, M., Z. KURTH-NELSON et A. D. REDISH. 2012, «Information processing in decision-making systems», *The Neuroscientist*, vol. 18, nº 4, p. 342–359. 66

VAN DER MEER, M. A., A. JOHNSON, N. C. SCHMITZER-TORBERT et A. D. REDISH. 2010, «Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task», *Neuron*, vol. 67, nº 1, p. 25–32. 27

MEYER, D. E. et R. W. SCHVANEVELDT. 1971, «Facilitation in recognizing pairs of words : evidence of a dependence between retrieval operations.», *Journal of experimental psychology*, vol. 90, nº 2, p. 227. 11

MILLER, E. K. et J. D. COHEN. 2001, «An integrative theory of prefrontal cortex function», *Annual review of neuroscience*, vol. 24, nº 1, p. 167–202. 66

MILLER, G. A. 1956, «The magical number seven, plus or minus two : Some limits on our capacity for processing information.», *Psychological review*, vol. 63, nº 2, p. 81. 66

MILLER, K., A. SHENHAV et E. LUDVIG. 2016, «Habits without values», *bioRxiv*, p. 067 603. 125

MILNER, B., S. CORKIN et H.-L. TEUBER. 1968, «Further analysis of the hippocampal amnesic syndrome : 14-year follow-up study of hm», *Neuropsychologia*, vol. 6, n° 3, p. 215–234. 10

MISHKIN, M. 1978, «Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus», . 12

MONTAGUE, P. R., P. DAYAN et T. J. SEJNOWSKI. 1996, «A framework for mesencephalic dopamine systems based on predictive hebbian learning», *The Journal of neuroscience*, vol. 16, n° 5, p. 1936–1947. 43, 53, 58

MOURET, J.-B. et S. DONCIEUX. 2010, «Sferes v2 : Evolvin' in the multi-core world», dans *WCCI 2010 IEEE World Congress on Computational Intelligence, Congress on Evolutionary Computation (CEC)*, Ieee, p. 4079–4086. 79, 91, 123

MURRAY, E. A. et M. MISHKIN. 1984, «Severe tactual as well as visual memory deficits follow combined removal of the amygdala and hippocampus in monkeys», *The Journal of neuroscience*, vol. 4, n° 10, p. 2565–2580. 13

NEATH, I. et A. SURPRENANT. 2003, *Human Memory : An Introduction to Research, Data, and Theory*, Thomson/Wadsworth, ISBN 9780534595623. 3, 9

NIV, Y. 2009, «Reinforcement learning in the brain», *Journal of Mathematical Psychology*, vol. 53, n° 3, p. 139–154. 40

NORWICH, K. 2003, *Information, Sensation and Perception*, Academic Press San Diego. 75, 126

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», *science*, vol. 304, n° 5669, p. 452–454. 59

OHMAE, S. et J. F. MEDINA. 2015, «Climbing fibers encode a temporal-difference prediction error during cerebellar learning in mice», *Nature neuroscience*. 120, 125

O'KEEFE, J. et L. NADEL. 1978, *The hippocampus as a cognitive map*, vol. 3, Clarendon Press Oxford. 8, 12, 15, 44

OLAFSDOTTIR, H. F., C. BARRY, A. B. SALEEM, D. HASSABIS et H. J. SPIERS. 2015, «Hippocampal place cells construct reward related sequences through unexplored space», *Elife*, vol. 4, p. e06 063. 127

OLTON, D. S., J. A. WALKER et F. H. GAGE. 1978, «Hippocampal connections and spatial discrimination», *Brain research*, vol. 139, n° 2, p. 295–308. 15

OVERMAN, W. H., G. ORMSBY et M. MISHKIN. 1990, «Picture recognition vs. picture discrimination learning in monkeys with medial temporal removals», *Experimental brain research*, vol. 79, n° 1, p. 18–24. 13

PACKARD, M. 2009, «Anxiety, cognition, and habit : a multiple memory systems perspective», *Brain Res*, vol. 1293, n° 0, p. 121–128. 66

PACKARD, M. G. 1999, «Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning», *Proceedings of the National Academy of Sciences*, vol. 96, n° 22, p. 12 881–12 886. 20, 22, 43

