Apprentissage et contrôle cognitif : une théorie computationnelle de la fonction exécutive préfontale humaine

Anne Collins

▶ **To cite this version:**

Anne Collins. Apprentissage et contrôle cognitif : une théorie computationnelle de la fonction exécutive préfontale humaine. Neurosciences. Université Pierre et Marie Curie - Paris VI, 2010. Français. NNT : 2010PA066124 . tel-00814840

HAL Id: tel-00814840

https://theses.hal.science/tel-00814840

Submitted on 17 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6
École doctorale : Cerveau, Cognition, Comportement

présentée par :
Anne Collins

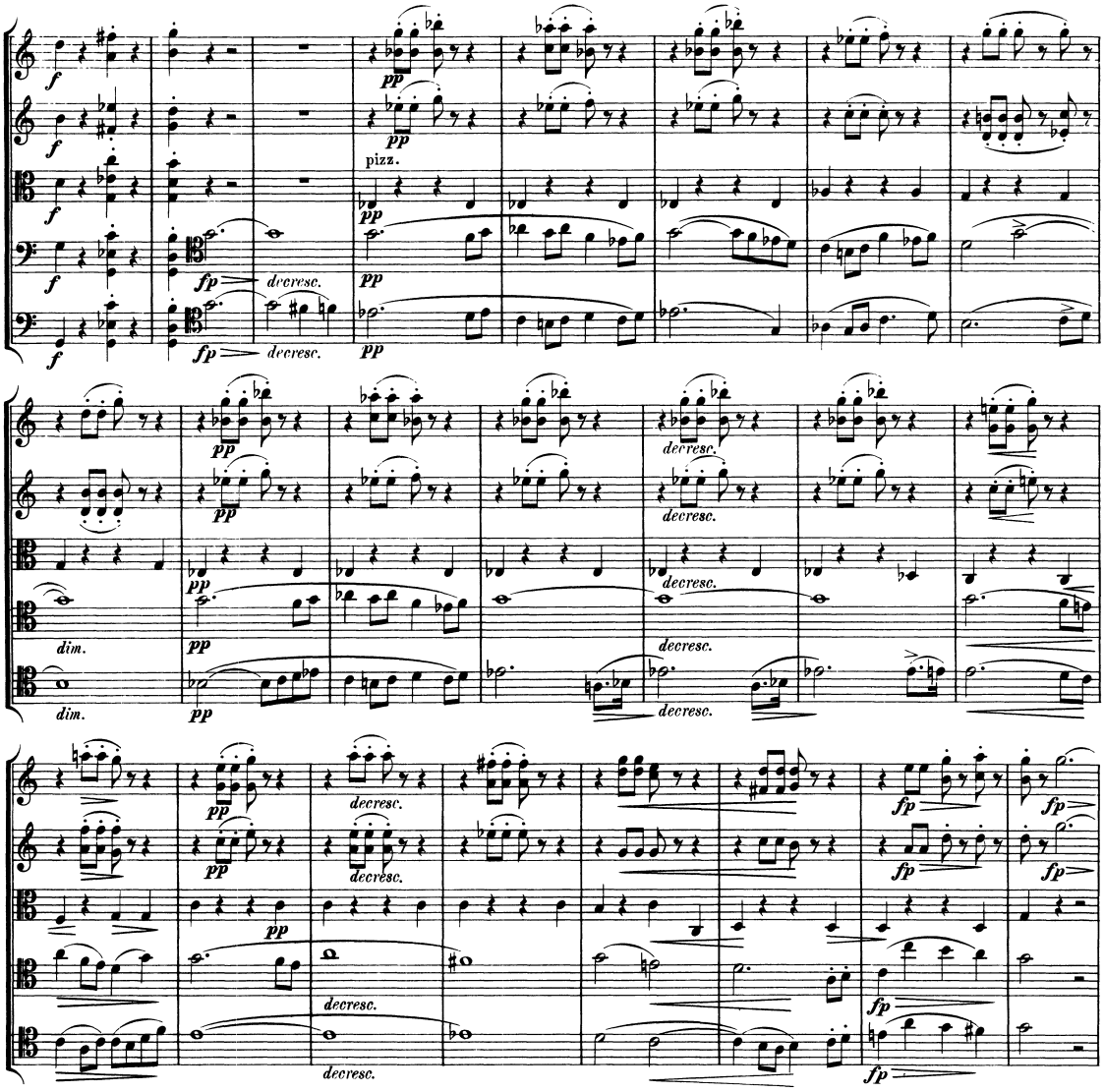
Pour obtenir le grade de DOCTEUR DE L'UNIVERSITE PARIS 6

APPRENTISSAGE ET CONTRÔLE COGNITIF : UNE THÉORIE COMPUTATIONNELLE DE LA FONCTION EXÉCUTIVE PRÉFRONTALE HUMAINE

Soutenue le 5 Janvier 2010

Devant le jury composé de :

Dr Etienne KOECHLIN	Directeur de thèse
Dr Alexandre POUGET	Rapporteur
Dr Emmanuel PROCYK	Rapporteur
Pr Richard LEVY	Examinateur
Dr Sophie DENÈVE	Examinateur
Dr Mathias PESSIGLIONE	Examinateur
Dr Jean-Claude DREHER	Examinateur



La musique, source de bonheur, source d'inspiration sur l'apprentissage, le contrôle, l'exploration, la complexité et la motivation [191]

Remerciements :

Je tiens ici à apporter ma reconnaissance à tous ceux qui m'ont soutenue, de différentes manières, pendant les quelques années qui ont conduit à cette thèse.

Tout d'abord, je tiens à remercier les rapporteurs et les examinateurs de mon jury, qui ont bien voulu évaluer ce travail : Dr E. Procyk, Pr A. Pouget, Pr R. Levy, Dr S. Denève, Dr J-C. Dreher, Dr M. Pessiglione.

Mes remerciements vont également à tous ceux qui ont contribué, scientifiquement, à l'avancée des idées présentées dans cette thèse. Tout particulièrement, je tiens à remercier Etienne Koechlin, qui a accepté de diriger cette thèse et a permis une interaction scientifique passionnante. Je remercie également les membres du Laboratoire de Neurosciences Cognitives de l'Ecole Normale Supérieure, ainsi que les anciens membres du laboratoire ANIM, en particulier Sylvain Charron, Maël Donoso, Beth Pavlicek, Alexandre Hyafil, Emmanuel Guigon, Stéphane Genet, Christopher Summerfield, Julie Grèzes et Sylvie Berthoz. Enfin, je remercie les chercheurs de laboratoires visités, qui ont apporté des suggestions intéressantes : Peter Dayan, Tim Behrens, Michael Frank et David Badre.

Ma reconnaissance va particulièrement aux relecteurs de cette thèse : Françoise, Vincent, Claire, Sylvain, Maël et Etienne.

Enfin, je tiens à témoigner ma reconnaissance à tous ceux qui m'ont soutenue pendant cette thèse : Vincent, Françoise, Marie, Claire, Emmanuel, Paul, Isabelle, Hélène, Sylvain, Benoît, Sonoko et la petite Aimée, Maud, Anne-Claire, Helmi, Michel, et beaucoup d'autres.

Pour finir, je tiens à témoigner ma gratitude à la famille Postel-Vinay, qui m'a permis de travailler, pendant cette thèse, dans des conditions inespérées.

Résumé

Le contrôle cognitif est la capacité à réagir à des stimuli de manière adaptée au contexte présent ou aux indices passés, en tenant compte de nos buts internes. Le contrôle cognitif et l'apprentissage entretiennent des liens profonds et réciproques. D'un côté, le contrôle cognitif requiert que nous ayons appris un répertoire de comportements ainsi que leur valeur dans différentes conditions, afin de les utiliser à bon escient. D'un autre côté, l'apprentissage d'un répertoire de comportements nécessite du contrôle cognitif, notamment pour réguler l'équilibre entre exploration et exploitation, mais également pour généraliser, décider d'un switch, induire une structure dans un problème, etc.. . .

Le contrôle cognitif et l'apprentissage sont donc indissociablement liés dans la flexibilité qui caractérise la fonction exécutive préfrontale humaine. Cependant, ce lien est actuellement mal compris et peu de travaux de psychologie ou neurosciences cognitives intègrent ces deux aspects. De même, les modèles computationnels d'apprentissage ou de décision existants ne rendent pas compte de leur interaction.

Dans ce travail de thèse, nous proposons une théorie mathématique reposant sur des mécanismes d'apprentissage par renforcement et d'inférence bayésienne, qui intègre l'apprentissage de répertoires de comportements (task-sets) dans un milieu incertain et le contrôle cognitif (task-switching) en présence ou en l'absence d'information contextuelle. Cette théorie permet de faire des prédictions spécifiques que nous avons testées dans le cadre de deux expériences comportementales. Celles-ci ont permis de valider les prédictions de la théorie et d'invalider d'autres modèles existants. De plus, la théorie proposée permet d'avancer un facteur explicatif des différences qualitatives de stratégies d'exploration observées entre différents individus.

La théorie proposée caractérise de façon intrinsèque des notions essentielles telles que le comportement par défaut, le switch et l'exploration. Elle permet de faire émerger naturellement un mécanisme de contrôle du compromis exploitation – exploration, ainsi que son facteur de pondération. Enfin, les résultats empiriques valident les prédictions et confirment les hypothèses du modèle. Celui-ci pourra être utilisé pour comprendre les computations effectuées par le cerveau dans des études d'imagerie fonctionnelle, avec le cortex préfrontal, les ganglions de la base et des neuromodulateurs (dopamine et norépinephrine) comme centres d'intérêt principaux.

Abstract

Cognitive control enables appropriate action selection according to stimuli, but also present context or past cues, while taking our internal goals into account. Cognitive control and learning are profoundly and reciprocally linked. On one side, cognitive control requires that a repertoire of behaviors be learnt, as well as their values in different conditions, for appropriate use. On the other side, cognitive control is needed for learning of a repertoire of behaviors, notably to regulate the exploration-exploitation trade-off, but also to generalize, decide to switch, infer a structure in a problem, etc. . ..

Thus, cognitive control and learning are strongly linked in the flexibility that characterizes human prefrontal executive function. However, this link is presently poorly understood and few psychological or cognitive neuroscience studies include both aspects. Moreover, existing computational models of learning and decision do not account for their interaction.

In this PhD thesis, we propose a mathematical theory combining reinforcement learning and Bayesian inference mechanisms. This model includes learning of repertoires of behaviors (task-sets) in an uncertain environment as well as cognitive control (task-switching) in presence or absence of contextual information. This model makes specific predictions that we tested in two behavioral experiments. They validate the predictions of the theory against other existing models. Moreover, the theory proposes an explanatory factor for qualitative differences in exploratory strategies that we observed across individuals.

The proposed theory intrinsically characterizes essential notions such as default behavior, switch and exploration. It allows for the natural emergence of a control mechanism of the exploitation-exploration trade-off, as well as its weighing factor. Lastly, empirical results validate the predictions and confirm the hypotheses of the model. The model may be used in functional imaging studies to understand computations executed in the brain, with prefrontal cortex, basal ganglia and neurotransmitters such as dopamine and norepinephrine as main points of interest.

Table des matières

Table des figures

Introduction

Le contrôle cognitif représente notre capacité à avoir un comportement intentionnel. Plus précisément, le comportement intentionnel se reflète dans le fait d'agir non pas seulement en réaction aux stimuli extérieurs, mais aussi en fonction de nos buts, de nos croyances, du contexte immédiat et d'indices passés. Le contrôle cognitif humain est caractérisé par la très grande flexibilité cognitive que nous avons et qui nous permet de nous adapter efficacement à de nouveaux contextes ou objectifs.

L'étude des patients présentant des lésions du cortex préfrontal a rapidement permis de mettre en valeur le cortex préfrontal comme le candidat principal pour l'implémentation du contrôle cognitif. Depuis, son implication a été confirmée dans de nombreuses études, comme nous le montrerons dans la première partie de cette thèse. Plus particulièrement, l'existence de représentations hiérarchiques des problèmes de contrôle cognitif a été mise en évidence dans le cortex préfrontal.

Cependant, le cortex préfrontal est également impliqué dans les problèmes d'apprentissage. Dans l'abstract de sa revue sur le rôle du cortex préfrontal dans le contrôle cognitif (2000 [147]), Miller écrit :

Nearly all intended behaviour is learned and so depends on a cognitive system that can acquire and implement the 'rules of the game' needed to achieve a given goal in a given situation. (Miller, 2000 [147])

Miller souligne ainsi que le système cortical qui implémente le contrôle cognitif doit nécessairement être capable d'apprentissage, puisque tout comportement intentionnel a été appris. Démontrer son rôle dans l'apprentissage des « règles du jeu » a d'ailleurs très largement contribué à démontrer le rôle du cortex préfrontal dans le contrôle cognitif. On voit donc que l'apprentissage et le contrôle cognitif sont liés par leurs bases neurales. Cela semble naturel, vue la dépendance réciproque qu'on peut mettre en évidence entre ces deux processus : si d'une part, l'apprentissage est nécessaire au contrôle cognitif pour l'acquisition des comportements intentionnels, le contrôle semble également nécessaire à l'apprentissage pour les décisions indispensables : exploration des différentes options, arrêt de l'apprentissage à la complétion d'un objectif, décision quant au niveau hiérarchique auquel il est

nécessaire d'apprendre (faut-il s'adapter localement, ou apprendre globalement un nouveau comportement ?)... Pourtant, bien que ces deux processus soient inextricablement liés, les principes d'intégration entre l'apprentissage et le contrôle cognitif dans le cortex préfrontal restent peu connus. En 2008, dans deux revues consacrées à l'organisation hiérarchique du cortex préfrontal, deux auteurs mettent en valeur, dans les questions scientifiques futures, notre absence de connaissance sur l'interaction entre l'apprentissage et le contrôle cognitif, dans le cadre de représentations hiérarchiques :

How are hierarchical representations of behavior learned ? How might such learning yield skills that are useful across tasks, supporting later learning ? What are the relevant neural mechanisms ? Are computational techniques for hierarchical reinforcement learning potentially relevant in addressing these questions ? (Botvinick, 2008 [24])

How are novel tasks and rule structures acquired within this architecture ? (Badre, 2008 [10])

Nous tentons, dans cette thèse, de répondre aux questions posées par ces auteurs : comment l'apprentissage et le contrôle cognitif interagissent-ils pour permettre de former les représentations hiérarchiques dont ils dépendent ? Quels modèles computationnels seraient utiles pour répondre à ces questions ?

Dans la première partie d'étude bibliographique, nous nous proposons de présenter les connaissances établies concernant l'apprentissage de règles et le contrôle cognitif dans le cortex préfrontal et dans les ganglions de la base, du point de vue des neurosciences cognitives et des neurosciences computationnelles. Nous présenterons tout d'abord rapidement les systèmes corticaux mis en jeu par la suite : le cortex préfrontal, les structures sous-corticales telles que les ganglions de la base, enfin les systèmes de neuromodulateurs. Dans un premier chapitre, nous parlerons de l'apprentissage par renforcement et de ses liens avec le contrôle cognitif. Ensuite, nous montrerons l'importance de la hiérarchie dans le contrôle cognitif et nous présenterons des modèles hiérarchiques existants d'apprentissage et de décision. Dans le troisième chapitre, nous définirons les notions de task-set et de task-switching comme une première brique d'une organisation hiérarchique du contrôle cognitif. Enfin, cela nous permettra d'établir les questions précises posées dans la thèse.

Dans la deuxième partie, nous présenterons la théorie computationnelle construite pour expliquer les mécanismes d'interaction entre l'apprentissage et le contrôle cognitif chez l'homme. En troisième partie, nous exposerons les deux études comportementales que nous avons menées pour tester les prédictions effectuées par la théorie. Enfin, les résultats présentés seront discutés et confrontés aux connaissances actuelles.

Première partie

Étude Bibliographique

Trois systèmes neuraux sont essentiels au contrôle cognitif et à l'apprentissage. Comme ceux-ci seront largement évoqués plus loin, nous commençons par une courte présentation du cortex préfrontal, des ganglions de la base et du système des neurotransmetteurs.

Le cortex Préfrontal (PFC) Le cortex préfrontal se situe dans le lobe frontal du cortex. Il regroupe les aires de Brodman (BA) 8,9,10,32,44,45,46 (Brodmann, 1909 [31], voir figure 1, page 16). Il représente chez l'homme environ un tiers de la surface totale du cortex, cette proportion cortex préfrontal/cortex étant la plus importante de l'ensemble des mammifères, à égalité avec les grands singes (Semendeferi et al, 2002 [195]). C'est probablement la région qui a évolué le plus tardivement au cours de l'évolution (Jerison et Zaidel, 1994 [117]). C'est aussi une des parties du cerveau qui continue à se développer le plus tard, jusqu'à 20 ans (Sowell et al, 1999 [198]).

Le cortex préfrontal est largement connecté avec les aires associatives, à l'exception de sa partie la plus rostrale, le cortex fronto-polaire, BA 10 (Petrides et Pandya, 2002 [169]). Il a également une très forte connectivité interne, locale ou entre différentes aires.

On distingue différentes aires fonctionnelles qui seront évoquées précisément plus tard. On propose ici simplement un résumé simplifié des rôles fonctionnels de différentes régions. Le cortex préfrontal dorso-latéral est impliqué dans la sélection de tâches (Miller et Cohen, 2001 [148]). Le cortex préfrontal médial, en particulier le cortex cingulaire antérieur, joue un rôle dans la motivation pour le contrôle (Ridderinkhof et al, 2004 [177], Rushworth et al, 2004 [180], 2007 [182]). Il est aussi connu pour réagir aux erreurs, à l'incertitude ou au conflit cognitif. Le cortex orbito-frontal évalue la valeur d'actions ou d'objets (O'Doherty et al, 2001 [157], Wallis et al, 2007 [210]). Le cortex frontopolaire permet la mise en attente d'une tâche pendant la réalisation d'une autre (Koechlin et al, 1999 [131]).

Les ganglions de la base et le thalamus Le cortex préfrontal est également connecté de manière essentielle avec les structures sous-corticales (voir figure 1, page 16 et figure 2, page 17). ll existe en particuliers cinq circuits parallèles partant d'une aire particulière du lobe frontal, se projetant sur un noyau du striatum, puis sur le pallidum et finalement sur le thalamus, avant de revenir en boucle sur le point de départ (Alexander et al, 1986

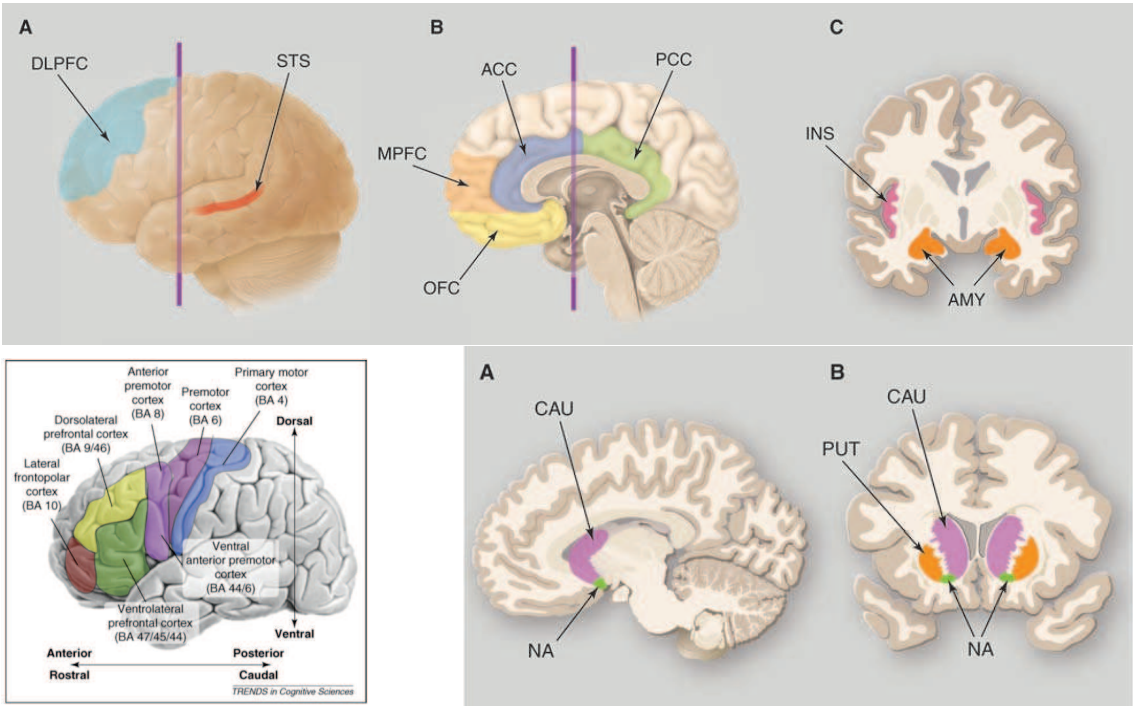


FIGURE 1 – Régions d'intérêt (corticales et sous corticales) pour le contrôle cognitif. Adapté de Badre et al, 2008 [10], Sanfey et al, 2007 [188]. DLPFC : cortex préfrontal dorso-latéral. MPFC : cortex préfrontal médial. OFC : cortex orbito-frontal. ACC : cortex cingulaire antérieur. INS : insula. AMY : amygdale. Noyaux du striatum : CAU : noyau caudé, PUT : putamen, NA : noyau accumbens.

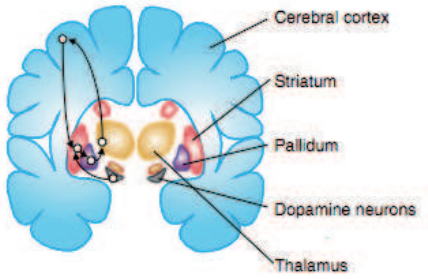


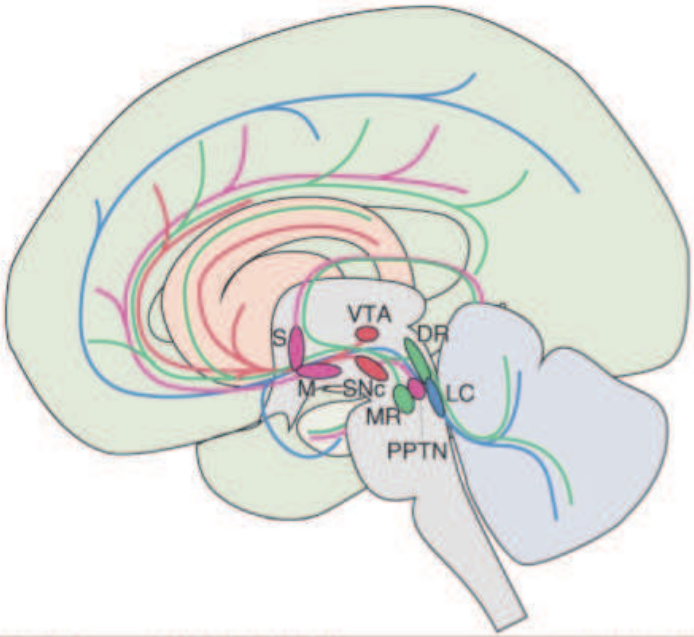
FIGURE 2 – Boucle Cortico-basale. Adapté de Doya, 2008 [74].

[4], 1990 [3], Cummings et al, 1995 [57]). Les ganglions de la base jouent donc un rôle essentiel, probablement d'intégration et de filtrage de l'information, en relation directe avec le PFC.

Les ganglions de la base comprennent plusieurs noyaux distincts, notamment le striatum (comprenant le noyau caudé, le putamen et le noyau accumbens), le pallidum, la substance noire, et le noyau sub-thalamique (Meininger, 1983 [146]).

Les neuromodulateurs On connaît quatre systèmes de neuromodulateurs agissant sur le cerveau (voir figure 3 page 18). Les neuromodulateurs sont produits par des neurones sous-corticaux ayant la spécificité de projections très vastes sur le cortex et les structures sous-corticales, permettant de supposer qu'ils sont impliqués dans des rôles peu spécifiques. Ces neuromodulateurs sont

- La Dopamine (DA). La dopamine est produite dans l'aire tegmentale ventrale (VTA) et la Substantia Nigra, pars compacta (SNc). Elle se projette essentiellement vers le striatum et la partie médiale du cortex préfrontal, mais également vers d'autres parties du cortex préfrontal. La dopamine est impliquée de manière essentielle dans le traitement de la récompense et dans l'apprentissage (Schultz et al, 1997 [192]).
- La Sérotonine (5-HT). La sérotonine est produite dans le noyau raphé et se projette vers l'ensemble du cortex, des structures sous-corticales et du cervelet. La sérotonine est impliquée dans le traitement des punitions et des délais de récompense (Doya, 2002 [73], Daw et Doya, 2006 [60])
- La Noradrénaline, ou Norépinéphrine (NE). La Norépinéphrine est produite dans le locus



neuromodulator	origin of projection	major target area
dopamine (DA)	substantia nigra, pars compacta (SNc) ventral tegmental area (VTA)	dorsal striatum ventral striatum frontal cortex
serotonin (5-HT)	dorsal raphe nucleus (DR) median raphe nucleus (MR)	cortex, striatum cerebellum hippocampus
noradrenaline (NA) (norepinephrine, NE)	locus coeruleus (LC)	cortex, hippocampus cerebellum
acetylcholine (ACh)	Meynert nucleus (M) medial septum (S) pedunculopontine tegmental nucleus (PPTN)	cortex, amygdala hippocampus SNc, thalamus superior colliculus

FIGURE 3 – Systèmes neuromodulateurs. Adapté de Doya, 2002 [73].

coeruleus et se projette vers l'ensemble du cortex, des structures sous-corticales et du cervelet. Elle est impliquée dans l'attention et le contrôle (Aston-Jones et Cohen, 2005[6]).
– L'acétylcholine (ACh). L'acétylcholine est produite par différents noyaux sous-corticaux, notamment le septum, et se projette vers l'ensemble du cortex et des structures sous-corticales. Elle est impliquée dans le contrôle de l'équilibre entre maintenir en mémoire ou mettre à jour la mémoire (Doya, 2002 [73]).

Structure de l'étude bibliographique

Le premier chapitre de l'étude bibliographique a pour objectif d'établir le lien existant entre le contrôle cognitif et l'apprentissage. A cet effet, nous montrerons que la théorie de l'apprentissage par renforcement, qui permet de rendre compte de nombreuses données empiriques, permet également de rendre compte de problèmes de prise de décision et de planification, montrant ainsi que les questions d'apprentissage et de contrôle cognitif sont théoriquement imbriquées en un seul problème. Nous montrerons également que, loin de n'être que théorique, cette imbrication est également biologique : en effet, nous verrons que les problèmes d'apprentissage impliquent de manière essentielle le cortex préfrontal qui est, comme nous l'avons vu plus haut, le siège du contrôle cognitif. L'apprentissage et le contrôle cognitif, dont on peut affirmer par simple observation de la vie quotidienne qu'ils sont nécessaires l'un à l'autre, sont donc effectivement fortement intégrés, théoriquement et biologiquement, dans les modèles de décision et dans le cortex préfrontal.

Nous montrerons dans le deuxième chapitre que la complexité joue un rôle important dans le degré d'interaction entre le contrôle cognitif et l'apprentissage. En effet, nous montrerons qu'entre des problèmes simples et des problèmes plus complexes de contrôle ou d'apprentissage, un saut qualitatif de traitement existe, notamment par l'introduction de structures hiérarchiques permettant de représenter les problèmes plus efficacement. Nous montrerons donc que l'apprentissage et le contrôle cognitif sont représentés hiérarchiquement dans le cerveau, permettant plus d'efficacité et de flexibilité, et contraignant les théories computationnelles souhaitant intégrer l'apprentissage et le contrôle cognitif.

Enfin, dans une dernière partie, nous nous concentrerons sur la notion de *task-set*, dont

nous montrerons qu'elle représente l'unité comportementale de base (au dessus des actions simples) sur laquelle la structure hiérarchique du contrôle cognitif est construite. Nous montrerons que le passage de l'utilisation a priori d'un task-set, à son abandon pour en sélectionner une autre (*switch* ou *task-switching*), est un point critique du contrôle cognitif, justifiant ainsi la nécessité de pouvoir apprendre quand switcher, ainsi que de pouvoir explorer les options disponibles.

Nous pourrons ainsi poser explicitement la question à laquelle nous tentons de répondre dans cette thèse sur les principes d'intégration de l'apprentissage, du contrôle cognitif, du switch et de l'exploration dans le cortex préfrontal humain.

Chapitre 1

Comment l'apprentissage par renforcement se rapproche du contrôle cognitif

Dans cette première partie de l'introduction, nous nous attacherons à démontrer quels liens existent entre les problèmes d'apprentissage et de prise de décision. Pour ce faire, nous commencerons par présenter les travaux théoriques sur l'apprentissage par renforcement, qui ont permis de faire avancer la compréhension des bases neurales de l'apprentissage. Si la dopamine et les ganglions de la base sont les aspects les plus importants pour valider la pertinence des modèles d'apprentissage par renforcement pour l'étude du cerveau, le cortex préfrontal apparaît également impliqué dans les représentations des valeurs de l'apprentissage par renforcement. On peut ainsi montrer que deux arguments parallèles, l'un théorique et l'autre biologique, nous poussent à étudier conjointement l'apprentissage et le contrôle cognitif. D'une part, du côté théorique, les modèles d'apprentissage par renforcement peuvent étendre leur utilité à des problèmes de planification, de décision. D'autre part, les corrélats neuronaux de l'apprentissage s'étendent au cortex préfrontal, qui est essentiel pour le contrôle cognitif. Nous justifions ainsi de manière théorique l'intrication complexe des questions d'apprentissage dans les questions de contrôle.

Nous commencerons donc par une présentation du cadre formel de l'apprentissage par renforcement et de son extension théorique aux problèmes de contrôle. Nous présenterons les corrélats neuronaux de l'apprentissage par renforcement (dopamine, ganglions de la base, cortex préfrontal) qui valident l'utilité de ces modèles. Dans une deuxième partie, nous détaillerons l'importance du cortex préfrontal dans les problèmes d'apprentissage au delà du cadre du renforcement, ainsi que les raisons pour lesquelles le contrôle et l'apprentissage s'imbriquent.

1.1 Apprentissage et ganglions de la base : le cadre de l'apprentissage par renforcement

Traditionnellement, les méthodes d'apprentissage sont divisées en trois grandes familles : l'apprentissage supervisé, l'apprentissage semi-supervisé et l'auto-apprentissage. Dans l'apprentissage supervisé, un « professeur » explique la règle ou donne des exemples justes de ce qu'il faut faire à l'apprenant, qui doit alors seulement implémenter la règle et la mémoriser. C'est une forme d'apprentissage très courante, non seulement dans la vie de tous les jours, mais aussi dans le domaine des réseaux de neurones. Dans l'apprentissage semi-supervisé, l'apprenant est seul face à l'environnement. Cependant, l'environnement réagit aux choix de l'apprenant, souvent sous la forme d'un renforcement (récompense ou punition) donnant une information sur le choix effectué. L'apprentissage par « la carotte et le bâton » est un bon exemple d'apprentissage semi-supervisé. Enfin, dans l'auto-apprentissage, aucune interaction avec l'environnement n'a lieu. Les problèmes de catégorisation ou de généralisation, par exemple, sont représentatifs de ce type d'apprentissage.

Une large part de nos comportements intentionnels est apprise d'une manière supervisée : durant l'enfance (tourner un robinet pour obtenir de l'eau), ou durant l'âge adulte (comment utiliser telle fonctionnalité de son ordinateur), souvent quelqu'un est présent pour nous donner l'exemple. Cet apprentissage, très spécifique à l'être humain, est alors à l'âge adulte si efficace qu'il est souvent quasi instantané. Peu étudié en tant qu'apprentissage (Krueger et Dayan, 2009 [138], Doll et al, 2009 [71]), il est essentiellement utilisé dans les paradigmes de flexibilité cognitive comme l'implémentation rapide d'une règle, plutôt que comme un

apprentissage.

Nous nous concentrons donc dans cette thèse sur l'apprentissage semi-supervisé, qui requiert une interaction avec l'environnement. En particulier, nous adoptons le cadre formel de l'apprentissage par renforcement, ou *reinforcement learning*, couramment noté RL par la suite.

1.1.1 Cadre formel

Sutton et Barto, dans [202], définissent l'apprentissage par renforcement comme « apprendre quoi faire – comment associer des actions à des situations – de sorte à maximiser un certain signal numérique de récompense ».

Donnons deux exemples de situations auxquelles l'apprentissage par renforcement peut s'appliquer.

Les loteries. Pour aller de chez moi au laboratoire, j'ai le choix entre un bus et un métro. Le bus va un peu plus vite, mais passe moins souvent. Le métro va un peu moins vite, mais passe plus souvent. Mon but est de minimiser le temps de trajet entre chez moi et le laboratoire. Si je connaissais parfaitement toutes les données du problème, je pourrais calculer l'espérance du temps de trajet global pour chacune des deux options et choisir celle qui minimise cette espérance. En l'absence de ces données exactes, je peux explorer les deux options et déduire petit à petit de mes observations – mon temps de trajet – quelle est la meilleure option. Dans cet exemple, à chaque essai, il y a un retour de l'environnement. Cependant, ce retour est stochastique. Il est donc nécessaire d'explorer les deux options afin de déterminer la meilleure option.

Le labyrinthe. Mon université est très grande, on s'y perd facilement et la cafétéria est loin. Je voudrais atteindre mon but (la cafétéria) et obtenir ma récompense (le thé) en minimisant l'énergie dépensée. La première fois, après s'être longtemps perdu, on finit par trouver la cafétéria. Au fur et à mesure des explorations suivantes, on trouve des endroits par lesquels on est passé les fois précédentes et qui permettent de mener au but. Petit à

petit, on réussit à construire toutes les étapes qui conduisent à l'objectif. Dans cet exemple, je n'ai pas de retour de l'environnement à chaque action (tourner à gauche, descendre un escalier,...), mais seulement lorsque j'ai atteint mon but. Il est donc nécessaire d'apprendre à sélectionner des actions en se basant sur l'état auquel elles mènent, et si celui-ci est plus proche du but ou non.

Ces deux exemples sont des archétypes du problème d'apprentissage par renforcement, montrant que ceux-ci permettent de résoudre des problèmes très variés : le premier pour des renforcements incertains, le deuxième pour des renforcements éloignés dans le temps ou dans l'espace.

Données du problème d'apprentissage par renforcement

Pour poser de manière formelle un problème d'apprentissage par renforcement, on doit définir les caractéristiques de l'environnement :

Espace d'états : L'environnement lui-même est décrit par un espace des états accessibles, qu'on note $\{s_i\}$. Ces états peuvent représenter des endroits physiques, comme dans l'exemple du labyrinthe ; ils peuvent représenter l'état d'un problème particulier, par exemple la configuration actuelle d'un jeu d'échec. Dans tous les cas, la détermination de l'espace d'états est cruciale. Il s'agit de représenter seulement les données pertinentes pour le problème. Nous ne poserons pas ce problème par la suite et supposerons que l'échantillonnage des états est déjà effectué.

Espace d'actions : L'interaction entre le sujet et l'environnement se fait par l'intermédiaire d'actions. En chaque état, un certain jeu d'actions est accessible au sujet, qui est un sous-ensemble de l'espace d'actions totales : $\{a_i\}$.

Effet des actions sur l'environnement : L'effet des actions choisies par le sujet sur l'environnement est décrit par une fonction de transition. La fonction de transition $p_{s_t,s_{t+1}}(a_t)$ décrit la probabilité que l'on se trouve dans l'état s_{t+1} au temps $t+1$ si on se trouvait dans l'état s_t au temps t et qu'on a choisi l'action a_t : $P(s_{t+1} = s2|s_t = s1, a_t) = p_{s_1,s_2}(a_t)$. Cette définition encapsule l'hypothèse que la propriété de Markov est vraie : mon prochain état ne dépend que de mon état et de l'action choisie à t, pas de mon passé. Notons par exemple que

si les actions sont des forces exercées, on ne peut pas décrire l'espace des états uniquement par une position à t. L'espace d'états doit être l'espace produit Position x Vitesse pour que la propriété de Markov soit vérifiée.

Effet de l'environnement sur le sujet : L'effet de l'environnement sur le sujet est défini par la fonction de renforcement dépendante de s_{t+1}, s_t, a_t : $P(r_{t+1}|s_t, a_t, s_{t+1})$, où r_t dénote le renforcement, positif (récompense) ou négatif (punition) reçu au temps t.

A tout instant t, s_t est observé, a_t est choisi et r_t observé par l'acteur. Toutes les autres données du problème (fonction de transition, fonction de renforcement) peuvent être soit parfaitement connues, soit complètement ignorées par l'acteur.

Reprenons les deux exemples donnés.

– Labyrinthe : L'espace des états est l'espace physique accessible dans le labyrinthe. L'espace d'actions est l'ensemble des déplacements disponibles à chaque endroit. La fonction de transition est déterministe : l'effet d'une action dans un état est toujours le même. La fonction de renforcement est nulle, sauf si s_{t+1} est le but du labyrinthe, auquel cas elle vaut une valeur positive représentant l'intérêt de ce but.
– Loteries. Il n'y a ici qu'un seul état qui est intéressant pour le problème : le point de départ. Il y a deux actions : prendre le bus ou prendre le métro. La fonction de renforcement est stochastique, anticorrélée à la longueur totale du trajet.

Définition du problème

L'objectif de l'acteur est de maximiser le renforcement cumulatif obtenu au cours du temps R_t. On définit $R_t = \sum_k \gamma^k r_{t+k}$ la récompense cumulée à partir du temps t, avec $0 \leqslant \gamma \leqslant 1$. Ce paramètre, appelé *facteur de discount*, mesure la dévaluation du renforcement avec le temps. Quand $\gamma = 0$, on a $R_t = r_t$; maximiser R_t revient à maximiser la récompense immédiate. Au contraire, quand $\gamma = 1$, chaque pas de temps compte autant dans la récompense cumulée. On choisit en général $0 < \gamma < 1$.

On définit la stratégie de l'acteur par la fonction $\pi(s, a) = P(a|s)$, la probabilité de sélectionner l'action a dans l'état s. Si cette fonction est déterministe, on peut poser plus simplement $\pi(s) = a$, l'action a sélectionnée dans l'état s.

Quand une stratégie est définie, on peut définir les fonctions de valeur suivantes :

- $V_\pi(s) = E_\pi(R_t|s_t = s)$. $V_\pi(s)$ est l'espérance de renforcement cumulé lorsque l'état est s, si la stratégie de sélection des actions après t est π. Elle représente la valeur attendue de l'état s dans le cadre de la stratégie π.
- $Q_\pi(s, a) = E_\pi(R_t|s_t = s, a_t = a)$. $Q_\pi(s, a)$ est l'espérance de renforcement cumulé lorsque l'état est s et l'action choisie a, si toutes les actions suivantes sont choisies selon la stratégie π. Elle représente la valeur attendue de la paire état s, action a dans le cadre de la stratégie π.

Le but du problème de renforcement peut alors se traduire de manière mathématique comme suit : trouver une stratégie π^* qui maximise pour tout s, $V^*(s) = V_{\pi^*}(s)$.

Les différents algorithmes d'apprentissage par renforcement proposent différentes méthodes de résolution de ce problème de maximisation. Celles-ci sont adaptées à diverses situations. Nous présentons dans la suite les plus représentatives.

Résolution du problème

i. Lien entre fonctions de valeurs et stratégies

Supposons qu'on connaisse $Q^*(s, a)$, la fonction de valeur optimale. Trouver une stratégie optimale est alors trivial : on l'appelle la stratégie *greedy*. Elle est définie comme suit :

- $\pi^*(s, a^*) = 1$ si $a^* = \mathbf{argmax}_a(Q^*(s, a))$
- $\pi^*(s, a) = 0$ sinon.

Cette stratégie choisit en tout état s l'action qui maximise la récompense future, elle répond donc au problème. Le problème peut donc se réduire à trouver la fonction de valeur optimale.

ii. L'équation de Bellman

On a pour toute stratégie π, par définition de $V_\pi(s)$,

$$V_\pi(s) = \sum_a \pi(s, a) Q_\pi(s, a)$$

Par ailleurs, on peut écrire $Q_\pi(s,a) = \sum_{s'} p_{s,s'}(a)[R_{s,s'}(a) + \gamma V_\pi(s')]$.

Notons n_A le nombre d'actions, n_S le nombre d'états. En reliant les deux équations précédentes, on peut donc écrire deux systèmes d'équations linéaires :

Un système de n_S équations à n_S inconnues (les valeurs $V_\pi(s)$) :

$$V_\pi(s) = \sum_a \pi(s,a) \sum_{s'} p_{s,s'}(a)[R_{s,s'}(a) + \gamma V_\pi(s')]$$

Un système de $n_S \times n_A$ équations à $n_S \times n_A$ inconnues (les valeurs $Q_\pi(s,a)$) :

$$Q_\pi(s,a) = \sum_{s'} p_{s,s'}(a)[R_{s,s'}(a) + \gamma(\sum_a \pi(s',a) Q_\pi(s',a))].$$

Ce sont les équations de Bellman (Bellman, 1952 [18], 1967 [19]).

Si on suppose que l'environnement est parfaitement connu, alors les coefficients de ces systèmes linéaires dépendent de valeurs connues : $p_{s,s'}(a)$, $R_{s,s'}(a)$ et $\pi(s,a)$. On peut alors simplement inverser ces systèmes pour obtenir les fonctions de valeur d'une stratégie.

De même, pour la fonction de valeurs optimale, on peut écrire $V^*(s) = \max_a Q^*(s,a)$ et

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} p_{s,s'}(a) V^*(s')$$

Ce qui permet de poser l'équation de Bellman optimale :

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} p_{s,s'}(a) \max_{a'} Q^*(s',a')$$

A nouveau, on a $n_S \times n_A$ équations avec le même nombre d'inconnues. Cependant, cette fois, les équations font intervenir un opérateur max qui implique une non linéarité dans le problème, rendant le système beaucoup plus difficile à résoudre de manière analytique.

Tous les algorithmes d'apprentissage par renforcement reposent largement sur la nécessité de consistance entre les estimations de la fonction de valeurs en deux points différents résumés par l'équation de Bellman.

iii. Une méthode de résolution model-based : dynamic programming

On va présenter ici une méthode *model-based* de résolution du problème par renforcement, la *value iteration algorithm* (Bertsekas, 1987 [21]). Le terme *model-based* signifie ici que cet algorithme repose sur l'hypothèse que l'acteur a un modèle parfait de l'environnement, notamment des fonctions de transition et de renforcement.

A chaque essai t, on estime une fonction de valeur au point (s_t, a_t) en approximant dans l'équation de Bellman $Q^*(s,a)$ par la valeur estimée en ce point $Q^t(s,a)$.

$$Q^{t+1}(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{s'=1}^{n_S} p_{s_t,s'}(a_t) \max_{a'} Q^t(s', a')$$

On peut alors démontrer, si $\gamma < 1$ (ce qui permet d'avoir une fonction contractante, donc d'utiliser un théorème du point fixe) et si chaque paire état-action est visitée une infinité de fois, que Q^t converge vers Q^*.

La condition demandant de visiter suffisamment chaque paire état-action est vérifiée par une stratégie ϵ-*greedy* : avec probabilité $1 - \epsilon$, on sélectionne l'action choisie par la stratégie *greedy*, soit l'action maximisant la fonction de valeur estimée ; avec probabilité ϵ, on choisit une autre action uniformément au hasard. Si ϵ est petit, cela permet d'assurer une exploration complète de l'espace, mais également une exploitation correcte de Q.

Cette méthode permet d'établir le premier rapprochement avec les problèmes de prise de décision, par l'intermédiaire de la notion de planification.

La planification, ou *goal directed behaviour*, est définie comme la tentative d'atteindre le plus facilement possible un objectif donné. On a dans ce cas une fonction de renforcement au moins partiellement connue : on peut regarder le fait d'atteindre d'un objectif comme une récompense, qu'il faut obtenir le moins péniblement possible, donc chaque pas ne conduisant pas à l'objectif est un effort. Le problème se réduit donc à choisir des actions qui maximisent le renforcement à long terme, la fonction de renforcement étant au moins partiellement connue.

Le type d'algorithme model-based décrit auparavant pourrait donc s'interpréter comme le fait de simuler virtuellement les différentes options disponibles suffisamment de fois, afin de pouvoir estimer correctement leur valeur optimale et de déterminer la meilleure stratégie

pour atteindre l'objectif fixé.

Ce type de d'algorithme répond aux problèmes dans lesquels les données du problème sont parfaitement connues, comme par exemple le problème archétype du voyageur de commerce : utiliser un réseau de transport pour visiter un certain nombre de villes, en minimisant le temps de trajet. Cependant, ce type d'algorithme ne peut pas être utilisé lorsqu'un modèle de l'environnement n'est pas connu.

iv. une méthode de résolution model-free : TD-learning

Contrairement aux méthodes *model-based*, les méthodes de résolution *model-free* ne font aucune supposition sur la connaissance de l'environnement par l'acteur. Au contraire, elles permettent, simplement en utilisant des informations obtenues localement et en les faisant circuler par proximité d'un état à son voisin ou suivant, d'estimer globalement les valeurs attendues optimales.

Comme précédemment, on estime approximativement $V^*(s)$ ou $Q^*(s,a)$ par des fonctions $V(s)$ ou $Q(s,a)$, qui sont corrigées en fonction de ce qu'on observe à chaque pas de temps.

Supposons qu'on avait estimé $V^*(s_t)$ par $V(s_t)$ et qu'on a reçu r_t à l'essai au temps t. Alors, on a une nouvelle estimation plus précise de $V(s_t)$ par la valeur : $r_t + \gamma V(s_{t+1})$. On peut alors mesurer l'inconsistance entre les deux prédictions :

$$\delta = r_t + \gamma V(s_{t+1}) - V(s_t),$$

soit la différence entre l'estimation au temps t et l'estimation au temps $t+1$. On appelle δ l'erreur de prédiction. L'équation de Bellman dit que l'erreur de prédiction doit être nulle. On va donc modifier l'estimation de $V(s_t)$ dans le sens de l'erreur commise : $V^{t+1}(s_t) = V^t(s_t) + \alpha\delta$.

On appelle le facteur de proportionnalité α $(0 < \alpha < 1)$, la vitesse d'apprentissage. Sous l'hypothèse, à nouveau, que toutes les paires (état, action) soient visitées une infinité de fois, et que α diminue à une vitesse appropriée, on peut alors montrer que cet algorithme converge vers V^*.

Nous verrons plus loin que δ, l'erreur de prédiction, est une valeur essentielle des problèmes d'apprentissage. Elle donne son nom à l'algorithme précédent, TD-learning pour *temporal difference learning*.

Watkins, 1992 [211], a proposé une version plus efficace de cet algorithme, *Q-learning*, qui repose directement sur les paires (états, actions), permettant un transfert plus simple vers la sélection de la stratégie. Cette fois encore, l'algorithme estime les valeurs des paires états-actions. Il met à jour ces estimations à l'aide de l'erreur de prédiction, qui est calculée comme $\delta = r_t + \gamma \max_a(Q(s_{t+1}, a)) - Q(s_t, a_t)$, de la manière suivante : $Q^{t+1}(s_t, a_t) = Q^t(s_t, a_t) + \alpha\delta$. Jaakkola et al, 1994 [114], ont montré la convergence de cet algorithme vers Q^*, avec les mêmes hypothèses que précédemment.

v. conclusions sur les algorithmes de RL

Notons que si la fonction de transition est uniforme et la fonction de renforcement indépendante de l'état à $t+1$, mais seulement dépendante de l'action choisie, alors maximiser $Q(s, a)$ devient indépendant de maximiser $Q(s', a)$. On peut alors s'intéresser uniquement au problème avec $\gamma = 0$, ce qui conduit à une simplification des algorithmes de type TD-learning. On a alors en effet $\delta = r - V$ ou $\delta = r - Q$ tout simplement, ce qui revient à la règle de Rescorla-Wagner, 1972 [175].

C'est ce cas qui nous intéressera le plus par la suite : il est en effet particulièrement adapté à la représentation des stimuli (les états) auxquels il faut répondre selon une certaine règle pour gagner des points, mais sans que cela influence quel sera le prochain stimulus présenté.

1.1.2 Dopamine et Ganglions de la Base

Dans ce paragraphe, nous présentons des résultats qui donnent leur légitimité aux algorithmes d'apprentissage par renforcement : le fait qu'ils semblent permettre de modéliser efficacement l'apprentissage instrumental ou pavlovien.

Dopamine

La dopamine (DA) est un neurotransmetteur produit par les neurones dopaminergiques de l'aire tegmentale ventrale (VTA) et de la partie réticulée de la substance noire (SNr), dans le mésencéphale (Doya, 2002 [73]). Ces neurones ont des projections très larges dans les ganglions de la base et le cortex, en particulier la partie médiale du cortex préfrontal (ACC, préSMA, vmPFC).

De nombreux arguments conduisent à s'intéresser à la dopamine dans le cadre de l'apprentissage basé sur des récompenses ou des punitions. Schultz et al, 1997 [192], cite en particulier des arguments pharmacologiques montrant que l'influence addictive de drogues telles que les amphétamines et la cocaïne est liée au fait qu'elles prolongent l'effet de la dopamine sur les neurones cible. Par opposition, des rats soumis à des produits bloquant les récepteurs de la dopamine apprennent moins vite à effectuer une action pour obtenir de la nourriture. De nombreuses expériences d'autostimulation sur des rats montrent également l'importance de la dopamine dans le traitement des récompenses : dans ces expériences, des rats implantés d'électrodes stimulatrices dans les noyaux dopaminergiques peuvent presser un levier pour stimuler eux-mêmes ces électrodes. Ils tendent à choisir cette action plutôt que celles menant à des récompenses primaires telles que la nourriture ou l'acte sexuel (Wise et Rompre, 1989 [212]).

Au-delà de l'importance de la dopamine dans le circuit de la récompense, de nombreux arguments soutiennent également son importance dans les questions d'apprentissage basé sur les récompenses. On sait en effet (Calabresi, 2000 [37]) que la dopamine influence les mécanismes de la plasticité neuronale des neurones du striatum, en particulier la LTP (*long term potentiation*, facilitation de la réponse neuronale) et LTD (*long term depotentiation*, effet inverse). La dopamine se présente donc comme ayant un rôle dans les mécanismes d'apprentissage.

L'utilisation des modèles d'apprentissage par renforcement en neurosciences a pris un essor tout particulier depuis la fin des années 90 quand ils ont permis d'éclairer le rôle plus précis de la dopamine dans l'apprentissage et la récompense.

Schultz, Montague et Dayan, 1997 [192], présentent une expérience de conditionnement

effectuée sur des singes. Les singes voient une lumière (conditionned stimulus CS) et doivent appuyer sur un levier pour obtenir une récompense (jus, unconditionned stimulus, US) après un certain délai. L'activité des neurones dopaminergiques de l'aire tegmentale ventrale (VTA) est mesurée. En début d'apprentissage, ceux-ci répondent de manière phasique à la récompense US (voir figure 1.1, page 33 en haut). Cependant, au cours de l'apprentissage, cette activation phasique se produit de plus en plus tôt avant US, tandis qu'on ne l'observe plus pour US. En fin d'apprentissage, l'activation phasique a lieu au moment de l'apparition de l'indice prédictif de la récompense (CS) (figure 1.1, page 33, milieu). Si la récompense est alors omise, on obtient une désactivation phasique des neurones dopaminergiques à l'instant prédit pour US (figure 1.1, page 33, en bas).

Ce pattern d'activations est très bien expliqué par l'hypothèse que la dopamine code une erreur de prédiction dans le cadre d'un modèle *Temporal Difference Learning* (Schultz et al, 1997 [192], Schultz, 2002 [193]).

Plus précisément, les auteurs modélisent l'espace des états comme une ligne temporelle incluant $t = 0$ (CS) ainsi que T (US), discrétisée en pas de temps court. Initialement, pour tout temps, $V(t)$ est faible : aucun renforcement particulier n'est attendu. Au début de l'apprentissage, à T, on a alors $\delta = r + \gamma V(T+1) - V(T) \approx r$, une grande erreur de prédiction, correspondant au pic d'activation face à la récompense. Au fur et à mesure de l'apprentissage, on va donc apprendre que $V(T) \approx r$ et l'erreur de prédiction deviendra nulle à T, ce qui correspond à la disparition du pic à la présentation de la récompense. L'erreur de prédiction va donc se propager en arrière jusqu'à être reportée sur le CS : au temps $T-1$ on a $\delta = 0 + \gamma V(T) - \gamma V(T-1) \approx \gamma V(T)$. On apprendra donc une valeur $V(CS) > 0$. Ainsi, quand CS est présenté après apprentissage, on a une erreur de prédiction positive : $\delta = 0 + \gamma V(CS) - \gamma V(CS-1) > 0$, ce qui correspond au pic d'activation observé pour CS. Enfin, si la récompense n'est pas présentée une fois le comportement appris, on aura une erreur de prédiction $\delta = 0 + \gamma V(T+1) - V(T) = -r$, ce qui correspond à la désactivation observée en cas d'omission de la récompense.

Ces études suggèrent donc que la dopamine code pour une erreur de prédiction.

Bayer confirme quantitativement (2005, [13]), et sans a priori de modèles, cette théorie dans une étude d'électrophysiologie où des singes doivent apprendre à quel moment ef-

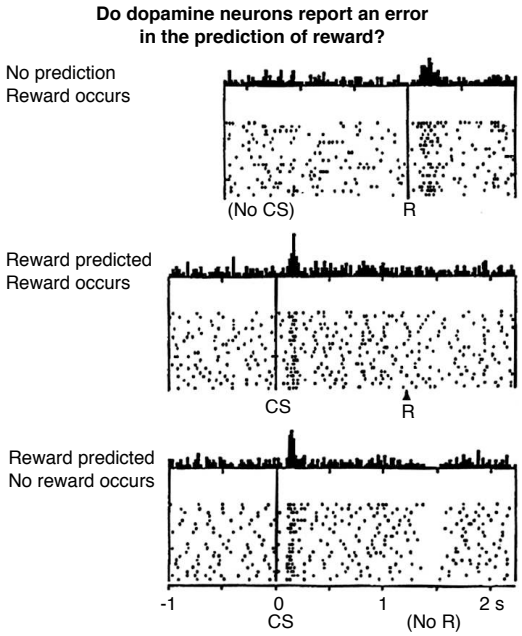


FIGURE 1.1 – Reproduit de Schultz et al, 1997 [192]. Signaux de sortie des neurones dopaminergiques face à une récompense non prédite (en haut), face à un CS suivi de la récompense prédite (au milieu), et face à un CS avec omission de la récompense prédite (en bas).

fectuer une saccade afin d'obtenir une récompense. Les auteurs effectuent une régression linéaire multiple entre les précédentes récompenses obtenues et le pattern d'activité de cinquante neurones dopaminergiques. Ils montrent que celui-ci est le mieux prédit par une différence entre la dernière récompense et une moyenne des récompenses précédentes, ce qui correspond effectivement à une erreur de prédiction de type δ.

Le rôle de la dopamine dans l'apprentissage a été depuis confirmé par d'autres études sur l'homme. L'analyse directe par électrophysiologie étant impossible, les démonstrations indirectes se font par l'intermédiaire d'études pharmacologiques et génétiques sur sujets sains, ainsi que par l'intermédiaire d'études sur des patients atteints de la maladie de Parkinson.

Pessiglione et al, 2006 [166], soumettent des sujets ayant reçu soit un agoniste de la dopamine (L-Dopa), soit un antagoniste (halopéridol), soit un placebo, à une expérience simple d'apprentissage de discrimination d'une paire avec probabilités $0,8 - 0,2$ d'une récompense fixée. Ils observent notamment que plus le niveau de dopamine présent est élevé, plus l'apprentissage de la meilleure option est rapide. Cet effet est parfaitement modélisé par un modèle d'apprentissage par renforcement incluant le rôle de la dopamine comme erreur de prédiction. Le paramètre d'échelle fixant la valeur arbitraire de la récompense est déterminé directement par le niveau de la réponse BOLD (*Blood Oxygen Level Dependent*, signal mesuré par le scanner) dans le striatum, une cible privilégiée des neurones dopaminergiques.

Critiques sur la dopamine et l'erreur de prédiction

Notons que, s'il est relativement bien accepté que la dopamine peut coder un signal d'erreur de prédiction, des critiques de ce modèle existent également. Parmi ces critiques, certaines se contentent de nuancer le type d'erreur de prédiction encodé par la dopamine (notamment, Dayan et Balleine, 2002 [62], Bertin et al, 2007 [20], et Daw, 2007 [58]). Par exemple, Bertin et al, 2007 [20], propose un modèle plus complexe d'apprentissage et montre que son erreur de prédiction correspond mieux au signal dopaminergique que l'erreur de prédiction obtenue par le modèle classique présenté ci-dessus. De même, Daw, 2007 [58],

montre que la dopamine répond parfois à une erreur de prédiction liée à l'action menant à la meilleure récompense, même si celle-ci n'était pas choisie, suggérant ainsi la nécessité d'adapter l'interprétation habituelle du rôle de la dopamine. La dopamine est également parfois citée comme répondant à la nouveauté (Redgrave et al, 2008 [173], 2006 [174]) ce qui peut être modélisé dans le cadre d'un modèle RL avec un bonus à la nouveauté (voir plus loin section 1.2.2, Kakade et al, 2002 [123]).

Cependant d'autres critiques reposent sur l'observation expérimentale de l'influence de la dopamine dans des situations non liées à l'apprentissage, en particulier dans des situations de contrôle cognitif. McNab et al, 2009 [144], montrent par exemple qu'améliorer sa capacité de mémoire de travail (maintien actif d'information dans le but de la manipuler) implique une augmentation de la densité des récepteurs D1 de la dopamine dans le préfrontal, rappelant ainsi que la dopamine joue un rôle important dans la mémoire de travail. Dans la même direction, Nagano-Saito et al, 2008 [153], montrent que la connectivité fonctionnelle fronto-striatale est réduite par une baisse du niveau de dopamine. Or le niveau de cette connectivité est corrélé avec l'efficacité du *task-switching* (passage de l'exécution d'une tâche à une autre nécessitant le contrôle cognitif). Ils suggèrent ainsi que la dopamine joue également un rôle important dans le task-switching, donc dans le contrôle cognitif.

Cohen et al (2002 [42], 1992 [41]) proposent que la dopamine augmente le gain de la fonction d'activation des neurones, améliorant ainsi le rapport signal sur bruit dans le préfrontal. Une autre théorie sur l'influence de la dopamine dans le contrôle cognitif est liée à son importance dans les boucles fronto-basales. Ce rôle est modélisé par l'équipe de R. O'Reilly et M. Frank (notamment dans 2005 [83]) et sera développé plus loin, section 3.2.1.

Ces critiques rappellent donc que le rôle de la dopamine n'est probablement pas limité à représenter l'erreur de prédiction. Malgré cela, le rôle de la dopamine dans le codage de l'erreur de prédiction reste largement accepté et nous présentons par la suite un modèle particulier d'apprentissage par renforcement utilisant cette hypothèse pour rendre compte de l'activité des ganglions de la base.

Le modèle acteur-critique des ganglions de la base.

le modèle acteur-critique

Le modèle acteur-critique d'apprentissage par renforcement est une variante particulièrement utilisée d'algorithme *model-free*, qui sépare d'une part la fonction d'apprentissage des valeurs (la critique) et d'autre part la fonction d'apprentissage d'une stratégie (l'acteur). En particulier dans cet algorithme, la valeur d'un état, d'une action ou d'un couple état - action est mise à jour exactement comme indiqué précédemment : $V^{t+1}(s) = V^t(s)+\alpha\delta$. Contrairement aux algorithmes précédents, la stratégie à chaque instant n'est pas inférée à partir des valeurs, mais elle est apprise de la même manière que les valeurs : $\pi_{t+1}(s,a) = \pi_t(s,a)+\alpha'\delta$. Plusieurs auteurs ont montré l'équivalence entre cet algorithme et d'autres algorithmes qui convergent, assurant sa fiabilité (Crites, Barto, 1985 [55]). Cet algorithme est proposé comme base de l'apprentissage par renforcement dans les ganglions de la base. Cette proposition initiale a été basée sur l'observation que la dopamine pouvait représenter un signal d'erreur de prédiction d'une part, et sur les effets de la dopamine sur la potentiation à long terme dans les ganglions de la base, d'autre part. Cela a conduit à interpréter les ganglions de la base comme entrainés par la dopamine.

rôle du striatum

Plusieurs études ont fourni des arguments en faveur de cette hypothèse. En électrophysiologie notamment, Samejima et Doya ont montré l'existence de neurones représentant des *action values* ou des *state values* dans le striatum de singes (Samejima et Doya, 2007 [186], Samejima et al, 2005 [187]). En imagerie par résonnance magnétique fonctionnelle (IRMf) dans le cerveau humain, Pessiglione et al, 2008 [165], montrent que l'activité du striatum corrèle avec des *action values*.

Dans une étude en IRMf, Seymour et al, 2004 [196]), utilisent une tâche simple de type labyrinthe (à quatre états, transitions non déterministes). Ils utilisent un modèle de type *TD-learning* qui leur permet de calculer un signal d'erreur de prédiction et de le corréler avec l'activité du cerveau. Ils montrent ainsi en particulier une activation du putamen et

du noyau caudé, ce qui est cohérent avec le fait que ces régions sont la cible de projections dopaminergiques.

La même équipe (O'Doherty et al) met également en évidence le rôle du striatum et de l'erreur de prédiction dans une autre étude (Schönberg et al, 2007 [190]). La variabilité inter-sujet est exploitée dans une tâche qu'à peu près la moitié des sujets parvient à apprendre et l'autre moitié non. Sont observées, d'une part, l'absence de signaux d'erreur de prédiction dans le striatum dorsal et ventral pour les non-apprenants contrairement aux apprenants ; d'autre part une corrélation entre la performance comportementale des sujets et la magnitude des signaux d'erreur de prédiction dans le striatum dorsal. Cette étude soutient la théorie selon laquelle l'erreur de prédiction joue un rôle crucial dans l'apprentissage, par l'intermédiaire du striatum.

L'étude précédemment indiquée de Pessiglione, 2008 [165], indique également un rôle du striatum (dorsal ou ventral) comme point d'utilisation du signal d'erreur de prédiction de la dopamine.

Si ces études montrent l'importance du striatum comme utilisateur de l'erreur de prédiction dans l'apprentissage par renforcement, d'autres études proposent de préciser son rôle dans le cadre du modèle acteur - critique.

Une étude menée par Kahnt et al, 2009 [122], montre une corrélation entre l'activité du striatum ventral et du striatum dorsal et un signal d'erreur de prédiction. Elle montre également d'une part l'existence d'une connectivité fonctionnelle entre la substance noire et le striatum dorsal, et d'autre part l'existence d'une connectivité fonctionnelle entre l'aire tegmentale ventrale et le striatum ventral. Ces résultats confirment le fait qu'un signal dopaminergique de type erreur de prédiction soit transmis au striatum. De manière cruciale, ils montrent également que seule la connectivité fonctionnelle entre le striatum dorsal et la substance noire est prédictive de l'impact du renforcement sur le comportement. Cela rappelle le résultat évoqué plus haut de Schönberg et al, 2007 [190], sur l'activité spécifique dans le striatum dorsal et son lien avec la performance comportementale des sujets, et suggère ainsi que le striatum dorsal jouerait un rôle d'acteur.

O'Doherty et al, 2004 [156], proposent explicitement une dissociation entre le rôle du stria-

tum ventral (critique) et celui du striatum dorsal (acteur) dans l'apprentissage instrumental. Dans une expérience en IRMf, ils montrent que le signal BOLD dans le striatum corrèle avec l'erreur de prédiction. De manière essentielle, ils contrastent deux tâches : dans l'une, les valeurs doivent être apprises mais aucun choix n'est effectué par le sujet ; dans l'autre, les valeurs doivent être apprises afin qu'un choix soit effectué par le sujet. Le contraste de ces deux tâches fait apparaître une activation plus forte du striatum dorsal, ce qui argumente en faveur de son rôle dans la sélection des actions et donc l'apprentissage de la stratégie.

critiques

Notons pour finir que, si le rôle des ganglions de la base et plus spécifiquement du striatum dans l'apprentissage est largement accepté, la validité du modèle acteur critique est largement controversée.

Joel, Niv et Ruppin, 2002 [118], montrent qu'il est difficile de l'implémenter dans un réseau de neurones biologiquement crédible. Ils passent en revue les différentes possibilités de rôle de la dopamine, argumentent que la dopamine sert également à effectuer une réduction dimensionnelle des entrées avant apprentissage.

Atallah et al, 2007 [7], ont proposé un autre modèle. Dans différentes expériences, ils entrainent des rats à un test de choix entre deux options. Des lésions chimiques réversibles sont effectuées dans le striatum ventral ou dorsal des rats, pendant les phases d'entraînement ou de test. On observe que les lésions du striatum ventral conduisent à une performance au niveau du hasard pendant l'entrainement et le test, indiquant que celui-ci est indispensable à l'apprentissage. Les lésions du striatum dorsal, par opposition, impliquent une performance au niveau du hasard seulement pendant la phase d'entraînement et non pendant la phase de test. Si la lésion est supprimée pendant la phase de test, la performance du rat redevient très vite proche de la performance contrôle. On en déduit que le striatum dorsal est essentiel pour la sélection des actions, mais pas pour l'apprentissage.

Les auteurs proposent donc de conserver le terme acteur pour le striatum dorsal, mais de ne pas conserver le terme critique pour le striatum ventral mais plutôt directeur. En effet,

les résultats observés favorisent plus un système où la stratégie est dirigée par les valeurs apprises, plutôt qu'indépendamment apprise par une critique *online*.

D'autres critiques (Dayan et Balleine, 2002 [62] ; Dayan et Niv, 2008 [61]) reposent sur des arguments de psychologie expérimentale. Ces auteurs insistent sur l'existence de deux systèmes d'apprentissage par renforcement : l'apprentissage pavlovien, ou apprentissage de valeurs d'états quand les actions sélectionnées par un animal n'ont pas d'incidence sur la prédiction ; et l'apprentissage instrumental où les contingences stimulus-actions-conséquences sont essentielles. Ils argumentent que ces deux types d'apprentissage ont lieu en parallèle et que le modèle acteur-critique ne peut rendre compte correctement de plusieurs phénomènes observés. Ils proposent un modèle *advantage Learning* qui repose sur une critique pavlovienne (valeurs de prédiction, erreur de prédiction habituelles), et un acteur instrumental reposant sur une compétition entre les avantages de chaque action. L'avantage est défini comme la différence entre la valeur de prédiction état-action et la valeur de prédiction état, soit l'avantage d'une action par rapport à la moyenne de toutes les autres. Elle peut être estimée par la mise à jour d'une autre erreur de prédiction : $\delta_a = \delta - Avantage$.

Ce modèle reflète de nombreuses observations expérimentales, notamment le transfert entre action apprise et habitude. En effet, lorsque l'avantage devient nul, la stratégie optimale a été apprise et le contrôle pavlovien prend le pas sur le contrôle instrumental. Ce modèle souligne en tout cas la complexité des notions d'erreur de prédiction et le rôle de la dopamine, ainsi que des limitations du modèle acteur critique.

Récompenses et punitions : voies Go-NoGo dans les ganglions de la base

Un autre point trop simplifié dans les modèles d'apprentissage par renforcement concerne les récompenses et punitions, traitées identiquement dans les modèles. Nous présentons ci-dessous des données indiquant une dissociation entre eux.

Pessiglione et al, 2006 [166], montrent que l'effet des renforcements positifs (récompenses) n'est pas modulé de la même manière par le niveau de la dopamine que l'effet des renforcements négatifs (punitions). L'équipe de Frank démontre également le rôle de la dopamine dans l'apprentissage par renforcement en insistant sur cette dissociation entre la carotte

et le bâton, dans une étude sur les patients parkinsoniens (2004 [88]) et dans une étude portant sur trois gènes de la dopamine (2007 [86]).

Dans cette étude, les auteurs montrent que les sujets améliorent leur capacité à apprendre par l'intermédiaire de récompenses lorsqu'ils prennent leurs médicaments augmentant le niveau de dopamine. Inversement, leur capacité à apprendre par l'intermédiaire de punitions est meilleure en l'absence de médicament. Le même effet de dissociation de l'apprentissage par les récompenses ou les punitions est montré chez des patients sains en fonction de leur niveau individuel de dopamine tonique par Cools et al, 2009 [48]. Cette dissociation est prévue par un modèle computationnel d'apprentissage dans les ganglions de la base.

Ce modèle repose sur un réseau de neurones représentant la structure biologique des ganglions de la base (figure 1.2 page 41). Ce modèle part de l'hypothèse que les ganglions de la base servent à sélectionner quelle action effectuer (action motrice ou de pensée) et à inhiber les autres. Ce filtrage s'effectue par l'intermédiaire de deux boucles :

– la boucle directe PFC → striatum → Pallidum (segment interne)/substance noire → thalamus → PFC. Cette boucle est une boucle Go qui, par une double inhibition entre le striatum et le thalamus, envoie un signal excitateur dans le cortex.
– la boucle indirecte PFC → striatum → Pallidum (segment externe) → Pallidum (segment interne)/subsance noire → thalamus → PFC. Cette boucle est une boucle NoGo qui, par une triple inhibition entre le striatum et le thalamus, envoie un signal inhibiteur dans le cortex.

Les récepteurs D1 de la dopamine, essentiellement situés dans le circuit direct, augmentent le contraste des cellules Go. Au contraire, les récepteurs D2 de la dopamine, essentiellement situés dans le circuit indirect, inhibent les neurones NoGo. Une augmentation phasique de la dopamine permet donc de diminuer l'inhibition globale (NoGo) et d'augmenter la sélectivité du signal (Go) en favorisant les neurones les plus actifs. Une diminution phasique présente l'effet contraire. Ce mécanisme permet d'expliquer la fonction de filtrage postulée.

Une nuance supplémentaire est introduite dans une étude génétique par Frank et al, 2007 [86]. Trois gènes sont explorés pour leur polymorphisme. Deux portent sur des récepteurs striataux de la dopamine, l'un portant sur les récepteurs D1 (correspondant dans le modèle de Frank à l'apprentissage par récompense), l'autre sur les récepteurs D2 (punitions).

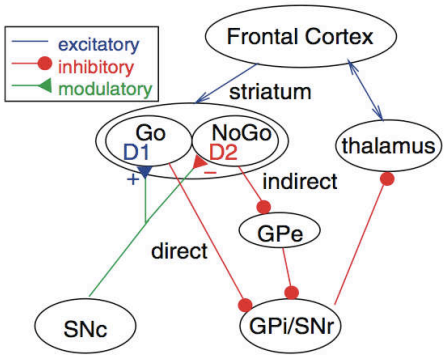


FIGURE 1.2 – Reproduit de Frank et al, 2004 [88]. Boucles directe et indirecte. GPe : globus pallidus segment externe. GPi : segment interne. SNr : partie réticulée de la substance noire. SNc : partie compacte de la substance noire (neurones dopaminergiques).

Le troisième porte sur les récepteurs préfrontaux de la dopamine. Les résultats comportementaux montrent l'effet attendu pour les deux premiers gènes (meilleur apprentissage par récompense ou punition, selon l'influence des allèles) et pas d'effet global pour le troisième gène. Deux paramètres sont obtenus par fitting d'un modèle d'apprentissage par renforcement : une vitesse d'apprentissage positif et une vitesse d'apprentissage négatif. Ces deux paramètres sont significativement différents en fonction des groupes définis pour les gènes D1 et D2, respectivement, soutenant les résultats exposés précédemment.

On voit donc que, si les modèles en réseau de neurones permettent plus de précision, on peut malgré tout adapter les algorithmes d'apprentissage par renforcement pour tenir compte des facteurs qu'ils ont mis en valeur, en particulier ici, la séparation de deux vitesses d'apprentissage, distinctes pour les erreurs de prédiction positives (récompense) et négatives (punition).

On a présenté les algorithmes d'apprentissage par renforcement et justifié leur pertinence par l'efficacité avec laquelle ils modélisent des effets d'apprentissage, aussi bien comportementalement que par rapport aux computations effectuées par la dopamine et les ganglions de la base. Dans la partie suivante, nous montrerons comment les problèmes d'apprentissage sont également implémentés par le cortex préfrontal.

1.2 Apprentissage et cortex préfrontal

Dans cette partie, on montrera l'implication du cortex préfrontal (PFC par la suite, pour *prefrontal cortex*) dans l'apprentissage, en insistant sur les liens réciproques entre apprentissage et contrôle cognitif. Pourquoi les situations d'apprentissage impliquent-elles également le cortex préfrontal ? Nous montrerons que l'implication du PFC peut avoir plusieurs raisons. Premièrement, l'apprentissage ou l'évaluation de la valeur des options est souvent nécessaire au contrôle cognitif pour la prise de décision, il est donc naturel de retrouver les valeurs d'apprentissage dans le PFC, siège du contrôle cognitif. Deuxièmement, l'exploration et la gestion de l'incertitude sont indispensables, à la fois à l'apprentissage et à la décision, et impliquent fortement le PFC. Enfin, nous montrerons que le contrôle cognitif peut être nécessaire à l'apprentissage dans des situations où celui-ci est, par exemple, plus complexe.

1.2.1 Cortex préfrontal et apprentissage par renforcement : l'apprentissage pour le contrôle

Le contrôle cognitif représente notre capacité à avoir un comportement adapté aux circonstances. Il semble aller sans dire qu'afin d'agir de manière adaptée, il est nécessaire de savoir, donc d'avoir appris, quelle est l'action adaptée aux circonstances. Il n'y a pas de contrôle cognitif sans apprentissage le précédant. C'est d'ailleurs flagrant dans les études portant sur le contrôle cognitif, qui impliquent presque toutes un fort entrainement des sujets précédant le protocole expérimental en lui-même.

On peut d'ailleurs percevoir cet aspect essentiel de l'apprentissage dans le contrôle cognitif par le fait que de nombreux signaux d'apprentissage, pouvant potentiellement être essentiels à des décisions, sont présents dans le cortex préfrontal, région indispensable au contrôle cognitif.

Dans cette partie, nous présentons tout d'abord les liens anatomiques et fonctionnels qui existent entre le cortex préfrontal et les ganglions de la base, qui permettent d'intriquer apprentissage et décision.

Par la suite, nous présentons de nombreux résultats montrant la présence de signaux d'apprentissage (notamment d'erreur de prédiction) dans le cortex préfrontal.

Boucles ganglions de la base-PFC

On sait depuis longtemps (par exemple Alexander, 1990 [3], 1986 [4], Cummings, 1995 [57]) que le lobe frontal et les ganglions de la base interagissent de manière privilégiée. Cette interaction se fait sous forme de nombreuses boucles parallèles ouvertes. En fonction de la région frontale concernée, on peut séparer ces boucles en cinq grandes familles, de structures très similaires : projection sur un noyau du striatum qui, lui, se projette sur le globus pallidus et la substance noire. Ces derniers sont connectés au thalamus qui se projette à son tour sur l'aire de départ. On parle de boucle ouverte car il est démontré que les points d'arrivée et de départ sont très proches, cependant des entrées d'autres régions extérieures à la boucle ont lieu à différents endroits de la boucle.

Plus précisément, les cinq circuits sont :
- Les circuits moteur et oculomoteur qui commencent dans l'aire motrice supplémentaire (SMA) et le *frontal eye field*, respectivement et projettent vers le putamen.
- Le circuit dorsolatéral préfrontal qui inclut le noyau caudé dorsal.
- Le circuit orbitofrontal qui inclut le noyau caudé ventral.
- Le circuit médial frontal qui inclut le noyau accumbens (striatum ventro-médial).

Beaucoup d'hypothèses sont proposées quant à l'intérêt fonctionnel de ces structures. On a vu une proposition dans le modèle d'apprentissage de Frank. Un rôle d'initiation de l'action est également reconnu à la boucle motrice. D'autres hypothèses ont également été posées, comme un mécanisme de filtrage de l'information pertinente (Joel et al, 2002 [118]) ou un mécanisme d'interruption du flux d'information (Frank, O'Reilly et al, par exemple [160], [159], [161]), qui sera exposé plus en détail dans les modèles de contrôle cognitif.

On voit en tout cas que les liens préfrontaux - sous corticaux sont importants, ce qui justifie en particulier de retrouver des signaux d'apprentissage par renforcement, présents dans les ganglions de la base, dans le cortex préfrontal.

Signaux d'apprentissage par renforcement présents dans le PFC

De nombreuses études ont utilisé le cadre formel de l'apprentissage par renforcement pour mettre en évidence des signaux d'apprentissage dans les boucles fronto-baso-thalamiques. Dans ces études, un modèle d'apprentissage par renforcement (par exemple, Q-learning) est *fitté* pour trouver les paramètres (par exemple, vitesse d'apprentissage) qui permettent au modèle d'expliquer au mieux le comportement des sujets (rats, singes ou humains). Cette procédure de fitting se fait en général en maximisant la vraisemblance (*likelihood*) du modèle, définie comme la somme des logarithmes de probabilités estimées par le modèle des choix effectués par le sujet. Ces paramètres sont ensuite utilisés pour générer, à partir des choix des sujets, les valeurs pertinentes du modèle : valeur estimée de l'état ou de l'action choisie, erreur de prédiction après obtention du renforcement, etc. Ces valeurs sont ensuite utilisées soit pour les corréler avec l'activité de neurones individuels (électrophysiologie animale), soit avec l'activité globale d'une zone du cerveau (humain), par la technique appelée *model-based* IRMf (détaillée par Corrado et Doya, 2007 [54], O'Doherty et al, 2007 [158]).

Ces techniques permettent de mettre en évidence la présence de signaux de type « erreur de prédiction » dans la boucle orbitofrontale. En particulier, Tanaka et al, 2004 [205], mais aussi d'autres études (Pasupathy et al, 2005 [163], Seymour et al, 2004 [196]), rapportent des activations correspondant à une erreur de prédiction dans l'OFC ou dans l'ACC, lors de tâches d'apprentissage légèrement complexes (tâche de type 'labyrinthe' pour Tanaka et al, tâche mélangeant incertitude et structure temporelle pour Seymour et al).

Dans le cadre de l'apprentissage par renforcement, a également été mise en évidence la représentation de signaux codant les valeurs attendues dans le cortex orbito-frontal (Kim et al 2006, [128] ; Tanaka et al, 2004 [205], 2006 [204], Valentin et al, 2007 [208]). Tanaka, qui effectue une tâche avec structure temporelle complexe, retrouve également des signaux de valeurs de prédictions dans le cortex préfrontal dorso-latéral et dans l'ACC.

O'Doherty et al, 2001 [157], montrent dans une tâche d'apprentissage (*probabilistic reversal Learning*, détaillée plus bas) une activation du cortex orbitofrontal corrélée à la valeur d'une récompense ou d'une punition, ce qui peut être relié au codage de valeurs dans

l'OFC, comme indiqué plus haut. Ce résultat est confirmé par Daw et al, en 2006 [60]. On détaillera plus loin cette étude, cependant, on peut noter également qu'elle met en évidence une corrélation positive entre la probabilité de l'action choisie et l'activité dans le cortex orbitofrontal et ventromédial. Cette probabilité étant fortement corrélée à la valeur de chaque option, ce résultat confirme les précédents.

D'autres modèles que des modèles d'apprentissage par renforcement sont également utilisés et produisent des valeurs comparables. Par exemple, Hampton, Bossaerts, et O'Doherty, 2006 [102], construisent un modèle bayésien qui leur permet d'estimer un signal comparable à l'erreur de prédiction : la mise à jour entre les probabilités a priori et a posteriori. Ce signal est corrélé à l'activité du striatum ventral et du vmPFC, conformément aux autres études. Par ailleurs, une estimation de la valeur attendue du choix effectué, la probabilité a priori que le choix soit correct, corrèle avec l'activité du cortex orbitofrontal et du cortex préfrontal médian.

De même, dans Boorman et al, 2009 [23], les auteurs utilisent un modèle bayésien et montrent que l'apprentissage de l'avantage lié à l'option actuelle est représenté dans le vmPFC tandis que l'apprentissage de l'avantage lié à l'option non sélectionnée est représenté dans le cortex frontopolaire.

Nous avons donc montré que de nombreuses études retrouvent dans le cortex préfrontal, et en particulier dans ses parties médiales ou ventrales (cortex orbitofrontal, cortex préfrontal ventromédial) qui sont des cibles préférentielles de la dopamine, des signaux proches de l'apprentissage par renforcement. Cela montre que l'apprentissage, nécessaire au cortex préfrontal pour les problèmes de prise de décision, est bien également en partie représenté dans le cortex préfrontal. Cela suggère donc une interaction forte entre l'apprentissage et le contrôle cognitif dans le cortex préfrontal.

1.2.2 Exploration et incertitude dans l'apprentissage

Nous avons jusqu'ici insisté sur la nécessité d'apprendre des valeurs pour prendre des décisions, soit l'apprentissage pour le contrôle cognitif. Nous montrons maintenant que des

aspects essentiels de l'apprentissage, notamment l'exploration et l'incertitude, impliquent le PFC, ce qui permet de penser que le contrôle cognitif peut également être nécessaire à l'apprentissage.

L'exploration et l'incertitude sont deux facteurs intrinsèques de l'apprentissage et de la prise de décision : l'exploration est indispensable pour former une représentation correcte des conséquences des actions, et l'incertitude est omniprésente en début d'apprentissage mais aussi pour les décisions dans des environnements bruités. Nous montrons dans cette section que la représentation de l'incertitude et la gestion de l'exploration, en situation d'apprentissage ou de prise de décision, impliquent le PFC médial et sont intimement liés.

Nous présentons tout d'abord les recherches théoriques liées au problème de l'exploration, avant d'évoquer les résultats empiriques impliquant le PFC et les notions d'incertitude dans les problèmes d'exploration.

Exploration, problèmes théoriques

Comme nous l'avons évoqué lorsque nous avons présenté les modèles d'apprentissage par renforcement, leur convergence n'est garantie que lorsque l'ensemble des états et des actions est testé un grand nombre de fois, afin que toutes les informations puissent être prises en compte. L'exploration est donc un facteur essentiel de l'apprentissage, et les hypothèses des théorèmes assurant la convergence des méthodes d'apprentissage par renforcement reflètent ce fait. Cependant, trop explorer implique de sélectionner des actions sous-optimales, ce qui rend la performance moins bonne qu'elle ne pourrait l'être. Se pose donc un dilemme : faut-il agir de manière la plus optimale possible, quitte à manquer des possibilités meilleures parce qu'on n'a pas pris le risque de les explorer, ou faut-il explorer plus, quitte à risquer de perdre de l'énergie sans compensation de résultats ?

Pour décrire ce problème, on utilise le terme de compromis (ou *trade-off*) exploration-exploitation : il s'agit de trouver le bon équilibre entre les comportements exploitant les informations connues et les comportements explorant les domaines encore pas assez bien connus. Ce processus d'exploration, de gestion du compromis, est stratégiquement et théoriquement complexe.

Le problème du compromis exploration-exploitation reste très largement non résolu. L'indice de Gittins (1974, [93]), seule solution optimale connue à un problème de compromis exploration-exploitation, n'est limité qu'à des cas très précis de problèmes (notamment les *bandits*) dans un environnement stable et pour un comportement à la limite. Cette solution n'est donc pas pertinente en général. Les autres propositions de solution, liées au cadre de l'apprentissage par renforcement, se concentrent sur la définition de la stratégie permettant cette exploration.

Nous avons déjà présenté une stratégie permettant l'exploration : la stratégie ϵ-greedy. Cette stratégie permet de choisir la plupart du temps la meilleure action, mais occasionnellement (avec probabilité petite ϵ), de sélectionner au hasard n'importe laquelle des autres actions possibles. Si cette stratégie assure en effet l'exploration finale de toutes les possibilités, elle semble ne pas être très subtile. En effet, toutes les autres options sont traitées uniformément, alors qu'il pourrait être intéressant de les dissocier. Notamment, si deux actions ont des valeurs très fortes et les autres sont toutes faibles, il est probablement plus intéressant d'explorer la $2^{\text{ème}}$ meilleure action plutôt que les autres, au cas où un complément d'information montrerait qu'elle est en fait la meilleure.

La stratégie nommée *softmax* (soit, maximum adouci) permet d'effectuer ce type d'exploration dirigée. Elle définit la probabilité de sélectionner une action a face à un stimulus s comme proportionnelle à la valeur $exp(\beta Q(s,a))$, soit $\frac{exp(\beta Q(s,a))}{\sum_{a'} exp(\beta Q(s,a'))}$. Dans la limite où β est très grand, cette valeur vaut 1 si a est l'action *greedy*, et 0 dans les autres cas. Avec des valeurs finies de β, cette valeur favorise la sélection des actions ayant les plus grandes valeurs attendues. Plus β est petit, plus les probabilités des autres options sont proches, biaisant le compromis exploration-exploitation en direction de l'exploration.

Cette stratégie est très largement utilisée et le paramètre β, souvent appelé *température inverse*, représente le degré d'exploration. Une combinaison des deux modèles précédents peut également être construite, permettant l'aspect aléatoire proche de la distraction du ϵ-greedy et l'aspect dirigé de l'exploration du softmax.

Cependant, on peut observer un problème dans ces stratégies d'exploration : en effet, imaginons une situation où deux solutions sont proches en valeur (par exemple, prendre le métro ou le bus pour aller de chez soi à son travail). Le softmax indique que, quel que

soit le nombre de fois où j'aurai testé les deux solutions, comme leurs valeurs sont proches, je continuerai à utiliser les deux, avec un léger biais vers la meilleure. Pourtant, heuristiquement, il semble évident que lorsque la connaissance du fait qu'une des deux méthodes s'avère légèrement meilleure est établie, alors on n'utilisera plus que celle-là, puisqu'il n'y a plus d'information à gagner à explorer la deuxième meilleure.

Plusieurs solutions peuvent être proposées pour résoudre ce problème. Tout d'abord, on pourrait imaginer que les paramètres de stratégie ne sont pas fixes : si β diminue, alors naturellement, l'exploration diminuerait au fur et à mesure du temps, conduisant à exploiter à long terme ce qui est connu. Cependant, il faudrait alors pouvoir déterminer à quelle vitesse β devrait diminuer, quand il devrait à nouveau augmenter (au cas où l'environnement change), etc. Cette solution ne semble donc pas possible sans implémenter des mesures de contrôle supplémentaires. Plusieurs autres méthodes ont été proposées pour encourager l'exploration, reposant généralement sur l'ajout d'un bonus à l'exploration (Dearden et al, 1998 [65]). Diverses techniques de calcul du bonus à l'exploration sont proposées. L'une des plus simple, nommée *novelty bonus*, propose simplement d'initialiser de manière optimiste les actions non connues, s'assurant ainsi qu'elles sont testées jusqu'à ce que le modèle ait appris leur valeur, ce qui assure ainsi une exploration minimale (Kakade et al, 2002 [123]). Cette méthode semble particulièrement pertinente en regard de l'observation d'activations phasiques de la dopamine face à la nouveauté, non expliquées par les modèles habituels de RL, mais qui pourraient l'être par l'introduction d'un bonus à la nouveauté (Kakade et al, 2002 [123], Redgrave et al, 2006 [174], Redgrave, 2008 [173]).

Nous présentons également une autre méthode de bonus à l'exploration qui nous semble particulièrement pertinente. La méthode d'exploration par *bonus à l'incertitude* propose de guider l'exploration en la dirigeant vers les actions permettant de diminuer l'incertitude, soit d'obtenir de l'information (Dayan et Sejnowski, 1996, [63]). Cette méthode demande de pouvoir mettre à jour à chaque essai, non pas seulement la moyenne d'une distribution, soit la valeur attendue comme dans les algorithmes habituels d'apprentissage par renforcement, mais également son écart type, afin d'avoir accès à l'incertitude sur la valeur attendue. Plusieurs méthodes peuvent être proposées. Une méthode proposée par Dearden et al, 1998 [65] et utilisée par Daw et al, 2005 ([59]), est la méthode de Bayesian Q-learning, proposant de mettre à jour non seulement $Q(s, a)$ pour chaque paire stimulus action, mais aussi la

distribution $P(Q(s,a) = q)$, par un processus d'inférence bayésienne. Une mesure très naturelle de l'incertitude est alors l'écart type de cette distribution.

Notons qu'il est démontré que le niveau de confiance (et donc d'incertitude) joue un rôle pour guider la prise de décision et l'adaptation, dans une étude portant sur des rats (Kepecs et al, 2008 [127]). Cela montre que la notion de confiance, liée à l'incertitude (omniprésente dans les décisions de la vie de tous les jours, Platt et Huettel, 2008 [170]) n'est pas nécessairement une variable de haut niveau cognitif et peut être calculée simplement.

Nous montrerons dans le paragraphe suivant les arguments cognitifs appuyant le rôle de l'incertitude dans l'exploration humaine.

Exploration et incertitude dans le PFC

Bien que peu d'études aient observé les corrélats neuronaux de l'exploration, nous montrons ici qu'elle implique de manière importante le PFC.

Hampton et al, 2006 [102], soumettent des sujets à une tâche de *probabilistic reversal learning*. Dans cette tâche, deux stimuli sont présents, l'un est gagnant avec probabilité p, l'autre avec probabilité $1-p$; le rôle des deux stimuli est périodiquement inversé. Le sujet doit apprendre à choisir le stimulus optimal. Les sujets sont familiarisés avec la structure de la tâche (niveau de stochasticité, probabilité de renversement). Cela permet de construire un modèle bayésien incluant la structure plus complexe de la tâche pour modéliser le comportement des sujets. Les auteurs observent alors une activation du PFC dorso-latéral (ainsi que de l'ACC) liée à une probabilité a priori incorrecte, soit une décision incertaine et exploratoire.

Dans une étude spécifiquement portée sur le problème de l'exploration, Daw, O'Doherty et al, 2006 [60], montrent que les décisions d'exploration impliquent des régions essentielles au contrôle cognitif. Dans cette étude, les sujets doivent choisir entre quatre actions. Chaque action rapporte des points tirés au hasard autour d'une moyenne ; la valeur moyenne de chacune des quatre actions change lentement au cours du temps. Le sujet doit donc en permanence apprendre la valeur de chaque option afin de sélectionner l'action la plus rémunératrice. Il est donc essentiel qu'il ne fasse pas qu'exploiter une seule action donnée,

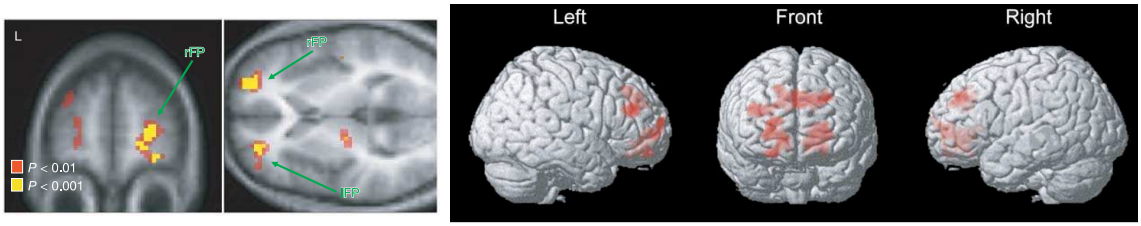


FIGURE 1.3 – Activations des mêmes régions du cortex préfrontal antérieur pour l'exploration (à gauche, issu de Daw et al, 2006 [60]), et l'incertitude (à droite, issu de Yoshida et Ishii, 2006 [216]).

mais explore les autres actions. Les auteurs définissent simplement la sélection d'une action exploratoire comme la sélection d'une action sous-optimale, d'après un modèle d'apprentissage par renforcement simple. Ils observent, spécifiquement pour les actions exploratoires, une activation du cortex préfrontal antérieur (voir figure 1.3, page 50), ainsi que du PFC dorso-latéral liée aux décisions exploratoires.

De manière cruciale, la même région du cortex préfrontal antérieur est rapportée comme codant pour l'incertitude, dans une étude menée par Yoshida et Ishii, 2006 [216] (voir figure 1.3, page 50). Dans cette étude, les sujets sont familiarisés avec la structure d'un labyrinthe et avec la navigation à l'intérieur de ce labyrinthe. Ils doivent par la suite, sans savoir quel est leur point de départ exact, atteindre un objectif situé à un point fixé du labyrinthe. Ils doivent donc tenter d'inférer, au cours de leur déplacement, leur position actuelle, et ce, en utilisant leur connaissance parfaite du labyrinthe et l'observation des conséquences des actions prises précédemment. Ce dessin expérimental favorise non seulement de forts niveaux d'incertitude, mais également la possibilité d'observer son évolution et sa réduction totale. Les auteurs utilisent un *hidden markov model* (HMM) ayant pour variables cachées la position actuelle dans le labyrinthe, mais également le 'mode actuel' du sujet, correspondant à sa stratégie d'inférence actuelle. En modélisant l'incertitude comme l'entropie sur les états cachés, les auteurs montrent une activation du cortex préfrontal antérieur, correspondant avec la région d'exploration proposée par Daw, O'Doherty et al, 2006 [60].

On a donc d'une part, une théorie (bonus à l'exploration) proposant qu'une méthode efficace d'exploration serait guidée par l'estimation de l'incertitude sur les différentes options, et d'autre part deux études étudiant indépendamment d'une part l'exploration, d'autre

part l'estimation de l'incertitude, montrant que ces deux processus sont pris en charge par la même région du cerveau. Si cela renforce l'hypothèse d'un lien entre exploration et incertitude, aucun lien direct n'est effectué entre les deux.

Frank et al, 2009 [85], proposent de faire le lien explicite, dans une étude comportementale, entre exploration et incertitude. Ils construisent un protocole expérimental permettant d'assurer un apprentissage correct malgré un grand nombre d'états, ceux-ci permettant de maintenir un certain niveau d'incertitude et d'exploration au cours de la tâche. Les auteurs modélisent le comportement des sujets en proposant que les sujets choisissent entre deux options (non précisément fixées et soumises à ajustements), et en estimant la distribution de probabilité, pour chacune de ces deux options, d'être liée à une chance de récompense supérieure à celle attendue. Cette formulation approximative permet une mise à jour bayésienne simple ainsi que l'estimation aisée de l'écart-type sur chacune des deux distributions, permettant d'encoder l'incertitude liée à chacune des deux options. Les auteurs montrent que, parmi un certain nombre de modèles d'apprentissage et de décisions incluant différentes manières d'explorer, le modèle utilisant l'incertitude pour guider l'exploration est celui qui correspond au mieux au comportement des sujets. Ils valident ainsi comportementalement le rôle de l'incertitude pour l'exploration.

Incertitude

On a montré que l'incertitude permettait de guider l'exploration dans l'apprentissage, et que ces processus impliquaient le cortex préfrontal ventro-médial. On montre dans cette section que l'incertitude permet également de guider d'autres aspects de l'apprentissage, par l'intermédiaire du cortex cingulaire antérieur (ACC).

De nombreux facteurs indirects peuvent influencer quelle devrait être la méthode optimale d'apprentissage. Par exemple, le poids relatif à donner à différentes informations devrait dépendre du degré d'incertitude lié à cette information ; l'horizon temporel des événements à prendre en compte devrait dépendre de la rapidité de changement de l'environnement, etc.

Deux propositions, non contradictoires, sont présentes dans la littérature sur la gestion de

ces informations supplémentaire dans le cerveau.

ACC et incertitude

La première proposition repose sur le rôle de l'ACC dans l'intégration des informations diverses pour l'apprentissage et la décision. Si les théories principales du rôle de l'ACC reposent sur la notion de conflit (Botvinick et al, 2004 [27]) ou d'erreur likelihood (Brown et al, 2005 [35]), d'autres études montrent des activités de l'ACC dans des situations ne correspondant pas à ces théories. Par exemple, Budhani et al, 2007 [36], effectuent une étude de *reversal learning* avec des singes. Ils observent des activations de neurones de l'ACC lors de réponses correctes pendant l'acquisition d'un problème, ce qui n'est ni une situation de conflit, ni une situation d'erreur.

Kennerley et al, en 2006 [126], proposent la théorie selon laquelle l'ACC ne sert pas à détecter des erreurs, mais à intégrer l'historique des renforcements afin d'apprendre la valeur des actions. Dans cette étude, des singes ayant subi une ablation de l'ACC ne parviennent pas à maintenir un comportement conduisant pourtant à un renforcement.

Allant dans le sens de cette étude, Hayden et al, en 2009 [105], montrent que des neurones de l'ACC de singes sont sensibles à des renforcements fictifs, qui sont utilisés pour l'apprentissage de l'action appropriée.

Behrens et al, en 2007 [17], démontrent que l'ACC contribue à évaluer la variabilité, donc l'incertitude intrinsèque, du problème et à régler en conséquence la vitesse d'apprentissage. Dans cette étude, les sujets effectuent une tâche de *probability tracking* dans laquelle un des deux stimuli présents à l'écran est récompensé avec probabilité p, l'autre avec probabilité $1 - p$. La valeur de cette probabilité p change plusieurs fois au cours de l'expérience. Crucialement, la durée de stabilité de p change au cours de l'expérience, représentant un environnement incertain changeant plus ou moins vite – de volatilité plus ou moins grande. Les auteurs de cette étude modélisent à l'aide d'un modèle bayésien l'apprentissage non seulement des probabilités p, mais aussi de la volatilité de l'environnement à tout instant. Cette volatilité guide la vitesse d'apprentissage de p : en effet, lorsqu'un environnement change fréquemment, il faut donner peu de poids aux événements anciens ; tandis que dans un

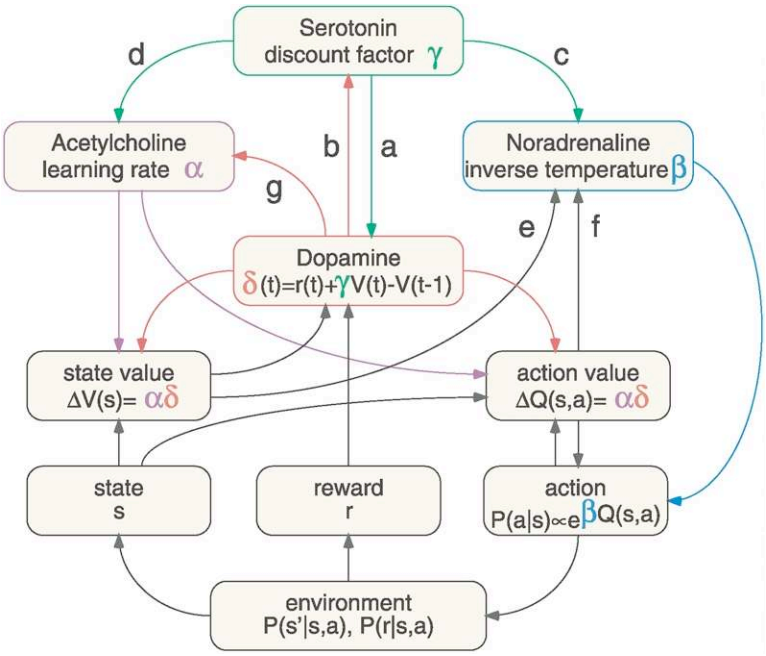


FIGURE 1.4 – Extrait de Doya, 2002 [73]. Schéma des interactions possibles entre les neuromodulators representant le signal d'apprentissage global et les métaparamètres α, β, γ ; l'expérience d'un agent sous la forme de fonctions de valeurs ; et l'état, l'action, et la récompense de l'environnement.

environnement stable, il faut limiter l'influence des événements immédiats. L'étude montre une corrélation de l'activité de l'ACC avec la volatilité perçue par les sujets, comme inférée par le modèle. L'étude conclut donc sur le rôle de l'ACC dans le guidage de l'apprentissage dans des situations complexes, en prenant notamment en compte l'incertitude intrinsèque de l'environnement.

Neuromodulateurs

La deuxième proposition repose sur le rôle de différents neuromodulateurs dans la modulation de l'apprentissage.

On a déjà vu que la dopamine semblait plus ou moins bien représenter un signal de type *erreur de prédiction* dans un modèle d'apprentissage par renforcement. Se reposant sur de nombreuses données existantes, Doya, en 2002 (*metalearning and neuromodulation* [73]) propose que chacun des trois autres neurotransmetteurs soit relié à un paramètre du modèle RL :

- γ (facteur de discount temporel) pour la sérotonine. En effet, plusieurs études montrent le rôle de la sérotonine dans le compromis délai/valeur de renforcement (Tanaka et al, 2004 [205] Schweighofer et al, 2007 [194]).
- α (vitesse d'apprentissage) pour l'acétylcholine, connue pour contrôler le stockage et la mise à jour de souvenirs.
- β (parametre d'exploration) pour la norépinephrine. Nous reviendrons largement dans la dernière partie sur les arguments liés à cette proposition, par exemple, ceux de Cohen, McClure et Yu, 2007 [43].

1.2.3 Cortex préfrontal et apprentissage de règles : le contrôle pour l'apprentissage

Dans la section précédente, nous avons montré que l'exploration et la gestion de l'incertitude dans l'apprentissage impliquaient le cortex préfrontal, essentiellement dans ces parties médiales. Si cela montre que le cortex préfrontal, région indispensable au contrôle cognitif, est nécessaire également à l'apprentissage, cela ne montre pas que le contrôle cognitif lui-même est nécessaire à l'apprentissage. Nous développons ce point dans cette partie, tout d'abord en montrant que le contrôle cognitif est nécessaire en début d'apprentissage, puis en montrant qu'il est également nécessaire lors de situations d'apprentissage plus complexes.

Dual processing theory : nécessité de contrôle en début d'apprentissage

Daw, Niv et Dayan, 2005 [59], proposent un modèle pour concilier les rôles respectifs à la fois des ganglions de la base et du cortex préfrontal dans l'apprentissage.

Les auteurs s'appuient sur des expériences de psychologie comportementale montrant qu'il y a non pas un seul, mais deux systèmes d'apprentissage et de contrôle : un système habituel,

peu flexible, et un système dirigé vers un but (*goal-directed*), plus flexible.

Les études démontrant l'existence de ces deux systèmes mettent en jeu l'apprentissage d'actions afin d'obtenir une récompense. Elles testent par la suite le comportement des sujets (rats, humains) lorsque la récompense est dévaluée (par exemple par satiété, Valentin et al, 2007 [208]). Deux types de comportement sont possibles (voir par exemple Rangel et al, 2008 [172]). On parle de comportement *habituel* lorsque le sujet continue à effectuer les actions apprises, malgré la dévaluation de la conséquence de ces actions. Ce comportement est peu flexible et insensible à l'issue des actions. On parle de comportement *goal-directed* (orienté vers un but) lorsque le comportement est sensible à l'issue des actions, i.e lorsque le sujet n'effectue plus les actions lorsque l'issue a été dévaluée.

Le comportement habituel est généralement observé lorsque les sujets sont surentrainés. Le comportement dirigé vers un but est observé lorsque les sujets sont modérément entrainés, mais aussi lorsque le problème à apprendre est plus complexe.

Par ailleurs, des études ont montré qu'un type d'apprentissage pouvait prendre le pas sur l'autre : en particulier, des rats dont l'effet de la dopamine sur le striatum était bloqué continuaient à présenter un comportement goal-directed même après un très long entrainement.

On en déduit que les deux systèmes d'apprentissage fonctionnent en parallèle et reposent sur des substrats neuronaux partiellement différents.

Daw, Niv et Dayan, 2005 [59], proposent d'associer l'apprentissage dirigé vers un but et l'apprentissage habituel à l'apprentissage par renforcement *model-based* et *model-free* respectivement. En effet, un algorithme d'apprentissage par renforcement *model-based* est immédiatement sensible à la dévaluation de l'issue, puisqu'il utilise le modèle de l'environnement pour effectuer des prédictions à chaque essai. Il reflète donc la flexibilité face au changement de renforcement du comportement dirigé vers un but. Dans l'apprentissage par renforcement *model-free*, par contre, la structure du problème est implicite dans les valeurs apprises, et toute modification de renforcement est lente à se propager. Cela reflète donc bien l'apprentissage habituel, peu sensible à la modification de la valeur de l'issue. Ces deux apprentissages ont lieu en parallèle. L'arbitrage entre les deux contrôleurs pour la sélection

des actions se fait par rapport à l'incertitude inhérente à chaque modèle. L'incertitude liée au *model-based RL* est liée à la difficulté d'utiliser parfaitement le modèle de l'environnement pour recalculer à chaque essai la valeur de chaque option, impliquant nécessairement des approximations. L'incertitude liée au *model-free RL* est liée à la dépendance de chaque estimation dans d'autres estimations.

Ce modèle souligne le fait que le système dirigé vers un but, associé au préfrontal et à la planification, donc proche du contrôle cognitif, est plus important en début d'apprentissage.

De nombreuses données d'IRMf confirment cet effet. Par exemple, dans une expérience de résolution de problème où les sujets doivent découvrir une règle grammaticale abstraite pour effectuer une catégorisation binaire, Strange et al, 2001 [201], observent une diminution de l'activité fronto-polaire relative à la résolution de problème au fur et à mesure de l'apprentissage. Koechlin et al, 2002 [135], observent une diminution de l'activité du PFC latéral lors de l'apprentissage de séquences motrices. De même Chein et al, 2005 [40], effectuent une expérience d'apprentissage complexe sur des paires de formes abstraites et observent une diminution de l'activité dans un réseau de régions incluant le PFC dorsolatéral, le PFC ventral, l'ACC et l'insula. Ce résultat est conforté dans le même article par une méta-analyse d'expériences portant sur un long apprentissage et montrant une désactivation progressive de ce réseau, liée à la pratique. Ce résultat est interprété, ainsi que Daw et al, 2005 [59], comme une transition d'un comportement contrôlé (ou *goal-directed, model-based*) à un comportement automatisé (ou *habituel, model-free*). Ce résultat est indépendant de la nature de ce qui est appris.

On trouve également des résultats d'électrophysiologie confortant cette théorie. En particulier, Asaad et al, 1998 [5], effectuent des expériences de *reversal learning* sur des singes en mesurant l'activité de neurones du PFC latéral. Ils montrent que l'activité préfrontale, lors de la sélection d'actions pour une paire de stimuli surentraînés, est extrêmement réduite par rapport à une paire de stimuli en cours d'apprentissage.

Toutes ces études sont convergentes pour dire qu'une raison de l'implication du cortex préfrontal latéral dans l'apprentissage est liée à la nécessité de contrôle cognitif en début d'apprentissage.

Nécessité de contrôle dans les problèmes d'apprentissage plus complexes

Différents types de complexité demandent l'implication de contrôle cognitif pour un apprentissage plus efficace. Nous montrons que le contrôle cognitif joue un rôle dans l'apprentissage lorsque celui-ci s'effectue à un niveau hiérarchique plus complexe ou lors d'un switch dans l'apprentissage.

Niveau de complexité

Daw, Niv et Dayan, 2005 [59], rappellent également que le comportement dirigé vers un but (lié au contrôle cognitif) persiste après surentrainement dans les cas où le problème à apprendre est plus complexe (plus grande séquence d'actions avant l'obtention d'une récompense, par exemple). Ce résultat, ajouté au fait que plus de contrôle est nécessaire en début d'apprentissage, semble montrer que l'implication du cortex préfrontal et du contrôle cognitif dans l'apprentissage dépend du niveau de complexité perçu du problème (un problème d'apprentissage est perçu plus difficile au début qu'à la fin).