PACKARD, M. G. et J. GOODMAN. 2013, «Factors that influence the relative use of multiple memory systems», *Hippocampus*, vol. 23, nº 11, p. 1044–1052. 9

PACKARD, M. G., R. HIRSH et N. M. WHITE. 1989, «Differential effects of fornix and caudate nucleus lesions on two radial maze tasks : evidence for multiple memory systems», *The Journal of neuroscience*, vol. 9, nº 5, p. 1465–1472. 12, 14, 15, 17, 19, 21, 22, 27, 43

PACKARD, M. G. et J. L. MCGAUGH. 1996, «Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning», *Neurobiology of learning and memory*, vol. 65, nº 1, p. 65–72. iii, 19, 20, 22, 24, 26, 27, 43

PEARCE, J. M., A. D. ROBERTS et M. GOOD. 1998, «Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors», *Nature*, vol. 396, nº 6706, p. 75–77. iii, 48, 49, 51

PESSIGLIONE, M., V. CZERNECKI, B. PILLON, B. DUBOIS, M. SCHÜPBACH, Y. AGID et L. TREMBLAY. 2005, «An effect of dopamine depletion on decision-making : the temporal coupling of deliberation and execution», *Journal of cognitive neuroscience*, vol. 17, nº 12, p. 1886–1896. 67

PESSIGLIONE, M., B. SEYMOUR, G. FLANDIN, R. J. DOLAN et C. D. FRITH. 2006, «Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans», *Nature*, vol. 442, nº 7106, p. 1042–1045. 59

PEZZULO, G., F. RIGOLI et F. CHERSI. 2013, «The mixed instrumental controller : using value of information to combine habitual choice and mental simulation», *Frontiers in psychology*, vol. 4, p. 92. 58

PFEIFFER, B. E. et D. J. FOSTER. 2013, «Hippocampal place-cell sequences depict future paths to remembered goals», *Nature*, vol. 497, nº 7447, p. 74–79. 102, 127

PHILLIPS, R. et J. LEDOUX. 1992, «Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning.», *Behavioral neuroscience*, vol. 106, nº 2, p. 274. 16, 19

PIRON, C., D. KASE, M. TOPALIDOU, M. GOILLANDEAU, H. ORIGNAC, T.-H. N'GUYEN, N. ROUGIER et T. BORAUD. 2016, «The globus pallidus pars interna in goal-oriented and routine behaviors : Resolving a long-standing paradox», *Movement Disorders*. 125

POLDRACK, R. A. et M. G. PACKARD. 2003, «Competition among multiple memory systems : converging evidence from animal and human brain studies», *Neuropsychologia*, vol. 41, nº 3, p. 245–251. 9

POLSTER, M. R., L. NADEL et D. L. SCHACTER. 1991, «Cognitive neuroscience analyses of memory : A historical perspective», *Journal of Cognitive Neuroscience*, vol. 3, nº 2, p. 95–116. 9

POSTLE, B. R. 2006, «Working memory as an emergent property of the mind and brain», *Neuroscience*, vol. 139, nº 1, p. 23–38. 67

PRADO-ALCALA, R., Z. GRINBERG, Z. ARDITTI, M. GARCIA, H. PRIETO et H. BRUST-CARMONA. 1975, «Learning deficits produced by chronic and reversible lesions of the corpus striatum in rats», *Physiology & behavior*, vol. 15, nº 3, p. 283–287. 15

PROCYK, E. et P. S. GOLDMAN-RAKIC. 2006, «Modulation of dorsolateral prefrontal delay activity during self-organized behavior», *The Journal of neuroscience*, vol. 26, n° 44, p. 11 313–11 323. 67

PROCYK, E., Y. TANAKA et J.-P. JOSEPH. 2000, «Anterior cingulate activity during routine and non-routine sequential behaviors in macaques», *Nature neuroscience*, vol. 3, n° 5, p. 502–508. 68, 97

QUILODRAN, R., M. ROTHE et E. PROCYK. 2008, «Behavioral shifts and action valuation in the anterior cingulate cortex», *Neuron*, vol. 57, n° 2, p. 314–325. 68, 91, 97

RAGOZZINO, K. E., S. LEUTGEB et S. J. MIZUMORI. 2001, «Dorsal striatal head direction and hippocampal place representations during spatial navigation», *Experimental Brain Research*, vol. 139, n° 3, p. 372–376. 27

RAGOZZINO, M. E. 2002, «The effects of dopamine d1 receptor blockade in the prelimbic–infralimbic areas on behavioral flexibility», *Learning & Memory*, vol. 9, n° 1, p. 18–28. 25

RAGOZZINO, M. E. 2007, «The contribution of the medial prefrontal cortex, orbitofrontal cortex, and dorsomedial striatum to behavioral flexibility», *Annals of the New York Academy of Sciences*, vol. 1121, n° 1, p. 355–375. 25

RAGOZZINO, M. E., J. KIM, D. HASSERT, N. MINNITI et C. KIANG. 2003, «The contribution of the rat prelimbic-infralimbic areas to different forms of task switching.», *Behavioral neuroscience*, vol. 117, n° 5, p. 1054. 25

REDDI, B. et R. CARPENTER. 2000, «The influence of urgency on decision time», *Nat Neurosci*, vol. 3, n° 8, p. 827–830. 75

REDISH, A. D. et D. S. TOURETZKY. 1997, «Navigating with landmarks : Computing goal locations from place codes», *Symbolic visual learning*, p. 325–351. 44

REDISH, D., S. JENSEN et A. JOHNSON. 2008, «A unified framework for addiction : vulnerabilities in the decision process», *Behav Brain Sci*, vol. 31, n° 4, p. 415–437. 66

RENAUDO, E., B. GIRARD, R. CHATILA et M. KHAMASSI. 2015, «Which criteria for autonomously shifting between goal-directed and habitual behaviors in robots ?», dans *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, IEEE, p. 254–260. 62, 64

RICH, E. L. et M. SHAPIRO. 2009, «Rat prefrontal cortical neurons selectively code strategy switches», *The Journal of Neuroscience*, vol. 29, n° 22, p. 7208–7219. 26

RICH, E. L. et M. L. SHAPIRO. 2007, «Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks», *The Journal of neuroscience*, vol. 27, n° 17, p. 4747–4755. 26, 27

RITCHIE, B., B. AESCHLIMAN et P. PIERCE. 1950, «Studies in spatial learning. viii. place performance and the acquisition of place dispositions.», *Journal of comparative and physiological psychology*, vol. 43, n° 2, p. 73. 20

ROEDIGER, H. L. 1980, «Memory metaphors in cognitive psychology», *Memory & Cognition*, vol. 8, n° 3, p. 231–246. 3

ROEDIGER, H. L., S. RAJARAM et K. SRINIVAS. 1990, «Specifying criteria for postulating memory systemsa», *Annals of the New York Academy of Sciences*, vol. 608, nº 1, p. 572–595. 8

RONDI-REIG, L., G. H. PETIT, C. TOBIN, S. TONEGAWA, J. MARIANI et A. BERTHOZ. 2006, «Impaired sequential egocentric and allocentric memories in forebrain-specific–nmda receptor knock-out mice during a new task dissociating strategies of navigation», *The journal of Neuroscience*, vol. 26, nº 15, p. 4071–4081. 102

ROTHÉ, M., R. QUILODRAN, J. SALLET et E. PROCYK. 2011, «Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation», *The Journal of Neuroscience*, vol. 31, nº 31, p. 11 110–11 117. 97

ROUGIER, N. P., D. C. NOELLE, T. S. BRAVER, J. D. COHEN et R. C. O'REILLY. 2005, «Prefrontal cortex and flexible cognitive control : Rules without symbols», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, nº 20, p. 7338–7343. 67, 127

RUMMERY, G. A. 1995, *Problem solving with reinforcement learning*, thèse de doctorat, University of Cambridge Ph. D. dissertation. 37

RUMMERY, G. A. et M. NIRANJAN. 1994, *On-line Q-learning using connectionist systems*, University of Cambridge, Department of Engineering. 37

SADACCA, B. F., J. L. JONES et G. SCHOENBAUM. 2016, «Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework», *eLife*, vol. 5, p. e13 665. 126

SALLET, J., N. CAMILLE et E. PROCYK. 2013, «Modulation of feedback-related negativity during trial-and-error exploration and encoding of behavioral shifts», *Frontiers in neuroscience*, vol. 7, p. 209. 97, 98