Koechlin et al, 2002 [135], montrent effectivement cet effet dans une étude en IRMf. Dans cette étude, les sujets doivent d'une part apprendre des séquences motrices, d'autre part des séquences de tâches à effectuer sur des stimuli. Ce contraste permet d'observer deux niveaux hiérarchiques de complexité dans l'apprentissage de séquences d'actions. On observe effectivement une activation du striatum ventral lié à l'apprentissage des séquences motrices, tandis qu'on observe une activation du cortex préfrontal médial antérieur lié à l'apprentissage des séquences de tâches.

De nombreuses autres études montrent l'implication du cortex préfrontal dans des apprentissages plus complexes. On peut citer une activation du cortex fronto-polaire dans une expérience de résolution de problème où les sujets devaient inférer une règle grammaticale abstraite pour apprendre à catégoriser des stimuli (Strange et al, 2001 [201]).

Switch

D'autres résultats montrent également que l'apprentissage n'est pas toujours un processus continu comme modélisé par l'apprentissage par renforcement, mais peut s'effectuer par paliers, avec certains essais jouant un rôle plus important que d'autres (Gallistel, 2004 [92]).

Dans des tâches d'acquisition de règles sensorimotrices, Brovelli et al, 2007 [32], montrent une activation spécifique du PFC dorso-latéral gauche lors du premier essai récompensé, reflétant peut-être l'encodage certain d'une association correcte pendant l'apprentissage. Cools et al, 2002 [49], obtiennent un résultat comparable dans une tâche de *probabilistic reversal learning*. Afin d'apprendre à choisir le stimulus optimal, le sujet doit interpréter correctement les erreurs observées, soit comme étant des erreurs aléatoires, soit comme le signe que le rôle s'est inversé. L'étude montre une activation spécifique du PFC ventro-latéral et du Striatum ventral lors du dernier essai précédant le changement comportemental du sujet qui commence alors à choisir l'autre stimulus. Cet effet ne peut pas être interprété comme un signal d'erreur de prédiction de type RL puisqu'il est indépendant du nombre d'erreurs persévératives précédentes. Il semble donc être plutôt représentatif d'un pallier, ou d'un switch.

On voit donc que les régions latérales du cortex préfrontal, qui sont les régions du contrôle cognitif (Koechlin et al, 2003 [136]) sont essentiellement impliquées dans l'apprentissage, notamment lorsque celui-ci est complexe : initialement, lorsqu'il faut repartir à zéro, mais peut-être également lorsqu'il convient de trancher un compromis exploration-exploitation (activation dlPFC pour une action exploratoire, Hampton et al, 2006 [102]).

Conclusion : On a donc montré l'implication du cortex préfrontal (en particulier dorso-latéral ou médial) dans les situations d'apprentissage particulières mais courantes, que ce soit lié à la complexité (initiale ou durable) du problème, à la nécessité de décisions d'exploration ou de switch, ou à la nécessité de réguler l'apprentissage en fonction de l'incertitude présente dans l'environnement. On a montré que l'apprentissage était nécessaire au contrôle cognitif et on a évoqué un faisceau d'arguments permettant de supposer l'importance du contrôle cognitif pour l'apprentissage.

Les problèmes d'apprentissage et de contrôle cognitif sont donc bien intégrés à l'intérieur du cortex préfrontal.

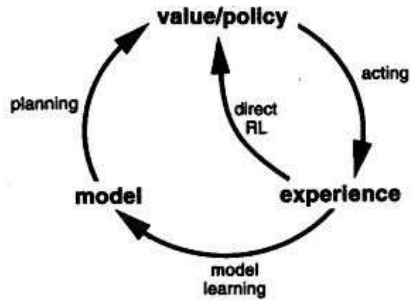


FIGURE 1.5 – Reproduit de Sutton et Barto, 1998 [202]. Liens entre apprentissage et contrôle, ou planification.

Pour conclure cette première partie bibliographique, deux schémas semblent adaptés à résumer l'interaction apprentissage – contrôle cognitif dans le cortex préfrontal et les ganglions de la base.

Tout d'abord, le schéma extrait de Sutton et Barto, 1998 [202] (figure 1.5) permet de résumer le fait que la planification peut être modélisée par l'apprentissage, rappelant que les comportements intentionnels requièrent un apprentissage pour pouvoir trancher sur l'opportunité des différentes options. L'apprentissage est donc indispensable au contrôle cognitif.

De même, de manière schématisée, la figure 1.4, page 53, extraite de Doya (2002 [73]), montre que l'apprentissage nécessite du contrôle, représenté ici par les systèmes régulateurs des neurotransmetteurs, mais effectué également par le cortex préfrontal.

Chapitre 2

Importance de la hiérarchie dans le contrôle cognitif et l'apprentissage

Nous avons vu dans le chapitre précédent que le contrôle cognitif et l'apprentissage étaient particulièrement imbriqués l'un dans l'autre lorsque les problèmes d'apprentissage augmentaient en complexité. Cette complexité se traduit souvent par une structure organisée sous-jacente au problème, qu'il convient d'inférer afin d'améliorer sa résolution, dans le cadre de l'apprentissage, ou d'utiliser au mieux dans le cadre du contrôle.

Nous mettrons en valeur dans ce chapitre le rôle d'une structure organisée, souvent de manière hiérarchique, dans l'optimisation de l'apprentissage et du contrôle cognitif. Nous montrerons tout d'abord que le cerveau est capable d'une grande flexibilité d'apprentissage : en fonction de la nécessité perçue dans l'environnement, il peut s'adapter pour apprendre (et oublier) vite ou lentement. Nous montrerons ensuite que l'homme utilise des structures hiérarchiques dans les problèmes de planification pour prévoir le contrôle séquentiel d'actions. Enfin, nous montrerons que le cerveau permet d'extraire des règles structurées de tâches cognitives pour optimiser nos décisions.

2.1 Différentes échelles de temps, différentes vitesses d'apprentissage

Lorsque l'environnement change fréquemment, il n'est pas bon d'utiliser de l'information obtenue longtemps auparavant pour prendre des décisions et il est indispensable de donner un fort impact aux informations obtenues récemment. A l'inverse, lorsque l'environnement est stable, il convient de donner peu de poids aux informations récentes afin d'éviter d'agir de manière impulsive. Dans le cadre de l'apprentissage par renforcement, contrôler cette importance relative donnée à l'information ancienne/récente se fait par l'intermédiaire du paramètre de discount γ, qui régule le poids relatif donné à la récompense d'un essai par rapport à celle de l'essai précédent.

Tanaka et al, 2004 [205], ont cherché à montrer d'une part que nous pouvions effectivement régler ce paramètre en fonction de la pertinence face au problème actuel et d'autre part comment notre cerveau encodait ce réglage. Ils soumettent leurs sujets à deux tâches d'apprentissages dans lesquelles les transitions entre trois états sont déterministes. Dans la condition SHORT, l'action a_1 apporte toujours la récompense r_1, l'action a_2 la punition $-r_1$. Dans la condition LONG, l'action a_1 apporte deux fois une faible punition mais permet d'accéder à une forte récompense, alors que l'action a_2 apporte deux fois une faible récompense mais permet d'accéder à une forte punition. Optimalement, dans les deux problèmes, il faut choisir tout le temps l'action a_1. Pour apprendre cela, il n'est nécessaire d'utiliser que de l'information à court terme pour la condition SHORT. Par contre, il est indispensable d'utiliser de l'information à long terme pour l'apprendre, dans la condition LONG, faute de quoi, impulsivement, on choisit les petites récompenses immédiates plutôt que la grande récompense plus tard.

L'analyse des données d'IRM fonctionnelle montre différentes régions d'activations pour différentes échelles temporelles : notamment le cortex orbitofrontal et le striatum pour une courte échelle (confirmant leur rôle dans le traitement des récompenses) ; le PFC dorso-latéral et le noyau raphé (producteur de la sérotonine) pour une longue échelle, confirmant le rôle du PFC dorso-latéral dans l'apprentissage plus complexe, en particulier dans la planification, ainsi que celui de la sérotonine dans la régulation de l'utilisation de l'information.

L'analyse model-based, s'appuyant sur de l'apprentissage par renforcement avec différentes valeurs du paramètre γ, permet de mettre en évidence un gradient d'activation dans l'insula et le striatum : le sens ventro-antérieur à dorso-postérieur encodait des erreurs de prédictions ou des récompenses prédites en tenant compte d'une longue échelle temporelle à une courte échelle. On en déduit que les boucles cortico-basales permettent d'encoder l'information observée à différents niveaux d'échelles de temps.

Fusi et al, 2007 [89], proposent également l'existence d'un modèle à deux échelles temporelles permettant d'expliquer des données obtenues chez le singe. Dans une expérience de *reversal Learning* (voir plus haut pour une description), les auteurs observent qu'après une erreur (essentiellement la première erreur liée au renversement de la règle), la performance des singes revient très rapidement (en un essai) au niveau du hasard, puis augmente lentement jusqu'à un niveau optimal. Ils observent cependant que pour une autre paire de stimuli n'étant jamais renversée, présentée aléatoirement entre la paire subissant des renversements, on n'observe pas cet effet de retour immédiat au niveau du hasard après une erreur.

Pour modéliser ce comportement, les auteurs utilisent deux populations de *spiking neurons* sélectives des deux réponses disponibles, conformément aux résultats d'électrophysiologie (Asaad et al, 1998 [5]). Ces deux populations sont en compétition par auto-excitation et inhibition mutuelle, de telle sorte qu'une des deux émerge plus ou moins rapidement, permettant une décision. Les quatre connexions entre les deux stimuli et les deux populations sont plastiques, dépendant de la récompense obtenue. Une récompense positive entraine un renforcement de la connexion s_t - a_t et une diminution des connexions s_t - {autre action}, à la vitesse $\alpha+$. Un renforcement négatif entraine une diminution de la force de connexion entre le stimulus et les deux populations, à la vitesse $\alpha-$. Fixer le paramètre $\alpha-$ vingt fois plus fort que $\alpha+$ permet d'expliquer l'effet de Switch immédiat observé après une erreur, et de réapprentissage graduel observé par la suite.

En plus de ce mécanisme de plasticité, les auteurs ajoutent un autre mécanisme de plasticité à plus longue échelle. Le principe est exactement le même, avec des vitesses d'apprentissage beaucoup plus faibles. De cette sorte, lorsque les associations changent souvent et sont équilibrées, comme dans le cas de la paire renversée, l'apprentissage à long terme annule ses propres effets. Par contre, lorsque les associations ne changent pas souvent, comme dans le

cas de l'autre paire, l'apprentissage à long terme implémente un biais suffisamment solide pour qu'une seule erreur ne suffise pas à remettre le singe au niveau du hasard, comme ce qui est observé expérimentalement.

Ce mécanisme permet donc de mieux tenir compte de la structure historique de la tâche pour adapter les vitesses d'apprentissage à la situation, rapide quand l'environnement change vite, lente dans le cas inverse.

Behrens et al, 2007 [17], formalisent cette tentative en construisant un *ideal observer* (observateur bayésien idéal) pour une tâche de type *probabilistic reversal Learning*.

Dans l'expérience décrite par Behrens, le sujet doit apprendre la probabilité d'obtenir une récompense liée à la paire de stimuli, qui peut changer, plus ou moins souvent, dépendant de ce qui est appelé la volatilité de l'environnement. Cette volatilité peut également changer, impliquant la nécessité pour le sujet d'adapter sa vitesse d'apprentissage (plus forte en cas de volatilité forte, plus faible dans un environnement stable).

L'observateur idéal bayésien doit donc, à partir de la variable d'observation récompense, inférer des probabilités sur un ensemble hiérarchiquement structuré de variables cachées : la variable cachée contrôlant le changement de la volatilité, la volatilité contrôlant le changement de taux de récompense, et le taux de récompense contrôlant la récompense. Cette inférence optimale peut être effectuée après chaque observation par propagation des croyances, étant donné le modèle génératif décrivant le problème et la dépendance hiérarchique entre les différentes variables.

Cette structure de problème permet d'apprendre à partir des récompenses observées, non seulement l'information dont on a besoin pour décider une action (le taux de récompense), mais également le poids à donner à chaque nouvelle observation, la valeur de l'information.

On voit donc que cela permet un apprentissage et des décisions à deux niveaux différents, hiérarchiquement liés entre eux. Tenir compte du deuxième niveau hiérarchique plutôt que de se contenter du premier (comme serait le cas d'un RL sans modification de la vitesse d'apprentissage) permet d'utiliser plus d'information présente dans l'historique de la tâche et d'obtenir de meilleures performances. En exhibant des activations de l'ACC qui corrèlent

avec la volatilité, les auteurs montrent que les sujets présentent cette flexibilité théorisée par le modèle, mettant ainsi à nouveau en valeur le rôle essentiel de l'apprentissage à différentes échelles hiérarchiques dans les problèmes de décision plus complexes.

2.2 Organisation hiérarchique temporelle de la planification

On a vu dans la première partie le lien théorique très fort entre l'apprentissage et la planification. Les modèles d'apprentissage par renforcement *model-based*, censés rendre compte de certains aspects de la planification, reposent beaucoup pour leur résolution sur la recherche dans des arbres de possibilité. En effet, l'utilisation d'un modèle de l'environnement permet de ne pas seulement tirer des conséquences pour les actions choisies et les conséquences observées, mais aussi sur toutes les possibilités d'actions et de conséquences au temps $t+1$, puis $t+2$, et ainsi à n'importe quel niveau de profondeur.

Cette exploration d'arbres de possibilités est soumise au problème de la dimensionnalité : dans les problèmes complexes impliquant plusieurs actions et états, la taille de l'arbre en fonction du niveau de profondeur devient très vite trop grande pour permettre une planification exacte. On peut facilement se représenter ce problème dans le cadre du jeu d'échecs, lorsqu'on tente de planifier à 2, 3, 4 voir 5 coups d'avance.

Il est donc naturel d'introduire dans les problèmes de planification des sous objectifs constituant chacun un problème séparé. Comme ces sous-objectifs ne contiennent qu'une partie des états ou actions du problème, ils permettent de réduire la dimensionnalité localement et d'appliquer en pratique l'axiome « diviser pour conquérir ». On voit ici qu'une structure hiérarchique est ainsi introduite dans le problème de planification afin de le simplifier. On montrera dans ce chapitre qu'une représentation hiérarchique de la planification, bien que débattue, est souvent observée chez les sujets. Nous montrerons différents modèles permettant d'expliquer comment cette hiérarchie peut être implémentée.

2.2.1 Un exemple : la tour de Londres

Une expérience particulièrement utilisée pour étudier la planification est l'expérience de la tour de Londres (Owen, 1997 [162]). Dans cette expérience, trois billes de couleurs différentes sont positionnées sur trois piques de taille une, deux ou trois billes. Le but est de déplacer les billes d'une position de départ à une position d'arrivée en utilisant le moins de transitions possibles.

Dehaene et Changeux (1997 [67], 2000 [68]) proposent un modèle hiérarchique de type réseau de neurones pour résoudre ce problème de planification. Ce modèle sépare deux systèmes hiérarchiques parallèles : un système d'évaluation et un système d'exécution. Au premier niveau hiérarchique se trouvent l'état actuel et l'objectif côté évaluation, les gestes basiques (sélection d'une bille, sélection d'un emplacement libre) du côté exécution. Au deuxième niveau hiérarchique (niveau opérations) se trouvent les objectifs (positions de billes) atteignables, les objectifs restants et les billes déplaçables du côté évaluation ; les opérations (codant pour une séquence de gestes basiques) de l'autre. Au dernier niveau hiérarchique se trouvent les récompenses du côté évaluation (motivation : distance à l'objectif positive ; correct : distance diminue ; erreur : distance augmente). La planification en tant que telle est effectuée au dernier niveau hiérarchique (niveau plan) côté exécution. A l'aide d'un système de mémoire de travail stockant l'état actuel et d'évaluation de la prochaine action (erreur ou correct) avant de l'effectuer, ce niveau permet de valider une action avant de l'effectuer en influençant les deux autres niveaux d'exécution sous son contrôle hiérarchique.

Ce modèle permet de répliquer les performances de sujets humains sains ou présentant des lésions, en modélisant une lésion du cortex préfrontal comme une impossibilité d'activer les unités du niveau plan.

Il est cependant très important de noter que ce modèle est construit *en dur* : toutes les connexions entre neurones sont implémentées pour encoder la structure de la tâche sans apprentissage impliqué. Cela pose la question de la validité de ce modèle, même s'il permet de souligner l'importance d'une influence top-down d'une région contrôlant hiérarchiquement (au niveau du plan) des actions plus simples.

Nous montrons dans la suite d'autres modèles tentant de résoudre ces problèmes.

2.2.2 Modèles hiérarchiques de la planification

De nombreuses études ont montré que, même dans les faits les plus simples de la vie de tous les jours, une structure hiérarchique peut être observée : chaque objectif est composé de sous-objectifs eux-mêmes décomposables à nouveau, etc. Un exemple particulièrement étudié (Cooper et Shallice, 2000 [51]) consiste en l'objectif de se préparer le matin. Celui-ci peut être décomposé dans les sous objectifs : se laver, s'habiller, déjeuner. L'objectif de déjeuner peut être décomposé en sous-objectifs préparer et consommer le café, préparer et consommer une tartine de confiture. A nouveau, préparer et consommer le café peut être décomposé en différent niveaux de sous objectifs jusqu'à atteindre le niveau d'actions élémentaires : prendre une cuiller, verser, boire, ... On voit ainsi qu'on peut décomposer sous forme d'un arbre un comportement dirigé vers un but. Notons que cet arbre n'est pas strict, dans le sens où plusieurs étapes pourraient être interverties ou effectuées en même temps ou d'une autre manière ; de même que d'autres comportements pourraient utiliser des sous-parties de cet arbre.

Bien que la structure hiérarchique de ces comportements soit largement acceptée par de nombreux auteurs (Cooper et Shallice, 2000 [51] ; Botvinick et Plaut, 2004 [27] par exemple), la question de l'implémentation de ce type de comportements hiérarchiques nourrit une large polémique, en particulier entre les deux équipes citées précédemment. Rapidement, Cooper et Shallice (2000 [51], 2006 [52] [53]) argumentent pour une représentation hiérarchique explicite des objectifs et sous-objectifs et de leur résolution à l'aide de schémas. Au contraire, Botvinick et Plaut (2004 [27], 2006 [29]) plaident pour une représentation implicite de la hiérarchie de ces séquences d'actions. L'outil essentiel d'évaluation des deux modèles concurrents (voir figure 2.1, page 69) est la capacité à répliquer les données comportementales observées, testées sur la tâche de faire du café ou du thé. Au delà de l'exécution correcte de ces tâches dans différents contextes et avec différents moyens, le type d'erreurs effectuées par le modèle bruité (pour représenter soit une distraction habituelle, soit un syndrome de désorganisation) est comparé à ce qui est observé empiriquement.

Le modèle proposé par Cooper et Shallice, 2000 [51], repose sur une représentation explicite de la structure hiérarchique d'une tâche. Cette représentation se fait sous forme d'arbre, dont les nœuds représentent soit des (sous-) objectifs, soit des schémas. Un schéma est défini

comme un moyen d'atteindre un objectif. Au niveau le plus faible, c'est une action simple (feuille de l'arbre) ; autrement, c'est un sous-arbre. La structure de l'arbre est définie *en dur* dans le modèle. L'activation de nœuds de l'arbre dépend de cinq types de connexions :

- influence top-down, de telle sorte qu'un schéma hiérarchiquement plus haut (faire du café) peut activer un schéma plus bas (faire chauffer de l'eau),
- influence de l'environnement, de telle sorte qu'un système de gating est en place : un schéma n'est activé que si certaines préconditions sont remplies et si l'objectif n'est pas atteint,
- influence latérale, de telle sorte à implémenter un 'winner take all' (une seule unité active au même niveau hiérarchique) : on observe alors des séquences d'actions, plutôt que plusieurs à la fois,
- auto influence, pour le maintien en mémoire de travail,
- bruit.

Les simulations montrent que ce modèle parvient à effectuer des séquences d'actions et à reproduire les erreurs observées empiriquement. Les limitations soulevées suggèrent l'absence d'implémentation plausible de ce modèle, notamment pour les *gating units* permettant de contrôler le début (conditions pour effectuer un schéma) et la fin (obtention de l'objectif) d'une séquence. Par ailleurs, le fait que le codage de la hiérarchie de la tache soit effectué à la main pose le problème de l'acquisition de cette hiérarchie : comment est-elle apprise ?

Botvinick et Plaut proposent en réponse à ce modèle, un modèle de type réseau de neurones (2004 [27]). Ce modèle est volontairement non hiérarchique. Il est muni d'une couche d'entrée (états actuels), connectée entièrement à une couche cachée de neurones très interconnectés, eux-mêmes connectés entièrement à une couche de sortie (actions) agissant sur l'environnement. Dans ce modèle, l'encodage de séquences d'actions est implicite et émergent. En effet, la présence de connexions récursives dans la couche cachée permet de faire persévérer des informations précédentes, encodant ainsi un contexte temporel ou environnemental nécessaire à l'implémentation d'une action au sein d'une séquence. Le réseau de neurones est soumis à un entrainement supervisé : séquences d'entrées et de sorties imposées, apprentissage des forces de connexions par backpropagation de l'erreur observée. Les séquences d'actions sont alors encodées sous forme d'attracteurs dynamiques du réseau de neurones (au lieu de schémas ponctuels comme plus haut), représentant essentiellement

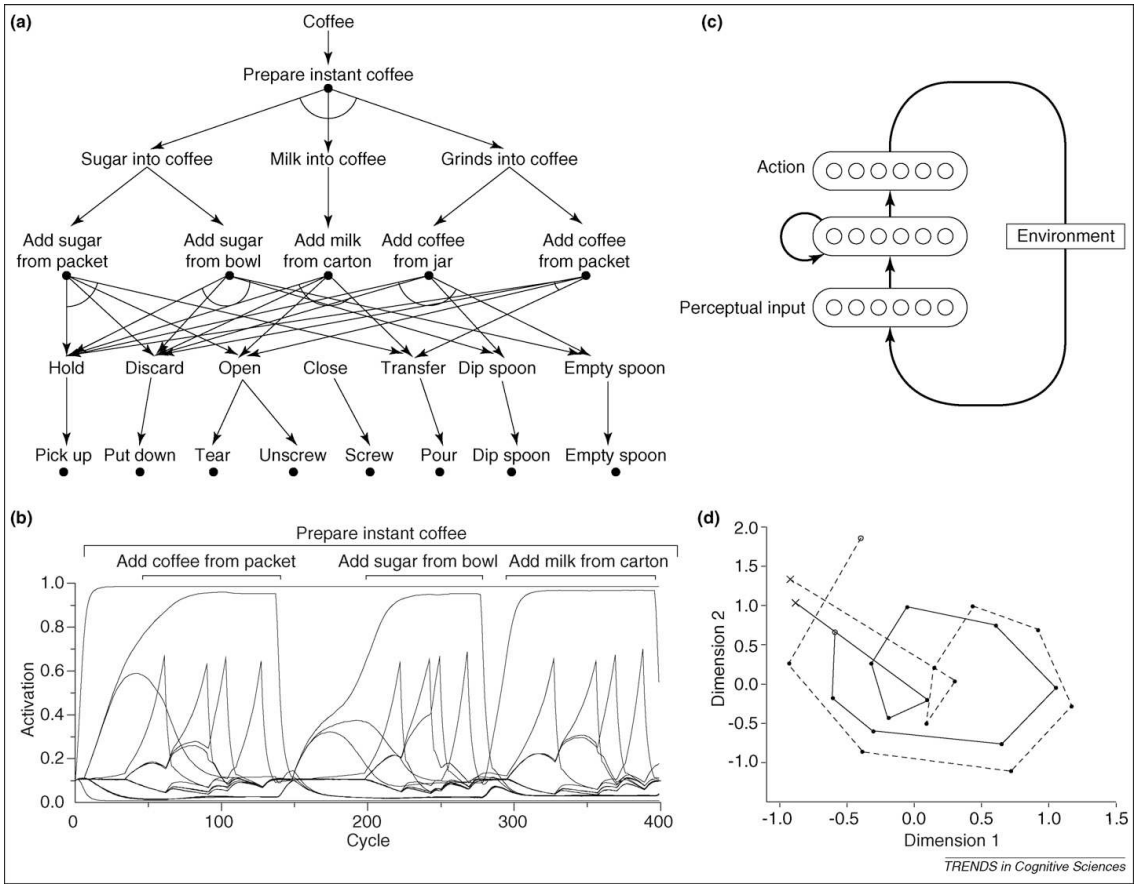


FIGURE 2.1 – Figure extraite de [28], [51], [52], [27]. A gauche **(a,b)**, le modèle de planification explicitement hiérarchique de Cooper et Shallice ([51], [52]). A droite **(c,d)** le modèle implicitement hiérarchique de Botvinick et Plaut ([28] [27]), reposant sur des attracteurs dynamiques internes (deux trajectoires pour deux tâches proches représentées en bas).

les statistiques de transition apprises lors de l'entraînement. De même que le précédent modèle, celui-ci permet d'effectuer des séquences d'actions. Il réplique certains résultats de production d'erreurs d'inattention ou du syndrome de désorganisation de l'action.

Si la polémique sur la nécessité de représenter ou non explicitement la structure hiérarchique de la planification du comportement n'est pas résolue, on peut en tout cas louer l'effort lié à la nécessité de pouvoir apprendre des hiérarchies fourni par Botvinick et Plaut.

On montrera par la suite plusieurs articles effectuant une tentative de rapprochement des aspects intéressants de chaque modèle – hiérarchie explicite, apprentissage possible, codage distribué plutôt que ponctuel.

En particulier, Botvinick, 2007 [28], modifie légèrement son précédent modèle pour tenir compte des données d'imagerie (voir section 2.3.2) indiquant l'existence de structures fonctionnelles hiérarchiques dans le cerveau, en particulier dans le cortex préfrontal. Il ajoute simplement une deuxième couche cachée de neurones, connectée de manière récurrente à elle-même et à l'autre couche cachée, mais non connectée à l'entrée et à la sortie. Il montre comme précédemment que, même sans encoder de structure hiérarchique explicite, le modèle parvient à effectuer une séquence de tâches (thé, café). Par contre, il montre que les neurones cachés sont plus sensibles au contexte de la tâche effectuée qu'au stimulus immédiat, et que cet effet est plus fort pour les neurones cachés les plus distaux (couche cachée non connectée aux entrées/sorties). Cette simulation propose une explication pour le rôle hiérarchique du PFC dans le comportement (représenté par la couche distale). Cependant, cette simulation montre également que même avec seulement un (2004 [27]) ou deux (2007 [28]) niveaux hiérarchiques permis dans la structure du réseau de neurones, on peut effectuer une tâche de structure hiérarchique de profondeur plus grande que deux. Si l'introduction d'une possibilité de hiérarchie implique effectivement que celle-ci sera naturellement utilisée pour encoder une hiérarchie, celle-ci n'est pas indispensable au comportement séquentiel.

2.2.3 Apprentissage par renforcement hiérachique

Les algorithmes d'apprentissage par renforcement que nous avons évoqués en première partie semblent particulièrement mal adaptés pour tenir compte d'une structure hiérarchique tem-

porelle ou séquentielle des problèmes d'apprentissage ou de planification. En effet, l'espace des états est plat, dans le sens où tous les états sont traités identiquement, indépendamment d'une quelconque structure. De même, les actions sont des unités indivisibles discrètes, représentant un unique pas de temps, de telle sorte que « faire le café »ne peut être divisé en sous actions sans faire perdre à « faire le café »son existence en tant qu'action.

De nombreux travaux ont été effectués pour introduire dans les modèles d'apprentissage par renforcement une structure hiérarchique, ou une possibilité d'abstraction temporelle (McGovern et al, 1998 [143] ; Sutton et al, 1999 [203]).

Un des travaux les plus intéressants est l'introduction de la notion d'options dans l'apprentissage par renforcement. Heuristiquement, une option peut être vue comme une routine représentant un tout cohérent, mais nécessitant d'être décomposée en actions elle-même, comme « faire le café ». Une option correspond plus ou moins à la notion de schéma développée par Cooper et Shallice : une méthode pour atteindre un sous-objectif. C'est donc une généralisation de la notion d'action basique.

Formellement, une option est définie par trois éléments : un ensemble d'initiation dans lequel l'option peut être sélectionnée, une stratégie π : Etats x Actions $\rightarrow [0, 1]$ et une condition de terminaison b : Etats $\rightarrow [0, 1]$, souvent déterministe, indiquant l'état représentant le sous-objectif visé par l'option. Une action peut être vue comme un cas particulier d'option. On peut alors définir une stratégie sur l'espace produit Etats x Options. Si une option non basique est choisie dans un état, elle est effectuée jusqu'à sa terminaison. Il est important de noter que le problème n'est plus markovien : l'état et l'action précédente ne suffisent pas à décrire l'historique du problème, le fait d'être dans une option ou non est également important. Il est démontré que ce processus est équivalent à un processus de décision semi-Markov (processus dont le changement d'état est régi par un processus de Markov, mais qui ajoute une notion de durée à chaque état).

Si les options sont données, au même titre que les actions, dans un problème, les méthodes habituelles d'apprentissage par renforcement peuvent être transférées immédiatement. On définit de manière analogue la valeur d'un état ou d'un couple (état, option) pour une stratégie sur des options. On peut également écrire des équations de Bellman simples et optimales, les utiliser pour de la planification dans le cadre du dynamic programming,

ou pour de l'apprentissage dans le cadre du Q-learning adapté. La mise à jour se fait à la terminaison d'une option, et en tenant compte du nombre de pas de temps discrets effectués à l'intérieur d'une option et du renforcement r cumulé (avec time discount) pendant l'exécution de l'option :

$$Q(s_t, option_t) \rightarrow Q(s_t, option_t) + \alpha[r + \gamma^k \max Q(s_{t+1}, options) - Q(s_t, option_t)]$$

Avec le même type de conditions que celles posées pour le Q-learning, cet algorithme converge vers Q* et permet donc de trouver une stratégie optimale sur l'espace des options.

Un exemple donné dans (Botvinick et al, 2008 [26], Sutton et al, 1999 [203], Mc Govern et al, 1998 [143]) permet de mettre en évidence la pertinence de l'utilisation d'options dans l'apprentissage par renforcement pour modéliser la structure hiérarchique de problèmes d'apprentissage et de planification. Dans ce problème, un espace *gridworld* carré est séparé en quatre pièces reliées l'une à l'autre par le milieu de chaque mur (voir figure 2.2 **A)**, page 73). Huit options sont ajoutées aux quatre actions basiques (gauche, droite, avant, arrière) : leur ensemble d'initiation est une des quatre pièces, leur condition de terminaison est déterministe à l'atteinte du sous-objectif « porte », leur stratégie est déterministe et menant par le chemin le plus court à une des deux portes présentes dans la pièce (voir figure 2.2 **B)** page 73). Les auteurs montrent que l'ajout de ces huit options permet une exploration beaucoup plus rapide de l'ensemble de l'espace du problème et donc d'atteindre un objectif quelconque beaucoup plus efficacement (voir figure 2.2 **D)** page 73).

Plusieurs atouts des options sont ainsi mis en valeurs :

– ils offrent une solution possible aux problèmes de dimensionnalité qui rendent les algorithmes de RL impuissants quand les espaces d'états sont trop grands
– ils permettent d'explorer plus efficacement
– ils permettent de transférer une connaissance déjà apprise (la stratégie d'une sous tâche : aller à la porte) à d'autres problèmes. Cette capacité est connue en psychologie et nommée transfert positif.

Notons cependant que différents problèmes sont également introduits. On observe par exemple un phénomène de transfert négatif. Lorsque les options introduites dans le problème ne sont pas pertinentes, l'apprentissage est au contraire ralenti par l'augmentation

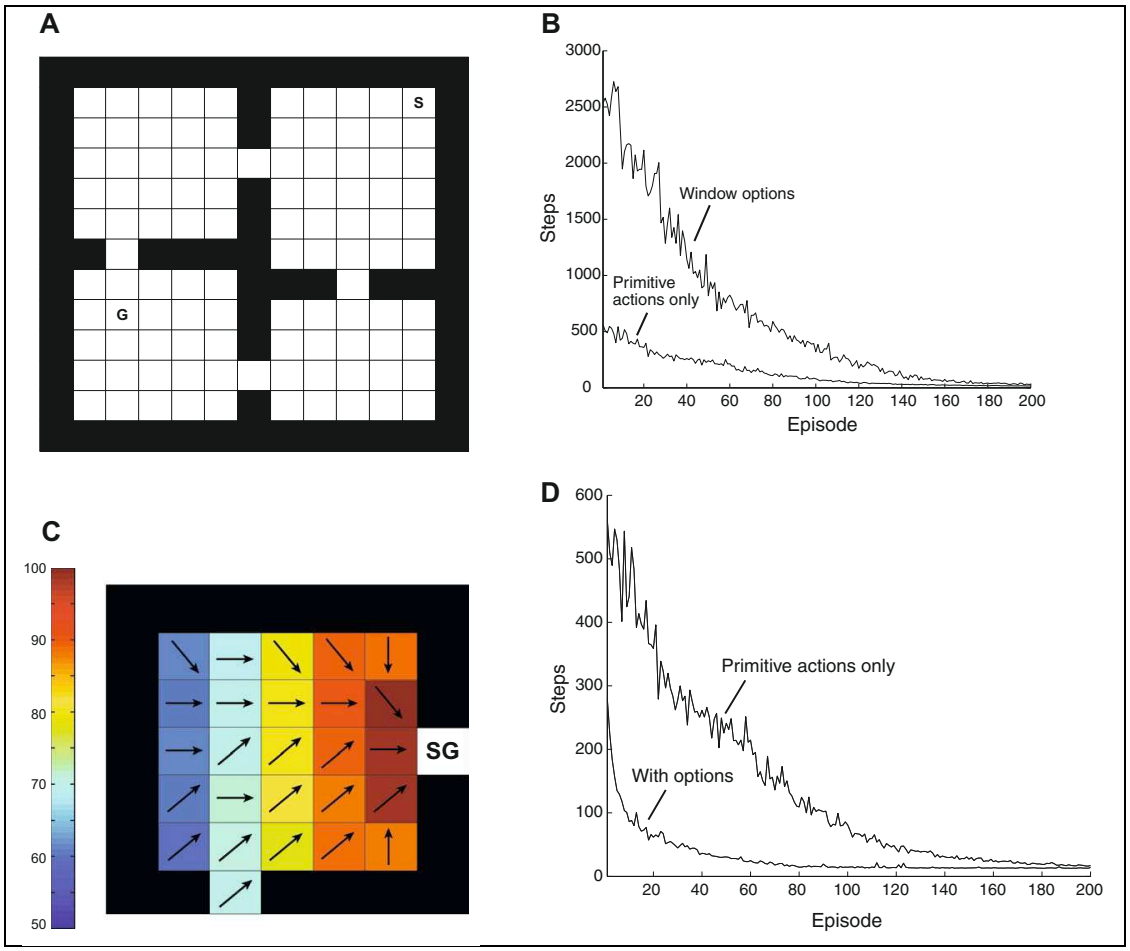


FIGURE 2.2 – Figure adaptée de Botvinick et al, 2008 [26]. A) Espace d'états du problème, avec un point de départ (S) et un objectif (G). B) Une des huit options, avec espace d'initiation dans la salle nord-ouest, sous-objectif (SG) la porte nord. Le code couleur représente la valeur de chaque état à l'intérieur de l'option, les flèches la stratégie (déterministe). D) Nombre d'actions nécessaires à atteindre l'objectif, en fonction du nombre de fois que le problème a été répété. Comparaison entre le modèle avec options *portes* déjà fournies (*With options*) et un modèle RL simple (*Primitive actions only*). C) Idem. Comparaison entre un modèle RL simple, et un modèle avec options *fenêtres* (*Window options*).

du nombre de choix fournis ne correspondant pas à une structure pertinente pour la tâche. Botvinick donne l'exemple d'options « fenêtre »au lieu d'options porte pour simuler cet effet de transfert négatif (voir figure 2.2 **B)**).

La question de la pertinence des options est donc cruciale. On peut étendre cette question à celle de la nécessité que les options pertinentes soient apprises et non seulement fournies.

Sutton et al, 1999 [203], proposent d'introduire une « valeur terminale de sous-objectif »jouant le rôle de récompense fictive pour l'apprentissage de la stratégie d'une option. Ainsi, au lieu de définir une option entière, on définit seulement son ensemble d'initiation et son sous-objectif lié à une récompense fictive. Un algorithme d'apprentissage par renforcement classique portant sur les états et les actions élémentaires et utilisant le renforcement fictif comme un renforcement normal permet alors d'apprendre la stratégie interne de l'option, même pendant un problème plus global. L'option ainsi apprise soit indépendamment d'un problème plus global, soit pendant un problème global, peut ensuite être réutilisée pour un autre problème.

Il est crucial de remarquer que le problème de la pertinence des options est avancé mais non complètement résolu par ce mécanisme d'apprentissage interne. En effet, le problème se reporte sur la définition des sous-objectifs. Plusieurs mécanismes possibles sont proposés pour cette étape. Des sous-objectifs d'intérêt particulier pourraient être des goulots d'étranglement dans l'espace du problème. Cependant, leur identification par cette méthode ne semble pas évidente ; pas plus que par l'acquisition d'une représentation causale du problème à résoudre. Un argument plus heuristique est lié au développement, impliquant la transformation de ce qui a été un problème complet précédemment (ouvrir une porte pour un enfant) en option disponible pour un problème plus tard.

La pertinence de la récompense virtuelle est argumentée neuroscientifiquement par Botvinick en faisant remarquer qu'en plus de s'activer pour des récompenses, les neurones dopaminergiques réagissent aux stimuli inattendus ou saillants (Redgrave et al, 2008 [173], Wittmann et al, 2008 [213]), ce qui pourrait être le cas de sous-objectifs atteints.

Botvinick propose d'étendre le modèle acteur-critique aux options de la manière suivante

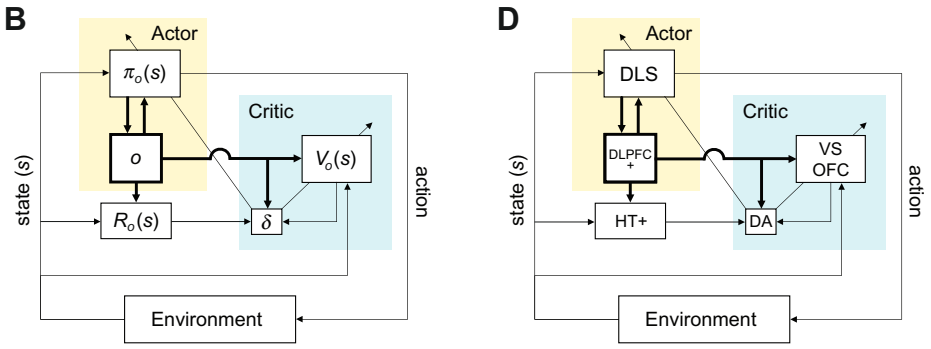


FIGURE 2.3 – Extrait de Botvinick et al, 2008 [26]. DLS : striatum dorso-latéral, acteur bas niveau. VS, OFC : striatum ventral, cortex orbito-frontal, critiques, calcul des valeurs d'états. DA : dopamine, signal d'erreur de prédiction δ. HT+ : hypothalamus et autres structures sous-corticales. DLPFC : cortex préfrontal dorsolatéral, maintien des options.

(voire figure 2.3). Il s'appuie sur les hypothèses suivantes, déjà exposées : rôle du striatum dorsal comme acteur, rôle du striatum ventral comme critique, rôle du PFC dorso-latéral dans la planification (donc maintien des options o), rôle de l'OFC dans l'évaluation avancée de valeurs (donc maintien de valeurs liées à une option, $V_o(s)$). Conséquemment, il propose que l'acteur de la stratégie sur les options soit lié au PFC dorso-latéral et celui de la stratégie interne à l'option au striatum dorsal ; tandis que la critique liée à la valeur des options serait liée à l'OFC et celle liée à la valeur des actions dans l'option au striatum ventral. Bien que cette théorie soit largement soutenue par de nombreuses données d'imagerie existantes, elle n'a pas encore été spécifiquement testée.

On voit en tout cas que l'importance de tenir compte d'une structure hiérarchique temporelle est essentielle dans la planification et largement reconnue. L'utilisation de divers modèles computationnels permet non seulement de proposer des mécanismes potentiels d'utilisation de cette hiérarchie, mais aussi d'explorer précisément le fonctionnement du cerveau face à ces problèmes.

2.3 Organisation hiérarchique structurelle de tâches

Nous avons montré la nécessité d'être flexible, de pouvoir apprendre à des échelles, des profondeurs ou des vitesses temporelles différentes. Nous avons également montré différents mécanismes permettant de prendre en compte la structure temporellement hiérarchique de la planification. Dans ce paragraphe, nous montrons que nous sommes également capables de tenir compte d'une organisation hiérarchique conditionnelle (plutôt que temporelle) de problèmes, que ce soit dans le domaine de l'apprentissage ou du contrôle cognitif. Nous montrons que l'architecture fonctionnelle du PFC permet une utilisation de cette hiérarchie. Nous présentons différents modèles proposant d'expliquer cette capacité.

2.3.1 Représentations hiérarchiques structurelles dans l'apprentissage

L'idée que nous sommes capables d'apprendre de manière plus profonde que par simple essai-erreur est très ancienne dans l'histoire de la psychologie. Dans une étude publiée en 1949 [104], Harlow démontre cette capacité. Il insiste également sur le fait que cette capacité à optimiser un problème n'est pas dépendante du langage, du raisonnement ou d'autres capacités spécifiquement humaines adultes, puisque ses sujets sont des singes (ainsi que des enfants de 2 à 5 ans). Harlow parle de *Learning sets* pour la capacité que nous (et les singes) avons d'apprendre à apprendre.

Dans une première expérience, Harlow soumet les singes à des problèmes de type *object discrimination* : deux objets sont présentés au singe dans une position (gauche-droite) aléatoire pendant un certain nombre d'essais (6 à 50). Le singe doit apprendre à sélectionner l'objet pour lequel il obtient une récompense. Cet exercice est effectué plusieurs centaines de fois, avec une nouvelle paire d'objets à chaque fois. Initialement, l'apprentissage est très lent et pourrait être modélisé par essai-erreur. A la fin par contre, le comportement est optimal : la performance après le premier essai (qui porte toute l'information) est proche de 100% correct. Harlow en déduit que les singes (et les enfants) acquièrent un *learning set*, soit une méthode pour apprendre les problèmes similaires. Il montre dans deux autres expériences que des Learning sets plus complexes peuvent également être appris : *reversal Learning set* (les sujets apprennent que le rôle des deux objets peut être échangé). Il montre

également que plusieurs *Learning set* interférents (passage de discrimination de l'objet à discrimination de la position) peuvent être appris relativement indépendamment.

On en conclut qu'une capacité élémentaire de l'apprentissage chez les primates est d'extraire une structure de l'histoire ou de l'environnement et de se construire une représentation interne de cette structure, pour améliorer l'apprentissage. Ici, cette structure est l'organisation hiérarchique de la tâche impliquant une « règle du jeu »qui pourrait être inférée par raisonnement, mais qui peut également être apprise par entrainement chez les singes.

Ce résultat a depuis été largement observé et affiné dans la littérature, par exemple par Landau et Gollin, 1966 [96], sur des enfants, en fonction du délai entre différents renversements. Il est depuis largement acquis, bien qu'aucun modèle ne soit proposé pour expliquer cette émergence d'une représentation, d'un modèle interne de la structure de la tâche, le *Learning set*, qui permet à l'apprentissage initialement laborieux de devenir optimal.

Aucun modèle n'a été proposé pour expliquer ces résultats d'apprentissage de *Learning sets*. Il est cependant important de noter qu'ils ne pourraient être expliqués par des algorithmes d'apprentissage par renforcement. En effet, les valeurs d'une paire de stimuli sont complètement indépendantes des autres paires de stimuli, de telle sorte qu'il ne peut pas y avoir d'amélioration au cours des différents problèmes.

Hampton et al, 2006 [102], montrent également que, plutôt que d'apprendre simplement par essai-erreur, nous utilisons un moyen plus optimal tenant compte de l'apprentissage d'une structure.

La tâche est une tâche de *probabilistic reversal* : un stimulus choisi apporte une récompense avec probabilité $0,7$, l'autre $0,4$. Leur rôle est renversé avec probabilité 0,25 à chaque essai après quatre choix corrects. Les sujets sont entrainés de telle sorte à acquérir un modèle interne de la structure de la tâche à laquelle ils participent. Ils effectuent tout d'abord une tâche simple de renversement où un stimulus est récompensé, l'autre non, avec le même critère de renversement. Cela permet d'acquérir l'anti-corrélation entre les stimuli (on peut apprendre quelque chose sur un stimulus même en sélectionnant l'autre) ainsi que la structure temporelle de renversement. En deuxième entrainement, ils sont soumis à une tâche de discrimination : choisir le stimulus qui rapporte le plus, avec un stimulus

récompensé avec $p = 0,7$, l'autre avec $p = 0,4$.

Les auteurs construisent un modèle d'apprentissage idéal bayésien qui incorpore l'identité du stimulus correct comme une variable cachée que le sujet doit inférer, et la structure de la tâche apprise par le sujet comme le modèle génératif des observations (soit en termes techniques, un *Bayesian Hidden state Markov model*, HMM). Plus précisément, l'état caché X est le fait que le stimulus choisi est correct ou non (inobservable, puisqu'il peut être récompensé ou non). Le modèle génératif reflète la structure supposée apprise de la tâche : $P(X_{t+1}|X_t,$ choix identique) est la probabilité de transition apprise en fonction du nombre d'essais corrects, $P(recompense|X_t)$ est également considéré comme appris par le deuxième entrainement. Par inférence Bayésienne, on peut donc inférer une probabilité $Prior(X_t = Correct)$ à partir du modèle génératif et du précédent renforcement obtenu afin de faire un choix.

Ce modèle permet d'utiliser de manière optimale la structure de la tâche et le renforcement obtenu à chaque essai, contrairement à un simple algorithme de type Q-learning qui traite indépendamment chaque stimulus et ne tient pas compte des connaissances sur le calendrier des renversements. Les auteurs montrent que l'algorithme HMM fitte mieux le comportement des sujets que l'algorithme RL. Par ailleurs, comme nous l'avons déjà évoqué plus haut, ils montrent des activations du cerveau corrélant avec des valeurs importantes du modèle HMM. On conclut de ces deux faits, la capacité à apprendre et à utiliser une structure lorsque celle-ci peut aider.

Krueger et Dayan, 2009 [138], montrent par ailleurs spécifiquement que l'apprentissage de la structure du problème avant l'apprentissage du problème complet permet un apprentissage global plus rapide que si tout était appris en même temps. Ils appelle cela le *shaping* : le fait d'extraire d'un problème des éléments hiérarchiquement plus simples, de les apprendre indépendamment, puis de revenir au problème complet. En utilisant des modèles en réseaux de neurones complexes, ils montrent que cette méthode accélère l'apprentissage d'une tâche structurée hiérarchiquement. Ce principe pédagogique de découpage de l'apprentissage en sous-tâches plus simples est appliqué en permanence : pour apprendre le grec ancien, on ne commence pas immédiatement en apprenant à lire Platon dans le texte, mais d'abord à lire l'alphabet, puis apprendre la grammaire et le vocabulaire, puis la syntaxe et enfin les

textes grecs.

2.3.2 Représentations hiérarchiques dans le contrôle cognitif

Dans le contrôle cognitif, la structure hiérarchique observée est souvent de type conditionnelle : si on est dans le contexte X, ou si on a vu l'instruction Y, ou si l'événement Z s'est produit, il faut appliquer une certaine règle. Cette structure est parfois implicite et limitée à la connaissance de l'existence de certaines règles et de la nécessité de découvrir et de maintenir laquelle est pertinente à un instant donné.

Nous montrons que le PFC est souvent modélisé comme un échelon hiérarchique unique ayant pour rôle de maintenir une règle donnée et d'influencer ses subordonnés effecteurs, qui eux encodent une association stimulus réponse (influence *top-down*).

Nous montrons ensuite qu'on peut représenter le rôle fonctionnel du PFC de manière plus fine dans le cadre d'une hiérarchie conditionnelle et temporelle à trois étages.

En dernier lieu, nous présentons une série de modèle bayésiens proposant une modélisation explicitement hiérarchique de contrôle du comportement.

Influence top down des règles

Une tâche paradigmatique du contrôle cognitif et très largement utilisée pour la détection de problèmes préfrontaux est la tâche de *Wisconsin Card Sorting Test* (WCST). Dans cette tâche, des cartes représentant des stimuli variant selon trois dimensions (forme, couleur, nombre) sont présentées au sujet. Celui-ci connaît les trois règles du jeu disponibles : trier les cartes selon une des trois dimensions. L'identification et le maintien d'une règle abstraite permettent donc d'influencer la décision de tri effectuée à chaque stimulus.

Si cette tâche semble particulièrement simple à effectuer, il est démontré qu'elle implique le contrôle par l'intermédiaire du cortex préfrontal. Les patients atteints par exemple de lésions préfrontales ou de maladies dégénératives effectuent des erreurs de deux types : persévération (maintien de l'utilisation d'une règle malgré des retours négatifs), distraction (échec du maintien d'une règle malgré des retours positifs).

Dehaene et Changeux, 1991 [66], proposent un modèle de type réseau de neurones capable d'effectuer une tâche de WCST. Leur modèle (noté par la suite DC91) permet en outre de rendre compte du type d'erreurs effectuées par des patients.

Leur modèle inclut des *rule coding clusters* représentant, maintenue en mémoire de travail, la règle actuelle à effectuer. Ces *rule coding clusters* exercent une influence top down sur les connexions entre des *memory clusters* représentant l'entrée selon ses trois dimensions, et les *intention clusters* représentant l'action préparée. Cette influence se fait par modulation de ces connexions. Le système est organisé en une boucle d'auto-évaluation : rule → stimulus Memory → action intention → error cluster → rule. Celle-ci permet de maintenir en mémoire de travail une règle tant que celle-ci est valide. Par ailleurs, l'activation d'une erreur permet une dépression de la précédente règle valide. La compétition latérale assure alors l'activation d'une nouvelle règle. En assurant une constante de rétablissement lente pour les *rule clusters* ayant subi une erreur, on permet au système d'avoir une mémoire épisodique afin de ne pas tester plusieurs fois une règle ayant déjà été prouvée incorrecte. L'action sélectionnée est ensuite déterminée par l'activation des 'intention clusters' lorsqu'un signal 'go' permet leur propagation vers la sortie motrice.

La plupart des principes présentés ici sont inspirés de la théorie attentionelle du contrôle cognitif (Miller et Cohen, 2001 [148]). Ils sont repris dans les modèles de réseaux de neurones modélisant le contrôle cognitif et notamment l'influence hiérarchique du préfrontal sur la sélection des actions. On peut citer en particulier les travaux de l'équipe de O'Reilly, Frank, Hazy etc. ([82], [106] [107] [108] [160] [159] [161]). Ces trois auteurs formalisent d'ailleurs les points critiques de ces modèles dans Hazy et al, 2007 [107], *Banishing the homunculus*. Avant de préciser ces points, notons que la notion d'homunculus résume l'importance de la hiérarchie dans le contrôle exécutif : il représente un centre de décision influençant tous ses subordonnés. L'effort computationnel présenté dans cette revue revient donc à parvenir à encoder une structure hiérarchique dans un modèle entièrement autonome, sans input extérieur (par exemple le signal go de Dehaene et Changeux, 1991 [66]).

Voici donc les « six besoins fonctionnels clé pour la mémoire de travail » :

– Existence de représentations multiples et séparées (comme les *rule coding clusters*) permettant de coder différentes règles.

– Maintenance robuste des représentations, robustesse aux distracteurs. Encodé par compétition latérale dans DC91.
– Mise à jour rapide des représentations (chez DC 91 par erreur), flexibilité cognitive.
– Processus indépendant de gating (barrage) des sorties pour le contrôle top-down. Implémenté chez DC 91 par la modulation des connexions, sans arguments de plausibilité biologique.
– Mise à jour sélective des représentations.
– Capacité d'apprendre quoi et quand bloquer ou autoriser le flux d'entrée ou de sortie de la mémoire de travail.

Les deux dernières exigences ne sont pas modélisées dans Dehaene et Changeux. Nous présenterons la proposition du groupe de O'Reilly plus loin.

Nous montrons dans le chapitre suivant que la structure hiérarchique du contrôle cognitif dans le cortex préfrontal peut être plus développée qu'un seul niveau d'influence top-down.

Architecture fonctionnelle du contrôle cognitif

Il est connu depuis longtemps qu'un gradient de complexité (Fuster, 2002 [91], Goldman-Rakic, 1996 [95]) ou de niveau d'abstraction (Grafman et al, 2002 [97]) existe dans l'organisation fonctionnelle du contrôle cognitif dans le cortex préfrontal latéral. Badre et D'Esposito, 2007 [11], proposent d'expliciter la notion de niveau d'abstraction et de montrer que l'organisation fonctionnelle hiérarchique du PFC latéral repose sur différents niveaux d'abstraction de représentations. Koechlin et al, 2003 [136], proposent au contraire une organisation fonctionnelle représentant différentes caractéristiques nécessaires à prendre en compte pour le contrôle.

Badre et D'Esposito (2007 [11]) proposent quatre niveaux hiérarchiques :
– sélection d'un mapping stimulus-réponse (abstraction du 1er ordre),
– sélection d'un ensemble de mappings stimulus-réponse (abstraction du 2e ordre),
– sélection d'un ensemble d'ensembles de mappings stimulus-réponse en fonction d'un indice perceptuel (abstraction du 3e ordre),

– sélection d'un ensemble d'ensembles d'ensembles de mappings stimulus-réponse en fonction d'une instruction passée (abstraction du 4e ordre).

Les sujets sont soumis à des expériences permettant de tester chacun de ces niveaux de hiérarchies. Les auteurs montrent un gradient d'activation dans le cortex préfrontal latéral : à l'ordre le plus bas, seules les régions caudales sont activées, tandis que l'introduction des autres ordres d'abstraction active, en plus, des régions de plus en plus rostrales.

Si cette méthode semble permettre de définir très indépendamment la notion de niveau d'abstraction, elle pose problème dans le sens où la nature de la tache effectuée par les sujets dépend du niveau d'abstraction proposé, ne permettant pas un parallèle strict entre chaque niveau de hiérarchie.

Koechlin et al, 2003 [136], proposent une structure hiérarchique permettant de comparer strictement parallèlement différents niveaux de contrôle, de manière indépendante de la quantité d'information maintenue en mémoire de travail :

– contrôle sensoriel : sélection d'une action en fonction d'un stimulus
– contrôle contextuel : sélection d'un ensemble d'associations stimulus-action (on nomme par la suite cet ensemble un *task-set*) en fonction d'un contexte présent.
– contrôle épisodique : sélection d'un ensemble d'associations contexte - task-set en fonction d'un événement (instruction) passé.

Les auteurs montrent une activation caudale pour le contrôle sensoriel. Ils montrent que cette activation s'étend de plus en plus rostralement avec les deux autres niveaux de contrôle sensoriel (voir figure 2.4 page 83). De manière critique, ils montrent que cette structure d'activation n'est pas dépendante du niveau de complexité de la tâche (puisque la quantité d'information nécessaire à effectuer la tâche est contrôlée), mais de son niveau hiérarchique. Par ailleurs les analyses de connectivité fonctionnelle montrent que cette organisation fonctionnelle s'effectue en cascade, avec une influence top-down des régions rostrales vers les régions caudales. Cette organisation fonctionnelle est confirmée par des études sur des patients préfrontaux (Azuar et al, 2009 [8]).

L'organisation fonctionnelle hiérarchique proposée par Koechlin et al n'est donc pas purement conditionnelle, comme proposé par Badre et al, mais sépare une dimension représentant une structure conditionnelle du contrôle cognitif (le contrôle contextuel) d'une

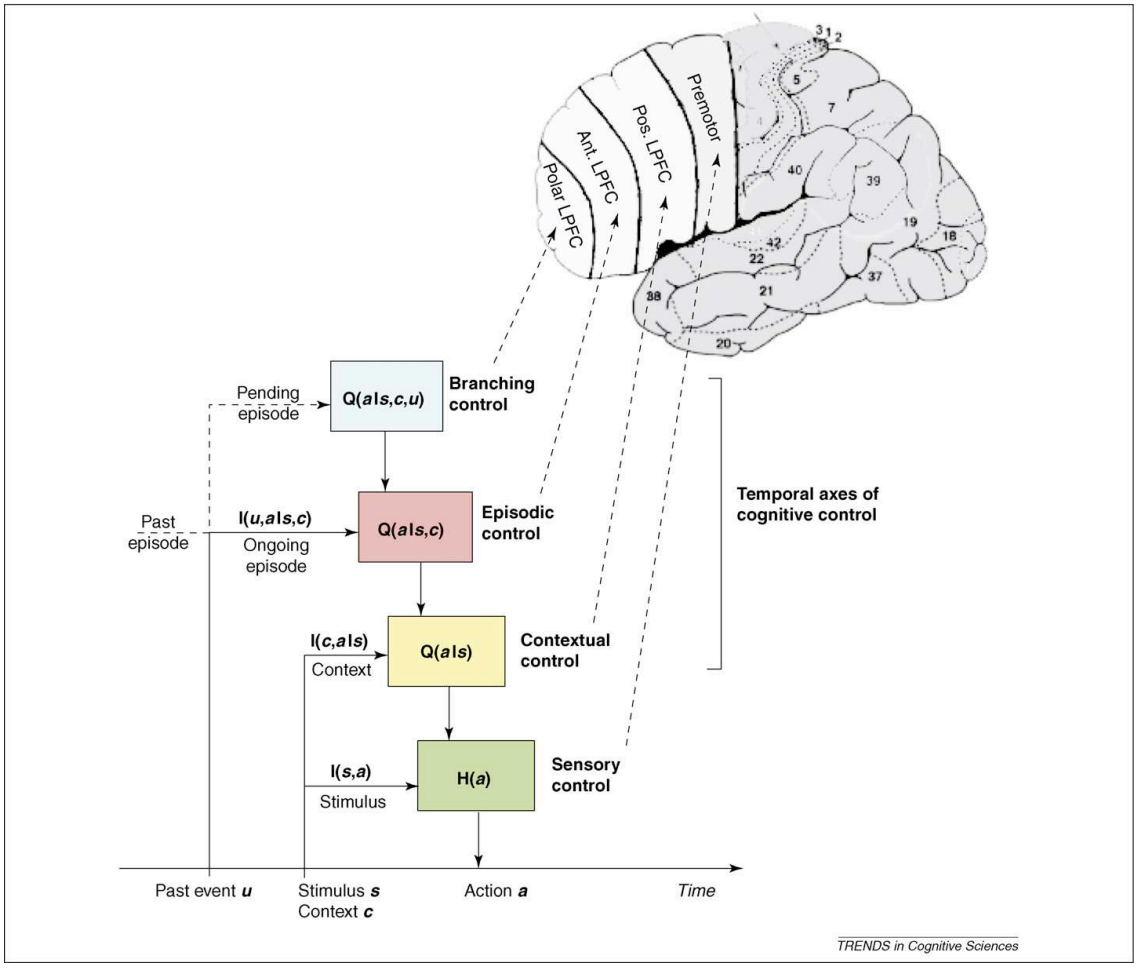


FIGURE 2.4 – Modèle en cascade de l'organisation fonctionnelle hiérarchique du cortex préfrontal latéral, extrait de Koechlin et Summerfield, 2007 [134].

dimension représentant une structure temporelle du contrôle cognitif (contrôle épisodique), au sein d'une même cascade hiérarchique.

Cette dimension temporelle se distingue de la hiérarchie séquentielle de la planification (malgré sa dimension temporelle), évoquée plus haut. En effet, la hiérarchie séquentielle de la planification permet plus de grouper à bas niveau des séquences d'actions pour produire des actions plus générales (options). La hiérarchie temporelle du contrôle cognitif (dimension épisodique), par contre, est bien une dimension temporelle liée à la structure de la tâche (donc de haut niveau) : elle correspond au maintien dans le temps d'une information passée nécessaire à la bonne exécution de la tâche.

Koechlin et al (1999 [131], Koechlin et Summerfield, 2007 [132], Koechlin et Hyafil, 2007 [134]) montrent qu'un dernier niveau de hiérarchie, temporel également, est encodé dans le cortex fronto-polaire dans le cadre du *branching*, soit la mise en attente d'une tâche pendant l'exécution d'une autre. Cette dernière dimension est multitemporelle, contrairement à la dimension épisodique du contrôle. L'importance de cette région fronto-polaire dans la mise en attente d'une tâche pendant qu'une autre est effectuée est confirmée par Boorman et al, 2009 [23], qui montrent que cette région encode la valeur d'une action non sélectionnée pendant que le PFC ventro-médial encode la valeur de l'action sélectionnée.

De nombreuses autres études mettent en évidence l'importance des structures hiérarchiques dans le cortex préfrontal et le contrôle cognitif. En particulier, Koechlin et Jubault, 2006 [133], montrent également une structure en cascade antéro-postérieure dans l'aire de Broca et son symétrique à droite pour le contrôle des séquences de tâches organisées hiérarchiquement. Kouneiher et al, 2009 [137], montrent également une dissociation entre deux niveaux hiérarchiques de motivation dans le cortex préfrontal médial.

Ces études montrent que le cerveau ne traite pas toutes les informations de manière égale, utilisant une représentation plate de l'espace des possibles, mais bien qu'il encode les relations existantes et les utilise pour optimiser le contrôle, séparant différents niveaux de relations et les utilisant pour influencer des décisions plus simples.

Nous présentons dans le prochain paragraphe des modèles de type bayésien cherchant à introduire cette structure dans les algorithmes d'apprentissage par renforcement.

Modèles d'apprentissage par renforcement hiérarchiques

Les modèles de type *mixture of experts* (Jacobs, Jordan, 1991-1994 [116] [115] [120], Wolpert et al, 1998 [214]) ont été proposés pour modéliser un contrôle tenant compte d'une structure hiérarchique cachée, non observable.

Nous présenterons ici plus précisément le modèle proposé par Doya et al, 2002 [75], appelé *multiple model-based reinforcement Learning model*, ou MMBRL. Ce modèle repose sur le principe de partager un problème complexe en plusieurs problèmes plus simples, idéalement en tenant compte de la structure existante du problème. Si ces modèles n'ont jamais été utilisés pour rendre compte de structures hiérarchiques dans le contrôle cognitif ou l'apprentissage, il nous semble intéressant de les mettre en valeur ici pour les applications qu'ils pourraient avoir dans ce domaine. Nous présentons donc ici ces modèles dans le cadre naturel où ils ont été élaborés – le contrôle moteur. Nous détaillerons plus loin dans la thèse leur utilisation possible pour le contrôle cognitif.

Le MMBRL (voir le schéma de principe figure 2.5, page 86) repose sur l'existence parallèle de multiples modules contrôleurs RL, chacun pouvant apprendre des valeurs attendues $V_i(s)$ et une stratégie. Dans le cadre du contrôle moteur, on parle de modèle inverse pour la stratégie (élaboration d'une commande motrice à partir d'un objectif de mouvement). Chaque module contrôleur est également muni d'un modèle direct prédisant les conséquences des commandes motrices qu'il élabore. La comparaison entre la prédiction de ce modèle direct et le nouvel état observé comme conséquence de la commande motrice effectuée, permet le calcul d'une erreur de prédiction différente de celle liée au RL.

Pour chaque module, on peut ainsi calculer un signal de responsabilité (*responsibility signal*) λ_i correspondant au softmax des erreurs de prédiction du modèle direct. Ce signal reflète la validité actuelle des prédictions effectuées par chaque module, et détermine ainsi la responsabilité qu'il devrait avoir dans la sortie motrice. Celle-ci est donc la somme des sorties motrices de chaque module, pondérée par les signaux de responsabilité. Ceux-ci servent également à moduler l'apprentissage de chacun des modules en pondérant les vitesses d'apprentissage. L'existence de différents modules dont l'apprentissage et la capacité de contrôle est modulée par leur responsabilité, permet au système l'accès à différents experts

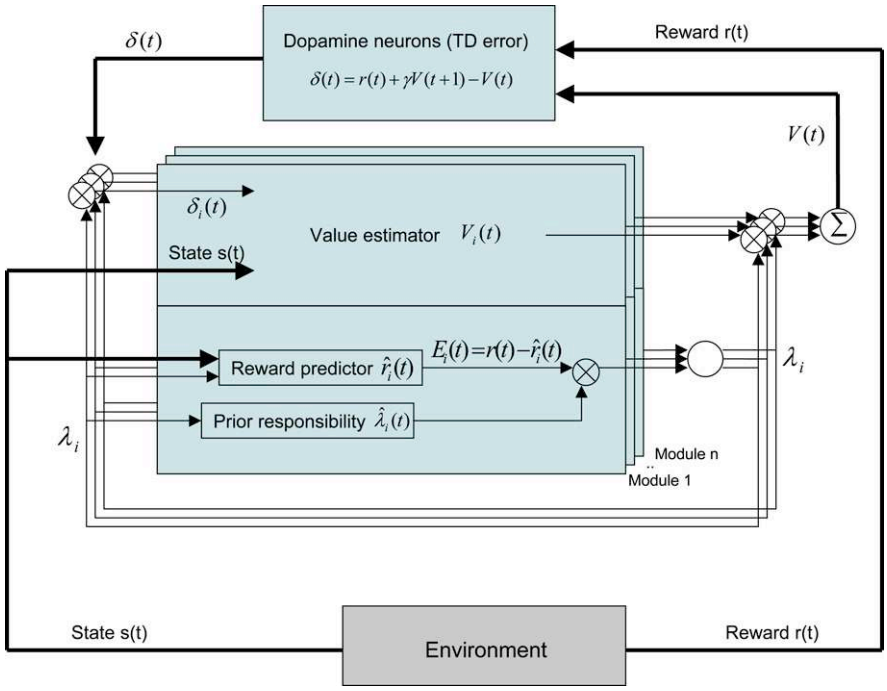


FIGURE 2.5 – Schéma de principe de MMBRL, issu de Bertin et al, 2007 [20]. Chaque feuillet a sa propre stratégie (représentée ici par le seul *Value estimator*) et un modèle interne prédictif (*Reward predictor*) qui permet le calcul d'un signal de responsabilité (λ_t). La stratégie globale appliquée par le modèle (représentée ici sous la forme de $V(t)$) est la somme pondérée des stratégies de chaque feuillet par leur responsabilité. L'apprentissage dans chaque feuillet est alors proportionnel à l'erreur de prédiction globale ($\delta(t)$) pondérée par la responsabilité précédente de chaque feuillet, soit $\delta_i(t)$.

auxquels il ne fait appel que lorsque ceux-ci sont dans leur domaine d'expertise.

Cette capacité peut être utilisée, comme dans Doya et al, 2002 [75], pour rendre moins complexe un problème, en le divisant en sous-problèmes sur la base de l'espace des états, chaque module étant responsable d'une partie de l'espace. Elle peut également être utilisée lorsque l'environnement est changeant, afin d'avoir différents experts correspondant à différentes situations possibles, comme dans Bertin et al, 2007 [20]. Dans cette étude, les différents modules prédisent différents délais de renforcement pour une étude de conditionnement. Bertin montre que les prédictions ainsi effectuées sur l'erreur de prédiction de renforcement coïncident avec des données observées sur la dopamine, non expliquées par un renforcement classique. Kawato, 1999 [125], propose également ce type de modèle dans le domaine du contrôle moteur avec variabilité de l'environnement. Il utilise ce modèle (*Multiple pairs of forward and inverse models*) avec différents modules pour porter un objet, spécialisés différemment selon le poids de l'objet, par exemple.

Ce modèle diffère des *mixture of experts models* dans le sens où il détermine de manière interne quel module doit être responsable de la sortie, au lieu d'être muni d'un *gating mechanism* externe sélectionnant les modules. Dans le cadre du contrôle moteur, Imamizu et al, 2004 [112], argumentent que les deux types de modèle peuvent représenter le fonctionnement de différentes parts du cerveau, le lobe frontal ayant plutôt une structure *mixture of experts* et le cervelet une structure MMBRL (nommé ici MOSAIC).

Imamizu et al (2007 [113], 2008 [111]) étendent ce modèle à la prise en compte d'information contextuelle pour ce qu'on pourrait appeler contrôle moteur contextuel, en parallèle avec le contrôle cognitif. L'information contextuelle est considérée comme un a priori sur les différents modules et utilisée, en plus de la vraisemblance de l'erreur de prédiction du modèle interne, pour calculer le signal de responsabilité. Cela permet au modèle d'avoir plus d'information pour savoir quel module il est pertinent d'utiliser actuellement, que les seules erreurs effectuées. On a ainsi des informations a posteriori et a priori. L'exemple utilisé est celui de deux rotations visuo-motrices. Les sujets doivent effectuer un mouvement partant du centre d'un cercle vers un point indiqué sur le cercle. Le retour visuel est soit correspondant au mouvement, soit perturbé dans le sens des aiguilles d'une montre, soit dans le sens inverse. On peut donc construire trois modules représentant ces circonstances

dans leurs modèles internes directs et indirects. L'ajout d'une information contextuelle indiquant le mode actuel de l'environnement permet une sélection initiale plus simple, mais une adaptation subséquente identique, comme prévu par le modèle. (Imamizu, Kawato et al., 2007, 2008 [112] [113] [111]).

On peut appeler les modèles de type MMBRL *hiérarchiques*, parce qu'ils introduisent une unité de comportement, représentée par les différents feuillets, qui est différente de l'unité des actions simple, mais qui influence celles-ci.

Conclusion : Nous avons montré que la connaissance de la structure d'une tâche est souvent apprise ou utilisée pour optimiser l'apprentissage, la planification ou la décision. Nous avons exposé plusieurs types de modèles, reposant sur de l'apprentissage bayésien, de l'apprentissage par renforcement ou sur des réseaux de neurones, pour mettre en œuvre des systèmes hiérarchiques de contrôle et d'apprentissage et expliquer notre capacité à former des représentations hiérarchiques des problèmes de décision ou d'apprentissage auxquels nous sommes soumis. Ces systèmes permettent d'apprendre à différentes échelles du problème, de stocker des aspects appris pour des décisions futures, d'utiliser plusieurs modèles différents pour adapter flexiblement son comportement dans un environnement changeant. Contrairement au cadre du contrôle moteur, dans le cadre du contrôle cognitif, adapter son comportement se fait souvent de manière binaire par l'intervention d'un task-switch permettant de passer instantanément d'une représentation à l'autre.

Dans le prochain chapitre, nous parlerons de l'importance du switch entre deux représentations hiérarchiquement avancées, les task-sets, pour le contrôle cognitif, et des modèles permettant de rendre compte de cette notion.

Chapitre 3

Les task-sets et le task-switching, première brique hiérarchique du contrôle cognitif

Dans ce chapitre, nous commençons par présenter la notion de task-set et de task-switching, établissant ainsi le rôle essentiel des task-sets comme première brique hiérarchique du contrôle cognitif. Nous proposons ensuite deux familles de théories proposant un mécanisme de switch : l'une propose un rôle de la noradrénaline dans le switch, l'autre des ganglions de la base. Enfin, le mécanisme de switch suppose le passage d'une tâche à une autre. En l'absence d'instructions, cela implique la nécessité d'explorer afin de décider de la nouvelle tâche pertinente. Nous présenterons les problèmes théoriques liés à la notion d'exploration, ainsi que les données expérimentales existantes.

3.1 Task-sets et mécanismes de task-switching

3.1.1 Task-sets

Sakai, en 2008, écrit dans sa revue extensive sur les Task-sets et le cortex préfrontal (Sakai, 2008 [184]) : *« a task-set is a configuration of cognitive processes that is actively maintained for subsequent task performance »*. Monsell en 2003 écrit *« each cognitive task [...] requires an appropriate configuration of mental resources, a procedural 'schema' or 'task-set' »* (Monsell, 2003 [150]). Ces deux citations nous fournissent la définition d'un task-set comme une configuration des ressources mentales qui, maintenue activement pour une tâche cognitive, permet d'effectuer celle-ci de manière appropriée. C'est donc un *mapping*, une fonction permettant de décider de manière non équivoque des réponses à sélectionner dans le cadre d'une tâche cognitive. Bien qu'il s'applique à différentes situations dans le cadre d'une tâche, un task-set est un tout, un seul ensemble approprié à une tâche cognitive précise.

Un task-set est également spécifiquement maintenu de manière active, dans ce qu'on appelle la mémoire de travail. On définit la mémoire de travail (*working memory*) comme un système actif de mémoire permettant de stocker temporairement et de manipuler de l'information nécessaire à l'exécution de taches cognitives complexes (Baddeley, 1992 [9]). On sait que le cortex préfrontal a un rôle essentiel dans la mémoire de travail (activité soutenue de neurones pendant le maintien en mémoire de travail, notamment Fuster et Alexander, 1971 [90], Asaad et al, 1998 [5], Constantinidis et Procyk, 2004 [45]). Dosenbach et al, 2006 [72], effectuent une large méta-analyse montrant que de larges parties du cortex préfrontal (mais également d'autres parties du cerveau et notamment le cortex pariétal) sont impliquées dans l'initiation, le maintien, et la mise à jour des task-sets. La figure 3.1 page 92 détaille ce système d'implémentation des task-sets.

Le passage d'une tâche à une autre, appelé task-switching, demande de reconfigurer le task-set actif et requiert donc un effort cognitif. Nous présenterons tout d'abord les différents paradigmes d'étude du task-switching et ce qu'ils impliquent en terme de *switch-cost*, l'effort cognitif effectué pour changer de tâche. Puis, nous montrerons l'importance de différentes notions de conflit et d'interférence impliquant de manières distinctes plusieurs aspects du

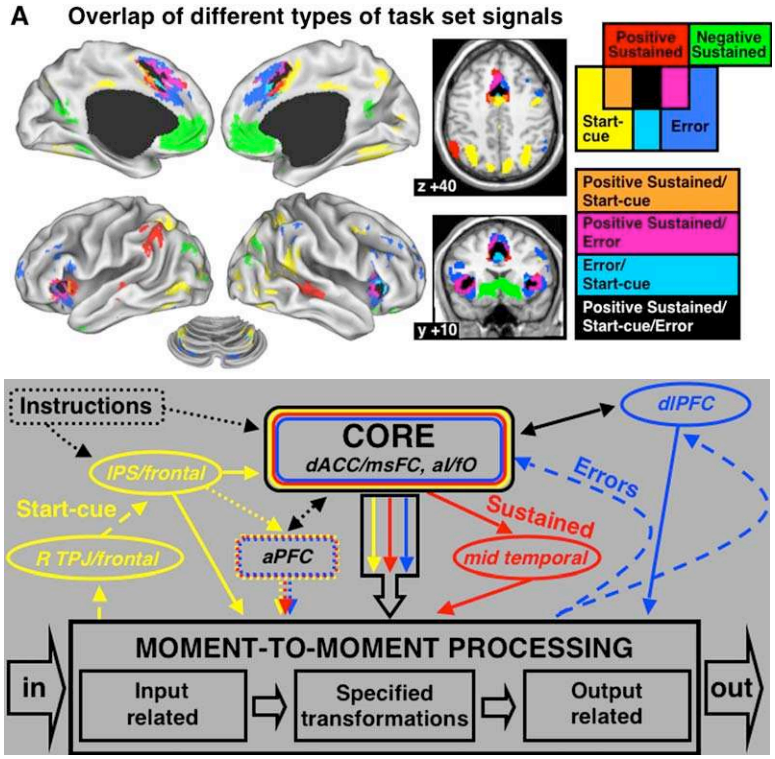


FIGURE 3.1 – Adapté de Dosenbach et al, 2006 [72]. IPS : sulcus pariétal inférieur. TPJ : jonction temporo-pariétale. aPFC : cortex préfrontal antérieur. dACC : cortex cingulaire antérieur dorsal. msFC : cortex frontal médial supérieur. aI : insula antérieure. fO : frontal operculum. dlPFC : cortex préfrontal dorso-latéral.

contrôle cognitif dans le task-switching.

3.1.2 Task-switching, switch-cost

Deux types de protocoles sont utilisés pour étudier le task-switching. Dans les protocoles par bloc, un task-set est valide pendant une série d'essais avant qu'une instruction n'indique le passage à un autre task-set pour une longue durée également. Ce type de protocole permet d'étudier le moment du switch mais aussi la période de maintien actif du task-set précédent un autre switch. Dans des protocoles de type *task-cueing*, un indice visuel (*cue*) est présenté à chaque essai, précédant ou accompagnant le stimulus. Cet indice visuel instruit le sujet sur la tâche à effectuer, impliquant ainsi un switch potentiel à chaque essai.

Les tâches utilisées dans les paradigmes de task-switching sont très variées, s'appuyant souvent sur des processus simples à apprendre : par exemple, discriminations entre nombres pairs / impairs, plus grands / plus petits que 5 ; paire identique / paire différente ; voyelle / consonne ; minuscule / majuscule ; mot abstrait / mot concret ; une syllabe / deux syllabes... Si des résultats spécifiques sont observés en fonction du domaine concerné par les stimuli ou le mode de réponse (moteur, oculomoteur, vocal...), on observe également des résultats très globaux, indépendants des spécificités des paradigmes.

En particulier, l'effet principal est connu sous le nom de *switch-cost* (voir figure 3.2 page 95), coût comportemental lié au changement de tâche. Cet effet se matérialise par une augmentation du temps de réaction et de la proportion d'erreurs effectuées dans un essai switch (changement de tâche) par rapport à un essai *stay*, où la tâche n'a pas changé. Cet effet est résistant à de nombreuses manipulations :

- temps de préparation. Si un long délai est laissé entre l'indice de changement de tâche et le premier stimulus auquel doit s'appliquer cette tâche, on observe une réduction du switch cost, mais pas sa disparition totale (Monsell, 2003 [150]).
- répétition. Dans des paradigmes incluant de longues périodes *stay*, la survenue d'une instruction, même si celle-ci n'implique pas un switch (même tâche que précédemment), provoque un coût comportemental (Monsell, 2003 [150]).
- informativité. Le *switch cost* persiste même lorsqu'il n'y a pas d'information apportée par l'instruction, c'est à dire lorsque l'action sélectionnée pour le stimulus serait identique,

quelle que soit la tâche. (Monsell et al, 2003 [150], Hyafil et al, 2009 [110]). On peut par exemple se référer aux deux barres de gauche dans la figure 3.2 page 95, montrant que même lorsque l'essai est congruent, le switch cost persiste.

On peut également noter que, même lorsque le switch intervient relativement peu fréquemment, on observe un effet à long terme du switch (mixing cost), en plus de l'effet transient (switch cost). En effet, les temps de réaction restent significativement plus lents (même pour les essais *stay*) que lorsqu'une seule tâche est effectuée pendant un bloc.

Différents cas particuliers du task-switching ont été étudiés plus spécifiquement. On peut parler en particulier de la notion de *set shifting* dans laquelle les task-sets sont définis en portant l'attention sur une dimension ou un aspect particulier des stimuli pour une tâche, sur une autre dimension pour une autre tâche. Un exemple connu de paradigme de type *extra-dimensional set-shifting* est le Wisconsin Card Sorting Test, dans lequel la règle associée à chaque task-set est définie par la dimension du stimulus à laquelle on s'intéresse (trier selon la forme, la couleur ou le nombre).

3.1.3 Processus de task-switching

Différents processus de *task-switching* sont impliqués pour expliquer ces phénomènes de *switch-cost*. En reprenant Monsell, 2003 [150], on peut définir le processus de *task-set reconfiguration* comme un processus qui inclut le fait de déplacer son attention vers un nouveau critère de tâche, de récupérer et de stocker en mémoire de travail les objectifs et conditions (règles d'action) de cette tâche, afin de permettre d'implémenter un nouveau jeu de réponses. Ce processus peut également inclure la nécessité d'inhiber le précédent task-set. Si ce processus semble pouvoir expliquer une partie du *switch-cost*, ainsi que le fait que celui-ci se réduit avec un temps de préparation suffisant laissé aux sujets, il peine à expliquer le coût résiduel qui ne disparaît pas, quel que soit le temps de préparation laissé aux sujets.

Un deuxième processus évoqué pour expliquer le *switch-cost* est lié à l'inertie transiente des task-sets. Ce processus est inspiré du fait qu'on observe, de manière surprenante, que le fait de switcher de la tâche la moins familière à la tâche la plus familière (par exemple, passer de l'action de nommer la couleur de l'encre à celle de lire un mot désignant une couleur dans une tâche de *stroop*, ou passer de la langue la moins naturelle à la plus naturelle

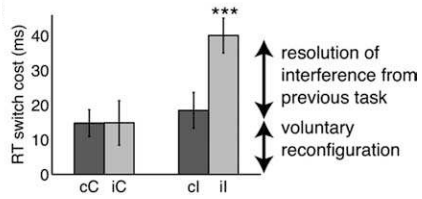


FIGURE 3.2 – Un expemple de switch cost, extrait de Hyafil et al, 2009 [110]. Différence entre les temps de réaction pour les essais *switch* et pour les essais *stay*, dans une tâche de stroop spatial. Lettres minuscules : congruence de l'essai précédent. Lettres majuscules : congruence de l'essai présent. c, C : congruent ; i, I : incongruent.

pour des bilingues) implique un switch cost plus élevé. Il semble difficile d'expliquer par le processus de *TS reconfiguration* qu'il serait plus difficile de reconfigurer vers le TS le plus familier. Certains auteurs ont donc proposé que ce TS le plus familier devait être inhibé plus fortement lors de la performance d'autres TS, ce qui ralentirait de manière transiente le switch. Ce phénomène peut également être observé dans le fait qu'il est plus difficile, après avoir quitté une tache 1 pour effectuer une tache 2, de revenir à la tâche 1 plutôt que de passer à une tâche 3 (Dreher et Berman, 2002 [76]).

Si les données expérimentales apportent à la fois des arguments pour et contre chacun des deux processus évoqués plus haut, la plupart des auteurs s'accordent pour dire que le switch cost est le fruit d'une combinaison de plusieurs processus distincts, et s'efforcent d'isoler ces différents processus, comme ayant différents rôles dans notre capacité de switcher d'une tâche à l'autre.

On peut citer en particulier les études de Crone, 2005 [56], Johnston et al, 2007 [119]) et Hyafil et al, 2009 [110], qui mettent en valeur des processus distincts intervenant dans le task-switching. Crone dissocie ainsi les concepts de *représentation* (récupérer, maintenir et implémenter des task-sets) et *reconfiguration* de task-sets. Hyafil et al, 2009 [110], proposent le fait qu'il y a une dissociation entre reconfigurer les priorités motivationnelles associées à chaque task-set et reconfigurer les task-sets pour contrer les interférences entre différents task-sets.

3.1.4 Théorie du conflit

La nécessité de switcher de manière flexible entre différentes règles est particulièrement évidente lorsque ces règles sont incongruentes, impliquant un conflit dans la sélection de la réponse appropriée. Ainsi, de nombreuses études cherchant à élucider les mécanismes et corrélats neuronaux du task-switching utilisent des paradigmes impliquant une variation du niveau de conflit présent et cherchent à mettre en avant son rôle.

Les paradigmes les plus utilisés sont les paradigmes de type *stroop*, *flanker* ou Simon. Dans les tâches de type *stroop*, un stimulus présentant deux dimensions est présenté, par exemple un nom de couleur écrit avec des lettres colorées. Deux task-sets peuvent être effectués, en fonction de l'instruction donnée au sujet : nommer la couleur des lettres, ou lire le mot. Le mot *rouge* écrit en rouge impliquera ainsi moins de conflit que le mot vert écrit en rouge. De nombreuses variantes de cette tâche existent (spatiales, numériques, etc.). Il est important de noter qu'il y a en général un déséquilibre entre les deux task-sets. En particulier, dans le cas décrit, le task-set lecture est plus automatique que le task-set couleur. Les tâches de type *flanker* manipulent le niveau de conflit en entourant un stimulus cible (par exemple une flèche $>$) d'autres stimuli identiques, ou conflictuels (par exemple $< <$). On observe dans toutes les tâches incluant du conflit, indépendamment de la présence ou de l'absence de switch, un « coût du conflit » semblable au *switch cost*, impliquant plus d'erreurs et des temps de réaction plus lents pour les essais dans lesquels il y a plus de conflit.

Les études croisant les facteurs de conflit et de *task-switching* montrent des effets séquentiels d'interaction entre les deux facteurs, en particulier une diminution du *switch-cost* après un essai incongruent (fort conflit). De cette observation relativement non controversée est issue la théorie selon laquelle un processus de détection du conflit provoquerait le processus distinct de résolution du conflit, c'est à dire l'augmentation du niveau de contrôle cognitif à exercer, signalée comme nécessaire par la détection du conflit (théorie soutenue notamment par Badre et Wagner, 2006 [12], Brown et al, 2007 [34], Botvinick et al, 2004 [25], Liston et al, 2006 [140], Wager et al, 2005 [209], Stemme et al, 2007 [200], [199]). S'appuyant sur des données d'IRM, les auteurs proposent que l'ACC (ou plus généralement des régions médiales du PFC) joue le rôle de détection du conflit, signale ainsi au PFC dorso-latéral la nécessité d'augmenter le niveau de contrôle top-down qu'il exerce, facilitant ainsi po-

tentiellement le prochain task-switch. En faveur de cette théorie, on trouve les nombreuses observations d'activations de l'ACC liées aux situations de conflit cognitif et l'observation qu'une activation de l'ACC à l'essai t-1 prédit une meilleure performance à l'effet t, etc. Cette théorie est également raffinée dans la proposition selon laquelle différentes formes de conflit (au niveau des stimuli, des tâches, ou des réponses) influencent différents types de contrôle dans des boucles parallèles (Egner 2008, [79]).

Cependant, cette théorie du *conflict monitoring* est également largement controversée. En particulier, les mêmes auteurs ont observé des activations de l'ACC indépendantes du conflit (Brown et Braver, 2005 [35]). Ils proposent que, plutôt que le conflit, ce soit un signal codant la prédiction de la vraisemblance d'observer des erreurs dans la situation présente qui soit représenté dans l'ACC, celui-ci ayant toujours un rôle d'alerte en direction du PFC latéral. Cependant, d'autres auteurs (Aarts et al, 2008 [1]) ont montré dans une étude combinant conflit et task-switching que les activations de l'ACC pouvaient être indépendantes du conflit ou de l'*error likelihood*, allant malgré tout dans le sens d'une alerte de la nécessité d'implémenter plus de contrôle cognitif. Ils montrent par ailleurs également un rôle du PFC dorso-latéral dans le codage du conflit. Cette observation est confirmée par Mansouri et al, 2009 [142], dans des études en électrophysiologie chez le singe. En effet, si aucun neurone codant le conflit n'a été observé dans l'ACC des singes, des neurones du PFC latéral correspondant à ce critère ont par contre été observés. Les études de lésions corroborent également ces résultats.

S'il semble donc acquis que des processus multiples (dépendants de substrats neuraux distincts) sont impliqués dans le task-switching, que l'ACC et le PFC latéral ont un rôle important, la controverse reste forte sur la dissociation exacte des différents processus et le rôle des différentes régions préfrontales.

On peut cependant retenir de ces études plusieurs points essentiels pour notre travail :
- l'existence de la notion de task-set, garantie par l'effet spécifique du passage d'un set à l'autre,
- la nécessité d'augmentation de contrôle cognitif liée au switch,
- la dissociation entre un phénomène de signalement du switch (probablement dans les régions médiales du PFC), et un phénomène de mise en place des ressources nécessaires

à effectuer le switch correctement (probablement dans les régions latérales du PFC).
Les études précédemment citées proposent souvent des modèles sous forme de réseaux de neurones plus ou moins développés, représentant les théories liées au conflit ([12] [33]) ou à l'*error likelihood* ([35] [34]), afin de modéliser un processus de task-switching et de reproduire des résultats expérimentaux observés. Par exemple, Badre et Wagner, 2006 [12], proposent un réseau composant une couche à deux unités représentant deux tâches, liée à une couche à quatre unités représentant les deux possibilités par tâche et une couche à deux unités représentant les réponses. Ils séparent ainsi une notion de conflit conceptuel (entropie dans la couche du milieu) d'une notion de conflit de réponse (entropie dans la couche de sortie). Cela leur permet de montrer que le premier est corrélé à l'activité dans le préfrontal latéral, le deuxième dans le pariétal, argumentant en faveur du rôle du PFC latéral pour réduire l'interférence entre deux tâches.

Cependant, ces modèles n'apportent généralement pas plus que l'exposé des hypothèses et théories liées au conflit ou au switch. Nous ne nous attarderons donc pas plus sur ces modèles.

Dans le chapitre suivant, deux types de modèles proposant des apports conceptuels au problème de modélisation du task-switching sont présentés.

3.2 Les mécanismes de switch

Nous avons montré précédemment l'importance des structures hiérarchiques dans l'apprentissage mais aussi essentiellement dans le contrôle cognitif. Le contrôle cognitif étant défini comme la capacité à adapter son comportement de manière flexible en fonction de contextes ou d'objectifs, il est très fondamentalement relié à la notion de task-sets et de switch. Ceux-ci constituent donc une brique élémentaire de l'étude du contrôle cognitif et de sa structure hiérarchique (Koechlin et al, 2003 [136], contrôle contextuel puis épisodique) et les mécanismes de switch représentent un challenge essentiel pour modéliser la fonction préfrontale.

Nous présenterons, en premier, une série d'études proposant une modélisation biologique-

ment plausible et globale du contrôle cognitif dans le préfrontal et les ganglions de la base. Nous exposerons ensuite une théorie plus abstraite et focalisée du switch et de son lien avec le neurotransmetteur noradrénaline.

3.2.1 Les modèles de *gating*

L'équipe de R. O'Reilly a publié depuis 2000 une série d'articles ([154] [82] [106] [107] [108] [160] [159] [161] [178]) visant à modéliser de plus en plus précisément les fonctions de switch et de mémoire de travail, qui prises ensemble, peuvent rendre compte d'une grande partie de la fonction exécutive frontale.

Ces études sont parties de deux constats : 1) que les ganglions de la base, qui ont longtemps été considérés comme une structure essentiellement motrice, jouaient, dans le cadre de la boucle motrice, un rôle de frein à main, bloquant ou autorisant la sélection d'actions par d'autres régions (motrices, prémotrices) ; 2) que les patients parkinsoniens, essentiellement affectés au niveau des ganglions de la base, présentaient malgré tout des troubles du comportement de type frontaux (Frank, 2005 [83]). Les auteurs proposent donc que les circuits préfrontal-ganglions de la base effectuent le même rôle de filtrage, ou *gating*, que les circuits moteurs-ganglions de la base. Ils argumentent ainsi pour leur rôle dans le maintien d'information en mémoire de travail (résistance aux distracteurs, portail fermé) ou la flexibilité (modification rapide des représentations, portail ouvert).

Le modèle global repose sur trois ensembles fonctionnellement distincts, représentés par trois grandes régions du cerveau :

- le cortex postérieur, qui gère les parties les plus automatiques et parallèles de la cognition : il apprend de manière lente et intégrée et représente bien les compétences et connaissances accumulées, sur lesquelles reposent les deux autres systèmes plus complexes ;
- le système hippocampique, capable d'apprendre rapidement ;
- le système PFC/BG, spécialisé dans la maintenance active d'information contextuelle (PFC), qui peut être mis à jour dynamiquement par les ganglions de la base (BG), ceux-ci apprenant à jouer leur rôle via l'action dopaminergique.

Les modèles présentés ne tentent pas de résoudre les deux premiers systèmes mais se contentent de représenter leur apport. Ils se concentrent sur les mécanismes liés au troisième

système, PFC/BG.

Les auteurs arguent que le contrôle cognitif et la mémoire de travail sont deux aspects d'un même phénomène, qui est le maintien contrôlé de représentations dans le PFC. Cet argument repose sur la théorie exposée par Miller et Cohen, 2001 [148], du contrôle cognitif comme l'influence top-down de représentations du PFC sur d'autres régions du cerveau ; la mémoire de travail représentant le maintien contrôlé des informations pertinentes pour une influence utile.

Le modèle (PBWM pour Prefrontal cortex, Basal ganglia in Working Memory), repose sur plusieurs mécanismes précis qui permettent d'implémenter les caractéristiques clés de la mémoire de travail :

– sa capacité de maintenir de manière robuste une représentation, mais aussi de la mettre à jour très rapidement est implémentée par un système de gating que nous détaillerons plus loin. Les connexions excitatrices récurrentes entre les différentes unités bi-stables du PFC permettent d'entretenir un pattern d'activation. Le système de gating, lorsqu'il est fermé, permet de stabiliser ce pattern, ou au contraire, lorsqu'il est ouvert, de le déstabiliser et de le laisser ouvert à d'autres influences (parties postérieures du cerveau) et ainsi de permettre l'émergence d'une autre représentation.
– son rôle dans le contrôle exécutif est modélisé par les connexions descendantes qui permettent aux représentations maintenues d'influencer la sélection des sorties (Miller, 2000 [147]), ainsi que dans de nombreux autres modèles du PFC en réseaux de neurones (ex : Dehaene et Changeux, 1991 [66])
– sa capacité de sélectivité, quant à ce qui doit être maintenu et ce qui doit être mis à jour, est modélisée par l'existence de nombreuses boucles parallèles PFC-BG, permettant ainsi un blocage de certaines à l'exclusion d'autres.
– enfin, la capacité d'auto-apprentissage du modèle est permise par des mécanismes impliquant l'action de la dopamine sur les BG dans la voie directe et indirecte.

Le mécanisme de gating (voir figure 3.3, page 101) repose sur la présence d'une boucle directe (Go ; double inhibition permettant un signal excitateur du thalamus vers le PFC) et d'une boucle indirecte (NoGo ; triple inhibition permettant un signal inhibiteur du thalamus vers le PFC), comme déjà exposé plus haut pour les mécanismes d'apprentissage. L'excitation par un stimulus de neurones du chemin Go implique un signal déstabilisant dans une zone

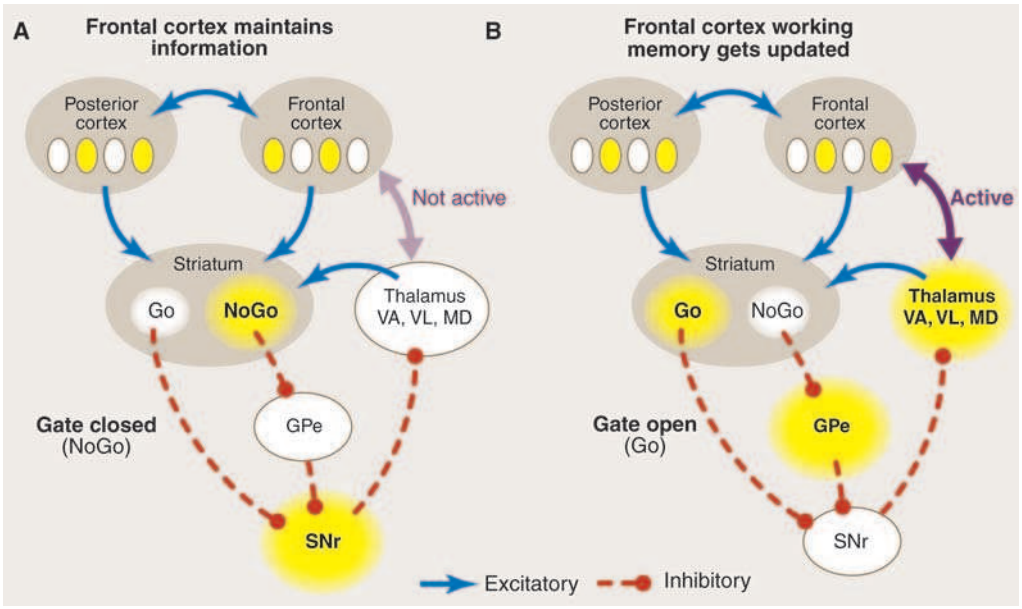


FIGURE 3.3 – Issu de O'Reilly, 2006 [159]. A) Portail fermé par une plus forte activité dans la boucle indirecte que dans la boucle directe, inhibant la connexion thalamus-préfrontal et assurant le maintien robuste d'une représentation dans le cortex préfrontal, qui est différente de la représentation d'entrée (*Posterior Cortex*). B) Portail ouvert par une plus forte activité dans la boucle directe, ce qui désinhibe le thalamus et lui permet de déstabiliser le cortex frontal, permettant la mise à jour de la représentation en mémoire de travail dans le préfrontal par l'influence des entrées sensorielles. GPe : pallidum, segment externe. SNr : partie réticulée de la substance noire.

correspondante du PFC, permettant (grâce à la séparation des boucles parallèles) à ce stimulus particulier d'influencer la représentation encodée dans cette zone du PFC, soit de le stocker en mémoire de travail. L'absence de signal Go dans une boucle implique un maintien robuste dans cette boucle de la représentation actuelle de cette zone (stripe) du PFC.

Il est donc crucial que le système puisse apprendre quand les neurones de la boucle Go doivent être activés et lesquels doivent l'être. Les auteurs proposent un mécanisme d'apprentissage faisant intervenir la dopamine comme critique et qui apprend la valeur attendue. Celle-ci permet d'entraîner l'acteur, le système de gating des ganglions de la base, de telle sorte que les associations stimulus - neurones Go soient renforcées par une coactivation et par la présence d'un signal dopaminergique, de manière à optimiser le renforcement futur. On a vu en effet que, lorsqu'un stimulus prédisait un futur renforcement, le signal dopaminergique se déplaçait du moment du renforcement au moment de l'apparition du stimulus, permettant ainsi de renforcer, au moment adéquat, l'association entre ce stimulus et le neurone Go correspondant à sa boucle.

Notons que ce rôle de la dopamine dans l'implémentation du mécanisme de gating, et particulièrement de maintien robuste ou de switch, justifie à la fois le fait qu'elle modélise une erreur de prédiction, mais aussi qu'elle joue malgré tout un rôle essentiel dans des processus de contrôle cognitif apparemment loin de l'apprentissage, comme évoqué dans la section 1.1.2 et mis en valeur par plusieurs études (McNab et al, 2009 [144] et Nagano-Saito et al, 2008 [153]). Cela explique notamment pourquoi les patients parkinsoniens, qui souffrent d'un déficit tonique de dopamine en l'absence de médicament et ont un niveau tonique élevé de dopamine lorsqu'ils prennent des médicaments, présentent également des dégradations de leurs performance de contrôle cognitif et pas seulement d'apprentissage (Frank 2005, [83]).

On a donc un modèle muni de mécanismes qui devraient lui permettre d'apprendre des task-sets (en utilisant des renforcements et la plasticité des boucles Go et No-Go), mais aussi d'apprendre quand passer d'un task-set à l'autre. Plus précisément, le modèle devrait pouvoir apprendre à quel moment il est nécessaire de maintenir une représentation en mémoire de travail ou de switcher. Cette performance s'effectue sans intervention extérieure, ce

qui fait dire aux auteurs qu'ils parviennent à modéliser un exécutif sans homonculus.

Les auteurs testent ce modèle sur une variété de tâches de contrôle cognitif ou de mémoire de travail exigeantes : stroop, Wisconsin card sorting test, task-switching, flanker, 12-AX (tâche à deux niveaux hiérarchiques, impliquant la détection d'un X après un A, dans le cas où le dernier chiffre était un 1 ; d'un Y après un B, dans le cas où le dernier chiffre était un 2). Ils répliquent ainsi un certain nombre de données comportementales (temps de réactions, erreurs) et de lésions. De manière essentielle, ils montrent que, sans aucune intervention d'un professeur (sans fournir au modèle aucune indication explicite sur la manière correcte d'effectuer la tâche), lorsque le modèle est entraîné de manière lente et variée (sur plusieurs types de tâches simultanément), les patterns de connexions dans le PFC montrent l'émergence de représentations abstraites de type « règles », ou « dimension perceptuelle » dans le cas où celles-ci ont une pertinence pour la résolution des tâches. Cet entraînement pourrait se comparer au développement enfant-adulte et justifier le temps très long nécessaire à l'acquisition du contrôle cognitif, celui-ci n'étant stabilisé qu'à l'âge adulte.

Ce travail très exhaustif semble permettre de modéliser très largement, non seulement des problèmes de contrôle cognitif, mais également ceux en lien avec l'apprentissage de ces tâches. Le système de *gating*, notamment, permet d'introduire un mécanisme à la fois robuste et flexible de *task-switching* en décidant, en fonction de ce qui a été appris, quand il est nécessaire d'appliquer une règle ou l'autre. Cependant, il se heurte aux limitations habituelles des modèles de type réseaux de neurones : complexification des architectures impliquant une limitation du pouvoir explicatif de chaque partie du réseau ; forte dépendance à la méthode d'entraînement ; manque d'arguments biologiques pour certaines composantes, bien que la structure soit basée sur des arguments biologiques (Cohen et Frank, 2009 [44] ; Reynolds et O'Reilly, 2009 [176]). Par ailleurs, il semble limité dans la capacité à effectuer de bonnes performances dans un cadre incertain : notamment, lorsque les contingences de l'environnement changent sans qu'un indice l'ait prévu, le modèle n'est pas capable de s'adapter, sans l'ajout d'un mécanisme ad-hoc impliquant une exploration au-delà d'un certain seuil d'erreurs. Enfin, notons qu'apprentissage et contrôle cognitif ne sont pas réellement en interaction ici : ces modèles résolvent plutôt un problème d'entraînement au contrôle cognitif.

Nous présentons dans le prochain chapitre, des travaux reposant sur des modèles moins exhaustifs quant à la représentation biologique dans un réseau de neurone des informations du modèle, mais plus simples et permettant de proposer des théories claires sur le rôle computationnel précis de certaines structures biologiques. En particulier, on se focalisera ici sur le rôle du neuromodulateur norépinephrine, aussi appelé noradrénaline.

3.2.2 La noradrénaline comme signal d'interrupteur

La noradrénaline, aussi appelée norépinephrine (NE), est un neuromodulateur produit par des neurones noradrénergiques, situé dans un noyau du pont, le locus coeruleus (LC) (Aston-Jones et Cohen, 2005 [6]). Le LC projette très largement dans toutes les zones du cerveau, à l'exception notable des ganglions de la base. Il reçoit des projections de l'amygdale et du cortex préfrontal, en particulier du cortex cingulaire antérieur (ACC) et du cortex orbito-frontal (OFC). La noradrénaline a été initialement liée au niveau d'éveil, dû à l'observation de sa corrélation avec les variations du niveau tonique de NE. Des études récentes plus poussées proposent son implication dans les fonctions cognitives supérieures, et nous montrerons notamment le lien entre les observations expérimentales et la notion de switch du comportement.

Dans leur revue sur les effets de la norépinephrine dans le cortex, Aston-Jones et Cohen, 2005 [6], rappellent que ceux-ci, observés en électrophysiologie chez le rat ou le singe, sont en apparence multiple : inhibition dans certains cas, renforcement de l'activation dans d'autres. Ils observent qu'en définitive, leurs effets peuvent être définis comme une augmentation du gain de la fonction d'activation des neurones cibles, impliquant un comportement de ces neurones plus binaire face aux entrées, donc probablement une augmentation du ratio signal sur bruit.

Ils rappellent également qu'on peut essentiellement regarder les patterns d'activations des neurones NE du LC comme bimodaux. Le premier mode, qu'ils appellent mode phasique, correspond à un niveau d'activation tonique faible et à la présence d'activations phasiques. Ce mode est largement corrélé avec un niveau de performance élevé des animaux. Le deuxième mode, appelé mode tonique, est caractérisé par un niveau d'activation tonique élevé et l'absence d'activations phasiques. Ce mode est largement corrélé à un faible niveau

de performance. La présence de l'un ou l'autre mode d'activation reflète donc le niveau d'engagement dans la tâche de l'animal. Les auteurs montrent un rôle causal du LC sur la performance, et non l'inverse, par des études pharmacologiques impliquant l'injection d'agonistes ou d'antagonistes de NE impliquant les effets comportementaux prévus.

Usher et al, 1999 [207], proposent un mécanisme expliquant la possible bimodalité du LC, reposant sur la modulation connue du couplage électrotonique des neurones du LC. Il n'indique pas de mécanisme de contrôle de cette modulation.

Le rôle du LC est investigué dans une tâche demandant d'effectuer une action afin d'obtenir une récompense après la détection d'une cible, et de s'abstenir de l'effectuer après la présentation d'un distracteur. On observe plus de fausses alarmes lorsque les neurones du LC sont en mode tonique. Dans le mode phasique, Aston-Jones et Cohen, 2005 [6], observent des activations phasiques sélectives en réponse aux cibles, et non aux distracteurs. Des manipulations sur le délai entre le stimulus et la récompense permettent de décorréler leur influence et semblent montrer un lien plus fort entre l'activation phasique et la réponse qu'entre l'activation phasique et la cible. Les activations phasiques sont, par ailleurs, indépendantes des erreurs et de l'action motrice en soi, comme le montre leur absence lorsque l'animal effectue l'action en l'absence d'un stimulus. Ces activations semblent donc être liées à la décision comportementale d'effectuer l'action. Par ailleurs, Bouret et Sara, 2005 [30], montrent que la présence d'un signal extérieur saillant attendu, mais non prédit de manière certaine, conduisant à une décision cognitive, a un rôle indispensable pour qu'on observe une activation phasique du LC. Enfin, Aston-Jones et Cohen, 2005 [6], soulignent le fait que cette activation est très dépendante de la tâche et que le LC est très plastique : en effet, dans des tâches de choix forcé à deux options avec reversal, où la cible devient distracteur et inversement, l'adaptation du pattern d'activation du LC est très rapide.

Les deux jeux d'auteurs déduisent de cet ensemble de caractéristiques que le rôle de la NE du LC est de promouvoir les shifts comportementaux induits par des stimuli extérieurs dans le cadre d'une tâche précise, contribuant ainsi à la cognition de haut niveau. Ce rôle est effectué par le mécanisme d'augmentation du gain des neurones ciblés, facilitant ainsi un reset d'un réseau de neurones sous l'influence des entrées sensorielles. Bouret et Sara, 2005 [30] et Sara, 2009 [189], argumentent d'ailleurs que ce phénomène a été observé très

précisément dans les réseaux de neurones du crabe, où un pic de NE conduit à l'interruption de l'activité globale du réseau et permet sa reconfiguration.

Nous insistons sur le fait que la notion de signal d'interruption provoquant un shift comportemental, implique le fait de sortir volontairement d'un comportement par défaut actuel, pour en sélectionner un autre pertinent, en fonction de l'environnement.

Cette théorie est concordante avec l'existence de deux modes corrélés avec des comportements différents. En effet, le mode tonique impliquerait ainsi une reconfiguration permanente du réseau de neurones cible, impliquant une moins grande stabilité et donc une plus grande distractibilité.

On peut se demander quel est le mécanisme de contrôle de ce signal phasique d'interruption envoyé par le LC, soit comment celui-ci détermine la nécessité d'envoyer ce signal. A cet effet, Aston-Jones et Cohen, 2005 [6], mettent en évidence le fait que les principales connexions entrantes du LC sont projetées depuis l'OFC et l'ACC, deux régions notablement engagées dans l'estimation de l'utilité et du coût des tâches cognitives. Ils argumentent que la nécessité d'effectuer une tâche cognitive en réponse à un stimulus est déterminée en fonction de sa valeur et de son coût et que le LC serait ainsi idéalement positionné pour effectuer la balance et influencer cette décision.