SALMON, D. P., S. ZOLA-MORGAN et L. R. SQUIRE. 1987, «Retrograde amnesia following combined hippocampus-amygdala lesions in monkeys», *Psychobiology*, vol. 15, nº 1, p. 37–47. 14

SAMEJIMA, K., Y. UEDA, K. DOYA et M. KIMURA. 2005, «Representation of action-specific reward values in the striatum», *Science*, vol. 310, nº 5752, p. 1337–1340. 59

SAMUELS, I. 1972, «Hippocampal lesions in the rat : Effects on spatial and visual habits», *Physiology & Behavior*, vol. 8, nº 6, p. 1093–1097. 15

SCHACTER, D. et E. TULVING. 1994, *Memory Systems 1994*, chap. What are the memory systems of 1994 ?, A Bradford book, Bradford Books, U. S., ISBN 9780262193504, p. 2–38. URL https://books.google.fr/books?id=4gdHL81eaQgC. 8

SCHACTER, D. L. et R. L. BUCKNER. 1998, «Priming and the brain», *Neuron*, vol. 20, nº 2, p. 185–195. 11

SCHACTER, D. L., M. VERFAELLIE et D. PRADERE. 1996, «The neuropsychology of memory illusions : False recall and recognition in amnesic patients», *Journal of Memory and Language*, vol. 35, nº 2, p. 319–334. 11

SCHULTZ, W., P. APICELLA et T. LJUNGBERG. 1993, «Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task», *The Journal of Neuroscience*, vol. 13, nº 3, p. 900–913. 43, 53

SCHULTZ, W., P. DAYAN et P. R. MONTAGUE. 1997, «A neural substrate of prediction and reward», *Science*, vol. 275, nº 5306, p. 1593–1599. 43, 53, 58, 67, 120, 124

SCHWARZ, G. 1978, «Estimating the dimension of a model», *Ann Stat*, vol. 6, nº 2, p. 461–464. 81

SCOVILLE, W. B. et B. MILNER. 1957, «Loss of recent memory after bilateral hippocampal lesions», *Journal of neurology, neurosurgery, and psychiatry*, vol. 20, nº 1, p. 11. 9, 10

SHAPIRO, M. L., P. J. KENNEDY et J. FERBINTEANU. 2006, «Representing episodes in the mammalian brain», *Current opinion in neurobiology*, vol. 16, nº 6, p. 701–709. 119

SHERRY, D. F. et D. L. SCHACTER. 1987, «The evolution of multiple memory systems.», *Psychological review*, vol. 94, nº 4, p. 439. 8

SILVA, D., T. FENG et D. J. FOSTER. 2015, «Trajectory events across hippocampal place cells require previous experience», *Nature neuroscience*, vol. 18, nº 12, p. 1772–1779. 120, 127

SILVEIRA, J. M. et D. P. KIMBLE. 1968, «Brightness discrimination and reversal in hippocampally-lesioned rats», *Physiology & Behavior*, vol. 3, nº 5, p. 625–630. 15

SQUIRE, L. R. 2004, «Memory systems of the brain : a brief history and current perspective», *Neurobiology of learning and memory*, vol. 82, nº 3, p. 171–177. iii, 9, 14, 15, 101

SQUIRE, L. R. et S. ZOLA-MORGAN. 1991, «The medial temporal lobe memory system», *Science*, vol. 253, nº 5026, p. 1380–1386. iii, 10, 13, 14

STEFANI, M. R., K. GROTH et B. MOGHADDAM. 2003, «Glutamate receptors in the rat medial prefrontal cortex regulate set-shifting ability.», *Behavioral neuroscience*, vol. 117, nº 4, p. 728. 25

SUTHERLAND, R. et R. MCDONALD. 1990, «Hippocampus, amygdala, and memory deficits in rats», *Behavioural brain research*, vol. 37, nº 1, p. 57–79. 16

SUTTON, R. S. 1990, «Integrated architectures for learning, planning, and reacting based on approximating dynamic programming», dans *Proceedings of the seventh international conference on machine learning*, p. 216–224. 39