Yu et Dayan, 2006 [64], proposent de spécifier plus précisément le rôle des activations phasiques de la NE comme un signal d'interruption lié aux événements inattendus. Ils construisent un modèle bayésien de la tâche de détection d'une cible. L'attente d'une cible incertaine est modélisée comme un a priori sur la probabilité de la cible, ajouté à une incertitude sur le moment d'arrivée d'un stimulus pendant la fixation initiale. L'observation du stimulus est sujette au bruit (confusion stimulus-distracteur), de telle sorte que la probabilité d'observer correctement le stimulus est fixée à un paramètre $\eta > 0,5$, fixant le niveau de difficulté de la tâche.

L'inférence bayésienne est alors possible de manière exacte et permet de calculer à chaque pas de temps à l'intérieur d'un essai individuel, l'incertitude sur l'état présent dans la tâche. Les auteurs définissent cette incertitude comme le quotient de la probabilité a posteriori (après observations à chaque pas de temps) d'observer une cible, par la probabilité a priori.

Un argument ici est essentiel pour justifier cette formulation de l'incertitude : l'hypothèse d'un comportement a priori, par défaut, incitant à ne pas sélectionner d'action. En effet, la séquence fixation → distracteur est la plus fréquente ; on peut donc considérer le fait de ne pas réagir à un stimulus comme le comportement par défaut. Cela justifie la dissymétrisation de l'incertitude vers la cible, plutôt que vers le distracteur.

Les auteurs proposent alors que la NE coderait pour cette incertitude qui permettrait ainsi, lorsqu'un événement est inattendu dans le cadre du comportement par défaut, d'interrompre ce comportement par défaut et de switcher vers le comportement adapté. Les simulations numériques effectuées dans le cadre de ce modèle parviennent à reproduire les effets qualitatifs observés dans plusieurs expériences et décrits plus haut, argumentant en faveur de la NE phasique comme un signal d'interruption lié aux événements incertains dans le cadre d'un comportement par défaut.

On voit donc qu'il existe un faisceau d'arguments connectant le rôle de la norépinephrine à la notion de switch comportemental. On notera cependant que le type de switch présenté ci-dessus est de relativement bas niveau, dans le sens où il implique plus un simple niveau stimulus- réponse que le niveau hiérarchique plus élevé du task-set, permettant de modéliser le task-switching.

Cependant, Yu et Dayan, 2005 [217], proposent que le NE, dans sa dimension tonique cette fois, implémente cet aspect du switch au niveau de la tâche, en codant cette fois non pas l'incertitude sur un état à l'intérieur d'une tâche, mais l'incertitude sur la validité d'un task-set global.

Yu et Dayan proposent de séparer deux notions d'incertitudes. La première, l'incertitude attendue, fait partie intégrante de la tâche, qui est définie non seulement par un task-set guidant sa stratégie, mais aussi par un modèle interne prédisant les conséquences de sa stratégie, ici en termes de fiabilité de la récompense. La deuxième est l'incertitude inattendue, soit 1 moins la confiance que le modèle a dans la validité du task-set présentement utilisé.

A nouveau, les auteurs proposent d'estimer de manière bayésienne la quantité qui nous intéresse. Le modèle interne estimant l'incertitude attendue est simplement mis à jour de

manière fréquentielle, après observation des récompenses à chaque essai. La probabilité a posteriori que chacun des task-sets disponibles soit correct, après observation des récompenses précédentes, est par contre calculée de manière bayésienne. Utilisant un argument lié à la capacité de codage des réseaux de neurones, les auteurs soutiennent qu'une inférence exacte serait impossible. Ils proposent donc d'établir le task-set utilisé à un instant donné comme le task-set par défaut, et d'approximer l'estimation de la confiance dans ce task-set par défaut en supposant une uniformité sur les autres task-sets possibles, rendant les calculs abordables.

Ils proposent alors qu'un task-switch aurait lieu lorsque la confiance dans le task-set par défaut devient plus faible que la confiance dans l'ensemble des autres task-sets.

Les auteurs émettent l'hypothèse d'un codage par l'acétylcholine (ACh) de l'incertitude attendue, et un codage par la NE de l'incertitude liée à la tâche, inattendue. Le critère de switch (lié à la confiance) se traduit par le simple dépassement d'un seuil par le niveau de NE, ce qui nous ramène à la notion de codage d'un task-switch par l'activation du mode tonique fort de la norépinephrine.

Les auteurs justifient la validité du modèle en mettant en avant les relations intriquées entre l'ACh et la NE dans ce modèle. En effet, plusieurs effets peuvent être prédits :

- une augmentation de la persévération (pas de switch lorsque la tâche change) lorsque le niveau de NE est faible, ne permettant pas une hausse suffisante de l'incertitude inattendue,
- une forte distractibilité (switch trop fréquent) lorsque le niveau de ACh est faible ; en effet, si l'incertitude attendue n'est pas encodée, elle est transférée sur l'incertitude inattendue, impliquant des switch plus fréquents,
- une réduction de ces deux effets lorsque le niveau d'ACh et de NE à la fois est faible, dû à leur interaction dans l'estimation de l'incertitude inattendue.

Des résultats allant dans les sens des effets prédits sont effectivement observés, soutenant ainsi la validité du modèle. On peut donc supposer un rôle essentiel de la norépinephrine dans le task-switching, probablement via l'estimation d'incertitude sur la tâche effectuée.

Nous montrerons dans la partie suivante que l'incertitude, le switch et l'exploration sont reliés fortement et que la norépinephrine semble être au cœur des mécanismes impliqués dans ces fonctions.

3.2.3 Switch et exploration, rôle de la noradrénaline

Avant de revenir au rôle de la norépinephrine dans l'exploration, ainsi qu'à son lien avec l'incertitude, revenons au problème du switch. Si nous avons exposé plus haut la notion fondamentale de task-set impliquant un effort cognitif pour sa désactivation, nous n'avons décrit pour le moment que des mécanismes de switch impliquant le fait d'interrompre le comportement précédent par défaut. Or, le switch implique également la nécessité de sélectionner un nouveau comportement par défaut. Cette sélection peut être triviale, en particulier lorsqu'un contexte ou une instruction sont à l'origine du switch, définissant le prochain task-set. Cependant, il arrive également que, soit le switch soit initié en l'absence d'instructions (par exemple à la suite d'erreurs), soit que l'instruction n'indique pas le prochain TS à appliquer. En présence d'au moins trois options, il revient alors au sujet de choisir parmi celles qu'il ne vient pas de quitter. Ce processus d'exploration est stratégiquement et théoriquement complexe, comme nous l'avons déjà exposé dans le premier chapitre.

Nous avons montré d'une part, le lien entre exploration et incertitude (dans la section 1.2.2), d'autre part, le lien entre incertitude et norépinéphrine (dans la section précédente). Nous pouvons donc maintenant argumenter sur le fait que la norépinéphrine joue un rôle essentiel dans l'exploration. Dans Cohen, McClure et Yu, 2007 [43], les auteurs se reposent sur le modèle proposé par Yu et Dayan, 2005 [217] pour proposer que le signal tonique de la norépinephrine, lorsqu'il devient fort, implique un switch, soit l'abandon du comportement par défaut présent. Il semblerait donc que le mode tonique élevé de la norépinephrine corresponde aux périodes d'exploration par opposition aux périodes d'exploitation. Aston-Jones et Cohen, 2005 [6], argumentent également dans cette direction, lorsqu'ils posent la question de l'intérêt de ce mode tonique. En effet, les résultats comportementaux observés impliquent plus d'erreurs et un moins grand engagement dans la tâche. Cette distractibilité pourrait permettre un comportement exploratoire dans le sens ou elle s'écarte de l'exploitation des actions connues comme correctes, et pourrait donc dans ce sens être intéressante

d'un point de vue évolutionnaire, malgré les erreurs qu'elle apporte. C'est d'ailleurs ce qui est observé (Dayan et Yu, 2006 [64]) : en effet, des rats ayant reçu un traitement de telle sorte que leur niveau de NE soit augmenté, parviennent à s'adapter plus rapidement à une nouvelle règle que les rats qui n'ont pas reçu ce traitement, indiquant plus de flexibilité, liée à une exploration plus importante.

Pour finir, il est très important de noter que le LC reçoit essentiellement des projections du cortex orbito-frontal et de l'ACC, deux régions impliquées dans l'évaluation des récompenses (OFC, ACC), du conflit, de la volatilité et des erreurs (notions proches de l'incertitude, ACC). Nous avons donc un faisceau convergent d'arguments pour un rôle de la NE, pour calculer (à partir d'entrées signalant l'utilité attendue d'une tâche et l'incertitude précédente) l'incertitude présente, la signaler à l'AFC et agir comme signal de switch pour entrer en période d'exploration.

Cet ensemble d'arguments liant exploration, incertitude, switch avec la norépinéphrine et le cortex ventro-médial tend à montrer que l'exploration n'est pas seulement un phénomène de bas niveau, mais qu'il peut s'agir d'une décision impliquant du contrôle cognitif, au même titre qu'un switch.

Chapitre 4

Question de thèse

Apprentissage et contrôle cognitif : données empiriques. Nous avons montré jusqu'ici que la recherche a établi la forte imbrication des problèmes d'apprentissage et de contrôle cognitif, dans les ganglions de la base et le cortex préfrontal. Nous avons montré que le cerveau était capable de découvrir et d'utiliser les structures sous-jacentes aux problèmes et qu'il était muni d'une structure fonctionnelle hiérarchique particulièrement bien adaptée à résoudre les problèmes de contrôle cognitif. Nous avons montré l'existence d'une première unité hiérarchique du contrôle cognitif, le task-set, et que la flexibilité du contrôle cognitif était bien représentée par le switch entre différents task-sets, processus permettant de passer de l'exploitation d'une tâche par défaut à l'exploration des options disponibles. Nous avons montré que ces processus de switch et d'exploration étaient crucialement dépendants des questions d'incertitude.

Apprentissage et contrôle cognitif : modèles. Nous avons montré que des modèles de réseau de neurones se basant sur une imitation précise des structures et connectivités connues permettaient de reproduire beaucoup d'observations empiriques, sans toujours être satisfaisants quant à leur pouvoir explicatif (Cohen et Frank, 2009 [44]). Nous avons montré que deux grandes classes de modèles théoriques sont utilisées – parfois de manière jointe – pour proposer des théories fonctionnelles de l'apprentissage et de la décision : l'apprentissage par renforcement et les modèles d'inférence bayésienne.

Programme de la thèse. Dans cette thèse, nous avons posé la question de l'apprentissage des task-sets en vue de leur utilisation par le contrôle cognitif. Nous voyons cette question comme englobant très largement les fonctions exécutives préfrontales : en effet, elle peut être vue comme un problème d'utilisation du contrôle cognitif, pour extraire une notion de task-set, dans un problème d'apprentissage. Elle peut également être vue comme un problème d'utilisation de l'apprentissage pour le contrôle cognitif : apprentissage de task-sets avant de pouvoir effectuer du task-switching entre ces task-sets.

Nous posons la question de l'apprentissage des task-sets en l'absence d'apprentissage dirigé, en présence de renforcement et en présence ou absence d'information contextuelle. Cette question est posée dans un environnement incertain et pour un minimum de trois task-sets : en effet, nous avons vu le rôle crucial joué par l'incertitude dans le switch, et nous argumentons qu'un switch entre deux options est qualitativement différent d'un switch en présence de trois options, puisque ce dernier met en jeu de l'exploration.

Nous proposons d'apporter une réponse à cette question en utilisant des modèles quantitatifs de type RL ou bayésiens plutôt que des réseaux de neurones, afin de pouvoir proposer simplement des mécanismes de manipulation de l'information qui pourraient être effectués dans le cerveau.

Nous soutenons qu'aucun modèle présenté ne résout ce problème, et passons rapidement en revue, ci-dessous, des modèles qui pourraient prétendre à le résoudre ainsi que leurs limitations.

RL, par exemple Q-learning. Si les modèles d'apprentissage par renforcement peuvent résoudre le problème de l'apprentissage d'un task-set, ils ne peuvent pas résoudre celui de l'apprentissage de plusieurs task-sets. En effet, l'apprentissage d'un nouveau task-set provoque l'effacement des valeurs apprises précédemment, donc l'oubli du task-set précédent. Pour permettre au RL de pouvoir résoudre le problème de l'apprentissage des task-sets, il faudrait ajouter un mécanisme de contrôle permettant de décider du stockage d'un task-set en mémoire au moment opportun, et de sa réutilisation le cas échéant.

Même en présence d'information contextuelle et en mettant dans l'espace des états non pas seulement les stimuli mais plutôt les paires (contexte, stimulus), le résultat n'est pas

satisfaisant. En effet, le RL peut, dans ce cas, apprendre différents TS dans différents contextes ; mais deux problèmes se posent. Tout d'abord, aucun transfert n'est possible si deux contextes sont valides pour un task-set. Ensuite, le passage d'un stimulus à l'autre n'est pas différent du passage d'un contexte à l'autre, donc d'un task-set à l'autre, niant ainsi la spécificité cognitive de la notion de tâche et de switch.

Unexpected uncertainty. Le modèle bayésien proposé par Yu et Dayan, 2005 [217], proposant l'incertitude inattendue comme signal de switch est satisfaisant dans sa capacité à maintenir un task-set par défaut en présence d'incertitude et à switcher quand nécessaire. Cependant, il suppose les task-sets déjà appris et ne propose pas de mécanisme d'exploration pour la sélection du prochain task-set après un switch.

Ce modèle pourrait être couplé à un modèle de RL pour apprendre des task-sets, le signal de switch permettant alors de signaler le moment de stocker ces task-sets et, soit d'en apprendre un nouveau, soit d'en réutiliser un (par exemple en utilisant de l'information contextuelle). Cependant, même ainsi modifié, il ne permet pas non plus d'explorer l'espace des task-sets après un switch.

MMBRL. Le modèle mixte RL, bayésien, proposé par Doya et al (MMBRL, 2002 [75]), adapté à un environnement discret convenant au contrôle cognitif, permettrait à la fois l'apprentissage et le stockage d'un nombre prédéfini de task-sets, l'exploration de l'espace des task-sets appris et le passage d'un task-set à l'autre. Cependant, nous argumentons qu'il représente mal l'interaction de l'apprentissage et du contrôle cognitif pour plusieurs raisons. Le première est la symétrie parfaite du rôle de chaque TS à tout temps dans le modèle, alors que nous avons vu qu'un comportement par défaut était sélectionné tant qu'il semblait valide. Par ailleurs, cette symétrie implique un ralentissement considérable des apprentissages initiaux. La deuxième est l'absence de notion de switch, puisque tous les TS prennent part à la décision à chaque essai. Enfin, ce modèle pose un problème de rigidité : il présuppose un nombre de TS à apprendre dès le début. Si le problème a un nombre effectif plus bas de TS à considérer, il y a gâchis d'énergie ; si, au contraire, ce nombre est plus grand, le modèle ne pourra pas résoudre le problème de manière adéquate.

Dans la prochaine partie, nous montrerons comment nous proposons de réunir les points positifs de ces différents modèles (structure hiérarchique et exploration de MMBRL, structure binaire de décision dans *unexpected uncertainty*) en un modèle permettant de proposer un principe d'acquisition de task-sets pour le contrôle cognitif.

Nous montrerons ensuite, dans deux expériences comportementales, que le modèle proposé permet, contrairement aux autres modèles, de rendre compte de la manière très variable dont des sujets humains intègrent l'apprentissage et le contrôle cognitifs.

Deuxième partie

Modèles d'apprentissage et de contrôle cognitif

Chapitre 5

Les modèles

Dans cette partie, nous présentons la théorie construite pour représenter l'interaction de l'apprentissage et du contrôle cognitif. Nous présenterons tout d'abord le cadre précis dans lequel nous étudions ce problème. Nous présenterons ensuite trois modèles adaptés de modèles existants auxquels nous comparerons le modèle proposé. Enfin, nous détaillerons les aspects spécifiques de notre modèle, tout d'abord sans information contextuelle, puis en présence d'information contextuelle. En dernière partie, nous concluerons sur les apports de notre modèle et les prédictions qu'il permet d'effectuer.

5.1 Définition du cadre

Nous nous plaçons dans un environnement doté d'un espace fini de stimuli $\{s_i\}_{i=1...n_S}$, ainsi que d'un espace fini d'actions $\{a_i\}_{i=1...n_A}$. Nous définissons l'interaction de l'environnement avec le sujet par l'existence de task-sets réels, qui seront notés par la suite TS*. TS* représente le comportement qui est correct à un moment donné, dans le sens où l'utilisation de ce comportement est optimale en terme de renforcements obtenus par l'acteur. TS* n'est pas observable par l'acteur, il est fixé par le monde extérieur. Les TS* sont en nombre fini, TS*$= 1, \ldots, n_{TS}$.

Nous supposons qu'initialement, TS* ne change pas trop fréquemment, afin de permettre

un apprentissage au moins approximatif d'un TS* avant qu'il ne change. Cette hypothèse est motivée par le fait que, lorsque des règles ne sont pas exclusives et sont apprises simultanément, elles tendent à ne former qu'un seul TS. Cela implique qu'il est nécessaire d'isoler un minimum dans le temps l'utilisation d'un task-set afin de l'apprendre, au moins par renforcement. Nous appelons un **épisode** une période de temps pendant laquelle un TS* ne change pas.

Bien que ce ne soit pas une limitation théorique de notre modèle, nous nous limitons par la suite à une fonction de renforcement stochastique binaire : après chaque action sélectionnée, l'acteur a soit gagné ($r = 1$) soit perdu ($r = 0$), sans implication d'un niveau de récompense. La probabilité de gagner lorsque l'action sélectionnée par l'acteur correspond à l'action indiquée par le TS* est γ, où γ est au moins supérieur à 0,5, et peut dépendre des stimuli et des TS*. En pratique, on fixe γ à au moins $0,75$. La probabilité de gagner lorsque l'action sélectionnée par l'acteur ne correspond pas à l'action indiquée par le TS* est $1 - \gamma$. γ représente donc la fiabilité de l'environnement, à quel point une réponse *« gagné »* (respectivement *« perdu »*) est indicative ou non d'une action correcte (respectivement incorrecte).

Notons que les task-sets sont en général représentés comme des ensembles d'associations stimulus – action. En particulier, nous supposons ici que, si l'acteur est soumis à une série de stimuli dans le but d'apprendre différents task-sets, il n'a pas d'effets sur l'environnement autre que de provoquer des renforcements : il ne provoque pas le prochain stimulus présenté par l'action sélectionnée (ce qui serait le cas dans une expérience de type labyrinthe, par exemple). Au contraire, la séquence de stimuli est aléatoirement déterminée par l'environnement. Autrement dit, la fonction de transition est uniforme, l'action choisie n'influence pas le prochain stimulus présenté, ce qui justifie, comme expliqué dans le paragraphe 1.1.1, d'apprendre indépendamment les valeurs de chaque stimuli. Par exemple, pour le RL, on posera $\gamma = 0$.

Nous étudions par la suite deux cas : dans un cas, aucune information contextuelle n'est disponible, dans l'autre en revanche, si. Nous appelons un **contexte** une deuxième dimension des entrées sensorielles, différente des stimuli. Nous supposons que les contextes sont informatifs sur les task-sets réels, TS*. Selon la terminologie de Koechlin et al, 2003 [136], nous

nous plaçons dans le cadre du contrôle contextuel et non dans le cadre du contrôle épisodique. Cela signifie que les associations contexte – task-set ne changent pas, le contexte est complètement informatif sur le TS* : un seul TS* peut être valide pour un contexte donné. La réciproque est fausse : plusieurs contextes peuvent être associés au même task-set.

5.2 Cas sans contexte : définition des trois autres modèles et de leurs points forts

En l'absence de contexte, la seule information présente afin d'apprendre les task-sets est le renforcement observé après la sélection par le modèle d'une action en réponse à un stimulus. Nous présentons ci-dessous trois modèles qui permettent d'agir plus ou moins bien face aux situations indiquées ci-dessous.

5.2.1 RL (apprentissage par renforcement simple)

Dans ce modèle, nous supposons simplement que le sujet estime les valeurs $Q(s, a)$ pour tous stimulus s, action a. La mise à jour de ces valeurs s'effectue comme indiqué dans le paragraphe 1.1.1, après l'observation du renforcement r au temps t :

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha(r - Q_t(s, a))$$

Nous avons vu plus haut que plusieurs règles de sélection des actions étaient possibles. Nous proposons ici une règle combinant le softmax et le ϵ-greedy :

$$P(a|s) = (1 - \epsilon)\frac{exp(\beta Q(s, a))}{\sum exp(\beta Q(s, a'))} + \frac{\epsilon}{nA} \tag{5.1}$$

Nous avons donc trois paramètres :
– α contrôle la vitesse d'apprentissage,
– β contrôle l'exploration dirigée,
– ϵ représente un bruit de sélection global.
Ce modèle est très efficace pour l'apprentissage d'un task-set initial. Il est également capable de s'adapter après un changement d'épisode pour apprendre un nouveau task-set.

Cependant, ce faisant, il efface le précédent task-set et doit donc le réapprendre au prochain épisode dans lequel ce task-set est valide. Par ailleurs, le passage d'un task-set à l'autre demande de désapprendre ce qui a été appris précédemment et implique donc une adaptation relativement lente, comme le montre la figure 5.2 a), page 132. Nous détaillerons les résultats présentés dans cette figure dans la section .

5.2.2 Modèle UU (unexpected-uncertainty, plus RL)

Un modèle incluant la détection du moment du changement d'épisode afin de pouvoir repartir sur un apprentissage nouveau permettrait de pallier ce problème de 'désapprentissage'. Nous proposons de combiner un modèle RL à un modèle estimant l'incertitude inattendue afin de pouvoir détecter la nécessité de switcher.

Plus précisément, dans ce modèle, l'acteur apprend comme ci-dessus un task-set, par mise à jour RL. Cependant, il apprend également à prédire l'incertitude inhérente à ce task-set. Il est muni d'un modèle interne de l'incertitude, soit la probabilité d'observer un *gagné* ($r = 1$) ou *perdu* ($r = 0$), pour une action a face à un stimulus s, dans le cadre de ce TS : $\gamma(s, a) = P(r|s, a, TS)$. Nous supposons que ce modèle est appris par observation fréquentielle : γ est estimé par le nombre de fois que la paire (s,a) a conduit à une récompense $r = 1$ divisé par le nombre de fois que la paire a été testée dans le cadre de ce task-set (n(s,a)). En pratique, la mise à jour s'effectue essai après essai par :

$$\begin{aligned} \gamma_{t+1}(s,a) &= \gamma_{t+1}(s,a) + \frac{1-\gamma_{t+1}(s,a)}{n(s,a)} \text{ pour } r_t = 1 \\ &= \gamma_{t+1}(s,a) - \frac{\gamma_{t+1}(s,a)}{n(s,a)} \text{ pour } r_t = 0 \end{aligned}$$

γ est un modèle interne direct de l'incertitude, car il permet à l'acteur de prédire la conséquence de ses actions : dans le cadre du task-set TS, si on sélectionne l'action a face au stimulus s, on s'attend à obtenir $r = 1$ avec probabilité $\gamma(s, a)$ et $r = 0$ avec probabilité $1 - \gamma(s, a)$. Il peut donc être utilisé pour une inférence bayésienne sur la variable cachée TS^*.

Conformément au modèle proposé par Yu et Dayan, 2005 [217], dans ce modèle, l'hypothèse est que le task-set présentement utilisé est correct par défaut, jusqu'à ce qu'on

observe la nécessité de switcher. Il est donc mesuré contre toutes les autres possibilités. On estime ainsi la confiance dans le task-set utilisé (TS) après chaque observation : $\lambda_{t+1} = P(TS^*_{t+1} = TS|r_t, histoire)$. Notons que λ_{t+1} est une confiance ex-ante : avant qu'aucune observation ne soit effectuée au temps $t+1$, on détermine la probabilité que le TS utilisé à t soit également valide à $t+1$.

La mise à jour se fait de manière approximative par inférence bayésienne, en introduisant de manière intermédiaire μ_t défini comme suit :

$$\mu_t = P(TS^*_t = TS|r_t) = \frac{P(r_t|TS^*_t = TS)P(TS^*_t = TS)}{P(r_t|TS^*_t = TS)P(TS^*_t = TS) + P(r_t|TS^* \neq TS)P(TS^*_t \neq TS)}$$

Toutes les probabilités sont conditionnées sur l'histoire des observations jusqu'à $t-1$. Cette formule se résume simplement par :

$$\mu_t = \frac{\gamma\lambda_t}{\gamma\lambda_t + \xi(1-\lambda_t)} \text{ si } r_t = 1 \quad (5.2)$$

$$= \frac{(1-\gamma)\lambda_t}{(1-\gamma)\lambda_t + (1-\xi)(1-\lambda_t)} \text{ si } r_t = 0 \quad (5.3)$$

Ici ξ représente la probabilité de gagner si ce TS n'est pas correct : $P(r_t = 1|TS^*_t \neq TS)$. La valeur de ξ est fixée au niveau du hasard, $1/n_A$.

Cette probabilité est une évaluation ex-post de la confiance sur le fait que le TS utilisé était correct, après avoir observé le renforcement conséquent. Pour en déduire λ_{t+1}, la confiance ex-ante pour l'essai $t+1$, on a besoin d'une probabilité de transition : en effet,

$$\begin{aligned}
\lambda_{t+1} = P(TS^*_{t+1} = TS_t|r_t) &= P(TS^*_{t+1} = TS_t, TS^*_t = TS_t|r_t) + P(TS^*_{t+1} = TS_t, TS^*_t \neq TS_t|r_t) \\
&= P(TS^*_{t+1} = TS_t|TS^*_t = TS_t)P(TS^*_t = TS_t|r_t) \\
&\quad + P(TS^*_{t+1} = TS_t|TS^*_t \neq TS_t)P(TS^*_t \neq TS_t|r_t) \\
&= \tau\mu_t + (1-\tau)(1-\mu_t)
\end{aligned}$$

La dernière ligne est approximée en supposant une probabilité de stabilité du TS^* fixe, $P(TS^*_{t+1} = TS^*_t) = \tau$, indépendante des TS, pour simplifier. Cette probabilité était fixée à $\tau = 1$ dans Yu et Dayan, 2005 [217], dans l'hypothèse de changement rare d'épisode, impliquant simplement $\lambda_{t+1} = \mu_t$. Nous ne reprenons pas cette possibilité. En effet la

distinction des rôles de μ (évaluation ex-post) et λ (décision ex-ante) joue un rôle très important, et sera particulièrement mise en valeur dans les modèles avec contextes [1].

Le modèle permet alors d'indiquer la nécessité de switcher : lorsque la probabilité que le task-set actuel soit correct devient plus faible que la probabilité qu'il soit incorrect, il est nécessaire de quitter le task-set par défaut et d'en utiliser un autre. Ce critère se traduit par $\lambda_{t+1} < 1 - \lambda_{t+1}$, soit $\lambda_{t+1} < 0,5$. On introduit donc naturellement un seuil pour la confiance dans le TS actuel, en dessous duquel il est nécessaire de switcher.

L'intervention de ce signal de switch permet de signaler que ce qui est observé n'est plus conforme à ce qui est attendu (par le modèle interne direct) dans le cadre d'utilisation du TS appris. Elle permet donc d'interrompre l'adaptation sur ce modèle et de pouvoir envisager le stockage du TS appris avant qu'il n'ait été modifié par apprentissage dans une nouvelle condition. C'est donc un mécanisme qui permet d'apprendre un TS et de déterminer à quel instant il faut interrompre son apprentissage avant qu'il ne soit dégradé. On voit sur la figure 5.2 b), page 132, le switch signalé par la chute de la confiance dans le task-set peu après le début de l'épisode, et on constate également que l'apprentissage des nouveaux task-sets est bien plus rapide avec le rajout du switch au RL. On commentera plus précisément cette figure dans la section 5.3.5.

Cependant, ce mécanisme de switch ne fournit aucun mécanisme naturel permettant de décider quel task-set appliquer après le switch. En effet, même si les task-sets précédemment appris sont stockés, comment décider lequel utiliser autrement qu'en en sélectionnant un au hasard et en l'essayant pendant un certain nombre d'essais (ce qui est proposé dans Yu et Dayan, 2005 [217]) ? Il est évident que ce processus est inefficace. De plus, il n'est pas satisfaisant, dans le sens où l'espace des task-sets à explorer doit être construit avant d'exister : après chaque switch une décision doit être prise entre deux processus, l'apprentissage d'un nouveau task-set ou la sélection d'un task-set précédemment appris. Aucune proposition n'existe pour arbitrer entre ces deux processus.

Nous nous limitons donc dans ce modèle au seul mécanisme qu'on peut implémenter natu-

1. Notons que garder $\tau < 1$ permet également d'éviter des problèmes de simulation. En effet, 1 est un attracteur de la fonction de mise à jour (5.2) et risque d'être atteint, par approximation de Matlab, lors des simulations. L'introduction de $\tau < 1$ assure que pour tout t, $\lambda_t < \tau$.

rellement dans son cadre, pour le problème d'apprentissage des task-sets. Après un switch, un nouveau task-set est systématiquement appris, à partir d'une table de Q-values initialisée pour une sélection uniforme des actions.

Ce modèle ne permet donc, pas plus que le précédent, de stocker des task-sets dans l'espoir de les réutiliser, en l'absence de moyen de savoir quel task-set réutiliser. Cependant, il permet d'introduire un switch comme la détection de la fin d'un épisode et la nécessité de passer à autre chose. Il évite donc le problème du modèle précédant, devant effacer ce qu'il a appris avant de pouvoir apprendre à nouveau.

5.2.3 MMBRL (Multiple Model-Based Reinforcement Learning)

Le troisième modèle adapté de modèles existants (Doya et al, 2002 [75]) permet de proposer une méthode de re-sélection des task-sets déjà appris.

Ce modèle est muni de N feuillets, pouvant chacun représenter un task-set, tous initialisés comme un comportement aléatoire. Chaque feuillet est muni d'un module d'apprentissage par renforcement interne, en particulier d'une table de Q-values qui lui est spécifique : $Q_i(s,a)$, pour TS_i, $i = 1, ..., N$; et d'une stratégie déterminée à partir de cette table Q_i comme pour le modèle RL, formule 5.1. Chaque feuillet est également muni de son propre modèle interne direct : $\gamma_i(s,a) = P(r = 1|s, a, TS^* = TS_i)$. Enfin, pour chaque feuillet, représentant un possible task-set, on calcule la confiance qu'on a dans ce task-set particulier, étant donné l'historique des observations effectuées : pour $i = 1, \ldots, N$, $\lambda_i(t+1) = P(TS^*_{t+1} = TS_i|r_t,\ histoire)$. Ce signal est appelé *responsibility signal*, ou signal de responsabilité, dans Doya et a, 2002 [75].

Les valeurs de responsabilité λ_i sont définies exactement comme la valeur λ du modèle UU ci-dessus pour le TS en cours. La différence essentielle est que dans le modèle UU, cette valeur n'est estimée que pour le task-set par défaut en cours (ce qui implique une approximation dans la mise à jour), alors qu'elle est estimée pour l'ensemble des N task-sets dans MMBRL.

La mise à jour de l'estimation de la confiance ex-ante dans chaque task-set s'effectue après observation d'un renforcement par la formule bayésienne suivante, en passant à nouveau

par l'intermédiaire ex-post μ :

$$\mu_i(t) = P(TS_t^* = TS_i|r_t) = \frac{P(r_t|TS_t^* = TS_i)P(TS_t^* = TS_i)}{\sum_{j=1}^{j=N} P(r_t|TS_t^* = TS_j)P(TS_t^* = TS_j)}$$

A nouveau, toutes les probabilités sont conditionnées sur l'histoire des observations jusqu'à $t-1$. Cette formule se résume donc simplement par :

$$\begin{aligned} \mu_i(t) &= \frac{\gamma_i \lambda_i(t)}{\sum_{j=1}^{N} \gamma_j \lambda_j(t)} \text{ si } r_t = 1 \\ &= \frac{(1-\gamma_i)\lambda_i(t)}{\sum_{j=1}^{N} (1-\gamma_j)\lambda_j(t)} \text{ si } r_t = 0 \end{aligned}$$

Comme précédemment, on déduit de cette confiance ex-post μ une confiance ex-ante λ grâce à la probabilité de stabilité τ par la formule : $\lambda_i(t+1) = \tau\mu_i(t) + \frac{1-\tau}{N-1}(1-\mu_i(t))$.

Les λ_i (*responsibility signals*) sont ainsi nommés parce qu'ils permettent la pondération entre les task-sets, sur la question de savoir quel task-set prend la responsabilité de l'action et de l'apprentissage. Ils sont utilisés de trois manières par le modèle :

– λ_i pondère la vitesse d'apprentissage du module RL du task-set correspondant ; en particulier, la formule de mise à jour des Q-values est pour tout $i = 1, \ldots, N$:

$$Q_{t+1}(s,a) = Q_t(s,a) + \lambda_i(t)\alpha(r - Q_t(s,a))$$

ainsi, un task-set n'est appris que lorsque le modèle est confiant quant au fait qu'il se trouve dans l'environnement correspondant à son application ;

– λ_i pondère également l'apprentissage du modèle interne de la même manière que ci-dessus ; ainsi, le modèle n'apprend à prédire les conséquences de ses actions que dans l'environnement auquel il est adapté ;

– enfin et essentiellement, λ_i sert à déterminer quel task-set est acteur à un instant donné. Plus précisément, à chaque essai, le task-set actif est sélectionné sur la base d'un softmax portant sur les λ_i : chaque task-set est sélectionné avec une probabilité proportionnelle à $\exp(\beta'\lambda_i(t))$. Le task-set sélectionné applique alors sa stratégie pour déterminer une action.

Nous avons donc deux paramètres supplémentaires : N et β'.

Ce modèle est capable d'apprendre plusieurs task-sets et de savoir quand les réutiliser. En effet, initialement, tous les feuillets sont vierges et approximativement à égalité et apprennent tous en parallèle. Cependant, la présence de bruit dans l'initialisation soit des λ_i soit des γ_i assure qu'un task-set aura une responsabilité légèrement plus haute que les autres. Pendant le premier épisode, la dynamique non linéaire de mise à jour des signaux de responsabilité assurera que ce task-set, apprenant légèrement plus vite son modèle interne, devient légèrement plus prédictif de ce qui est observé et finira donc par émerger (relativement rapidement) comme le seul task-set représentatif du premier épisode (voir figure 5.2 c, page 132). Cette simulation sera détaillée dans la section 5.3.5.

A un changement d'épisode correspondant à la présentation, pour la première fois, d'un nouveau task-set, le même phénomène d'émergence aura lieu entre les feuillets restants encore vierges. A un changement d'épisode correspondant à la présentation d'un task-set déjà appris lors d'un épisode précédent, le modèle peut faire re-émerger ce task-set grâce à la mise à jour permanente des λ_i de tous les task-sets (voir courbe magenta, figure 5.2, d), page 132, cette simulation sera détaillée dans la section 5.3.5). En effet, alors, le task-set déjà appris est le plus prédictif de ce qui est observé et devient donc le plus responsable.

Les limitations de ce modèle sont les suivantes :

- Le fait de devoir fixer initialement le nombre N de task-sets pouvant être stockés et appris par le modèle. Si le nombre effectif de task-sets à apprendre est plus grand que N, le modèle échoue à apprendre et à utiliser une partie des task-sets. S'il est plus petit que N, le modèle effectue beaucoup de calculs n'apportant rien à l'apprentissage et le ralentissant potentiellement.
- L'apprentissage parallèle sur tous les task-sets ralentit leur apprentissage initial, comme on voit sur la courbe noire de la figure 5.2 c), page 132).
- La sélection d'un task-set à chaque essai ne semble pas correspondre au comportement humain. Nous avons en effet vu que nous tendions à avoir un comportement par défaut et à switcher et explorer uniquement lorsque c'était nécessaire.

5.2.4 Résumé

Nous avons donc vu que les modèles existants, même adaptés naturellement au problème que nous posons, ne semblent pas en mesure d'expliquer notre capacité à lier le contrôle cognitif et l'apprentissage. Cependant, chacun de ces modèles apporte une pierre à l'édifice en proposant l'un, la méthode de base d'apprentissage à partir des renforcements, l'autre la méthode de switch après un comportement par défaut pour apprendre un nouveau task-set, le troisième, la méthode d'exploration pour la réutilisation des task-sets.

Dans la partie suivante, nous décrivons le modèle que nous proposons, qui combine les différents points forts des modèles présentés ci-dessus.

5.3 Modèle proposé

Ainsi que le MMBRL, notre modèle présente, à tout instant t, plusieurs « feuillets »(appelés par la suite TS), qu'il surveille de manière permanente, munis chacun d'un module d'apprentissage par renforcement et d'un modèle interne qui lui sont propres.

Contrairement au MMBRL, le nombre de TS (n) n'est pas fixé à l'avance. Au cours de l'apprentissage, n peut augmenter. Nous décrirons plus loin le processus de création de TS.

De plus, contrairement au MMBRL, dans notre modèle, tous les TS n'apprennent pas à la fois. Nous définissons une notion de **task-set acteur** (task-set par défaut ou task-set test) étant le seul à pouvoir sélectionner les actions, à apprendre une stratégie et un modèle interne. Comme dans le modèle UU, un critère de seuil simple permet de déterminer la nécessité d'un switch et de quitter le task-set par défaut.

Nous décrivons ci-dessous en détail les différentes composantes du modèle proposé.

5.3.1 Stratégie et modèle interne d'un task-set

Les modules d'apprentissage par renforcement et du modèle interne sont semblables à ceux présentés plus haut.

Module d'apprentissage par renforcement :

- Pour tout TS_i, $Q_i(s,a)$ représente la valeur attendue quand l'action a est sélectionnée face au stimulus s et que le modèle pense que $TS^* = TS_i$
- L'erreur de prédiction se calcule après chaque essai par $\delta = r - Q_i(s,a)$
- Elle est utilisée pour mettre à jour $Q_i(s,a)$ d'une valeur proportionnelle à δ par la vitesse d'apprentissage : $\alpha\delta$. Le calcul de δ et cette mise à jour s'effectuent **exclusivement** pour le TS acteur (voir plus loin).
- Les $Q_i(s,a)$ sont utilisées pour représenter la stratégie, quand le TS est acteur : $\pi_i(s,a)$ proportionnel à $\exp(\beta Q_i(s,a))$, à un bruit général près contrôlé par ϵ :

$$\pi_i(s,a) = (1-\epsilon)\frac{exp(\beta Q(s,a))}{\sum exp(\beta Q(s,a'))} + \frac{\epsilon}{nA}$$

Modèle interne de l'incertitude :

- $\gamma_i(r,s,a) = P(r_t = r|s,a,TSi)$ représente la probabilité d'observer le renforcement r au temps t, dans le cadre de TS_i, pour l'action a choisie face au stimulus s. γ est le modèle interne de ce qu'on peut s'attendre à observer comme réaction du monde lorsqu'on est dans l'environnement dans lequel TS_i est valide.
- Lorsque le TS est acteur (et uniquement dans ce cas, contrairement à MMBRL), cette valeur est mise à jour par fréquence, comme décrit dans le modèle UU, pour le renforcement r :

$$\gamma_i^{(t+1)}(r,s,a) = \gamma_i^{(t)}(r,s,a) + \frac{r - \gamma_i(t)(r,s,a)}{n(s,a)}$$

5.3.2 Confiance dans les task-sets

On définit la **confiance ex-ante** dans un TS par la formule suivante habituelle : $\lambda_i(t+1) = P(TS^*_{t+1} = TS_i|r_t, H_t)$, avec r_t le renforcement obtenu au temps t ; H_t l'histoire des stimuli, actions et renforcements, jusqu'au stimulus t.

Notons n le nombre de TS actuellement en jeu dans le modèle. La mise à jour des λ_i s'effectue de manière bayésienne : après avoir reçu r_t, on peut estimer, pour tout TS_i (et non seulement pour le TS acteur) $\mu_i(t)$ (l'introduction de μ, confiance ex-post, comme intermédiaire de calcul, a été discuté en détail dans la section 5.2.2 et s'applique ici également.) :

$$\begin{aligned}\mu_i(t) = P(TS_t^* = TS_i|r_t) &= \frac{P(r_t|TS_t^* = TS_i)P(TS_t^* = TS_i)}{\sum_{j=0}^{n} P(r_t|TS_t^* = TS_j)P(TS_t^* = TS_j)} \\ &= \frac{\gamma_i(r_t, s_t, a_t)\lambda_i(t)}{\sum_{j=0}^{n} \gamma_j(r_t, s_t, a_t)\lambda_j(t)}\end{aligned}$$

On en déduit $\lambda_i(t+1)$ en supposant une probabilité de stabilité du TS^* fixe, $P(TS_{t+1}^* = TS_t^*) = \tau$, indépendante des TS, pour simplifier (voir section suivante pour plus de détails) :

$$\lambda_i(t+1) = \tau\mu_i(t) + \frac{1-\tau}{n-1}(1-\mu_i(t))$$

Bien que les λ_i soient calculés pour tous les TS à chaque essai, comme dans le MMBRL, contrairement à ce modèle, les λ_i ne sont pas utilisés pour pondérer un apprentissage ou pour décider la sélection d'un task-set. Par contre, de manière identique au modèle UU, ils sont utilisés pour déterminer un switch.

Si un TS_i, $i > 0$, vérifie $\lambda_i > 0,5$, cela signifie qu'on a plus confiance dans le fait qu'il soit valide plutôt que non valide. Ce critère définit la notion de **TS par défaut** [2]. Lorsqu'un TS par défaut existe, on dit qu'on est dans une **période d'exploitation**. Le TS par défaut est alors le seul TS acteur, avec apprentissage actif. Tous les autres task-sets sont alors simplement conservés sans apprentissage, bien que la confiance du modèle dans ces TS continue à être estimée.

La fin d'une période d'exploitation définit un **TS-switch**. Deux cas de figure peuvent se présenter. Si un autre TS émerge immédiatement comme TS par défaut, on passe directement à une nouvelle période d'exploitation ; sinon, on passe à une période d'exploration, qu'on définit dans le paragraphe suivant.

2. Notons bien que le TS par défaut est unique. En effet, comme $\sum_{i=0}^{n} \lambda_i = 1$, il ne peut y avoir qu'un seul TS vérifiant $\lambda > 0,5$

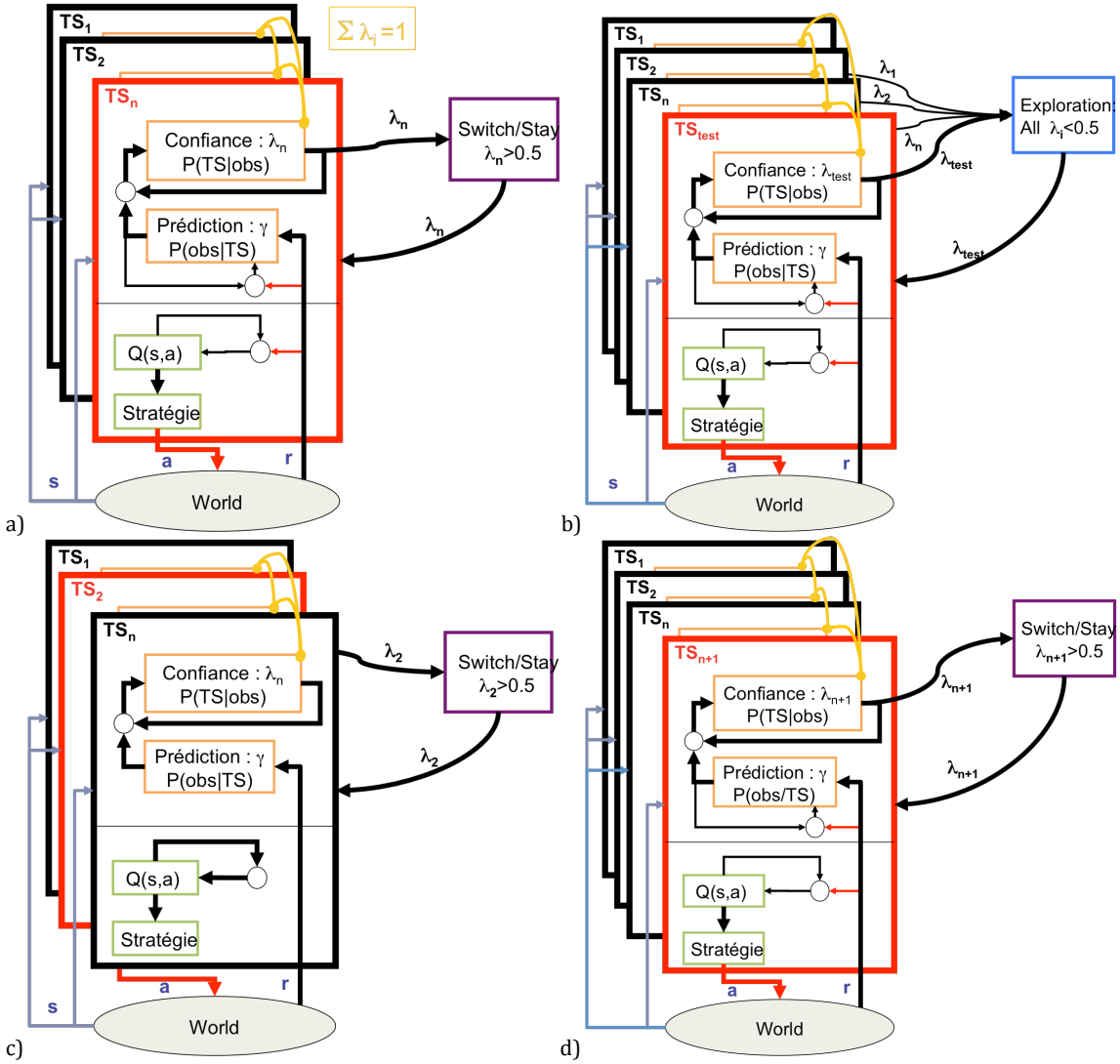


FIGURE 5.1 – Modèle proposé. Rouge : TS par défaut (a,c,d) ou TS test (b). a) période d'exploitation, TS_n comme task-set par défaut : pour ce seul TS, la récompense r est utilisée pour mettre à jour les *Q-values* et le modèle interne γ, pour l'action a qui a été choisie par ce TS face au stimulus s. La confiance λ est calculée pour l'ensemble des TS, mais TS_n reste le TS par défaut tant qu'il ne vérifie pas le critère de switch. b) Après un switch, période d'exploration : TS_{test} est le seul à apprendre, et sélectionne a. La période d'exploration se termine soit par l'émergence d'un TS stocké comme TS par défaut (c), soit par l'émergence du TS test comme TS par défaut, conduisant à l'élargissement de l'espace des TS (d).

5.3.3 Exploration

Lors d'une **période d'exploration**, par définition, pour tout $i = 1, \ldots, n$, on a $\lambda_i < 0,5$. Aucun TS ne vérifie alors la propriété : on a plus confiance dans le fait qu'il soit valide plutôt qu'invalide. On introduit alors un $n+1$ème TS, que l'on nomme TS_{test}, qui est, pendant la période d'exploration, le seul à apprendre. Il est également responsable de la sélection des actions. Il entre en compétition avec les n autres TS pour le calcul de la confiance.

TS_{test} a donc pendant une période d'exploration le même rôle que le TS par défaut a pendant une période d'exploitation. On a donc à tout instant un seul **TS acteur**, soit le TS par défaut (exploitation), soit le TS_{test} (exploration), seul responsable de la sélection des actions et de l'apprentissage.

La période d'exploration se termine à l'émergence d'un nouveau TS par défaut, $\lambda_d > 0,5$. L'émergence de TS_{test} comme nouveau TS par défaut ($\lambda_{test} > 0,5$) indique que le TS en cours d'apprentissage prédit mieux ce qui est observé qu'aucun des TS précédemment appris. Le modèle a donc perçu la nécessité de créer un nouveau TS, ce qui conduit à l'agrandissement de l'espace des TS. En revanche, l'émergence d'un autre TS que TS_{test} comme nouveau TS par défaut indique que le modèle a perçu le fait qu'un TS déjà appris était pertinent dans la situation actuelle, qu'il n'y avait pas nécessité de créer un nouveau TS, ce qui conduit donc à la suppression de TS_{test} et à la réutilisation d'un TS précédemment appris.

Cette méthode d'exploration permet de gérer l'arbitrage ou *trade-off* entre deux situations possibles : nécessité d'apprendre un nouveau TS, ou possibilité de réutiliser un TS déjà appris. Nous avons vu précédemment que la présence du switch dans UU permettait un apprentissage de nouveaux comportements efficace, au détriment de la réutilisation de task-sets appris, tandis que la présence de plusieurs TS parallèles dans MMBRL permettait la réutilisation de task-sets appris, au détriment de l'apprentissage efficace de nouveaux task-sets. En intégrant les deux outils de ces modèles, nous permettons au modèle proposé de se situer entre les deux extrêmes de l'arbitrage représentés par UU et MMBRL.

Par ailleurs, cette méthode d'exploration permet non seulement de décider de l'opportunité d'apprendre un nouveau task-set, mais également d'explorer la validité des TS déjà appris

de manière implicite, en parallèle, afin de pouvoir sélectionner celui qui est approprié si le nouvel apprentissage n'est pas nécessaire.

Cependant, ce trade-off est extrêmement dépendant de deux conditions initiales :
- la valeur initiale de confiance accordée au TS_{test}. Si la valeur naturelle à accorder est $1/(n+1)$, on peut biaiser en faveur de la réutilisation de TS ($\lambda_{test}(t_0)$ proche de 0) ou en faveur d'un nouvel apprentissage ($\lambda_{test}(t_0)$ grand). En effet, plus la valeur de $\lambda_{test}(t_0)$ est proche de $0,5$ plus TS_{test} risque d'émerger comme nouveau TS par défaut ($\lambda_{test} > 0,5$), même si un autre TS conviendrait. Le task-set est alors ré-appris comme un nouveau task-set plutôt que d'être simplement réutilisé. Le paramètre $\mathbf{p_{test}}$ règle ce biais : une valeur de p_{test} proche de 1 signifie un biais en faveur de la réutilisation des task-sets appris ($\lambda_{test}(t_0)$ proche de 0), tandis qu'une valeur de p_{test} proche de -1 signifie un biais en faveur de l'apprentissage des task-sets.
- la stratégie initiale de TS_{test}. Naturellement, celle-ci devrait être uniforme. Cependant, dans une logique de réutilisation de TS, il semble plus efficace que les associations déjà utilisées avec succès soient favorisées. A nouveau, le paramètre p_{test} règle le biais entre ces deux options : on initialise Q_{test} comme la moyenne pondérée entre $Q_{uniforme}$ et la moyenne des Q_i, $i = 1 \dots n$, avec p_{test} comme facteur de pondération.

5.3.4 Détails computationnels

En plus des paramètres communs aux modèles précédents (α : vitesse d'apprentissage ; β : température inverse ; ϵ : bruit) ainsi qu'au paramètre commun à UU et MMBRL (τ : stabilité des épisodes), nous rajoutons donc un paramètre critique : p_{test}, qui représente le biais initial lié à la période d'exploration.

Notons que nous pourrions également implémenter un biais de ce type dans le modèle UU. Puisqu'il n'y a pas de compétition entre TS, il est naturel d'initialiser la confiance du task-set en construction à une valeur haute, ce task-set étant la seule option, évaluée contre un comportement aléatoire. Par contre on pourrait imaginer un biais de bas niveau sur les paires stimulus-actions qui ont été historiquement le plus récompensées, ce qui se traduirait par une initialisation différente de la stratégie du nouveau TS (par l'intermédiaire de la table des *Q-values*), ainsi que dans le modèle proposé. Ce biais pourrait également

être dépendant d'un paramètre réglant l'équilibre décrit précédemment. Il n'y a donc pas fondamentalement de paramètre en plus dans notre modèle par rapport à UU.

Le paramètre $1 - \tau$ des modèles UU, MMBRL et du modèle proposé représente la vitesse de changement des épisodes et est donc dépendant de la durée de stabilité d'un task-set. C'est donc une mesure très proche de la volatilité comme proposée par Behrens et al, 2007 [17]. Elle peut donc être estimée facilement par l'observation des changements de task-sets décidés par le modèle, plutôt qu'être fixée comme paramètre. Nous avons testé les deux méthodes. La méthode d'estimation présente, logiquement, des résultats meilleurs lorsque la longueur des épisodes est variable. Cependant, par souci de simplicité, nous effectuons la suite de nos simulations dans des environnements à longueur d'épisode peu variable et nous nous contentons donc de fixer τ comme un paramètre.

Notons que le modèle proposé a toujours au minimum $n = 2$ feuillets. L'un de ces deux feuillets, que nous noterons le TS_0, représente le hasard, un comportement aléatoire. Aucun apprentissage n'a lieu pour TS_0, son modèle interne est fixé : pour tout stimulus s, action a, $\gamma_0(s, a) = P(r = 1|s, a, TS_0) = 1/n_A$, où n_A est le nombre d'actions possibles. Initialement, $n = 2$, avec (TS_0) et un TS représentant une place libre pour apprendre un premier task-set.

La présence du TS_0 semble peu importante. Cependant, elle permet crucialement l'initialisation du modèle. En effet, au début, la confiance du task-set en train d'être appris est comparée à ce que prédirait un task-set agissant au hasard. Elle assure la possibilité du premier switch lorsqu'agir au hasard est plus prédictif que continuer à utiliser le premier task-set par défaut. Il est important de noter qu'on a toujours $\sum_{i=0}^{n} \lambda_i = 1$, ce qui se traduit initialement par $\lambda_0 + \lambda_1 = 1$. En particulier, pour que λ_1 passe le seuil de switch, $0,5$, il faut que λ_0 soit devenu supérieur à $0,5$ lui même. TS_0 a donc un rôle particulier, puisqu'il ne peut pas être sélectionné comme un task-set acteur même s'il dépasse le critère. Il permet de représenter globalement ce que le modèle attend de l'environnement, comme le paramètre ξ du modèle UU.

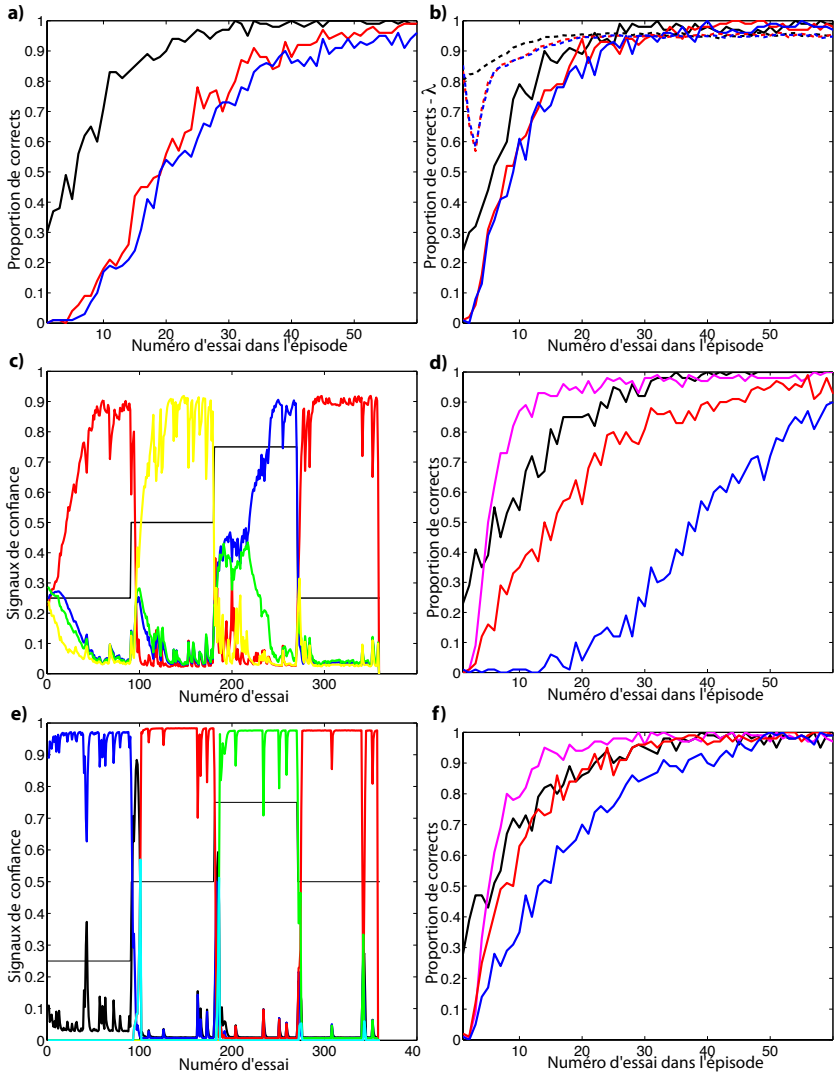


FIGURE 5.2 – Simulation des modèles. **a)** RL, **b)** UU, **c), d)** MMBRL, **e), f)** Modèle proposé. **a), b), d), f)** Proportions (sur 100 simulations) de réponses correctes en fonction du numéro d'essai après un début d'épisode. Noir : premier épisode. Bleu : épisode à nouveau task-set. Rouge : épisode à task-set répété, première répétition. Magenta : épisode à task-set répété, deuxième répétition. Voir section 5.3.5 pour plus d'information sur les figures. **c), e)** Noir en escalier : numéro du TS valide en fonction du numéro d'essai. Bleu, rouge, vert, jaune, signaux de responsabilité de quatre task-sets. Noir, signal de responsabilité de TS_0.

5.3.5 Simulations

Nous avons conduit des simulations des différents modèles proposés afin de mettre en valeur leurs différentes caractéristiques. Nous avons fixé des paramètres permettant d'optimiser les fonctions que chaque modèle prétend pouvoir exécuter. Les résultats sont présentés dans la figure 5.2, page 132.

Nous soumettons les modèles à 3, 4 ou 6 épisodes de 90 essais. Il y a $n_S = 3$ stimuli, $n_A = 4$ actions, et 2 ou 3 task-sets réels en fonction des simulations. Les task-sets sont fixés de telle sorte qu'il y ait une incongruence maximale entre eux, comme le montre la table des actions correctes en fonction des task-sets et des stimuli :

Actions	S_1	S_2	S_3
TS_1	a_1	a_2	a_3
TS_2	a_2	a_3	a_4
TS_3	a_3	a_4	a_1

La fonction de renforcement renforce les actions correspondant au task-set réel de l'épisode, avec une fiabilité de $\gamma = 90\%$. Nous testons la capacité des modèles à apprendre un task-set à l'intérieur de chaque épisode, à apprendre des nouveaux task-sets, et à éventuellement les réutiliser. Nous comparons donc plusieurs conditions : apprentissage initial du premier task-set, apprentissage dans un nouvel épisode d'un nouveau task-set, apprentissage dans un nouvel épisode d'un task-set précédemment utilisé. La mesure de performance des modèles est la vitesse d'adaptation après un changement d'épisode. Nous calculons donc, pour une série de 100 simulations, la proportion de réponses correctes pour chaque numéro d'essai après le changement d'épisode pertinent.

Pour chaque modèle, nous avons fixé des paramètres leur permettant une performance aussi bonne que possible.

Les figures **a)** et **b)** sont les simulations du modèle RL et UU, respectivement. Les paramètres communs sont $\alpha = 0,4$, $\beta = 30$, $\epsilon = 0$. La condition testant l'apprentissage d'un nouveau task-set, correspondant au dernier épisode d'une simulation à trois épisodes, $[TS_1, TS_2, TS_3]$, est tracée en bleu. La condition testant l'apprentissage d'un task-set déjà vu, correspondant au dernier épisode d'une simulation à trois épisodes, $[TS_1, TS_2, TS_1]$, est tracée en rouge. La condition testant l'apprentissage initial du premier task-set correspond

au premier épisode de ces simulations et est tracée en noir. Les lignes pleines représentent la proportion de réponses correctes, en fonction du numéro d'essai, après le début de l'épisode pertinent. Pour la figure **b)**, les lignes pointillées représentent la valeur moyenne de λ, dans chacune des conditions.

On voit que l'apprentissage initial (noir) est identique pour les deux modèles : la performance part environ au niveau du hasard (1/4) et progresse en environ 30 essais vers une performance asymptotique optimale, proche de 1. Par contre, l'apprentissage des task-sets du troisième épisode (rouge et bleu) est beaucoup plus efficace pour le modèle UU que pour le modèle RL. On voit en effet pour le modèle RL, qu'après un changement d'épisode, il faut 15 essais pour que la performance passe de 0 (persévération sur le task-set précédent) au niveau du hasard. Par contre, pour le modèle UU, après un changement d'épisode, la performance passe de 0 au niveau du hasard, en environ 5 essais. Après avoir atteint le niveau du hasard, l'apprentissage est approximativement identique entre les deux modèles, et aussi rapide que l'apprentissage initial : il faut environ une trentaine d'essais supplémentaires pour atteindre la performance asymptotique optimale (atteinte autour de l'essai 45 pour le RL, de l'essai 35 pour UU). Notez que les courbes bleue et rouge sont parallèles à la courbe noire à partir environ de l'essai 5, dans la figure **b)**, à partir environ de l'essai 15, dans la figure **a)**.

Cette adaptation plus rapide de UU que RL reflète le fait que le switch a introduit une interruption et a relancé à zéro l'apprentissage dans UU, alors que dans RL, le modèle doit désapprendre, ralentissant ainsi son adaptation. On voit d'ailleurs dans la figure **b)** pour les courbes de λ (pointillés bleu et rouge) que la confiance dans le task-set par défaut chute peu après le début de l'épisode (entre le 2e et 5e essai).Cette chute indique le moment du switch peu après le début de l'épisode et signale donc la fin de la courte période de persévération sur l'ancien task-set par défaut et le retour à la performance au niveau du hasard (1/4), suivie d'une vitesse d'apprentissage identique à celle initiale.

On voit par contre qu'il n'y a aucune différence entre les conditions rouge et bleue : aucun de ces modèles ne parvient à réutiliser la connaissance du fait que le TS_1 a déjà été appris précédemment, dans la condition rouge.

Les figures **c)** à **f)** sont les simulations du modèle MMBRL et du modèle proposé. Pour

MMBRL, on a fixé $N = 4$, pour le modèle proposé, $p_{test} = 0.9$ (soit $\lambda_{test}(t_0)$ faible).

Figures c) et e) Les modèles sont simulés pour **c)** (MMBRL) et **e)** (modèle proposé) sur quatre épisodes : $[TS1, TS2, TS3]$ suivis de la répétition de TS_1 ou TS_2. Dans ces figures, la courbe noire en escalier représente la valeur du task-set réel en fonction du numéro d'essai. On a tracé pour les différents task-sets les valeurs λ_i en fonction du numéro d'essai (rouge, jaune, vert, bleu).

Globalement, on voit que les deux modèles parviennent à identifier différents épisodes : pour chacun des quatre épisodes, un seul task-set émerge avec sa valeur de confiance λ_i haute, et la confiance dans le task-set qui a émergé chute rapidement, juste après un changement d'épisode. On voit également que les deux modèles parviennent à réactiver un task-set appris précédemment. Dans la figure **c)** pour le MMBRL, le quatrième épisode reprend le TS du premier épisode, qui a été encodé par le modèle dans le task-set rouge initalement. La confiance dans le task-set rouge croît effectivement très rapidement au début de l'épisode quatre, indiquant que le modèle reconnaît le domaine d'application du task-set qu'il a appris lors du premier épisode. Dans la figure **d)**, pour le modèle proposé, on observe de même que le task-set rouge (qui a encodé le task-set réel TS_2 du deuxième épisode), émerge également par défaut (confiance $> 0,5$ peu après le changement d'épisode) dans le quatrième épisode qui correspond au même task-set réel TS_2 que le deuxième épisode.

On voit cependant pour le MMBRL, dans la figure **c)**, que la spécialisation de chaque task-set est un processus lent : au début de chaque épisode, les λ_i des task-sets non encore utilisés sont en concurrence à, approximativement, la même valeur et tardent à se séparer, jusqu'à ce qu'un des TS finisse par émerger lentement. C'est particulièrement flagrant ici dans le troisième épisode, entre les courbes représentant les signaux de responsabilités de deux task-sets encore vierges (courbes bleue et verte). Ce processus ralentit l'apprentissage, comme on le montrera plus loin.

Par opposition, pour le modèle proposé (figure **e)**), la coupure après un épisode est très nette et permet l'émergence rapide d'un seul task-set acteur. On a également tracé, dans cette figure, la valeur λ_0 associée au TS_0 (noir ; on voit ainsi son rôle dans le premier switch) ainsi que la valeur λ_{test} associée au task-set test, lors de périodes d'exploration

(cyan). Cette valeur est non nulle pendant les périodes d'exploration suivant un switch. On voit après les premier et deuxième épisode que λ_{test} est devenu supérieur à $0,5$, assurant la création du deuxième (courbe rouge) puis du troisième task-set (courbe verte). Par contre, pendant la période d'exploration après le troisième épisode, λ_{test} reste faible et au lieu de créer un nouveau task-set, le modèle est capable d'utiliser cette période d'exploration pour faire émerger à nouveau TS_2.

Figures d) et f) Le modèle MMBRL et le modèle proposé, respectivement, sont simulés sur six épisodes : $[TS_1,\ TS_2,\ TS_3,\ TS_1,\ TS_2,\ TS_1]$. Cela permet de tester les conditions précédemment testées dans les figures **a** et **b**, plus une :

- apprentissage initial d'un task-set (noir).
- apprentissage d'un nouveau task-set : TS_3 dans le troisième épisode (bleu).
- premier épisode de réutilisation d'un task-set : TS_1 dans le quatrième épisode (rouge).
- deuxième épisode de réutilisation d'un task-set : TS_1 dans le sixième épisode (magenta). Cette condition permet de regarder à plus long terme la réutilisation des task-sets. Elle n'était pas pertinente plus haut, puisqu'il n'y a pas d'apprentissage à long terme possible.

Comme plus haut, nous traçons la proportion de réponses correctes en fonction du numéro d'essai dans l'épisode pertinent, sur 100 simulations.

MMBRL (d) On observe que l'apprentissage du task-set initial est identique à RL, UU et au modèle proposé : performance initiale autour du niveau du hasard, performance asymptotique atteinte en une trentaine d'essais. On voit par contre que l'apprentissage d'un nouveau task-set (courbe bleue) est extrêmement ralenti par rapport aux trois autres modèles : la performance est nulle pendant une quinzaine d'essais et n'atteint le niveau du hasard qu'en 30 essais environ, après lesquels la vitesse d'apprentissage se normalise (30 essais supplémentaires pour atteindre 90% de réponses correctes). Cela est dû au problème de compétition entre les deux task-sets vierges restants (déjà évoqué plus haut, figure **c)** troisième épisodes, courbes bleue et verte).

Par contre, on voit qu'à long terme, ce modèle est extrêmement efficace pour la réutilisation des task-sets (magenta). En effet, après une courte persévération (plateau à 0 initial) utile à la détermination du TS approprié, celui-ci est réactivé très rapidement : 60% de réponses

correctes dès le 5ème essai, performance asymptotique (>90%) dès le 12ème essai. Cela est dû au fait qu'aucun apprentissage n'est alors nécessaire, puisque le TS est en mémoire et qu'il suffit de l'appliquer, expliquant le fait que la courbe magenta passe au dessus de la courbe noire. Pouvoir réutiliser les task-sets réduit donc le temps d'adaptation au nouvel épisode (passage de 0 à une performance asymptotique) de 35 essais pour le modèle UU à 12 essais pour le MMBRL.

Modèle proposé (f) A nouveau, le modèle permet de mettre en valeur la distinction entre capacité de réutiliser (rouge et magenta) et nécessité d'apprendre un nouveau task-set (bleu). On voit que dès la première opportunité (rouge), mais encore plus à la deuxième (magenta), le modèle tire profit de la connaissance du task-set pour s'adapter très rapidement au nouvel épisode. En effet, malgré le temps nécessaire à décider de switcher, la performance rejoint rapidement (dès l'essai 12, courbe rouge) ou dépasse (dès l'essai 5, courbe magenta) celle de l'apprentissage initial. A la deuxième réutilisation du TS_1 (courbe magenta), une performance asymptotique est atteinte en environ 15 essais, ce qui met bien en valeur l'efficacité de réutiliser un task-set déjà appris plutôt que de le réapprendre afin d'optimiser le temps d'adaptation à un nouvel épisode.

On voit par ailleurs que cette capacité de réutiliser les task-set se fait légèrement au détriment de l'apprentissage de nouveaux task-sets. En effet, la courbe bleue de la figure **f** est en dessous de celle de la figure **b** et la performance asymptotique est atteinte au bout de 45 essais, au lieu de 35 pour le modèle UU, même si la performance à l'essai 35 est déjà supérieure à 85% de réponses correctes. Cependant, cette légère pénalisation de l'apprentissage de nouveaux task-sets n'est en aucun cas au niveau du MMBRL ou même simplement du RL. Notons par ailleurs que dans cette simulation, le paramètre de biais p_{test} a été fixé en faveur de la réutilisation de task-sets appris. Fixer ce paramètre à $p_{test} = 0,01$, en faveur de l'apprentissage, conduit à des courbes d'adaptation superposables à la figure **b** : aucun avantage à la condition d'apparition d'un TS connu plutôt que d'un nouveau TS dans un nouvel épisode, mais apprentissage très rapide après un switch efficace.

5.3.6 Conclusion

Nous avons donc proposé un modèle qui repose sur trois points cruciaux :

- Un mécanisme simple d'apprentissage par renforcement des stratégies, ainsi que l'apprentissage de modèles internes prédictifs, permettant l'évaluation de la validité de chaque task-set.
- Un mécanisme de switch permettant de passer efficacement de l'exploitation d'un task-set par défaut à l'exploration des options disponibles : la création d'un nouveau task-set ou la réutilisation d'un task-set précédemment appris.
- Un mécanisme d'évaluation du degré de validité des différentes options, à la base de l'exploration efficace des options disponibles.

La combinaison de ces trois points cruciaux permet au modèle

- de construire un répertoire de task-set (et d'apprendre leur stratégie et leurs modèles de prédictions) en fonction de ses besoins, ni plus, ni moins.
- de switcher rapidement quand nécessaire, à la détection d'un changement d'épisode, de manière à la fois flexible et robuste.
- de trancher entre la nécessité de créer un nouveau task-set ou d'en réutiliser un, et dans ce dernier cas, d'explorer implicitement et efficacement les task-sets déjà appris afin d'en sélectionner un rapidement.

5.4 Contrôle contextuel, a priori, a posteriori

Nous avons présenté jusqu'à maintenant le modèle proposé pour faire interagir le contrôle cognitif et l'apprentissage, seulement en présence d'information a posteriori : les renforcements. Cela implique que l'acquisition de task-sets ainsi que le switch sont possibles, mais seulement en réaction à des erreurs, après coup. Or, un aspect important du contrôle cognitif, le contrôle contextuel, consiste à être capable de switcher a priori en fonction du contexte dans lequel on se trouve : il serait dangereux d'attendre la première erreur pour se rendre compte, lorsqu'on arrive de France en Grande-Bretagne, que ce n'est plus à gauche, mais à droite, qu'il faut regarder avant de traverser la route.

Nous présentons donc maintenant l'extension du modèle déjà présenté, qui tient compte en

plus de l'information contextuelle présente et qui permet l'apprentissage des associations entre contexte et task-sets et l'utilisation de ces associations pour permettre un switch a priori.

5.4.1 Description du modèle

On peut représenter la chaîne d'observations, de décisions et d'évaluations du modèle comme suit :

$$\begin{array}{ccccccccccc} \cdots \rightarrow (C_t,\ s_t) & \longrightarrow & a_t & \longrightarrow & r_t & \longrightarrow & (C_{t+1},\ s_{t+1}) & \longrightarrow & \cdots \\ & \uparrow & & \uparrow & & \uparrow & & \uparrow & \\ & \lambda_i(t) & & TS_t^* & & \mu_i(t) & & \lambda_i(t+1) & \end{array}$$

Après observation d'une entrée sensorielle (le stimulus seul précédemment, joint ici à un contexte), le modèle utilise la probabilité a priori (**confiance ex-ante**) qu'un TS_i soit pertinent, $\lambda_i(t)$, pour sélectionner le task-set à utiliser $TS(t)$, et ainsi décider d'une action a_t. L'environnement, par l'intermédiaire du task-set actuellement correct, TS_t^*, produit un renforcement r_t. Le modèle utilise alors ce renforcement pour évaluer, a posteriori, la validité de son choix de TS précédent (ainsi que des choix qui n'ont pas été effectués), sous la forme de la probabilité que $TS_t^* = i$ sachant ce qui a été observé ; cette probabilité est notée $\mu_i(t)$ (**confiance ex-post**). En l'absence de contexte, le passage à l'essai suivant n'apporte pas d'information, et la probabilité a priori que TS_i soit pertinent pour le prochain choix ($\lambda_i(t+1)$) est simplement $\mu_i(t)$ pondéré par la possibilité que TS^* ait changé. En présence de contexte par contre, entre l'évaluation du choix précédent et la décision suivante, l'information portée par le contexte C_{t+1} est disponible, ce qui modifie la mise à jour de la confiance ex-ante, $\lambda_i(t+1)$.

La modification apportée au modèle dans le cadre contextuel permet au modèle de tenir compte de cette différence pour intégrer l'information contextuelle dans la confiance ex-ante. Le reste du modèle reste similaire, notamment l'utilisation des λ_i pour déterminer des **périodes d'exploitation** lorsqu'un task-set par défaut existe et des périodes d'exploration avec un $\mathbf{TS_{test}}$ acteur.

On a à nouveau $\lambda_i(t)$ qui sert de probabilité a priori pour la mise à jour Bayésienne de $\mu_i(t)$

grâce à l'observation du renforcement, et $\mu_i(t)$ qui sert de probabilité a priori pour la mise à jour Bayésienne de $\lambda_i(t+1)$ grâce à l'observation du contexte. La première mise à jour ne change donc pas : on a toujours

$$\begin{aligned}\mu_i(t) = P(TS_t^* = TS_i|r_t) &= \frac{P(r_t|TS_t^* = TS_i)P(TS_t^* = TS_i)}{\sum_{j=0}^{n_TS} P(r_t|TS_t^* = TS_j)P(TS_t^* = TS_j)} \\ &= \frac{\gamma_i(r_t, s_t, a_t)\lambda_i(t)}{\sum_{j=0}^{n_TS} \gamma_j(r_t, s_t, a_t)\lambda_j(t)}\end{aligned}$$

Par contre, la mise à jour de $\lambda_i(t+1)$ tient compte de l'information contextuelle :

$$\lambda_i(t+1) = \sum_{j=0}^{n_{TS}} P(TS_{t+1}^* = i|TS_t^* = j, C_{t+1})P(TS_t^* = j|r_t) \tag{5.4}$$

$$= \sum_{j=0}^{n_{TS}} \mathrm{Tr}(j, i, C_{t+1})\mu_j(t) \tag{5.5}$$

On note $\mathrm{Tr}(j, i, C_{t+1})$ ici la probabilité de transition de TS^* de TS_j à t TS_i à $t+1$ sachant que le contexte à $t+1$ est C_{t+1}, indépendamment de la récompense à t. En supposant que C_t est une observation dépendante uniquement de l'état caché TS_t^*, on peut alors calculer cette probabilité de transition par inférence Bayésienne sous la forme suivante :

$$P(TS_{t+1}^* = i|TS_t^* = j, C_{t+1}) = \frac{P(TS_{t+1}^* = i|C_{t+1})\tau_{(j,i)}}{\sum_k P(TS_{t+1}^* = k|C_{t+1})\tau_{(j,k)}} \tag{5.6}$$

Les valeurs $\tau_{(j,k)}$ représentent simplement les probabilités de transition inconditionnées entre task-sets. On peut soit, comme précédemment, fixer un paramètre τ tel que $\tau_{(j,j)} = \tau$ pour tout j et $\tau_{(j,k)} = (1-\tau)/(n_{TS}-1)$ si $j \neq k$, soit estimer ces valeurs comme reliées à une volatilité.

Les valeurs $P(TS^* = i|C)$ sont par contre extrêmement importantes : elles représentent les valeurs d'association entre contexte et task-set. Nous proposons de noter ces valeurs $Q(C, i)$ et que celles-ci soient apprises par renforcement par le modèle, avec la **confiance ex-post μ_i comme valeur de renforcement**. Cette proposition se base sur plusieurs arguments théoriques :

– Il existe un parallèle hiérarchique entre la sélection d'actions en réponse à un stimulus (contrôle sensoriel) et la sélection d'un task-set en réponse à un contexte (contrôle contextuel). On peut donc proposer que, parallèlement aux valeurs d'états-actions $Q(s, a)$, on ait des valeurs de contextes-TS, $Q(C, TS)$.

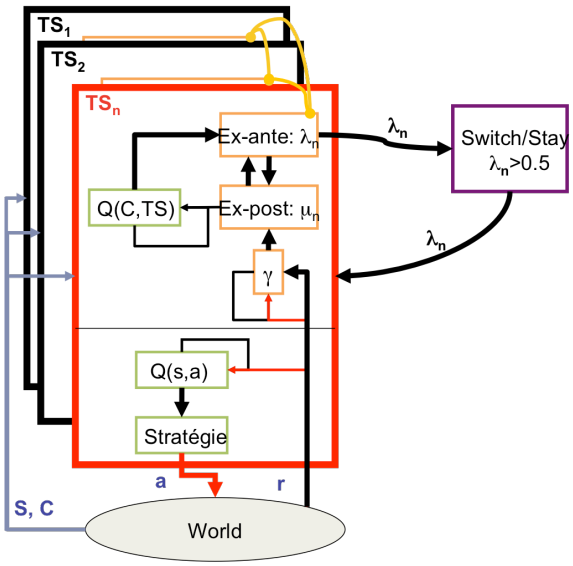


FIGURE 5.3 – Modèle proposé. Rouge : le task-set par défaut est le seul à utiliser le renforcement r pour mettre à jour ses *Q-values* et son modèle interne γ, en fonction de l'action a qu'il a choisie face au stimulus s. La confiance ex-post μ et ex-ante λ est calculée pour tous les task-sets. μ est utilisé pour tous les task-sets, pour mettre à jour leur valeur d'association avec le contexte présent, C. Celle-ci sert à calculer le prochain λ, qui décide de continuer une période d'exploitation ($\lambda_d > 0,5$), de switcher ($\lambda_d < 0,5$) ou de maintenir une période d'exploration (tous $\lambda_i < 0,5$).

– Dans un contexte $C_t = C$, $\mu_i(t)$ est une évaluation de la probabilité $P(TS^* = TS_i | C_t = C)$, sachant le renforcement obtenu. Il semble donc naturel de renforcer $Q(C, i)$, voué à estimer $P(TS^* = i|C)$ par $\mu_i(t)$. On renforce ainsi les associations contexte-TS par la confiance qu'on a dans la validité d'utiliser ce TS dans ce contexte ; cela signifie qu'on tente d'estimer le degré de validité d'un task-set dans un contexte donné, par l'estimation du degré de validité de ce task-set dans le contexte présent, en fonction de l'historique des actions et des renforcements.

Le modèle prenant en compte l'information contextuelle ne diffère donc du modèle décrit sans contexte que dans la mise à jour de $\lambda_i(t)$. En plus de l'apprentissage des stratégie et modèle interne du task-set par défaut ou du task-set test, après chaque renforcement, il est

nécessaire de mettre à jour les assoctions contextes-TS : pour tout i,

$$Q_{t+1}(C_t, TS_i) = Q_t(C_t, TS_i) + \alpha_C(\mu_i(t) - Q_t(C_t, TS_i)).$$

Ces valeurs sont ensuite utilisées pour calculer $\lambda_i(t+1)$, afin d'effectuer les décisions de switch, exploration, exploitation comme décrit précédemment. Notons l'apparition d'un nouveau paramètre essentiel, α_C, la vitesse d'apprentissage des associations contexte - task-set.

5.4.2 Contextes

Un point délicat de la notion de contrôle contextuel est celui de la définition du contexte. En effet, si on peut dire des contextes qu'à l'instar des simuli, ce sont des entrées sensorielles présentes, on ne peut pas dire ce qui les différencie spécifiquement des stimuli. Typiquement, dans les expériences de contrôle cognitif proposées par Koechlin et al, les entrées sensorielles sont généralement bi-dimensionnelles : des lettres de couleur. La couleur, qui joue le rôle de contexte, n'est pas a priori fondamentalement différente de la lettre, qui joue le rôle de stimulus.

Notre modèle permet de proposer une définition fonctionnelle de la notion de contexte. Nous argumentons dans cette section qu'un contexte peut être défini comme une partie de l'entrée sensorielle présente qui apporte de l'information pour la sélection d'un task-set.

Nous partons de l'hypothèse que le modèle Bayésien ne sait pas, a priori, extraire la dimension portant le contexte de l'entrée sensorielle. Il ne peut donc pas spécifiquement renforcer des associations $Q(C, TS)$, puisqu'il ne sait pas identifier C. Il peut, par contre, renforcer les associations entre les entrées sensorielles et les task-sets. Les entrées sensorielles sont constituées d'au moins trois types de signaux distincts, du point de vue du monde :

- les stimuli (s_t), pris dans l'ensemble $\{s_i\}_{i=1...n_S}$.
- les contextes (C_t), pris dans l'ensemble $\{C_i\}_{i=1...n_C}$.
- des distracteurs (d_t, tous les signaux présents non pertinents pour le problème), pris dans l'ensemble $\{d_i\}_{i=1...n_D}$.

On aura donc, après chaque essai, la possibilité de renforcer pour tout TS_i, $i = 1, \ldots, n$, les valeurs $Q_i([s_t, C_t, d_t], TS_i)$ par la confiance ex-post, $\mu_i(t)$. Ces valeurs sont alors utilisées

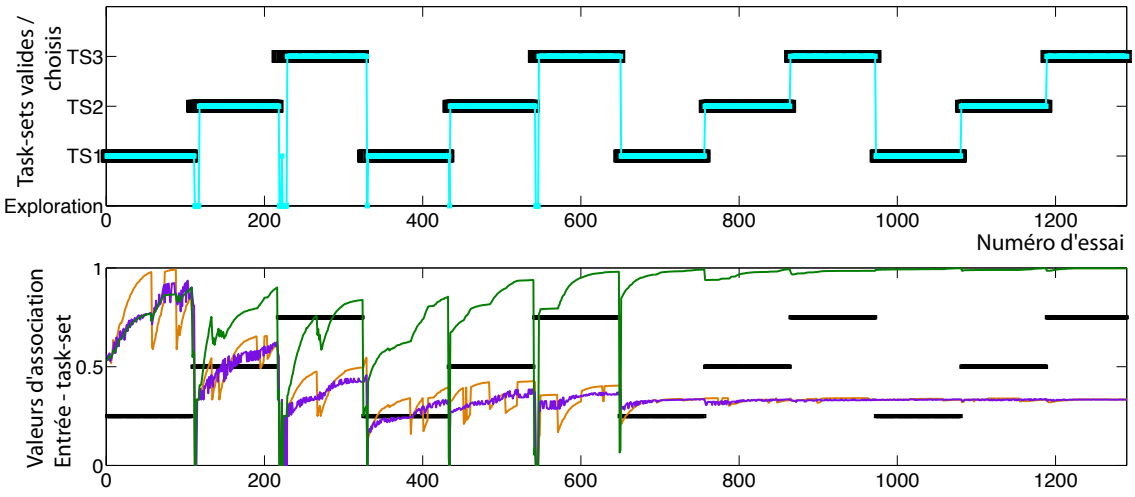


FIGURE 5.4 – Simulation du modèle, extraction d'information contextuelle. **Haut.** En noir, le TS^* valide (associé à un contexte par TS^*). En cyan, le TS utilisé par le modèle. **Bas** Courbes de valeur moyenne d'association entre une dimension d'entrée et le TS utilisé par le modèle. Vert : dimension contextes. Violet : dimension stimuli. Orange : dimension distracteurs. Mêmes paramètres que dans la figure 5.2, $\alpha_C = 0,2$. On voit que les valeurs d'association des contextes au TS utilisé tendent vers 1, tandis que celles des distracteurs et des stimuli tendent vers 1/(nombre de TS).

pour calculer la confiance ex-ante, λ_i.

Pour une valeur d'un stimulus ou d'un distracteur donné, tous les task-sets peuvent avoir soit une haute valeur de confiance, soit une faible valeur de confiance à un moment donné, en fonction de l'épisode dans lequel ce stimulus ou distracteur est présenté. Ainsi, si on regarde les valeurs d'association moyennes selon la dimension objective *stimulus* ou *distracteur*, celles-ci devraient tendre à s'uniformiser vers le niveau du hasard.

Par contre, pendant un apprentissage efficace, pour un contexte donné, le modèle ne devrait avoir une haute valeur de confiance essentiellement que pour un task-set, celui correspondant au TS* associé à ce contexte. Ainsi, si on regarde les valeurs d'association moyennes selon la dimension objective *contexte*, celles-ci devraient tendre à une valeur haute uniquement pour les paires contexte - task-set pertinentes.

La simulation représentée dans la figure 5.4 montre que le modèle est en effet capable

d'effectuer l'extraction des contextes. On trace, pour tout essai à $[s_t, C_t, d_t]$, avec utilisation de TS_t par le modèle, les valeurs suivantes :

- $Q(s_t, TS_t)$ défini par $Q(s_t, TS_t) = \sum_{i=1}^{n_C} \sum_{j=1}^{n_D} Q([s_t, C_i, d_j], TS_t)/(n_C * n_D)$. Cette valeur représente la valeur moyenne d'association stimulus - TS pour le stimulus et le task-set au temps t.
- $Q(C_t, TS_t)$ défini par $Q(C_t, TS_t) = \sum_{i=1}^{n_S} \sum_{j=1}^{n_D} Q([s_i, C_t, d_j], TS_t)/(n_S * n_D)$. Cette valeur représente la valeur moyenne d'association contexte - TS pour le contexte et le task-set au temps t.
- $Q(d_t, TS_t)$ défini par $Q(d_t, TS_t) = \sum_{i=1}^{n_S} \sum_{j=1}^{n_C} Q([s_i, C_j, d_t], TS_t)/(n_S * n_C)$. Cette valeur représente la valeur moyenne d'association distracteurs - TS pour le bruit d'entrée et le task-set au temps t.

Ces valeurs représentent l'information spécifique apportée par chacune des dimensions d'entrée à la sélection a priori de task-sets. On voit qu'initialement, pendant le premier épisode, les informations portées par chaque dimension sont identiques, un seul task-set ayant été utilisé. Dès l'apparition des deuxième et troisième task-set par contre, les valeurs d'associations des stimuli et distracteurs se positionnent autour du hasard (1/2 car deux task-sets au deuxième épisode, 1/3 car trois task-sets existants au troisième épisode), alors que les valeurs d'associations des contextes se démarquent. Au fur et à mesure de l'apprentissage, les valeurs tendent à refléter le fait que les stimuli et les distracteurs sont autant associés à chaque TS ($Q \to 1/3$) et donc ne sont pas prédictifs du task-set à utiliser. Par contre, pour les contextes, les valeurs d'associations tendent à refléter la corrélation parfaite entre task-sets et contextes : $Q \to 1$.

Le modèle parvient donc bien à extraire spécifiquement de l'entrée la dimension contextuelle qui apporte de l'information sur la sélection de la tâche. On voit par ailleurs, dans la figure du haut, que le modèle apprend à utiliser cette information a priori et devient capable de switcher au premier essai présentant un nouveau contexte : aux changements d'épisode, il ne passe plus par une période d'exploration après un switch, mais passe directement d'un task-set à l'autre. Il est donc capable, après apprentissage, d'effectuer du contrôle contextuel au sens de Koechlin et collègues : sélectionner la tâche appropriée, à chaque essai, en fonction du contexte changeant.

Nous avons donc montré que nous pouvions définir la notion de contexte comme une entrée

sensorielle informative quant à la sélection d'un task-set et que notre modèle était capable d'apprendre indépendamment à extraire les contextes.

Cependant, par la suite, par souci de simplicité, nous simulons notre modèle en supposant la dimension contextuelle connue, comme proposé avant cette section. En effet, l'ajout des stimuli et du bruit augmente la taille de la table d'associations contextes-TS et ralentit l'apprentissage de ces valeurs.

5.4.3 Détails computationnels

Lorsque l'on bloque l'apprentissage des associations contextes - task-set, en posant par exemple $\alpha_C = 0$, ou une valeur très faible, on a des valeurs $Q(C, TS)$ uniformes à tout temps. Il est alors aisé de voir que les formules de mise à jour de $\lambda_i(t+1)$ à partir de $\mu_i(t)$ du modèle avec contextes se réduisent à celles du modèle sans contextes.

Lorsque α_C est non nul, les valeurs d'associations des contextes - task-sets sont apprises. Notons qu'à l'apparition d'un nouveau contexte, la table $Q(C, TS)$ doit être initialisée pour ce contexte. Celle-ci représentant l'estimation d'une probabilité sur l'espace des task-sets, nous initialisons ces valeurs à $1/n$, où n est le nombre actuel de task-sets du modèle. On a ainsi pour tout $i = 1, \ldots, n$, les valeurs identiques $Q(C, TS_i) = 1/n$.

Il est essentiel de remarquer que lorsque ces valeurs sont uniformes, aucune information n'est apportée. La mise à jour $\lambda_i(t+1)$ à partir de $\mu_i(t)$ est alors équivalente à celle effectuée par le modèle en l'absence de contextes. Cela signifie qu'en l'absence d'information contextuelle, notamment à l'arrivée d'un nouveau contexte, seule l'information portée par le renforcement est utilisée, a posteriori, pour switcher. Par défaut, on continue d'utiliser le task-set acteur.

Pourtant, on pourrait argumenter qu'un changement de contexte est une observation en soi, qui pourrait inciter à changer de comportement. Les neurosciences montrent en effet que l'arrivée d'un changement dans le champ sensoriel peut provoquer un déplacement de l'attention et une interruption du comportement par défaut. Nous proposons donc d'introduire l'observation ΔC_t (valant 1 pour $C_{t-1} \neq C_t$, 0 sinon) comme source d'information supplémentaire à l'inférence de la variable cachée TS^*. Celle-ci intervient alors exclusive-

ment dans la mise à jour de $\lambda_i(t+1)$ au niveau du calcul des transitions entre task-sets. On modifie donc les équations 5.4 et 5.6 pour tenir compte de ΔC comme suit :

$$\begin{aligned} \lambda_i(t+1) &= \sum_{j=0}^{n_{TS}} \mathrm{Tr}(j,i,C_{t+1},\Delta C_{t+1})\mu_j(t) \\ \mathrm{Tr}(j,i,C_{t+1},\Delta C_{t+1}) &= P(TS^*_{t+1}=i|TS^*_t=j,C_{t+1}) = \frac{Q(C_{t+1},i)\tau_{(j,i)}(\Delta C_{t+1})}{\sum_k Q(C_{t+1},k)\tau_{(j,k)}(\Delta C_{t+1})} \end{aligned}$$

Ainsi, les valeurs $\tau_{(j,k)}(\Delta C_{t+1})$ représentent les probabilités de transition entre task-sets conditionnées simplement par l'observation ou non d'un changement de contexte. A nouveau, plusieurs degrés d'approximation sont possibles pour estimer ces valeurs. Celles-ci pourrait être apprises. Nous proposons la solution suivante, simple mais approximative : lorsque $\Delta C = 0$, $\tau_{(j,k)}(\Delta C) = \tau_{(j,k)}$; sinon $\tau_{(j,k)}(\Delta C) = \tau_\Delta \tau_{(j,k)}$. Ici, $\tau_{(j,k)}$ est estimé comme précédemment et τ_Δ est un paramètre représentant l'influence de l'observation d'un changement de contexte sur la probabilité de changement de task-set.

Lorsque le paramètre τ_Δ est différent de 1, on donne donc ainsi un rôle spécifique et ponctuel aux changements de contextes, indépendamment du rôle des contextes, qui permet une moins faible probabilité de switch à l'observation d'un changement de contextes. Nous verrons plus loin que cela semble pertinent en rapport au comportement des sujets.

5.4.4 Autres possibilités, simulations

UU et MMBRL

On pourrait introduire dans les modèles UU et MMBRL des adaptations similaires à celles effectuées pour notre modèle afin que ceux-ci prennent également en compte les contextes. Nous ne détaillerons pas ici ces modifications et signalons seulement les conséquences de ces adaptations sur ces modèles :

– Avantages pour UU : l'introduction d'information contextuelle dans le modèle UU permet de proposer une méthode pour pouvoir, après un switch, utiliser à nouveau des task-sets déjà appris dans un contexte déjà vu, en utilisant le valeurs d'associations $Q(C,TS)$ dans un softmax. L'apprentissage de nouveaux task-sets serait donc limité aux contextes nouveaux.

– Inconvénient pour UU : un task-set ne peut être sélectionné à nouveau que s'il apparaît dans le cadre d'un contexte déjà connu. Pourtant, nous somme capables d'utiliser un comportement appris dans un contexte nouveau, si celui-ci nous semble approprié.
– Inconvénients pour MMBRL : l'introduction des associations contextes - task-sets ralentit énormément l'apprentissage des TS par le MMBRL. En effet, lors de l'apprentissage initial, non seulement tous les task-sets vierges sont aussi prédictifs de ce qui est observé, compliquant leur séparation, mais le contexte est lié aussi fortement à chaque task-set jusqu'à ce qu'une séparation des rôles renforce plus une association que l'autre. Une inertie supplémentaire est donc rajoutée, rendant l'apprentissage laborieux, même si après un très fort entraînement, la performance devient optimale.

RL et contextes

Un modèle extrêmement simple et naturel pour traiter le problème de l'apprentissage de multiples TS en présence de contextes serait un modèle basique d'apprentissage par renforcement, prenant pour espace d'état non plus seulement l'espace des stimuli, mais plus globalement l'espace des entrées sensorielles, comprenant au moins les deux dimensions contexte et stimulus. Ainsi posé, ce modèle est capable d'apprendre des task-sets et de ne pas les oublier, puisqu'ils sont appris dans le cadre de leur contexte et qu'on a fait l'hypothèse qu'un seul TS est possible dans un contexte donné. Cependant, ce modèle pèche pour cause de la réciproque : plusieurs contextes peuvent être liés à un même task-set. Or, à l'apparition d'un nouveau contexte, ce modèle ne reconnait pas le même stimulus, mais voit un état différent. Il doit donc apprendre un nouveau task-set et ne peut réutiliser celui appris.

Simulations du modèle proposé et du RL.

Nous effectuons une série de 100 simulations de 7 épisodes de 90 essais sur le modèle RL et le modèle proposé afin de démontrer les effets spécifiques de notre modèle. A nouveau, on trace la proportion de réponses correctes pour chaque numéro d'essai après un changement d'épisode, ou de contexte dans un épisode. A nouveau, pour ces simulations, nous avons

fixé des paramètres permettant aux modèles une performance aussi bonne que possible.Les paramètres utilisés sont :

– paramètres communs : $\alpha = 0,4$, $\beta = 30$, $\epsilon = 0$
– paramètres spécifiques de notre modèle : $p_{test} = 0,9$, $\alpha_C = 0,1$. On ne tient pas compte ici de l'observation spécifique des changements de contexte.

La série d'épisodes et de contextes est

$$\begin{array}{cccccccc} TS_1 & TS_1 & TS_2 & TS_3 & TS_2 & TS_3 & TS_2 & TS_3 \\ C_1 & C'_1 & C_2 & C_3 & C_2 & C_3 & C_2 & C'_3 \end{array}$$

(les deux premiers épisodes n'en forment en fait qu'un de 2×45 essais). Nous pouvons ainsi tester les conditions suivantes :

– apprentissage initial : épisode 1, TS_1, C_1 (courbe noire).
– pas de changement de task-set, mais changement de contexte : épisode 2, TS_1, C'_1 (courbe bleue).
– nouveau task-set, nouveau contexte : épisode 3, TS_2, C_2 (courbe cyan).
– task-set, contexte déjà valides une fois : épisode 5, TS_2, C_2 (courbe rouge).
– task-set, contexte déjà valides deux fois : épisode 7, TS_2, C_2 (courbe magenta).
– nouveau contexte, task-set déjà valide une fois : épisode 8, TS_3, C'_3 (courbe verte).

On observe dans la figure 5.4.4, gauche, que pour le modèle RL, les comportements se divisent strictement en deux catégories. La première concerne les épisodes où le contexte a déjà été vu (rouge, magenta) ; le comportement est alors logiquement parfait : performance proche de 100% du début à la fin de l'épisode. La deuxième catégorie concerne les épisodes pour lesquels le contexte n'a jamais été vu (bleu, cyan, vert) ; le comportement est alors strictement identique à l'apprentissage initial du premier task-set : performance initiale au niveau du hasard (1/4), performance asymptotique atteinte en environ 30 essais.

La figure 5.4.4, droite, montre les mêmes courbes pour le modèle proposé. On observe cette fois un comportement plus complexe.

Tout d'abord, on voit sur la courbe bleue qu'après un changement de contexte mais non de task-set, le modèle est immédiatement à performance optimale (proche de 100%). En effet, par défaut, il continue d'utiliser le task-set présent jusqu'à une nécessité de switcher. Au contraire, le modèle RL devait réapprendre depuis le niveau du hasard dans cette

condition.

On voit également que le modèle est capable de réutiliser un task-set dans un nouveau contexte (courbe verte). En effet, dans cette condition, seuls 15 essais sont nécessaires pour passer d'une performance nulle (persévération sur le précédent contexte en l'absence d'informations sur la nouvelle couleur) à une performance asymptotique optimale (plus de 90% de réponses correctes), ce qui est bien plus rapide que le passage d'une performance aléatoire (1/4) à une performance asymptotique optimale en 30 essais observée lors de l'apprentissage du premier task-set (courbe noire).

Remarquons par contre que l'apprentissage d'un nouveau task-set dans un nouveau contexte (courbe cyan) est lui légèrement pénalisé par le modèle. En effet, le modèle doit détecter la nécessité d'un switch avant de pouvoir commencer à apprendre le nouveau modèle : il a ainsi besoin d'environ 5 essais pour passer d'une performance nulle à une performance au niveau du hasard. Cela n'est pas le cas du RL qui commence à apprendre dès que le nouveau contexte est présent, puisque la paire contexte-stimulus est vue comme un nouvel état. Le déficit est cependant léger, puisque notre modèle atteint malgré tout une performance asymptotique optimale en environ 30 essais.

Enfin, notons que, si au premier retour d'un épisode contexte - task-set connus (courbe rouge), la performance du modèle n'est pas aussi bonne que celle du RL (qui est optimale), elle est néanmoins meilleure que dans le cas task-set connu, nouveau contexte. En effet, on atteint une performance optimale en environ 8 essais dans le cas du contexte connu (rouge), contre 15 essais dans le cas du nouveau contexte (vert). Cela montre déjà non seulement une réactivation du task-set mais également une influence du contexte dans sa réactivation. Par ailleurs, dès le deuxième retour d'un contexte - task-set (courbe magenta), on observe une performance optimale dès le premier essai, ce qui prouve que le modèle fait du contrôle a priori et non seulement a posteriori.

5.4.5 Conclusion du modèle contextuel.

Nous avons donc montré que nous pouvions étendre notre modèle pour inclure l'information contextuelle a priori en plus de l'information du renforcement a posteriori. Cela se fait en

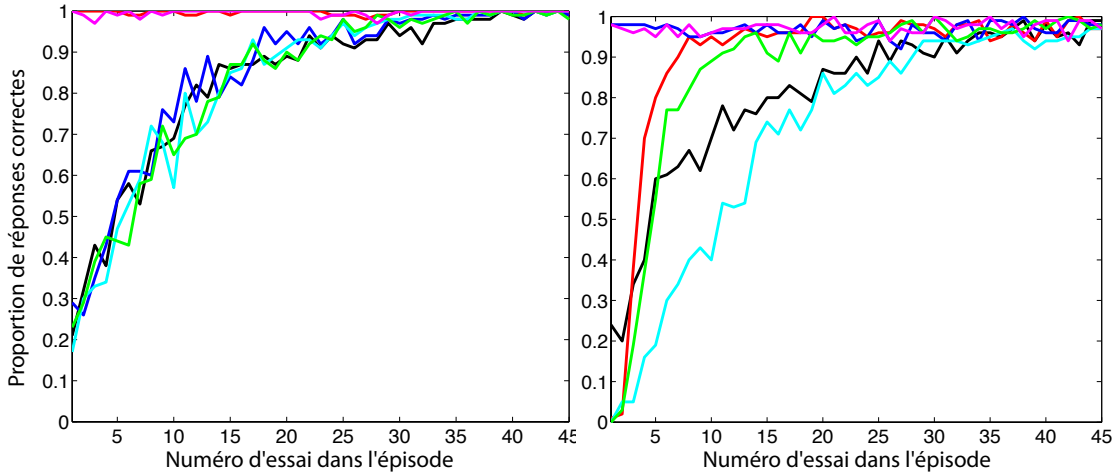


FIGURE 5.5 – Apprentissage de task-sets en présence de contextes. Proportions de réponses correctes (100 simulations) en fonction du numéro d'essai après le changement d'épisode. Gauche : RL. Droite : modèle proposé. Noir : premier épisode. Rouge : épisode à contexte et task-sets répétés, première répétition. Magenta : épisode à contexte et task-sets répétés, deuxième répétition. Cyan : épisode à nouveau contexte et task-set. Bleu : épisode où le contexte a changé, mais pas le task-set. Vert : épisode à contexte nouveau, mais à task-set répété. Voir section 5.4.4 pour plus de détails.

dissociant la confiance ex-post de la confiance ex-ante, en utilisant la confiance ex-post pour l'apprentissage de valeurs d'associations contextes - task-sets et en utilisant ces dernières pour le calcul de la confiance ex-ante, laquelle est utilisée pour les décisions d'action, de switch, d'exploration ou de création de nouveaux task-sets.

Nous avons montré qu'ainsi construit, notre modèle est capable d'apprendre plusieurs task-sets, d'apprendre à les associer à des contextes, à les réutiliser a priori en fonction des contextes ou a posteriori dans des nouveaux contextes. Après apprentissage, la réutilisation a priori peut se faire extrêmement efficacement en un seul essai, ce qui est un critère essentiel puisque requis par le contrôle cognitif. Le modèle inclut ainsi toutes les étapes de l'apprentissage initial de task-sets, au contrôle contextuel impliquant à chaque essai la sélection d'une tâche en fonction d'un contexte, a priori.

5.5 Conclusions

5.5.1 Deux limitations du modèle

Schéma de renforcement Nous nous sommes limités à un schéma de renforcement extrêmement basique, binaire. Cependant, il est fréquent que, plutôt qu'un simple *gagné* ou *perdu*, le retour de l'environnement soit plus continu et complexe (nombre de points, argent, quantité de plaisir, etc.). Cela n'est pas une limitation théorique de notre modèle : en effet, ce qui est important est que le sujet puisse prédire l'observation, c'est-à-dire estimer la probabilité d'observer le retour de l'environnement, afin de pouvoir estimer la confiance qu'il a dans son modèle. Il doit également pouvoir apprendre son modèle interne de prédiction des effets de l'environnement. Cet apprentissage (par fréquence) et cette prédiction sont particulièrement simples dans le cas binaire incertain, ce qui a motivé notre choix. Cependant, cela pourrait également être effectué dans un espace de renforcements continus avec plus de complications techniques.

Structure temporelle Nous nous sommes placés dans un cadre où le sujet n'interagit avec l'environnement que par l'intermédiaire des renforcements obtenus. Cependant, dans

de nombreux cas dans la vie de tous les jours, l'action sélectionnée a non seulement des conséquences sur ce qu'on en obtient (récompense, punition), mais également sur le nouvel état dans lequel on est. A nouveau, ce point n'est pas une limitation théorique de notre modèle. En effet, la prédiction des conséquences sur l'espace des états des actions dans le cadre des task-sets pourrait faire partie du modèle interne de prédiction des task-sets et être ainsi intégrée à la mesure de confiance, donc aux décisions de contrôle cognitif. C'est d'ailleurs exactement ce qui est fait dans les modèles MMBRL de contrôle moteur ([113]). A nouveau, nous avons ignoré cela afin de simplifier les composantes du modèle, et de se concentrer sur ses aspects critiques.

Niveau épisodique Pour l'intégration de l'information contextuelle, nous avons supposé que les associations contextes - task-sets étaient bien un mapping, un contexte ne pouvant pas être associé à plus d'un task-set. Notre modèle n'inclut donc pas la notion de contrôle épisodique, dans laquelle une instruction passée détermine ce mapping, permettant de passer entre deux mappings différents, associant en particulier un contexte à deux task-sets différents en fonction de l'indice épisodique. C'est une limitation théorique de notre modèle. Nous parlerons plus en détail de cette limitation dans la discussion de la thèse.

5.5.2 Conclusion, prédictions

Afin d'intégrer l'apprentissage et le contrôle cognitif, nous avons construit un modèle basé à la fois sur les techniques d'apprentissage par renforcement et celles d'inférence bayésienne d'un état caché, le TS^* actuellement valide. Notre modèle repose essentiellement sur l'hypothèse que nous avons un task-set par défaut. Cette hypothèse suppose que la sélection d'un task-set a un coût et que nous ne l'effectuons pas à chaque essai, mais plutôt que, a priori, nous continuons à utiliser un task-set qui semble valide et n'effectuons un switch que lorsque cela est nécessaire.

Notre modèle repose sur le fait que nous avons un **répertoire de task-sets**, pour lesquels nous pouvons à tout instant estimer la **confiance** que nous avons dans le fait qu'ils soient valides. Pendant une période d'**exploitation**, un task-set semble plus valide que l'ensemble des autres et il est donc singularisé : c'est le task-set par défaut, le seul activé en termes

d'apprentissage et de sélection des actions. Un **switch** a lieu lorsque le task-set par défaut ne semble plus être plus pertinent qu'incorrect. En l'absence de task-set par défaut, une période d'**exploration** a lieu. Puisque aucun task-set du répertoire ne peut être singularisé comme meilleur que tous les autres, on utilise un nouveau task-set test comme seul actif en terme d'apprentissage et de sélection des actions. Cela permet une exploration diffuse et parallèle de la confiance dans les task-sets du répertoire et permet soit la réutilisation comme task-set par défaut d'un de ces task-sets, soit l'**agrandissement du répertoire** par l'ajout du task-set test appris au répertoire. On arbitre ainsi la nécessité d'apprendre un nouveau comportement, ou d'en réutiliser un précédemment appris. Enfin, on sépare le rôle de l'évaluation **ex-post** de la confiance comme signal de renforcement des associations contexte-TS et le rôle de l'évaluation **ex-ante** de la confiance dans un task-set comme facteur de contrôle cognitif.

Le modèle ainsi construit est capable d'apprendre un répertoire de task-sets en fonction des besoins, éventuellement d'extraire une information contextuelle présente, d'apprendre son association avec les task-sets ainsi que de réutiliser ces task-sets de manière pertinente. Il se distingue ainsi des autres modèles existants, qui manquent de flexibilité sur la construction du répertoire de task-sets (MMBRL), ne permettent pas la réutilisation de task-sets appris (UU) ou ne permettent pas l'apprentissage d'un répertoire de task-sets (RL). Il permet donc d'effectuer des prédictions spécifiques sur le comportement de sujets, qui seront largement détaillées dans le prochain chapitre.