SUTTON, R. S. et A. G. BARTO. 1998, *Reinforcement learning : An introduction*, vol. 1, MIT press Cambridge. 31, 35, 37, 38, 39, 40, 127

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», *Nature neuroscience*, vol. 7, nº 8, p. 887–893. 59

THORNDIKE, E. L. 1933, «A proof of the law of effect.», *Science*. 9

TODD, M. T., Y. NIV et J. D. COHEN. 2009, «Learning to use working memory in partially observable environments through dopaminergic reinforcement», dans *Advances in neural information processing systems*, p. 1689–1696. 67, 127

TOLMAN, E. C. 1948, «Cognitive maps in rats and men.», *Psychological review*, vol. 55, nº 4, p. 189. 9, 19

TOPALIDOU, M., D. KASE, T. BORAUD et N. P. ROUGIER. 2016, «Dissociation of reinforcement and hebbian learning induces covert acquisition of value in the basal ganglia», *bioRxiv*, p. 060 236. 125, 126

TREMBLAY, L. et W. SCHULTZ. 1999, «Relative reward preference in primate orbitofrontal cortex», *Nature*, vol. 398, nº 6729, p. 704–708. 24

TRICOMI, E., B. W. BALLEINE et J. P. O'DOHERTY. 2009, «A specific role for posterior dorsolateral striatum in human habit learning», *European Journal of Neuroscience*, vol. 29, nº 11, p. 2225–2232. 124

TRULLIER, O. et J.-A. MEYER. 1997, «Biomimetic navigation models and strategies in animats», *AI communications*, vol. 10, nº 2, p. 79–92. 44

TULVING, E. 1972, «Episodic and semantic memory 1», *Organization of Memory. London : Academic*, vol. 381, nº 4, p. 382–404. 101

TULVING, E. 1985, «Elements of episodic memory», . 101

TULVING, E. 2002, «Episodic memory : From mind to brain», *Annual review of psychology*, vol. 53, nº 1, p. 1–25. 101, 102

TULVING, E. et W. DONALDSON. 1972, «Organization of memory.», . 125

TULVING, E. et D. L. SCHACTER. 1990, «Priming and human memory systems», *Science*, vol. 247, nº 4940, p. 301–306. 11

VIEJO, G., B. GIRARD et M. KHAMASSI. 2016, «[re] speed/accuracy trade-off between the habitual and the goal-directed process», *ReScience*, vol. 2, nº 1, doi :10.5281/zenodo.45852, p. NA. 54, 57, 58, 122

VIEJO, G., M. KHAMASSI, A. BROVELLI et B. GIRARD. 2015, «Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning», *Frontiers in behavioral neuroscience*, vol. 9. 54, 59

VILLETTE, V., A. MALVACHE, T. TRESSARD, N. DUPUY et R. COSSART. 2015, «Internally recurring hippocampal sequences as a population template of spatiotemporal information», *Neuron*, vol. 88, nº 2, p. 357–366. 120

VOERMANS, N. C., K. M. PETERSSON, L. DAUDEY, B. WEBER, K. P. VAN SPAENDONCK, H. P. KREMER et G. FERNÁNDEZ. 2004, «Interaction between the human hippocampus and the caudate nucleus during route recognition», *Neuron*, vol. 43, nº 3, p. 427–435. 27

WANG, Y., S. ROMANI, B. LUSTIG, A. LEONARDO et E. PASTALKOVA. 2015, «Theta sequences are essential for internally generated hippocampal firing fields», *Nature neuroscience*, vol. 18, nº 2, p. 282–288. 120

WARRINGTON, E. K. et L. WEISKRANTZ. 1968, «A study of learning and retention in amnesic patients», *Neuropsychologia*, vol. 6, nº 3, p. 283–291. 10

WARRINGTON, E. K. et L. WEISKRANTZ. 1970, «Amnesic syndrome : Consolidation or retrieval ?», *Nature*. 10

WATKINS, C. J. et P. DAYAN. 1992, «Q-learning», *Machine learning*, vol. 8, n° 3-4, p. 279–292. 71

WATKINS, C. J. C. H. 1989, *Learning from delayed rewards*, thèse de doctorat, University of Cambridge England. 37