Le modèle proposé permet bien de faire interagir l'apprentissage et le contrôle cognitif dans les deux directions dans lesquelles il interagit dans le comportement humain. L'apprentissage est utile au contrôle, puisqu'il permet la formation du répertoire de task-sets et éventuellement de leur association aux contextes, deux éléments indispensables au contrôle cognitif. Réciproquement, le contrôle cognitif est utile à l'apprentissage, puisqu'il permet de prendre les décisions de switch, d'exploration ou de sélection nécessaires à réguler l'apprentissage.

Troisième partie

Expériences comportementales

Nous avons exposé une théorie computationnelle basée sur un modèle bayésien hiérarchique d'apprentissage par renforcement qui propose d'expliquer comment l'apprentissage et le contrôle cognitif coopèrent. Dans cette partie, nous présentons les expériences comportementales conçues pour tester la validité de cette théorie par rapport au comportement humain.

Dans le chapitre 6, nous présentons une première expérience portant sur l'acquisition de task-sets en l'absence de signaux contextuels. Dans le chapitre 7, nous présentons une deuxième expérience portant sur l'acquisition de task-sets et de leur association à des indices contextuels. Dans les deux expériences, nous comparons les résultats expérimentaux aux prédictions effectuées par les modèles et validons ainsi les hypothèses effectuées dans le modèle proposé.

Chapitre 6

Expérience comportementale 1 : Apprentissage de task-sets sans contextes

Dans cette première expérience comportementale, nous testons les prédictions suivantes spécifiquement effectuées par notre modèle.

– Dans un environnement incertain, l'homme est capable de détecter un changement de contingences de l'environnement nécessitant de modifier son comportement. Il peut utiliser cette détection pour effectuer un switch abrupt de comportement, au niveau hiérarchique des task-sets, plutôt que de s'adapter localement, au niveau hiérarchique des stimuli.
– Les hommes sont capables d'utiliser cette détection de changement et ce switch pour stocker en mémoire des task-sets appris avant qu'ils ne soient effacés par adaptation, et pour pouvoir ainsi envisager de les réutiliser.
– Les hommes explorent de manière diffuse et parallèle le répertoire de task-sets déjà appris, et sont ainsi capables de réactiver un task-set appris précédemment, ou de décider de l'opportunité d'apprendre un nouveau task-set.

Afin de tester ces prédictions, nous avons soumis des sujets à une expérience d'apprentissage, par essai-erreur, de contingences stimulus-actions. Les sujets devaient s'adapter aux

changements de contingences au cours de l'expérience. Deux conditions permettaient de tester la possibilité de réutiliser des contingences déjà apprises contre la nécessité d'apprendre un nouveau comportement à chaque changement.

Nous présentons tout d'abord le protocole expérimental dessiné pour tester ces prédictions. Nous présentons ensuite les résultats comportementaux obtenus, en parallèle avec les prédictions de différents modèles : RL, UU, MMBRL (voir section 5.2, page 118) et le modèle proposé (noté LCI par la suite : *learning and control integrated*).

6.1 Matériel et méthodes

6.1.1 Participants

Pour cette première expérience, nous avons testé 26 sujets, âgés entre 18 et 35 ans. Les sujets ont été recrutés par des annonces affichées dans des universités. Les sujets ont été soumis à un un entretien médical afin de s'assurer qu'ils ne possédaient pas d'antécédents psychiatriques ou neurologiques susceptibles d'influencer leur participation à l'expérience, ni de problèmes de vision des couleurs ou d'audition. Les sujets ont tous été informés des conditions de participation aux expériences au moyen d'une notice écrite d'information légale, validée par le comité d'éthique de l'INSERM et ont signé en connaissance de cause, un consentement écrit avant leur inclusion dans le protocole. Les sujets ont été indemnisés pour leur participation (20 euros par session).

Parmi les 26 sujets, quatre n'ont pas été inclus dans les analyses pour cause de panne informatique sur une des sessions expérimentales. Nous avons donc 22 sujets (femmes 13, hommes 9) dans nos analyses.

6.1.2 Protocole

L'expérience se divise en deux sessions d'environ une heure, avec une condition expérimentale différente par session. Chaque session expérimentale se déroule dans une salle isolée,

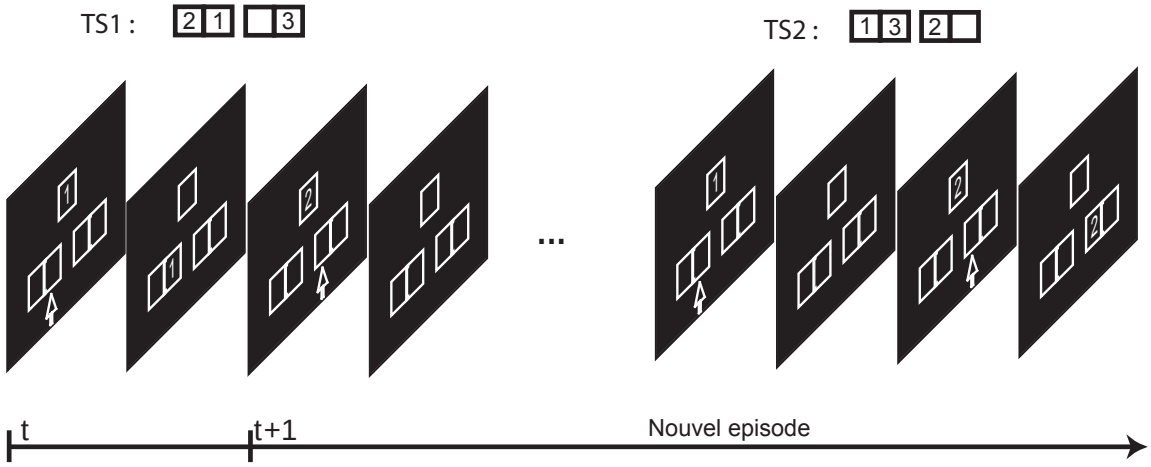


FIGURE 6.1 – Protocole expérimental. Le stimulus est présenté dans un carré central. Sur le schéma, la flèche indique l'action sélectionnée par le sujet. Si celle-ci correspond au task-set actuellement valide (représenté ici au dessus comme TS1), avec probabilité 0,9, le sujet reçoit un feedback sonore positif, et le stimulus se place dans la case correspondante (2e figure en partant de la gauche). Sinon, avec probabilité 0,9, le sujet reçoit un feeback sonore négatif et le stimulus disparaît simplement (4e écran). Rien ne signale un changement d'épisode (écrans 5 à 8), le sujet doit interpréter ses erreurs et découvrir le nouveau task-set valide (TS2).

sur un ordinateur, avec un casque audio. Les deux sessions sont séparées au minimum de 24h, au maximum d'une semaine.

A chaque essai, un stimulus (chiffre arabe blanc) est présenté au sujet à l'intérieur d'un cadre carré de couleur blanche situé au centre d'un écran noir. Il y a trois stimuli par session, pris dans l'ensemble $\{1, 3, 5\}$ pour l'une des sessions, $\{2, 4, 6\}$ pour l'autre session, afin d'éviter une interférence entre les deux sessions. Les sujets répondent à ces stimuli en actionnant l'une des quatre touches $[d, f, j, k]$ d'un clavier azerty, selon la règle d'actions suivante (favorisant une position naturelle fixe des mains) :

– d : majeur gauche
– f : index gauche
– j : index droit
– k : majeur droit

Ces quatre touches sont représentées à l'écran par quatre carrés disposés comme les touches, sous le domaine de présentation du stimulus. Après la sélection d'une action, le sujet reçoit un retour indiquant s'il a gagné ou perdu pour cet essai sous trois formes simultanément :

– Visuel : dans le cas *gagné*, le stimulus disparait du carré central et vient se placer dans la case correspondant à la touche sur laquelle le sujet a appuyé. Dans le cas perdu, le stimulus disparait simplement. Ce feedback visuel a pour but d'aider le sujet à se souvenir de l'association entre le stimulus et l'action.
– Visuel : sous les quatre cases représentant les actions, une jauge indique l'accumulation du gain. Elle avance dans le cas *gagné*, recule dans le cas *perdu*. Lorsqu'elle atteint le maximum, elle revient à zéro. En pratique, cela arrive plusieurs fois par session. Ce feedback a pour but de motiver le sujet.
– Auditif : dans le cas *gagné*, le sujet entend un son montant rapide, dans le cas perdu, un son descendant.

Chaque essai dure 2 secondes. Le stimulus est présenté pendant 1,5 seconde, pendant laquelle le sujet doit répondre. Le retour s'effectue 100ms après l'action effectuée par le sujet. Si le sujet ne répond pas à temps, aucun retour n'est fourni.

Chaque session se divise en 25 épisodes comprenant 36 à 54 essais, pour un total de 1134 essais par session. Pendant chaque épisode, un task-set précis est valide. Le retour correspondant à *gagné* ou *perdu* est donné en fonction de l'action prédite par ce task-set, avec un bruit de 0,1 : dans 10% des cas, le retour opposé à celui qui devrait être fourni est donné au sujet. Aucun signal n'indique au sujet les changements d'épisode.

Les sujets ont pour instruction de tenter de gagner le plus souvent possible, sans manquer d'essais [1]. Ils ont à leur disposition 6 pauses par session, qu'ils peuvent prendre quand ils le souhaitent.

Les deux sessions, testant deux conditions expérimentales distinctes, diffèrent par l'agencement des task-sets dans les épisodes.

1. Instruction exacte : *« Vous devez essayer de répondre le mieux et le plus vite possible. Votre indemnité dépendra de votre performance globale. »*

Condition principale Dans cette condition, trois task-sets seulement sont utilisés pendant toute l'expérience. Ces task-sets, décrits dans le tableau ci-dessous sont totalement incongruents :

Actions	S_1	S_2	S_3
TS_1	a_1	a_2	a_3
TS_2	a_2	a_3	a_4
TS_3	a_3	a_4	a_1

Les task-sets TS_2 et TS_3 sont chacun valides dans 8 épisodes, le task-set TS_1 dans le premier épisode puis dans 8 autres épisodes.

Cette condition permet de tester la réutilisation de task-sets par les sujets.

Condition contrôle Dans cette condition, dans chaque épisode, un task-set différent est valide. Afin de pouvoir comparer la condition contrôle à la condition principale, nous fixons les contraintes suivantes pour les 24 premiers task-sets :
– deux stimuli ne sont associés à la même action dans aucun task-set.
– les task-sets de deux épisodes successifs sont incongruents (les associations stimulus-action changent pour tous les stimuli).
Ces contraintes limitent à 24 le nombre de task-sets possible, impliquant qu'un des task-set est répété une fois. Nous éloignons au maximum dans le temps les deux épisodes associés au même task-set.

Dans cette condition, les sujets ne peuvent utiliser une information passée pour découvrir le nouveau task-set correct. Ils doivent donc strictement s'adapter en apprenant un nouveau task-set, alors que la condition principale leur donne l'opportunité de réutiliser un task-set qui semble adapté plutôt que de l'apprendre à nouveau.

Avant la première session, les sujets lisent les instructions expliquant la tâche et l'expérimentateur s'assure qu'ils ont compris ces instructions. Pour ce faire, les sujets sont soumis à deux épisodes construits selon le même principe que la tâche, mais limités à deux stimuli (ronds de couleur) et deux actions. Cela leur permet d'être informés du principe d'incertitude dans les récompenses et de la possibilité de changement de la règle à appliquer, ainsi

que de se familiariser avec le retour fourni par l'ordinateur. Les sujets ne sont pas au courant des deux conditions expérimentales.

Afin de motiver les sujets à apprendre du mieux qu'ils peuvent, ils sont avertis qu'ils peuvent à chaque session doubler le montant de leur indemnisation en fonction de leur performance pendant l'expérience.

Après chaque session, le sujet est soumis à un post-test passif. Pendant ce post-test, 6 task-sets sont présentés à l'écran au sujet, qui est invité à répondre sur une échelle de 0 à 4 à la question de savoir s'il pense avoir utilisé ce task-set avec succès pendant l'expérience. La note 4 indique qu'il est sûr de l'avoir utilisé, 3 qu'il pense l'avoir utilisé, 0 qu'il est sûr de ne pas l'avoir utilisé, 1 qu'il pense ne pas l'avoir utilisé et 2 qu'il ne sait pas. Pour l'expérience principale, 3 des 6 task-sets sont les task-sets valides pendant l'expérience. Pour l'expérience contrôle, les 6 task-sets sont valides pendant l'expérience. Un débriefing informel est également effectué.

L'ordre de passage des deux conditions entre les deux sessions est contrebalancé entre les sujets, ainsi que l'association entre un des deux jeux de stimuli et les sessions. Le débriefing est contrebalancé entre les sujets de telle sorte que, pour la condition principale, les 3 TS non valides d'un sujet correspondent au trois TS valides d'un autre sujet, et inversement. De même, pour la condition contrôle, les 6 TS d'un sujet correspondent aux 6 TS présentés à un autre sujet dans le débriefing de sa condition principale.

Les associations entre les stimuli d'une session et $\{S_1, S_2, S_3\}$ ainsi que celles entre les $\{a_1, a_2, a_3, a_4\}$ et les quatre touches sont tirées au hasard de manière à générer, dans la condition principale, à partir des trois task-sets présentés plus haut, tous les triplés de task-sets possibles respectant les conditions d'incongruence. Ainsi, les chances que les trois task-sets se déduisent l'un de l'autre par simple permutation circulaire sont faibles, et l'ordre des stimuli de gauche à droite sur l'écran n'est pas nécessairement monotone, ni identique d'un task-set à l'autre, limitant les risques de généralisation d'un task-set à l'autre.

Le bruit dans le renforcement fourni est pseudo-randomisé de manière à s'assurer que la répartition n'est pas déséquilibrée en faveur du début ou de la fin de la session.

L'expérience est présentée sur un ordinateur Mac, avec l'outil *Psychtoolbox* (http ://psychtoolbox.org) de *Matlab* (http ://www.mathworks.fr).

Les résultats de l'expérience sont analysés sur Matlab et SPSS (http ://www.spss.com/fr/).

Les modèles sont testés sur le même protocole que les sujets, afin de pouvoir comparer les performances des sujets aux prédictions effectuées par les modèles.

6.2 Résultats expérimentaux

6.2.1 Groupe entier

Quantités mesurées

Les prédictions théoriques portent sur les méthodes d'adaptation après un changement d'épisode. Afin de comparer les résultats expérimentaux aux prédictions effectuées par les différents modèles, nous mesurons trois quantités pertinentes pour les prédictions et regardons leur pattern d'évolution après un changement d'épisode. Ces trois quantités sont :

– la proportion de réponses correctes. Elle permet d'observer l'apprentissage ou l'utilisation du task-set au cours d'un épisode. Nous traçons la proportion de réponses correctes sur l'ensemble des sujets (ou simulations) et sur les épisodes 2-25, en fonction du numéro d'essai t après le changement d'épisode, pour $t = 1, \ldots, 36$. Le premier épisode est supprimé à cause de sa spécificité : il ne teste pas l'adaptation à un changement, mais l'apprentissage initial.

– l'exploration. Les réponses données par le sujet (ou les modèles) peuvent être classées comme correctes, persévératives (si elles sont correctes selon le task-set correspondant à l'épisode précédent) ou exploratoires, si elles ne sont ni correctes ni persévératives. L'exploration est, selon cette définition, la proportion de réponses exploratoires. Nous traçons la proportion de réponses exploratoires sur l'ensemble des sujets (ou simulations) et sur les épisodes 2-25, en fonction du numéro d'essai t après le changement d'épisode, pour $t = 1, \ldots, 36$.

– l'information mutuelle entre deux essais successifs. L'information mutuelle est définie par la formule :

$$\sum_{i=0}^{1}\sum_{j=0}^{1} P(Cor_t = i, Cor_{t+1} = j) \log\left(\frac{P(Cor_t = i, Cor_{t+1} = j)}{P(Cor_t = i)P(Cor_{t+1} = j)}\right)$$

Cette quantité représente la dépendance entre la distribution de probabilité Cor (le fait d'avoir choisi l'action correcte, $Cor = 1$, ou non, $Cor = 0$) à un essai t et à l'essai suivant $t+1$. Plus précisément, l'information mutuelle évalue les répercussions que peuvent avoir le fait d'avoir correct ou non à un essai, sur la stratégie à l'essai suivant. Lorsque la stratégie est connue, ces répercussions sont faibles : dans la limite de l'incertitude attendue, avoir correct ou incorrect ne modifie pas la stratégie. En période d'apprentissage de la stratégie, ces répercussions peuvent être très importantes si l'acteur utilise des tasks-sets. En effet, alors, trouver la bonne action pour un seul stimulus devrait permettre d'en déduire la bonne action pour les deux autres stimuli, impliquant une information mutuelle forte. Au contraire, si l'acteur apprend stimulus par stimulus, les conséquences d'une réponse correcte ne se répercutent que sur le stimulus testé, impliquant une information mutuelle faible. C'est donc une mesure essentielle pour tester l'hypothèse de l'utilisation de task-sets par l'acteur, plutôt que seulement des paires stimulus-actions indépendantes.

Pour l'information mutuelle, afin d'estimer de manière fréquentielle les probabilités nécessaires, nous avons besoin de suffisamment de points de données. Nous regroupons donc par 5 essais consécutifs les paires de stimuli et utilisons une fenêtre glissante de taille 5, nous permettant d'avoir 30 points pour représenter l'évolution de l'information mutuelle du début à la fin des épisodes de 36 essais. Le choix de groupes de 5 essais est arbitraire, mais permet d'avoir suffisamment de données pour estimer les probabilités. Nous nous limitons à la deuxième moitié de l'expérience (douze derniers épisodes) pour cette mesure, puisqu'elle est utilisée pour mesurer l'utilisation de task-sets, donc après apprentissage suffisant.

Les prédictions des modèles et les résultats expérimentaux sont résumés dans la figure 6.2, page 164. Les paramètres des modèles sont optimisés pour obtenir la plus grande proportion de bonnes réponses possibles sur les deux sessions, et sont résumés dans la table ci-dessous.

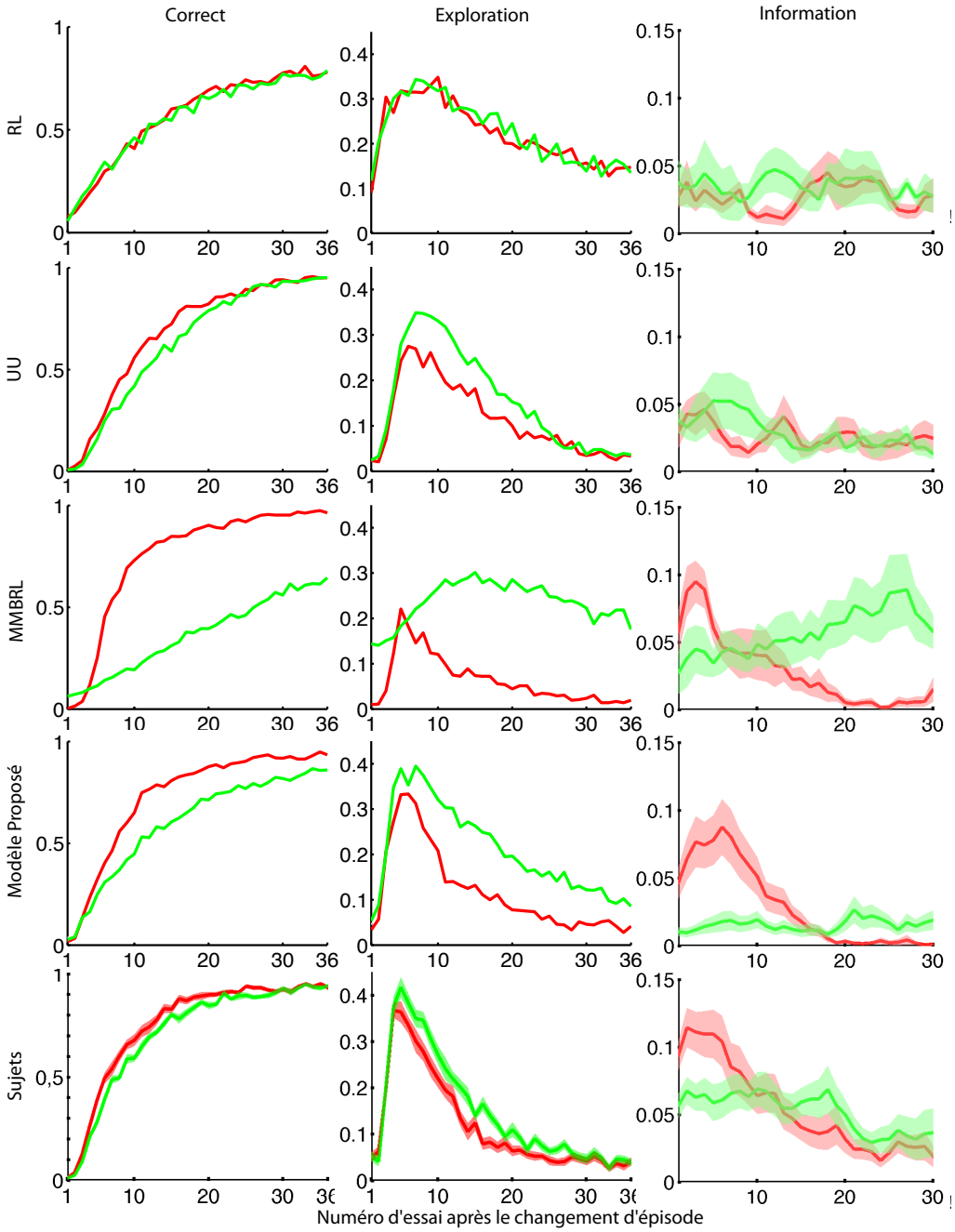


FIGURE 6.2 – Prédictions des modèles et résultats expérimentaux. Lignes, de haut en bas : modèles RL, UU, MMBRL, modèle proposé (LCI), résultats expérimentaux. Colonnes, de gauche à droite : proportion de réponses correctes, exploration, information mutuelle. Rouge : condition principale. Vert : condition contrôle. Les aires autour des courbes représentent l'erreur standard de la moyenne. L'abscisse est le numéro d'essai après un changement d'épisode.

	β	α	ϵ	N	p_{test}	τ
RL	15	0,8	0	-	-	-
UU	19	0,5	0	-	1	0,03
MMBRL	19	0,3	0	4	-	0,03
LCI	19	0,4	0,03	-	0,5	0,03

Notons que le paramètre de vitesse d'apprentissage choisi pour le modèle RL est particulièrement élevé (0,8), impliquant un apprentissage rapide en début d'épisode (voir figure 6.2, en haut à droite), mais une performance asymptotique faible (en effet, beaucoup de poids est donné aux observations les plus récentes, ainsi, une erreur due au bruit est interprétée comme un changement de contingences). Un paramètre plus faible aurait pu donner une performance globale équivalente en modifiant les prédictions qualitatives du modèle, avec une performance asymptotique meilleure, mais un apprentissage plus lent.

Résultats observés

Performance : adaptation efficace. Les données expérimentales montrent que les sujets parviennent à apprendre un task-set au sein d'un épisode : ils atteignent une performance asymptotique forte (>90%) dans les deux conditions expérimentales. Les modèles UU, MMBRL et LCI proposé atteignent cette performance asymptotique dans la condition principale (valeurs). Dans la condition contrôle, les modèles UU et LCI prédisent cette performance.

Le modèle MMBRL n'atteint pas de performance asymptotique dans la condition contrôle. En effet, comme $N = 4$, il est ralenti pour deux raisons : 1) l'apprentissage est très parallèle puisqu'aucun des 4 TS stockés ne correspond dans un nouvel épisode, impliquant qu'aucun TS n'émerge clairement 2) le modèle doit désapprendre un (ou plusieurs) des 4 TS stockés pour pouvoir en apprendre un nouveau. Même avec $N = 24$, la performance dans la condition contrôle est mauvaise : en effet, cela augmente le parallélisme de l'apprentissage et donc le ralentit.

Le modèle RL atteint une performance asymptotique relativement faible dans les deux conditions. Cela est dû au fait qu'afin d'optimiser la performance globale, on a favorisé un

apprentissage rapide (α fort), qui impose plus d'erreurs, dû au bruit présent dans le renforcement. Avec une valeur de α plus faible, on pourrait atteindre une valeur asymptotique identique à celle des sujets, mais avec un apprentissage en début d'épisode, très lent.

Switch. La présence d'un comportement de type *switch* après le changement d'épisode se matérialise par l'existence d'un court plateau de performance faible au début des épisodes (persévération sur le task-set précédent), suivi d'une amélioration brutale de la performance. On observe ce switch dans les données expérimentales, pour les deux conditions (contrôle et principale). Ce comportement est prédit à la fois par le modèle UU et le modèle LCI, conformément à la structure des ces modèles qui incorpore une décision de switch.

Ce comportement n'est pas prédit par RL, dans aucune des deux conditions : on observe en effet une évolution sans point d'inflexion de la proportion des réponses correctes. Il n'est prédit par MMBRL que dans la condition principale. L'échec du MMBRL à prédire un switch dans la condition contrôle est lié à son incapacité à apprendre correctement au sein d'un épisode, le switch ne pouvant avoir lieu qu'après un apprentissage correct.

Principal/contrôle On observe une différence dans la vitesse d'adaptation entre les conditions, dans les données expérimentales : les sujets atteignent une performance asymptotique plus rapidement dans la condition principale que dans la condition contrôle. De manière significative, la performance est meilleure pour la condition principale que pour la condition contrôle (différence des performances moyennes entre conditions signicativement positive, $t = 3,41$, $p < 0,005$).

Cette différence n'est pas prédite par le RL, pour lequel l'adaptation après un changement d'épisode est strictement identique selon les deux conditions. Elle est prédite par le MMBRL, de manière exagérée, à cause notamment de l'absence de performance correcte dans la condition contrôle.

Cette différence est prédite à la fois par le modèle UU et par le modèle proposé. Dans le modèle UU, cela reflète le fait que, dans la condition contrôle, pour un stimulus, trois des quatre actions sont associées à une récompense plus souvent que la quatrième. L'introduction d'un biais à l'initialisation du TS après un switch profite donc de ce léger biais pour

accélérer légèrement l'apprentissage, par rapport à la condition contrôle.

Dans le modèle proposé, cette différence reflète à la fois le biais de bas niveau décrit ci-dessus pour le modèle UU, et la capacité du modèle à utiliser des task-sets déjà appris (comme démontré dans la partie simulations des modèles).

L'analyse des courbes de performance n'est donc pas conclusive puisque les observations expérimentales sont prédites à la fois par le modèle UU et le modèle proposé. Elle tend néanmoins à disqualifier les modèles RL et MMBRL.

Exploration La différence entre les deux conditions observée dans les réponses correctes se retrouve dans les courbes d'exploration ($t = 4,6849$, $p < 10^{-4}$). On observe dans les données expérimentales, après une légère persévération, un pic d'exploration, reflétant le comportement de switch décrit plus haut. Ce pic est suivi d'une diminution rapide de l'exploration, reflétant l'apprentissage du task-set au sein de l'épisode. Ce pic est légèrement plus tôt et significativement moins haut dans la condition principale, indiquant une exploration plus efficace pour découvrir le TS correct dans cette situation. A nouveau, ce phénomène n'est prédit que par le modèle UU et le modèle proposé.

Information mutuelle L'information mutuelle est faible, proche de 0, lorsque le fait d'avoir un essai correct n'apporte pas d'information pour l'action correcte à l'essai suivant, soit que la sélection des actions pour deux stimuli différents est indépendante. Cela est vrai dans deux cas :

- pendant une phase d'apprentissage, si l'apprentissage se fait indépendamment stimulus par stimulus.
- après une phase d'apprentissage, lorsque le task-set est parfaitement appris. En effet, dans ce cas, les probabilités $P(Cor_t = 1, Cor_{t+1} = 1)$, $P(Cor_t = 1)$ et $P(Cor_{t+1} = 1)$ sont toutes proches de 1, impliquant que le terme à l'intérieur du logarithme est proche de 1, le logarithme lui même proche de 0. Les autres termes de la somme sont également proches de 0 puis pour $(i, j) \neq (1, 1)$, $P(Cor_t = i, Cor_{t+1} = j)$ est proche de 0. En d'autres termes, il n'y a pas de transfert d'information d'un essai au suivant parce que l'information est déjà connue.

Notons que, pendant une phase d'apprentissage, l'indépendance ne peut pas être parfaite : en effet, si on voit deux fois de suite le même stimulus, il y a nécessairement une faible dépendance entre les actions sélectionnées.

Au contraire, l'information mutuelle est forte lorsqu'il y a une dépendance importante entre la correction dans deux essais successifs, lorsque le fait d'être correct ou non à un essai est fortement informatif pour agir de telle sorte à être correct ou non à l'essai suivant.

Les modèles RL et UU prédisent une information mutuelle faible et approximativement constante pendant un épisode. En effet, l'apprentissage de l'action correcte pour un stimulus est strictement indépendant de celui pour les autres stimuli. Le MMBRL, par contre, prédit, pour la condition contrôle, une information mutuelle forte au début de l'épisode (indiquant l'utilisation de task-sets et donc une dépendance forte entre les actions sélectionnées pour différents stimuli), suivie d'une diminution de cette information lorsque le task-set est appris. Dans la condition contrôle, l'augmentation progressive de l'information mutuelle (MMBRL) reflète également une relation de dépendance entre les essais t et $t+1$ liée au fait que le modèle tente d'utiliser des task-sets déjà appris pour déterminer s'ils sont ou non adaptés à l'épisode présent. Comme l'apprentissage du MMBRL ne se termine pas au cours de l'épisode dans cette condition, l'information ne diminue pas.

Le modèle proposé reflète le même schéma d'information mutuelle que le MMBRL dans la condition principale. Initialement, l'information mutuelle est forte, reflétant la réutilisation de task-sets appris pour déduire d'une réponse correcte, les réponses correctes aux essais suivants (d'où transfert important d'information), puis diminution liée à une performance correcte à tous les essais et à l'absence d'information supplémentaire apportée par les essais corrects ou incorrects. Par contre, dans la condition contrôle, pour le modèle proposé, l'information mutuelle est faible et stable, comme pour RL ou UU, reflétant la possibilité du modèle d'apprendre des nouveaux task-sets, donc indépendamment, stimulus par stimulus.

On observe dans les données des sujets, le pattern prédit par notre modèle : dans la condition principale, l'information mutuelle initiale est forte puis diminue significativement vers 0. Dans la condition contrôle, l'information mutuelle est stable. L'information mutuelle initiale est significativement plus forte dans la condition principale que dans la condition

contrôle.

Conclusions

Les données expérimentales portant sur l'information mutuelle montrent bien que la différence observée entre les conditions (adaptation et exploration plus rapides dans la condition principale) sont bien dues à la capacité des sujets de réutiliser des task-sets déjà appris, plus qu'à un biais de bas niveau sur la fréquence d'utilisation des différentes actions.

Le modèle proposé est le seul à rendre compte du comportement d'apprentissage et de contrôle des task-sets par les sujets. En effet, ensemble, les effets expérimentaux observés (capacité d'adaptation efficace, switch, utilisation des task-sets dans la condition principale conduisant à une performance meilleure et une exploration plus rapide) ne sont prédits que par le modèle proposé dans cette thèse. Nous confirmons ainsi la validité du modèle dans le but de représenter l'apprentissage et le contrôle de task-sets par les sujets humains. Par ailleurs, ces résultats permettent de soutenir les hypothèses sur lesquelles repose notre modèle : les sujets sont capables d'apprendre des task-sets et de les réutiliser, dans un environnement incertain et sans indices contextuels.

6.2.2 Fitting du modèle aux données des sujets, debriefing

Si les résultats du groupe valident les prédictions de notre modèle, nous avons observé une variabilité importante dans la réaction des sujets à notre expérience. Nous détaillons tout d'abord cette variabilité indiquée par le debriefing. Afin d' étudier plus précisément les données expérimentales par rapport aux prédictions de notre modèle, nous avons effectué une procédure de fitting du modèle sur les données des sujets, afin de déterminer pour chaque sujet les paramètres qui expliquent le mieux ces données.

Debriefing

Il ressort du debriefing informel que la tâche est considérée comme difficile par les sujets. En particulier, ils ont nettement tendance à surestimer le niveau de bruit dans le renforcement

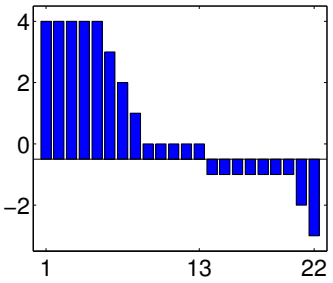


FIGURE 6.3 – Critère de debriefing, par ordre décroissant pour l'ensemble des sujets.

fourni par l'expérience.

Crucialement, seul un sous-groupe des sujets nous a rapporté correctement la différence essentielle entre les deux sessions expérimentales, indiquant qu'ils ont perçu le fait que seul un nombre limité de task-sets était utilisé dans la condition principale et qu'ils pouvaient donc activement réutiliser ces task-sets afin d'améliorer leur performance. L'autre partie des sujets ne se souvient pas d'une différence de structure entre les deux sessions, et en particulier, dit n'avoir jamais fait d'effort actif pour réutiliser des task-sets précédemment appris.

En utilisant le post-test, nous séparons deux groupes de sujets selon le critère suivant (voir figure 6.3) : si les notes qui ont été attribuées aux trois task-sets valides pendant la condition principale sont toutes meilleures que les notes qui ont été attribuées aux troix autres task-sets, le sujet est classé dans le groupe *Conservateur* (ainsi nommé parce qu'ils utilisent à nouveau des task-sets déjà appris plutôt que de les apprendre à nouveau). Sinon, il est classé dans le groupe *Réformateur*(ainsi nommé parce qu'ils s'adaptent en apprenant un nouveau comportement, plutôt que d'en utiliser un précédemment appris). Le groupe conservateur est constitué de 13 personnes, le groupe réformateur de 9 personnes. Ces groupes ne recoupent pas parfaitement les résultats du debriefing informel.

Il est par ailleurs important de noter que ni le debriefing informel, ni les groupes issus du post-test, ne semblent corrélés avec aucun facteur simple que nous avons pu tester a posteriori (âge, sexe, niveau d'éducation, ordre de passage des sessions).

Fitting

Notre modèle est muni d'un paramètre, p_{test}, gérant le biais entre deux types de comportement : apprendre un nouveau task-set, ou tenter d'en réutiliser un précédemment appris. Ce paramètre semble représenter idéalement la distinction entre les deux groupes observés à partir du debriefing, comme décrite ci-dessus. Afin d'étudier le lien possible entre ce paramètre et le comportement des sujets quantitativement, nous effectuons un fit du modèle sur les données des sujets.

Pour un sujet et un jeu de paramètres fixé, nous calculons une distance qui représente à quel point le modèle représente bien les données des sujets. Pour calculer cette distance, nous soumettons le modèle à la séquence précise de stimuli et de renforcements observés par le sujet et imposons ses choix d'actions par les actions du sujet. Pour chaque essai, nous pouvons calculer la probabilité (selon la règle ϵ-softmax du task-set actuellement utilisé, d'après le modèle) que le modèle choisisse l'action correcte sachant ce qu'il a observé jusqu'à cet essai et sous l'hypothèse que le sujet utilise notre modèle pour effectuer ses choix.

On peut alors dire que le modèle, avec ce jeu de paramètre, représente bien le comportement des sujets si les probabilités prédites par le modèle sont proches des fréquences observées. Nous calculons donc la proportion d'essais corrects, en fonction de la probabilité prédite par le modèle. Les fréquences (proportions d'essais corrects) sont calculées sur des ensembles d'essais homogènes quant à la probabilités de choisir l'action correcte prédite par le modèle (par exemple, sur l'ensemble des essais pour lequel le modèle prédit une probabilité de choisir une réponse correcte dans l'intervalle $[00, 1]$). Nous calculons alors la distance comme la somme, pour les différents groupes, du carré de la différence entre la probabilité moyenne de ces essais (fournie par le modèle) et la fréquence observée chez le sujet. On trace dans la figure 6.4 des exemples de graphes probabilités - fréquences par sujets.

Pour un sujet, nous explorons l'espace des paramètres pour calculer cette distance. Nous pouvons ainsi trouver le jeu de paramètres qui minimise cette distance pour un sujet, soit le jeu de paramètres pour lequel le modèle explique au mieux le comportement du sujet.

L'algorithme du modèle utilisé pour le fitting est identique à ce qui est décrit dans la partie sans contextes du chapitre présentant les modèles. Cependant, par souci de plausibilité,

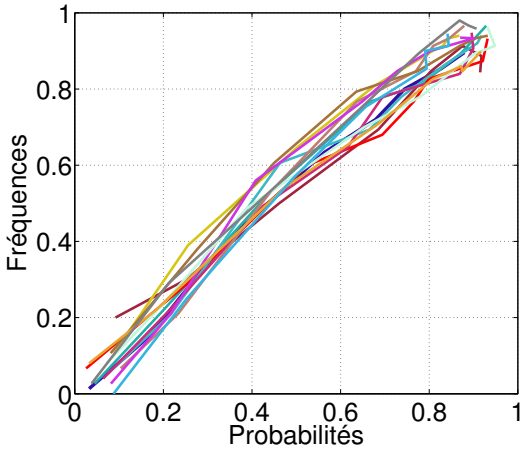


FIGURE 6.4 – Graphes de fitting probabilité - fréquence. En fonction de la probabilité de répondre correct prédite par le modèle fitté sur les données d'un sujet, on trace la fréquence de réponses correctes de ce sujet. Chaque couleur représente un sujet différent. Pour plus de clarté, seule une partie des sujets est tracée ici.

nous limitons la capacité de mémoire de notre modèle imposant un nombre limite de task-sets dont le modèle peut se souvenir. En effet, si l'ordinateur a une mémoire parfaite, ce n'est pas le cas des sujets. Il y a donc un risque dans les simulations, d'observer des effets d'interférence entre plusieurs task-sets appris à grande distance qui n'existe pas chez les sujets, pour cause de mémoire imparfaite. Si plus de task-sets que le nombre limite arbitraire imposé sont créés par le modèle, le TS appris il y a plus longtemps est oublié. Notons que cet ajout est fondamentalement différent du nombre limité de task-sets dans le MMBRL. En effet, le modèle commence quand même avec seulement 2 task-sets et peut ne jamais atteindre la limite. Par ailleurs, une fois la limite atteinte, de nouveaux TS peuvent être créés, même si cela se fait au prix de l'oubli d'un autre TS.

Résultats du fitting : La qualité du fitting n'est pas significativement différente entre les deux groupes. Nous n'observons pas de différence significative entre les deux groupes pour quatre des paramètres : β, α, ϵ, τ ($p > 0,4$). Par contre, nous observons une différence significative (Mann-Wittney, p$<$.004) pour le paramètre p_{test}, rapportée dans la figure

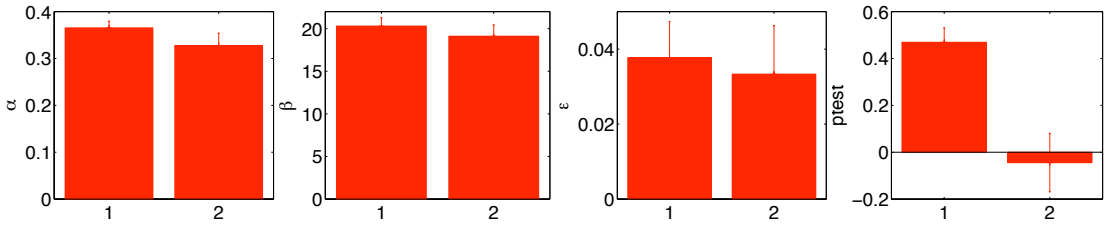


FIGURE 6.5 – Paramètres α, β, ϵ et p_{test} fittés pour le groupe conservateur (1) et réformateur (2). Aucune différence significative pour les trois premiers (p>0,4).

6.5.

Cela permet de confirmer la validité des groupes effectués ainsi que le rôle du paramètre p_{test} dans le biais d'apprendre à nouveau ou de tenter d'utiliser à nouveau des task-sets précédemment appris.

Dans le paragraphe suivant, nous analysons les données en séparant les deux groupes et en les comparant à notre modèle, en utilisant les paramètres fittés

6.2.3 Variabilité interindividuelle

La séparation en deux groupes des sujets permet de mettre en valeur un effet d'apprentissage à long terme dans le groupe conservateur. Cet effet n'est observé ni pour le groupe réformateur, ni pour l'ensemble du groupe.

On trace la performance moyenne (proportion de réponses correctes pendant un épisode pour les sujets d'un groupe) en fonction du numéro d'épisode dans la session principale, dans la figure 6.6. On effectue une régression linéaire afin de tester l'apprentissage à long terme au cours de l'expérience. Pour les sujets réformateurs, on observe une tendance (non significative, $r = -0,07$, $p > 0,3$) négative, reflétant probablement un effet de fatigue en l'absence d'effet d'apprentissage à long terme. Pour les sujets conservateurs par contre, on observe un coefficient de corrélation significativement supérieur à 0 ($r = 0,17$, $p < 0,003$). Cela indique que, bien qu'ils soient également soumis à la fatigue, les sujets progressent au fil de l'expérience. Cela permet de mettre en évidence un apprentissage à long terme,

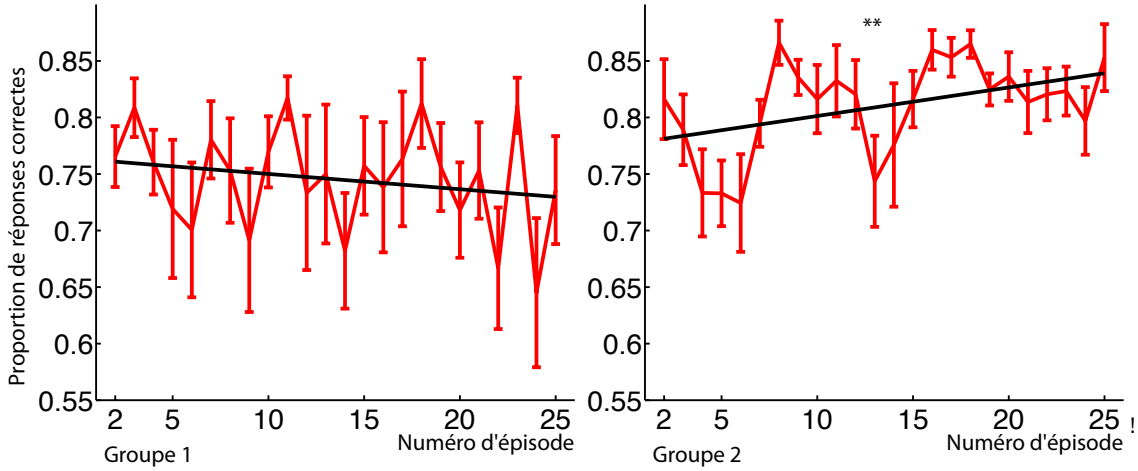


FIGURE 6.6 – Proportion de réponses correctes par épisode pour le groupe conservateur (droite) et réformateur (gauche). Barres d'erreur : erreur standard de la moyenne. En noir : droite de régression linéaire. On observe une augmentation significative de la performance pour le groupe conservateur, une diminution non significative pour le groupe réformateur.

ce qui est concordant avec l'hypothèse que ce groupe parvient à acquérir et réutiliser des task-sets.

Afin de montrer plus explicitement la différence entre les deux groupes quant à la capacité à stocker et réutiliser des task-sets, nous traçons à nouveau les courbes de performance et d'exploration, pour les deux groupes, dans la figure 6.7.

On voit que, pour le groupe conservateur, la performance dans la condition principale est très significativement supérieure à la performance dans la condition contrôle ($t = 4,487$, $p < 0,001$). L'exploration est également plus rapide ($t = -5,73$, $p < 10^{-4}$). Par opposition, pour le groupe réformateur, les deux conditions sont confondues (correct : $t = 0,7$, $p > 0,5$; exploration : $t = 1,7$, $p > 0,1$). Ces observations sont confirmées par l'information mutuelle pour les deux groupes. En effet, on voit dans la figure 6.8, pour le groupe conservateur, que l'information mutuelle est significativement plus grande au début de l'épisode pour la condition principale que pour la condition contrôle ; alors que celle-ci diminue significativement pour la condition principale, elle reste stable pour la condition contrôle. Cela est indicatif,

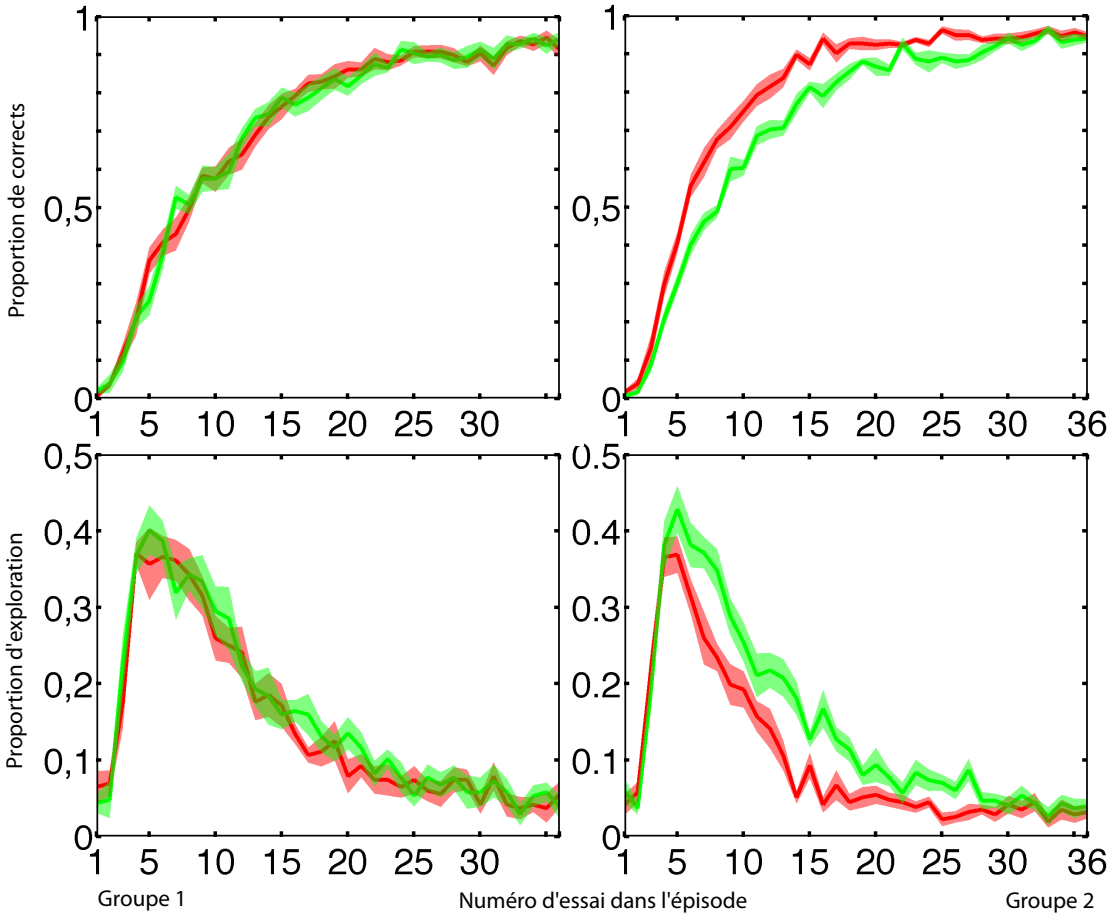


FIGURE 6.7 – Proportion de réponses correctes (haut) et exploration (bas) pour le groupe conservateur (droite) et réformateur (gauche) ; en fonction du numéro d'essai à partir du changement d'épisode. Rouge : condition principale. Vert : condition contrôle. Aires : erreur standard de la moyenne. On observe des différences très significatives pour le groupe conservateur, pas de différence significative pour le groupe réformateur.

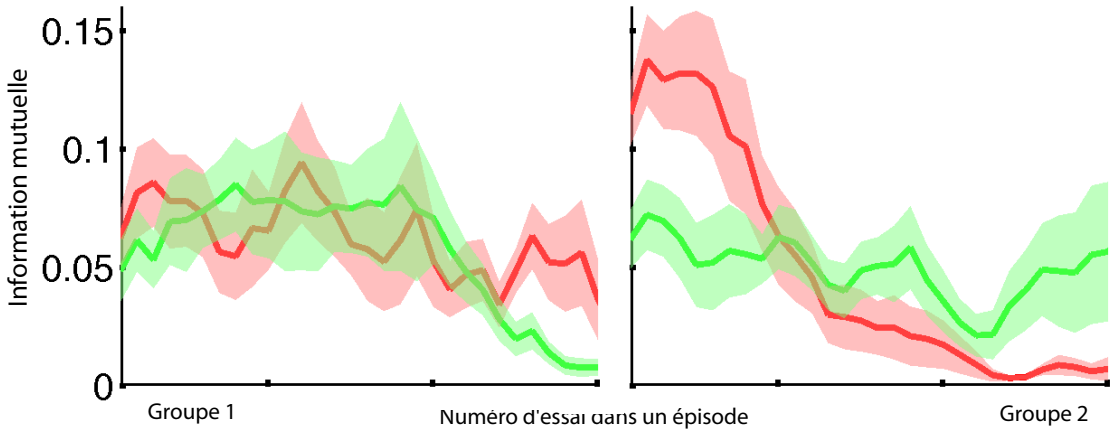


FIGURE 6.8 – Evolution de l'information mutuelle au cours des épisodes de la deuxième moitié de l'expérience, pour le groupe conservateur (droite) et le groupe réformateur (gauche), en fonction du numéro d'essai à partir du changement d'épisode. Rouge : condition principale. Vert : condition contrôle.

comme prédit par le modèle, de la réutilisation des task-sets. Par contre, pour le groupe réformateur, on n'observe pas de différence significative entre les deux conditions.

Si on trace sur un même graphe (figure 6.9 à gauche) les performances des deux groupes, on voit que trois des courbes sont strictement confondues, indiquant ainsi que le groupe conservateur profite de sa capacité à réutiliser des task-sets seulement lorsque cette capacité est opportune.

On a également tracé (figure 6.9, droite) les performances du modèle sur les paramètres fittés pour les deux groupes : $\beta = 19$, $\alpha = 0,5$, $\epsilon = 0,03$, $\tau = 0,03$ et $p_{test} = 0,49$ pour le groupe conservateur, $p_{test} = -0,04$ pour le groupe réformateur.

On voit donc que le modèle reproduit les comportements variables des sujets, en fonction du paramètre contrôlant le biais d'adaptation, p_{test}. En effet, pour la valeur de p_{test} proche de 0, correspondant au groupe réformateur, on n'observe pas de différence entre les deux conditions. Par contre, pour la valeur de p_{test} qui implémente un biais vers la réutilisation de task-sets, correspondant au groupe conservateur, on observe une forte différence entre les deux conditions. On voit par ailleurs que, comme pour les données expérimentales, la

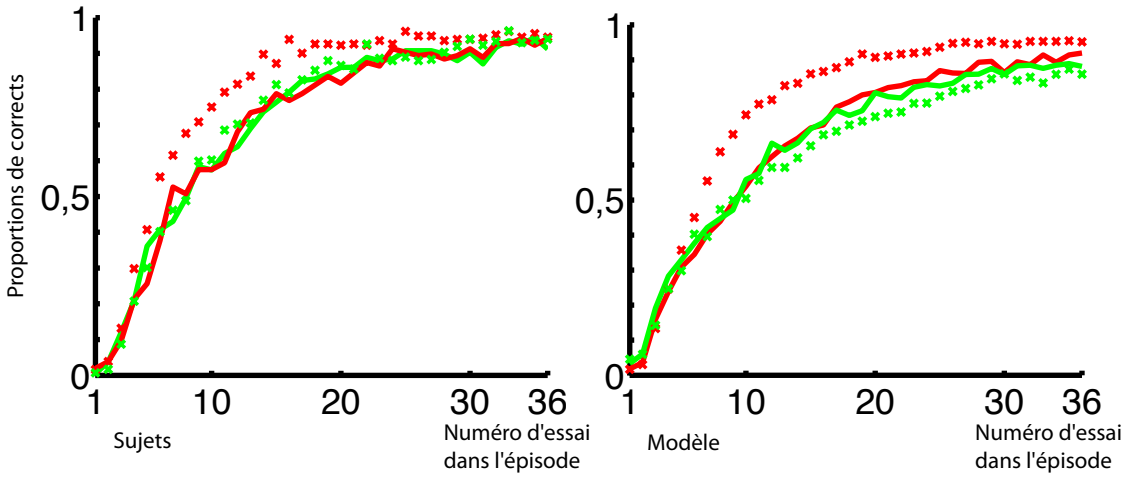


FIGURE 6.9 – Comparaison des performances des deux groupes et des prédictions du modèle. Gauche. Proportion de réponses correctes pour le groupe conservateur (croix) et réformateur (trait plain). Droite. Simulation des modèles avec les paramètres fittés sur les deux groupes. Rouge : condition principale, Vert : condition contrôle.

différence profite essentiellement à la condition principale et que les trois autres courbes sont très proches, même si l'adaptation est légèrement moins bonne pour la condition contrôle du paramètre conservateur.

Notons que, malgré le soin apporté dans la construction du protocole expérimental pour tenter d'éviter toute généralisation à la place d'apprentissage, les sujets utilisent encore une généralisation qui n'est pas à la disposition du modèle. Ils observent en effet que deux stimuli ne sont jamais associés à la même action, ce qui leur permet parfois, lors de l'apprentissage, de réduire le nombre d'actions à tester pour un stimulus, parce qu'une action est déjà « prise »pour un autre stimulus. Nous n'avons pas modélisé ce facteur. Cela peut expliquer une performance légèrement moins bonne du modèle par rapport aux sujets dans certaines conditions.