WHISHAW, I. Q., G. MITTLEMAN, S. T. BUNCH et S. B. DUNNETT. 1987, «Impairments in the acquisition, retention and selection of spatial navigation strategies after medial caudate-putamen lesions in rats», *Behavioural brain research*, vol. 24, n° 2, p. 125–138. 15

WHITE, N. M. et R. J. MCDONALD. 2002, «Multiple parallel memory systems in the brain of the rat», *Neurobiology of learning and memory*, vol. 77, n° 2, p. 125–184. iii, 18, 19

WHITE, N. M., M. G. PACKARD et R. J. MCDONALD. 2013, «Dissociation of memory systems : The story unfolds.», *Behavioral neuroscience*, vol. 127, n° 6, p. 813. 9

WICKELGREN, W. A. 1968, «Sparing of short-term memory in an amnesic patient : implications for strength theory of memory», *Neuropsychologia*, vol. 6, n° 3, p. 235–244. 10

WIERZBICKI, A. 1986, «On the completeness and constructiveness of parametric characterizations to vector optimization problems», *OR Spektrum*, vol. 8, p. 73–87. II

WITTEN, I. H. 1977, «An adaptive optimal controller for discrete-time markov environments», *Information and control*, vol. 34, n° 4, p. 286–295. 38

WOOD, W. et D. RÜNGER. 2016, «Psychology of habit», *Psychology*, vol. 67. 126

YAGER, R. 2004, «Generalized owa aggregation operators», *Fuzzy Optim Decis Ma*, vol. 3, n° 1, p. 93–107. II

YIN, H., S. OSTLUND et B. BALLEINE. 2008, «Reward-guided learning beyond dopamine in the nucleus accumbens : the integrative functions of cortico-basal ganglia networks», *Eur J Neurosci*, vol. 28, n° 8, p. 1437–1448. 66

YIN, H. H. et B. J. KNOWLTON. 2004, «Contributions of striatal subregions to place and response learning», *Learning & Memory*, vol. 11, n° 4, p. 459–463. 23

YIN, H. H. et B. J. KNOWLTON. 2006, «The role of the basal ganglia in habit formation», *Nature Reviews Neuroscience*, vol. 7, n° 6, p. 464–476. 9, 21, 22, 66

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», *European journal of neuroscience*, vol. 19, n° 1, p. 181–189. 23, 27

YIN, H. H., B. J. KNOWLTON et B. W. BALLEINE. 2005, «Blockade of nmda receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning», *European Journal of Neuroscience*, vol. 22, n° 2, p. 505–512. 23

YIN, H. H., B. J. KNOWLTON et B. W. BALLEINE. 2006, «Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning», *Behavioural brain research*, vol. 166, n° 2, p. 189–196. 23

ZILLI, E. A. et M. E. HASSELMO. 2008, «Modeling the role of working memory and episodic memory in behavioral tasks», *Hippocampus*, vol. 18, n° 2, p. 193–209. 107

ZITZLER, E. et L. THIELE. 1999, «Multiobjective evolutionary algorithms : a comparative case study and the strength pareto approach», *IEEE Trans. Evol. Comput*, vol. 3, n° 4, p. 257–271. 84

ZOLA-MORGAN, S., L. SQUIRE et M. MISHKIN. 1982, «The neuroanatomy of amnesia : amygdala-hippocampus versus temporal stem», *Science*, vol. 218, n° 4579, doi :10.1126/science.6890713, p. 1337–1339, ISSN 0036-8075. URL http://science.sciencemag.org/content/218/4579/1337. iii, 12, 13, 16

ZOLA-MORGAN, S. et L. R. SQUIRE. 1984, «Preserved learning in monkeys with medial temporal lesions : sparing of motor and cognitive skills», *The Journal of Neuroscience*, vol. 4, n° 4, p. 1072–1085. 13

ZOLA-MORGAN, S. et L. R. SQUIRE. 1985, «Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia.», *Behavioral neuroscience*, vol. 99, n° 1, p. 22. 13, 16

ZOLA-MORGAN, S., L. R. SQUIRE et D. G. AMARAL. 1986, «Human amnesia and the medial temporal region : enduring memory impairment following a bilateral lesion limited to field ca1 of the hippocampus», *The Journal of Neuroscience*, vol. 6, n° 10, p. 2950–2967. 10