6.3 Conclusions de l'expérience sans contexte

Les résultats expérimentaux confirment les prédictions de notre modèle et permettent d'infirmer les autres modèles étudiés. En effet, on observe que les sujets sont capables d'apprendre les task-sets dans un environnement incertain et de les réutiliser et qu'un mécanisme de contrôle sous la forme d'un switch intervient pour accélérer l'adaptation à de nouvelles contingences.

Par ailleurs, le modèle permet de capturer, par l'intermédiaire d'un paramètre de biais, une partie importante et qualitative de la variabilité comportementale observée. En effet, on peut séparer, à partir du débriefing, deux groupes dont on montre ensuite qu'ils ont des comportements qualitativement différents : l'un favorise l'adaptation par l'exploration et la réutilisation de task-sets appris, l'autre l'adaptation par apprentissage, ce qui implique une performance moins bonne quand les task-sets pourraient être réutilisés. Ces deux groupes présentent des paramètres de biais fittés significativement différents, validant ainsi son rôle dans l'explication de la variabilité comportementale. Avec les paramètres fittés, les simulations du modèle reproduisent les observations expérimentales.

Notons que la séparation en deux groupes n'est pas une séparation entre des *bons* et des *mauvais* sujets. Il se trouve que dans l'expérience présentée ici, effectivement, le groupe conservateur est avantagé en terme de performance, puisque sa capacité à réutiliser les task-sets est opportune dans la condition principale. Cependant, ce ne serait pas forcément le cas dans d'autres expériences, par exemple avec un environnement plus changeant, où les sujets réformateurs pourraient être avantagés par rapport aux sujets conservateurs par leur biais à apprendre du nouveau plutôt qu'à réutiliser de l'ancien. Ce sera d'ailleurs observé dans la prochaine expérience.

Chapitre 7

Expérience comportementale 2 : Apprentissage de task-sets en présence de contextes

Dans cette deuxième expérience comportementale, nous testons les prédictions suivantes spécifiquement effectuées par notre modèle.

- Dans un environnement incertain, nous sommes capables d'utiliser l'information observée pour détecter la fin d'un épisode correspondant à un comportement. Nous pouvons utiliser cette détection pour effectuer un switch abrupt de comportement, au niveau hiérarchique des task-sets, plutôt que de s'adapter localement, au niveau hiérarchique des stimuli.
- Nous sommes capables d'utiliser cette faculté de détecter la nécessité de switcher pour stocker en mémoire des task-sets appris avant qu'ils ne soient effacés par adaptation et pour pouvoir ainsi envisager de les réutiliser.
- Nous explorons de manière diffuse et parallèle le répertoire de task-sets déjà appris, sommes ainsi capables de réactiver un task-set appris précédemment, ou de décider de l'opportunité d'apprendre un nouveau task-set.
- Nous sommes capables de découvrir qu'une entrée sensorielle porte une information contextuelle utile à la sélection a priori d'un task-set.
- Nous sommes capables d'utiliser cette information contextuelle ex-ante pour sélectionner

un task-set approprié en fonction du contexte.
– Nous avons un task-set par défaut, que nous ne quittons pas avant qu'il ne soit établi qu'il est nécessaire de switcher.

Afin de tester ces prédictions, nous soumettons des sujets à une expérience reposant sur le même principe que celui de la condition principale de la première expérience (apprentissage d'un nombre restreint de task-sets en milieu incertain), en ajoutant la présence d'information contextuelle qui peut permettre aux sujets qui découvrent son rôle, de sélectionner plus facilement les task-sets.

Nous présentons tout d'abord le protocole expérimental dessiné pour tester ces prédictions. Nous présentons ensuite les résultats comportementaux obtenus, en parallèle avec les prédictions de différents modèles.

7.1 Matériel et méthodes

7.1.1 Participants

Pour cette deuxième expérience, nous avons testé 57 sujets, âgés entre 18 et 35 ans ; les sujets ont été recrutés par des annonces affichées dans des universités. Les sujets ont été soumis à un entretien médical afin de s'assurer qu'ils ne possédaient pas d'antécédents psychiatriques ou neurologiques susceptibles d'influencer leur participation à l'expérience, ni de problèmes de vision des couleurs ou d'audition. Les sujets ont tous été informés des conditions de participation aux expériences au moyen une notice écrite d'information légale, validée par le comité d'éthique de l'INSERM et ont signé, en connaissance de cause, un consentement écrit avant leur inclusion dans le protocole. Les sujets ont été indemnisés pour leur participation (20 euros par session).

Parmi les 57 sujets, cinq n'ont pas été inclus dans les analyses pour cause d'erreur de manipulation sur une des sessions expérimentales. Trois n'ont pas été admis à passer la deuxième session pour cause de non compréhension des instructions à l'issue de la première. Nous avons donc 49 sujets (25 femmes , 24 hommes) dans nos analyses.

Il n'y a pas de recoupement entre les sujets de la première expérience et ceux de la deuxième

expérience.

7.1.2 Protocole

L'expérience se divise en deux sessions d'environ une heure chacune. Chaque session expérimentale se déroule dans une salle isolée, sur un ordinateur, avec un casque audio. Les deux sessions sont effectuées sur deux jours consécutifs.

A chaque essai, un stimulus (chiffre arabe) est présenté au sujet à l'intérieur d'un cadre carré situé au centre d'un écran. Il y a trois stimuli, pris dans l'ensemble $\{1, 2, 3\}$.

Les sujets répondent à ces stimuli en actionnant l'une des quatre touches $[d, f, j, k]$ d'un clavier azerty, selon la règle d'action suivante (favorisant une position naturelle fixe des mains) :
– d : majeur gauche
– f : index gauche
– j : index droit
– k : majeur droit
Ces quatre touches sont représentées à l'écran par quatre carrés disposés comme les touches, sous le domaine de présentation du stimulus.

En plus des stimuli, une information contextuelle, codée par une couleur, est présentée au sujet pendant chaque essai (voir figure 7.1, page 182). Pour une moitié des sujets, cette information contextuelle est présente par le fond d'écran coloré par une couleur parmi huit possibles. Pour l'autre moitié, le fond d'écran est noir, mais tous les traits (cadre central, touches, stimuli) sont de couleur, au lieu d'être blancs comme dans l'expérience précédente. Ces deux conditions mettent en avant deux heuristiques différentes sur la notion de contexte : l'une voyant un contexte comme un aspect sensoriel à large champ, comme le fond d'écran, mais qu'il est donc facile de négliger ; l'autre voyant le contexte comme un aspect quelconque de l'entrée sensoriel, sur lequel notre attention est facilement portée, comme dans beaucoup d'expériences de contrôle contextuel (Koechlin et al, 2003 [136]).

Après la sélection d'une action, le sujet reçoit un retour indiquant s'il a gagné ou perdu pour cet essai, sous trois formes simultanément :

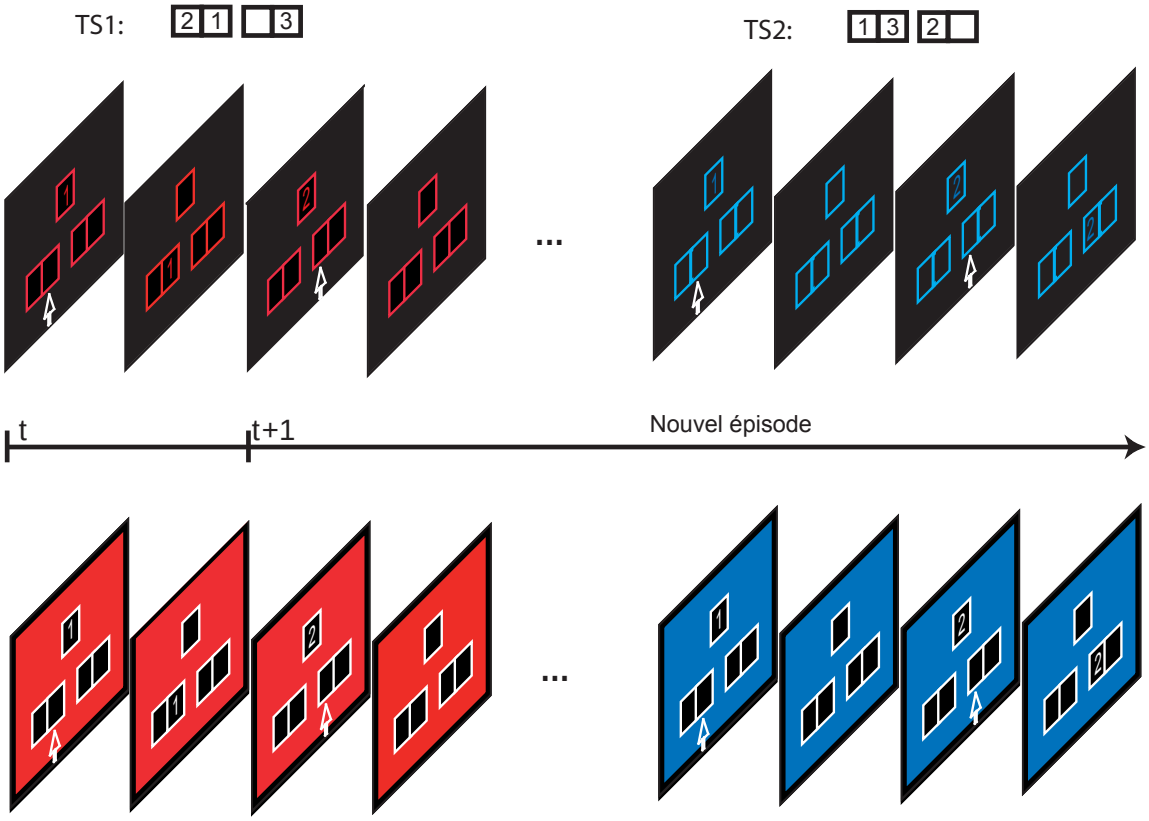


FIGURE 7.1 – Protocole expérimental. Le stimulus est présenté dans un carré central. Sur le schéma, la flèche indique l'action sélectionnée par le sujet. Si celle-ci correspond au task-set actuellement valide (représenté ici au dessus comme TS1), avec probabilité 0,9, le sujet reçoit un feedback sonore positif, et le stimulus se place dans la case correspondante (2e figure en partant de la gauche). Sinon, avec probabilité 0,9, le sujet reçoit un feeback sonore négatif et le stimulus disparaît simplement (4e écran). Une information contextuelle est fournie sous forme de couleur, soit par la couleur des traits (en haut), soit par la couleur du fond d'écran, en bas. Chaque couleur est associée à un unique task-set, mais un task-set peut être associé à plusieurs couleurs.

– Visuel : dans le cas *gagné*, le stimulus disparaît du carré central et vient se placer dans la case correspondant à la touche sur laquelle le sujet a appuyé. Dans le cas perdu, le stimulus disparait simplement. Ce feedback visuel a pour but d'aider le sujet à se souvenir de l'association entre le stimulus et l'action.
– Visuel : sous les quatre cases représentant les actions, une jauge indique l'accumulation du gain. Elle avance dans le cas *gagné*, recule dans le cas *perdu*. Lorsqu'elle atteint le maximum, elle revient à zéro. En pratique, cela arrive plusieurs fois par session. Ce feedback a pour but de motiver le sujet.
– Auditif : dans le cas *gagné*, le sujet entend un son montant rapide, dans le cas perdu, un son descendant.

Chaque essai dure 2 secondes. Le stimulus est présenté pendant 1,5 seconde, durant laquelle le sujet doit répondre. Le retour s'effectue 100ms après l'action effectuée par le sujet. Si le sujet ne répond pas à temps, aucun retour n'est fourni.

Chaque session se divise en 25 épisodes comprenant 36 à 54 essais, pour un total de 1134 essais par session. Pendant chaque épisode, un task-set précis est valide. Le retour correspondant à *gagné* ou *perdu* est donné en fonction de l'action prédite par ce task-set, avec un bruit de 10%, pseudo-randomisé de manière identique à la première expérience.

Les sujets ont pour instruction de tenter de gagner le plus souvent possible, sans manquer d'essais [1]. Ils ont à leur disposition 6 pauses par session, qu'ils peuvent prendre quand ils le souhaitent.

Les deux sessions sont dans la continuité l'une de l'autre.

Première session

Dans cette session, trois task-sets seulement sont utilisés pendant toute l'expérience. Ces task-sets, décrits dans le tableau ci-dessous sont totalement incongruents (identiques à la première expérience) :

1. Instruction exacte : *« Vous devez essayer de répondre le mieux et le plus vite possible. Votre indemnité dépendra de votre performance globale. »*

Actions	S_1	S_2	S_3
TS_1	a_1	a_2	a_3
TS_2	a_2	a_3	a_4
TS_3	a_3	a_4	a_1

Les task-sets TS_2 et TS_3 sont, chacun, valides dans 8 épisodes, le task-set TS_1 dans le premier épisode puis dans 8 autres épisodes.

Pendant la première session, quatre couleurs sont utilisées comme contextes. Pour les épisodes liés aux task-sets 2 et 3, on a une bijection contexte - task-set : le contexte est toujours la couleur C_2 pendant tous les essais des épisodes TS_2, toujours C_3 pendant tous les essais des épisodes TS_3.

Deux couleurs-contextes différentes peuvent être associées au TS_1. Plus précisément, pendant le premier épisode, la couleur C_1 est présentée pendant la moitié des essais puis la couleur C_1' pendant l'autre moitié des essais. Pour les 8 autres épisodes liés à TS_1, deux ont la couleur C_1 pendant la totalité de l'épisode, deux ont la couleur C_1' pendant la totalité de l'épisode, deux ont la couleur C_1 pendant la première moitié de l'épisode et la couleur C_1' pendant la première moitié de l'épisode, enfin, deux ont la couleur C_1' pendant la première moitié de l'épisode et la couleur C_1 pendant la première moitié de l'épisode.

Ce dessin permet de s'assurer que les sujets ne font pas de généralisation immédiate entre changement de couleur et changement d'épisode, puisque le premier changement de contexte n'est pas associé à un changement d'épisode.

A nouveau, les associations entre les stimuli d'une session et $\{S_1, S_2, S_3\}$ ainsi que celles entre les $\{a_1, a_2, a_3, a_4\}$ et les quatre touches sont tirées au hasard de manière à générer, dans la condition principale, à partir des trois task-sets présentés plus haut, tous les triplés de task-sets possibles respectant des conditions d'incongruence.

Les couleurs sont également tirées au hasard parmi un jeu de huit couleurs qui a été présenté au sujet avant l'expérience. Les sujets sont soumis au même entraînement préalable à la première session que pour la première expérience. Ils sont informés de la présence de couleurs et de la possibilité que la couleur change, mais pas du fait que celle-ci peut leur apporter une information. Ils sont soumis au même principe de motivation que pour la

première expérience : possibilité de doubler le montant de son indemnité en fonction de la performance.

A la fin de la première session, les sujets sont soumis à un débriefing actif : ils sont invités à rapporter les task-sets qu'ils se rappellent avoir utilisés avec succès. Ils sont également invités à indiquer quelles couleurs ils ont vues et si celles-ci avaient ou non, à leur avis, un rôle. Ils sont informés que la deuxième session continue sur la lancée de la première comme s'il n'y avait pas eu d'interruption.

Deuxième session

La deuxième session est effectivement dans la continuité exacte de la première session. Avant de débuter, les sujets sont simplement invités à relire les instructions de la veille et à prendre un instant pour se remémorer ce qu'ils ont fait. La première moitié de la deuxième session (13 premiers épisodes) est de structure complètement identique à la première session (les 3 mêmes task-sets, exactement ceux que les sujets ont vus pendant la première session, les 4 mêmes couleurs, les mêmes associations couleurs - task-sets).

Par la suite, nous appelons **période d'apprentissage** la période comprenant la première session et la première moitié de la deuxième session.

La deuxième moitié est une **période de test** permettant de mettre à l'épreuve les prédictions de notre modèle. Pour ce faire, les 12 épisodes de la deuxième moitié se séparent en 3 conditions :

- Condition contrôle, Ancien TS, Ancien Contexte (**TSaCa**) : pour quatre des 12 épisodes, le task-set TS_1 est valide, dans le contexte habituel pour TS_1. Les quatre épisodes sont ainsi $[TS_1, C_1]$, $[TS_1, C_1']$, $[TS_1, (C_1, C_1')]$ et $[TS_1, (C_1', C_1)]$. Cette condition permet de tester l'acquisition du task-set et de son association avec les contextes pendant la session et demie précédente.
- Condition Ancien TS, Nouveau Contexte (**TSaCn**) : pour six épisodes, les task-sets TS_2 et TS_3 sont valides, dans un nouveau contexte. Ainsi, les six épisodes sont $[TS_2, C_2']$ deux fois, $[TS_2, C_2'']$ deux fois et $[TS_3, C_3']$ deux fois. Cette condition permet de tester la capacité des sujets à réutiliser un task-set appris dans un nouveau contexte.

– Condition Nouveau TS, Nouveau Contexte (**TSnCn**) : pour deux épisodes, le nouveau task-set TS_4 est valide, dans un nouveau contexte : $[TS_4, C_4]$ à deux reprises. Cette condition permet de tester la capacité des sujets à apprendre des nouveaux task-sets, en plus du répertoire appris. Elle fournit également une ligne de base pour comparer l'avantage à l'adaptation fourni dans les deux autres conditions par la connaissance du task-set valide ou du contexte.

La période test utilise donc quatre task-sets (présentés dans la table ci-dessous) et 4 nouvelles couleurs.

Actions	S_1	S_2	S_3
TS_1	a_1	a_2	a_3
TS_2	a_2	a_3	a_4
TS_3	a_3	a_4	a_1
TS_4	a_4	a_1	a_2

A la fin de l'expérience, à nouveau, les sujets sont soumis à un post-test actif les invitant à rapporter les task-sets qu'ils se rappellent avoir utilisés avec succès pendant la session, ainsi que les couleurs qu'ils ont vues et leur lien, s'ils en ont perçu un, avec les task-sets. Ils sont également soumis à un débriefing informel.

L'expérience est présentée sur un ordinateur Mac, avec l'outil *Psychtoolbox* (http ://psychtoolbox.org) de *Matlab* (http ://www.mathworks.fr).

Les résultats de l'expérience sont analysés sur Matlab et SPSS (http ://www.spss.com/fr/).

Le modèle contextuel est testé sur le même protocole que les sujets, afin de pouvoir comparer les performances des sujets aux prédictions du modèle.

7.2 Résultats expérimentaux

7.2.1 Groupe entier

Période d'apprentissage, effets des contextes

La première session de cette expérience ne diffère de la condition principale de l'expérience précédente que par l'ajout d'information contextuelle. Nous commençons donc tout d'abord

par étudier l'effet qu'a l'ajout de cette information contextuelle sur la performance des sujets.

Notons tout d'abord que la performance globale n'est pas significativement améliorée ($p = 0.2$) par l'information contextuelle : les sujets font 74% de choix corrects dans la première session de l'expérience contextuelle, 71% de choix corrects dans la condition principale, première session, de l'expérience sans contextes. Afin d'observer si la légère différence (non significative) observée pourrait être un signe de l'utilisation de contextes, nous regardons plus spécifiquement, en fonction du numéro d'essai dans un épisode, la proportion de réponses correctes. On trace ces courbes dans la colonne de gauche de la figure 7.2. Pour la condition principale de l'expérience sans contextes, nous nous restreignons aux sujets qui ont effectué cette expérience en première session, afin de comparer des sujets au niveau de familiarité identique pour les deux expériences (ce qui explique la différence de taille dans l'erreur standard de la moyenne).

On observe une différence significative entre les deux courbes pendant environ les dix premiers essais après un changement d'épisode (et donc un changement de contexte). On observe également que le contexte n'a pas d'effet sur la performance asymptotique des sujets. Ces deux résultats sont cohérents avec le fait que, contexte présent ou absent, les sujets sont capables d'apprendre correctement un task-set au sein d'un épisode. Si le contexte apporte un effet, il est donc logique qu'il soit limité au début de l'épisode, et non à la fin, où la performance asymptotique est atteinte dans tous les cas.

On étudie, sur la période d'apprentissage de l'expérience contexte (session 1 et première moitié de session 2) l'effet à long terme d'évolution de la performance, en traçant la proportion de réponses correctes par épisode et par sujet en fonction du numéro d'épisode (figure 7.2, en haut à droite). On observe une augmentation significative de la performance au cours de la première session, passant d'environ 70% de bonnes réponses à presque 80% de bonnes réponses ($r = 0,19$, $t = 6,6$, $p < 10^{-7}$) indicative d'un effet d'apprentissage à long terme. Rappelons que l'on n'avait pas observé, pour la condition principale de l'expérience sans contexte, d'effets d'apprentissage à long terme.

On déduit de ces comparaisons entre deux expériences identiques à part l'ajout des contextes que, bien qu'il n'y ait pas d'effet des contextes sur la performance globale, ils semblent

malgré tout favoriser l'apprentissage à long terme et la vitesse d'adaptation initiale après un changement d'épisode.

Notons par ailleurs que ces résultats, ainsi que l'ensemble des résultats présentés ci-après, sont indépendant du type de contexte utilisé, contexte de type environnemental (fond d'écran) ou lié au stimulus (couleur des traits, voir figure 7.1, page 182). Nous analysons tous les résultats sans différencier le type de contexte

Changement de contexte

On regarde maintenant, indépendamment de l'expérience sans contexte, les effets spécifiques de changements de contexte.

En particulier, on peut observer l'effet d'un changement de contexte en l'absence de changements d'épisodes, ce qui permet de mesurer l'influence d'un changement dans une dimension sensorielle stable sur un changement de comportement. On trace dans la figure 7.2, en bas à droite, la proportion de réponses correctes spécifiquement dans le premier épisode TS_1 comprenant un changement de couleur (passage de C_1 à C_1' entre les essais 28 et 29, à Δ sur la figure.). On observe une chute de la performance de 90% à 71% de bonnes réponses, suivie d'une augmentation graduelle parallèle à l'apprentissage précédent.

Plus précisément, on compare la performance lors des trois essais précédents le changement de contexte (26-28, 89%), le premier essai de nouveau contexte (29, 71%) et les deux essais suivant le changement de contexte (30-31, 84%). La différence observée sur les 49 sujets est significative entre les trois essais précédant le changement et le premier essai du nouveau contexte ($t = 3,032$, $p = 0,003$), mais pas entre les essais précédant et suivant le changement ($t = 1,23$, $p = 0,2$).

Ce résultat permet d'infirmer fortement les prédictions du modèle RL. En effet, celui-ci prédit un retour de la performance au niveau initial à l'arrivée d'un nouveau contexte, ce qui est loin d'être le cas. Ce résultat permet également de confirmer en partie la prédiction de notre modèle portant sur le comportement par défaut. En effet, on voit que la performance après l'essai de changement de couleur n'est pas significativement moins bonne qu'avant, indiquant que les sujets ont continué à utiliser le task-set qu'ils utilisaient.

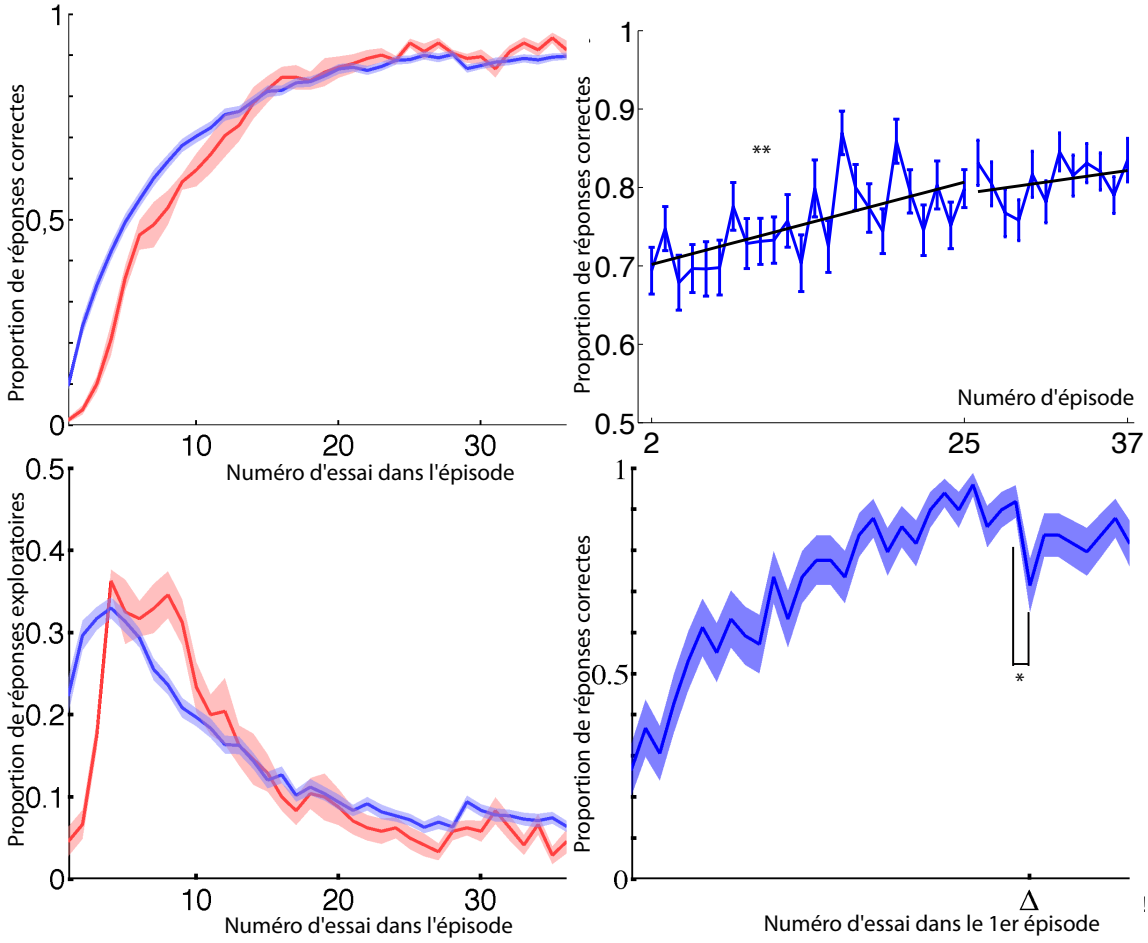


FIGURE 7.2 – Effets du contexte. Gauche : proportion d'essais corrects (haut) et exploratoires (bas) pour la première session de l'expérience contextes (bleu) et pour la condition principale de la première expérience (rouge) ; en fonction du numéro d'essai à partir du changement d'épisode. Droite, Haut : Performance moyenne par épisode, pour l'ensemble des sujets (bleu), première session et première moitié de la deuxième session, en fonction du numéro d'épisodes. Barres d'erreurs : erreur standard de la moyenne. Noir : droite de régression linéaire. Le coefficient est significativement positif pour la première session. Droite, bas : effet d'un changement de contexte, au cours du premier épisode, sur la proportion de réponses correctes, en fonction du numéro d'essai. L'aire autour de la courbe représente l'erreur standard de la moyenne.

Cependant, ce résultat met malgré tout en valeur un effet du changement de contexte sur le comportement, puisque l'essai de changement de contexte a une performance plus faible. Puisque la suite de la performance n'est pas altérée, il semble plus être révélateur d'un déplacement de l'attention sur le contexte impliquant une performance momentanément diminuée, que d'une décision de switcher après le changement de couleur.

Ce phénomène est à mettre en lien avec l'effet spécifique des changements de contextes mis en avant dans la description de notre modèle, notamment l'introduction du paramètre τ_Δ, qui introduit une hausse momentanée de la probabilité de transition d'un task-set à un autre et rend donc le modèle momentanément plus sensible à l'observation de retours inattendus, permettant éventuellement un switch plus rapide. On observe d'ailleurs dans la figure 7.2 à gauche, que contrairement à l'expérience sans contextes, il n'y a pas de plateau de persévération au début de l'épisode, indiquant que le switch se fait plus rapidement, suite au changement de contexte.

Période Test

La période test permet de tester deux prédictions indépendantes de notre modèle :
- Capacité à réutiliser des task-sets appris. Le modèle prédit que les sujets devraient s'adapter plus rapidement dans la condition TSaCn que dans la condition TSnCn, puisque dans un nouveau contexte, le modèle est capable de réutiliser un task-set déjà appris.
- Capacité à apprendre des associations contexte task-sets. Le modèle prédit que les sujets devraient s'adapter plus rapidement dans la condition TSaCa que dans la condition TSaCn, puisque le modèle peut utiliser l'information contextuelle pour découvrir plus efficacement quel task-set est valide.

On trace dans la figure 7.3 la proportion de réponses correctes et de réponses exploratoires en fonction du temps. On observe bien les deux effets prédits significativement, sur l'ensemble du groupe :
- Effet task-sets. La performance est très significativement ($p < 10^{-7}$) moins bonne dans la condition TSnCn (courbe violette) que dans la condition TSaCn (courbe jaune). L'exploration est également très significativement plus grande ($p < 10^{-7}$).
- Effet contextes. La performance moyenne est significativement ($p < 2.10^{-4}$) moins bonne

dans la condition TSaCn (courbe jaune) que dans la condition TSaCa (courbe bleue). L'exploration est également significativement plus grande ($p < 5.10^{-5}$).

On observe cependant que les effets sont nettement plus faibles que ceux qui seraient prédits par un modèle optimal. En effet, on pourrait attendre une performance asymptotique atteinte très rapidement après le changement d'épisode dans la condition TSaCa (bleue), si les sujets parvenaient à encoder parfaitement les associations contextes - task-sets, alors que ce n'est pas du tout le cas.

Notons par ailleurs que la proportion de réponses correctes au premier essai de la condition TSaCn est significativement plus grande que la proportion de réponses correctes au premier essai de la condition TSnCn. Pourtant, du point de vue du sujet, ces deux séries de premiers essais sont identiques : dans les deux cas, la couleur présentée est nouvelle. Cela confirme deux choses :

– l'effet du changement de contextes. La performance asymptotique est d'environ 93%, indiquant 7% d'erreurs, ou de distractions du sujet malgré un task-set appris. La proportion de réponses correctes au premier essai de la condition TSaCn est d'environ 17%, parmi lesquels dix points ne peuvent être attribués à la distraction. Cela confirme donc que le changement de contexte (en particulier l'apparition d'un nouveau contexte) a un effet supplémentaire sur le comportement, ce qui peut être modélisé par une valeur $\tau_\Delta < 1$.
– l'existence d'un biais de bas niveau. La proportion de réponses correctes au premier essai de la condition TSnCn est nulle, alors qu'elle est de 17% pour le premier essai de la condition TSaCn, bien que ces deux essais soient identiques du point de vue du sujet. Cela montre que les sujets testent plus les actions qui ont déjà été récompensées pour un stimulus que les autres. Cela valide l'introduction d'un biais de bas niveau dans l'initialisation du TS_{test} dans notre modèle.

Afin de préciser l'effet contexte et l'effet task-set observés faiblement, nous effectuons, comme précédemment, une division de notre groupe de sujets en plusieurs sous-groupes.

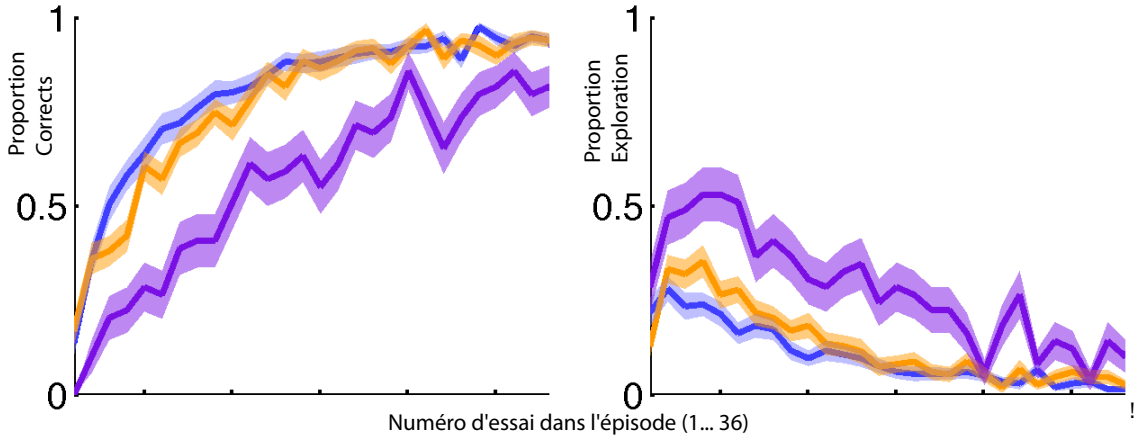


FIGURE 7.3 – Proportion de réponses correctes (gauche) ou exploratoires (droite) en fonction du numéro d'essai (1 à 28) après un changement d'épisode. Bleu : condition TSaCa (4 épisodes). Jaune : condition TSaCn (3 épisodes). Violet : condition TSnCn (1 épisode).

7.2.2 Fitting du modèle aux données de sujets, debriefing

Debriefing

A nouveau, les debriefing informels sont extrêmement indicatifs d'une variabilité comportementale très forte entre les sujets. Après la première session, ce qu'ont perçu les sujets de l'expérience va d'un extrême où ils ont correctement identifié trois task-sets, quatre couleurs et leurs associations, à l'autre extrême où ils ne se sont pas rendu compte que seul un nombre limité de task-sets était utilisé ni que les couleurs étaient informatives pour l'expérience, et même où certains sujets ne rapportent pas correctement les couleurs qu'ils ont vues.

Après la deuxième session, si une majorité des sujets (mais pas tous) a perçu le fait que les changements de contextes étaient fréquemment associés aux changements d'épisodes, la variabilité dans la compréhension de la structure de la tâche reste grande. Quelques sujets ont rapporté parfaitement les quatre task-sets, huit couleurs et leurs associations correctes. A l'autre bout du spectre, d'autres ne peuvent pas rapporter un seul task-set qu'ils ont utilisé et n'ont aucune idée quant au rôle des couleurs, ou au contraire, rapportent une

dizaine de task-sets qu'ils pensent avoir utilisé avec succès.

L'impression qui ressort aux expérimentateurs après debriefing des sujets est le fait qu'on puisse séparer deux types d'effets. Le premier porte sur l'apprentissage des task-sets, comme dans la première expérience : certains sujets parviennent à apprendre et stocker et utiliser les task-sets, d'autres non. Le deuxième porte sur l'apprentissage des contextes : parmi les sujets qui parviennent à apprendre les task-sets, certains parviennent à apprendre à les associer aux contextes, d'autres non.

Afin de séparer ces effets, nous utilisons le post-test pour séparer les sujets en trois groupes : un groupe qui acquiert les contextes et leurs associations aux task-sets (groupe 1), un groupe qui acquiert les task-sets mais pas les contextes (groupe 2) et un groupe qui n'acquiert pas les task-sets (groupe 3).

Le critère de séparation du groupe 1 est l'acquisition de l'association testée dans la condition TSaCa de la période test (barres grises dans la figure 7.4). Le critère de séparation du groupe 3 est une mauvaise acquisition de task-sets (moins de deux task-sets rapportés, ou nombre de task-sets faux rapportés important, barres marron dans la figure 7.4). Entre les deux le groupe 2 a une acquisition correcte des task-sets (au moins trois), mais pas des associations contexte - task-set.

Notons que la séparation en trois groupes ne présume pas du fait que cette division est optimale, ni du fait que la répartition est probablement plus continue. Cette séparation est un parti pris permettant d'isoler deux effets distincts, l'effet task-set et l'effet contexte.

Notons par ailleurs, qu'aucune corrélation n'a été trouvée entre la répartition des sujets dans les groupes et différents facteurs simples disponibles pour vérification a posteriori (âge, sexe, niveau d'éducation, type de contexte présenté).

Fittings

A nouveau, nous fittons, sujet par sujet, le modèle proposé sur les données expérimentales. Les paramètres variables sont α, β, ϵ, τ_Δ, p_{test} et α_C. La procédure de fitting est identique à celle décrite dans l'expérience sans contextes.

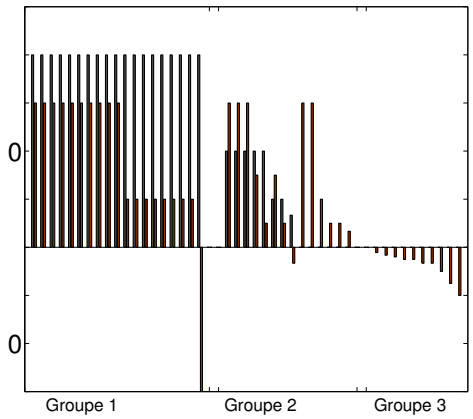


FIGURE 7.4 – Distribution des notes par critère (notes de contexte en gris, notes de task-sets en marron) pour les sujets des trois groupes.

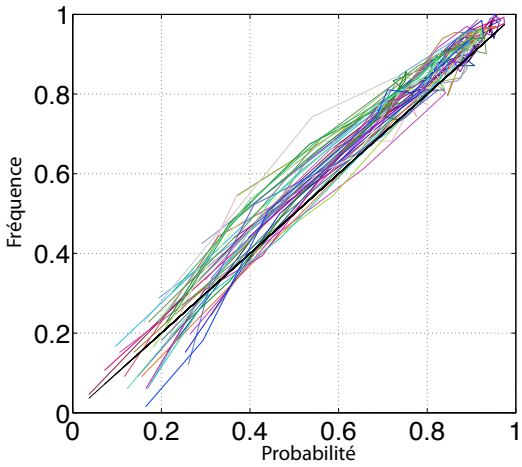


FIGURE 7.5 – Graphes de fitting probabilité - fréquence. En fonction de la probabilité de répondre correct prédite par le modèle fitté sur les données d'un sujet, on trace la fréquence de réponses correctes de ce sujet. Chaque couleur représente un sujet différent.

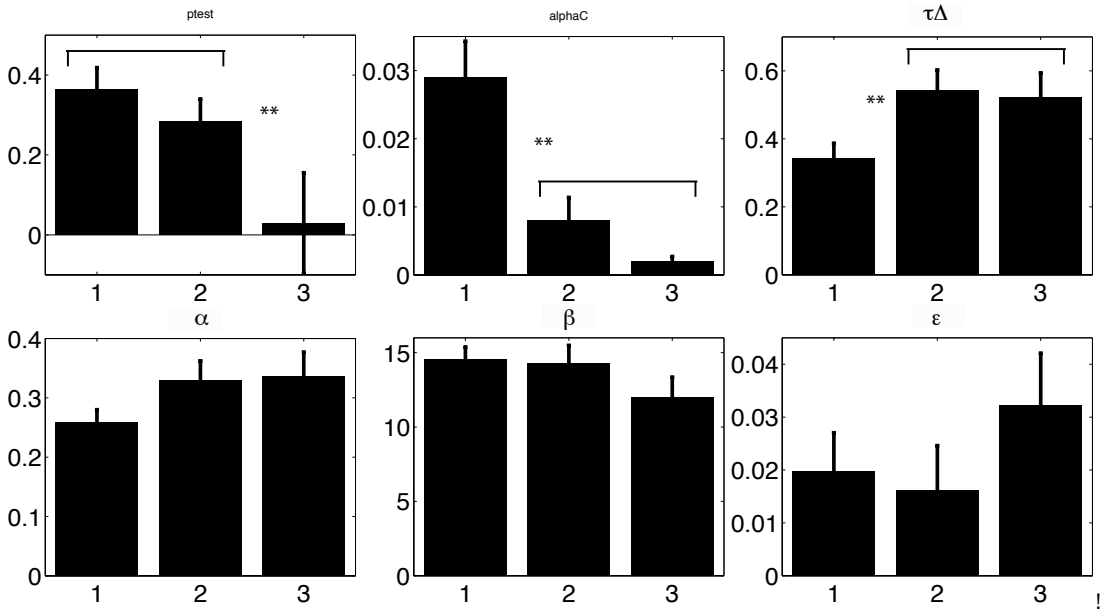


FIGURE 7.6 – Valeur des paramètres fittés, par numéro de groupe. Les barres d'erreur représentent l'erreur standard de la moyenne. Première ligne : p_{test}, α_C, τ_Δ. Deuxième ligne : α, β, ϵ.On observe une différence significative entre le groupe 1-2 et le groupe 3 pour p_{test}, entre le groupe 1 et le groupe 2-3 pour les paramètres α_C et τ_Δ.

Dans la figure 7.6, nous traçons la valeur des paramètres fittés en fonction des groupes. Nous n'observons pas de différence significative entre les trois groupes pour les paramètres (β et ϵ).

Pour le paramètre p_{test} (biais initial sur la valeur de confiance du TS_{test}), nous observons une différence significative entre les groupes 1-2 et le groupe 3 ($p = 0,039$). Cela confirme, comme dans l'expérience précédente, son rôle essentiel dans le biais entre le fait de stocker et utiliser des task-sets (groupes 1-2) ou de réapprendre à chaque épisode un nouveau task-set (groupe 3).

Pour le paramètre α_C (vitesse d'apprentissage des associations contexte-task-set), nous observons des différences significatives entre les trois groupes ($p < 0,03$). En particulier, le groupe 1, qui rapporte correctement des associations contexte - task-set pendant le post-test, correspond à une valeur de α_C plus élevée, ce qui permet de valider la pertinence des groupes effectués et des paramètres du modèle.

Pour le paramètre τ_Δ, nous observons une différence significative entre le groupe 1 et les groupes 2 et 3 ($p = 0,01$), avec une valeur plus faible pour le groupe 1. Cela est pertinent avec l'idée que ce groupe est le seul à percevoir le rôle des contextes correctement, et en particulier le rôle des changements de contextes qui sont dans cette expérience fréquemment associés à des changements d'épisodes.

Pour le paramètre α, nous n'observons pas de différence significative entre les groupes 1 et 2, et 2 et 3, mais une différence significative entre les groupes 1 et 3 ($p = 0,028$, Mann-Wittney). Ce résultat est cohérent avec l'analyse selon laquelle les sujets des groupes 2 et plus particulièrement 3, préfèrent utiliser une structure plus complexe pour résoudre le problème posé par l'expérience plutôt que de réapprendre et ont donc besoin de moins d'efficacité d'apprentissage local ; tandis que les sujets du groupe 3 n'utilisent pas de structure hiérarchique (task-sets) et ont donc besoin d'un apprentissage local plus efficace pour obtenir une performance comparable.

Les groupes effectués par debriefing semblent donc cohérents avec le fitting du modèle effectué, le paramètre p_{test} étant responsable de l'effet task-sets, le paramètre α_C de l'effet contextes. On montre dans la section suivante que cela se retrouve effectivement dans les

données comportementales des sujets.

7.2.3 Variabilité interindividuelle

Dans cette section, nous analysons les données en tenant compte des trois groupes effectués et les comparons à notre modèle, en utilisant les paramètres fittés.

Apprentissage à long terme

Tout d'abord, on observe l'évolution à long terme de la performance par épisode en traçant, pour chaque groupe, en fonction du numéro d'épisode dans la période d'apprentissage, la proportion de réponses correctes pendant l'épisode et pour les sujets du groupe. On voit dans la figure 7.7, page 198, que le pattern est différent en fonction du groupe. Pour le groupe 1, on a une augmentation significative de la performance pendant la première session ($r = 0,25$, $p < 3.10^{-5}$) et pendant la première moitié de la deuxième session ($r = 0,23$, $p < 0,002$). Pour le groupe 2, on a également une augmentation significative de la performance pendant la première session ($r = 0,23$, $p < 5.10^{-4}$), mais pas pendant la deuxième session ($p > 0,4$). Enfin, pour le groupe 3, on n'a aucune progression significative ($p > 0,1$, $p > 0,3$).

Ce pattern reflète le fait qu'une progression à long terme nécessite d'acquérir des informations au cours de l'expérience, ce que font dans une certaine mesure les sujets du groupe 2 et plus encore ceux du groupe 1, mais pas ceux du groupe 3.

Période test

La période test est spécifiquement conçue pour pouvoir tester les deux effets (acquisition des task-sets et apprentissage des contextes) pour lesquels on a mis en évidence, dans le débriefing, le post-test et le fitting des paramètres du modèle, une forte variabilité interindividuelle. Nous traçons donc dans la figure 7.8 la performance et la proportion d'exploration de chaque groupe dans chacune des conditions de la période test :

– TSaCa : ancien task-set, ancien contexte. Quatre épisodes sur le task-set TS_1, avec les contextes appris C_1 ou C'_1.

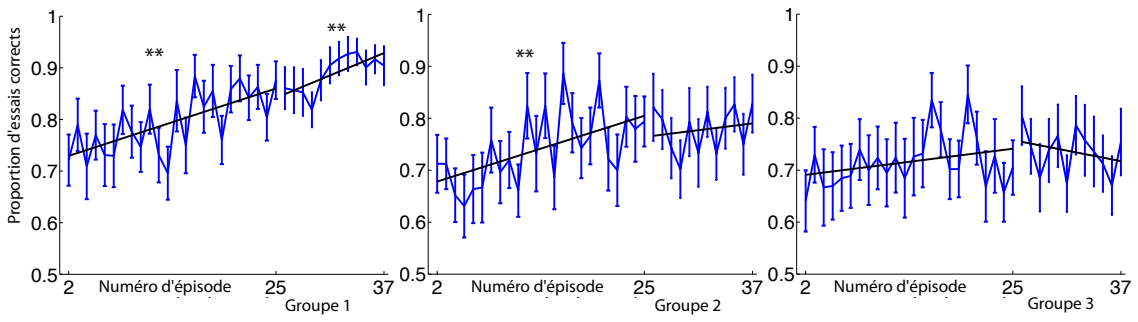


FIGURE 7.7 – Evolution de la performance moyenne par épisode pendant la période d'apprentissage, en fonction du numéro d'épisode. Les barres d'erreur représentent l'erreur standard de la moyenne. De gauche à droite, groupes 1, 2 et 3.

– TSaCn : ancien task-set, nouveau contexte. Trois épisodes sur les task-sets TS_2 ou TS_3, avec trois nouvelle couleurs-contextes.
– TSnCn : nouveau task-set, nouveau contexte. Un épisode sur le task-set TS_4 avec une nouvelle couleur-contexte.

On observe bien la dissociation comportementale correspondant au debriefing et prévue par le modèle, étant donnés les paramètres fittés :

– Pour le groupe 1, on observe un effet significatif de l'acquisition des task-sets (TSaCn $>$ TSnCn) pour la performance ($t = 8,9$, $p < 10^{-7}$) et pour l'exploration ($t = 5,3$, $p < 5.10^{-5}$). On observe également un effet des contextes (TSaCa $>$ TSaCn) sur la performance ($t = 5,6$, $p < 3.10^{-5}$) et sur l'exploration ($t = 4,6$, $p < 2.10^{-4}$).
– Pour le groupe 2, on observe un effet significatif de l'acquisition des task-sets (TSaCn $>$ TSnCn) pour la performance ($t = 8,55$, $p < 10^{-6}$) et pour l'exploration ($t = 5,72$, $p < 5.10^{-5}$). On n'observe pas d'effet des contextes sur la performance ($t = 1,2$, $p > 0,2$) ni sur l'exploration ($t = 1,42$, $p > 0,3$).
– Pour le groupe 3, les résultats sont légèrement moins clairs. Si, essai par essai, il ne semble pas y avoir d'avantage d'une condition sur l'autre, on trouve malgré tout des effets globaux. On observe notamment un effet significatif *négatif* des contextes (TSaCa $<$ TSaCn) pour la performance ($t = 3,2$) ; il n'est pas significatif pour l'exploration ($t = 1,8$). Cela indique en tout cas que l'information contextuelle n'aide pas ces sujets. Quant à l'effet task-set, on l'observe très légèrement significatif ($t = 2,19$, $p = 0,05$) pour

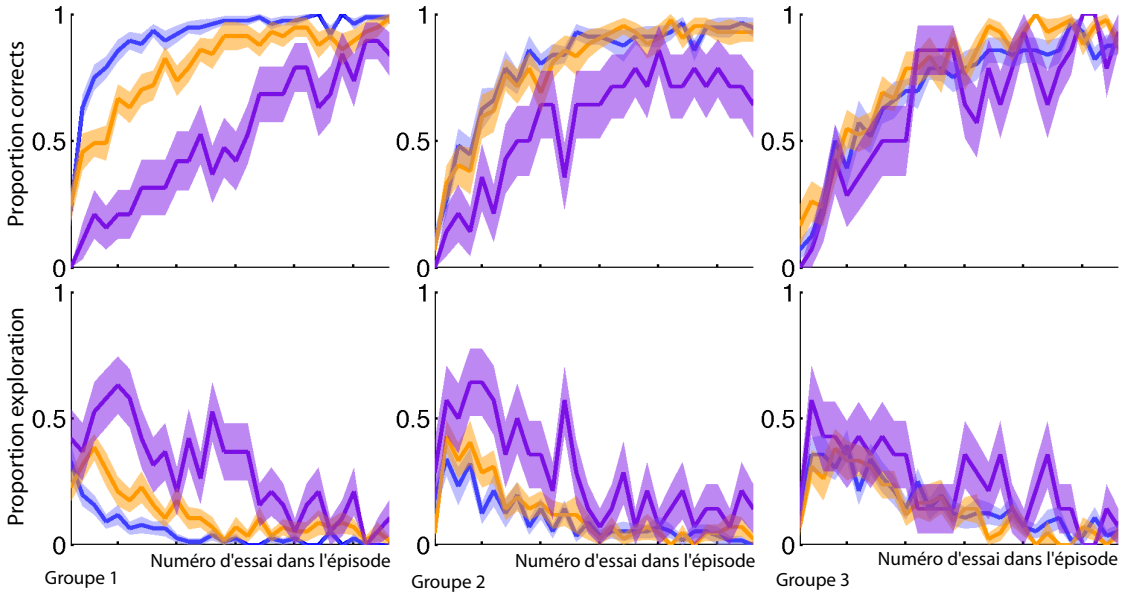


FIGURE 7.8 – Proportion de réponses correctes (haut) et d'exploration (bas) pour les groupes 1, 2 et 3, de gauche à droite. Abscisse : numéro d'essai après le changement d'épisode, [1 : 28]. Aires : erreur standard de la moyenne. Bleu : condition TSaCa. Jaune : condition TSaCn. Violet : condition TSnCn.

l'exploration, mais non significatif pour la performance ($t = 1, 4$).

Validation du modèle sur le comportement des sujets

On compare la performance des sujets à celle du modèle, en utilisant les paramètres fittés pour chacun des trois groupes, dans la figure 7.9. On reproduit avec les simulations les effets observés comportementalement, ce qui valide les prédictions de notre modèle.

On voit ainsi que, essentiellement, les paramètres du modèle α_C (apprentissage des associations contexte-task-set) et p_{test} (biais sur l'acquisition et l'utilisation d'un répertoire de task-sets, contre l'apprentissage à nouveau) permettent de rendre compte de la variabilité comportementale observée sur les sujets, qui se résume en deux effets principaux : capacité d'acquérir et de réutiliser des task-sets, capacité d'apprendre des associations contexte - task-sets. En effet, on voit dans la figure 7.10 que, aux autres paramètres fixés, varier α_C

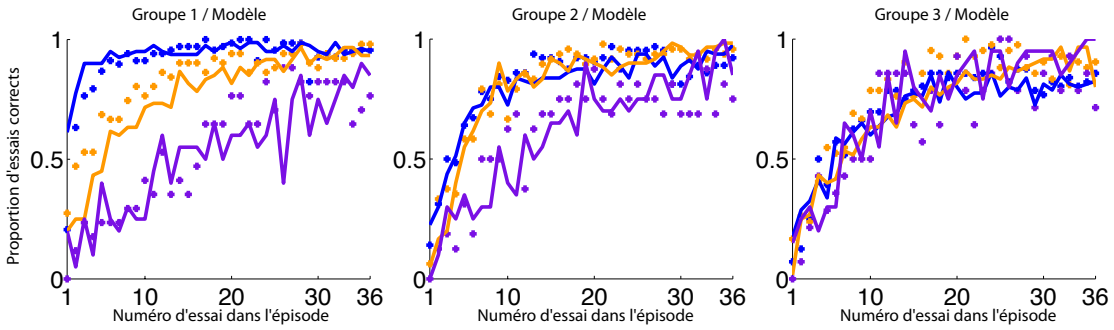


FIGURE 7.9 – Croix : Données comportementales, pour les groupes 1, 2 et 3 de gauche à droite. Trait plein : simulations du modèle pour les paramètres fittés sur les données des trois groupes. Abscisse : numéro d'essai après le changement d'épisode. Ordonnée : proportion d'essais corrects pour les conditions TSaCa (bleu), TSaCn (jaune) et TSnCn (violet).

et p_{test} seuls permet de reproduire les effets observés.

7.3 Conclusions de l'expérience avec contextes

Cette expérience permet donc de valider les prédictions du modèle. Elle montre que nous utilisons effectivement un task-set par défaut dont nous ne switchons qu'en cas de nécessité, même si des éléments surprenants dans le champ sensoriel peuvent faire diminuer l'attention et sensibiliser à la possibilité de switcher. Elle montre que nous pouvons apprendre un répertoire de task-sets et en tirer profit pour les réutiliser en cas de besoin, même dans un nouveau contexte. Elle montre que nous sommes capables de découvrir qu'une dimension sensorielle apporte de l'information contextuelle, et d'apprendre à utiliser cette information contextuelle a priori, pour la sélection des task-sets.

Par ailleurs, le modèle permet de capturer dans deux paramètres la variabilité comportementale observée entre les sujets. L'un des deux paramètres correspond au même effet que celui mis en évidence dans l'expérience précédente : acquisition des task-sets, par l'intermédiaire d'un biais gérant l'équilibre entre la tendance à réessayer des task-sets précédemment appris, ou à apprendre quelque chose de nouveau. Le deuxième paramètre correspond simplement à l'apprentissage de l'information portée par les contextes.

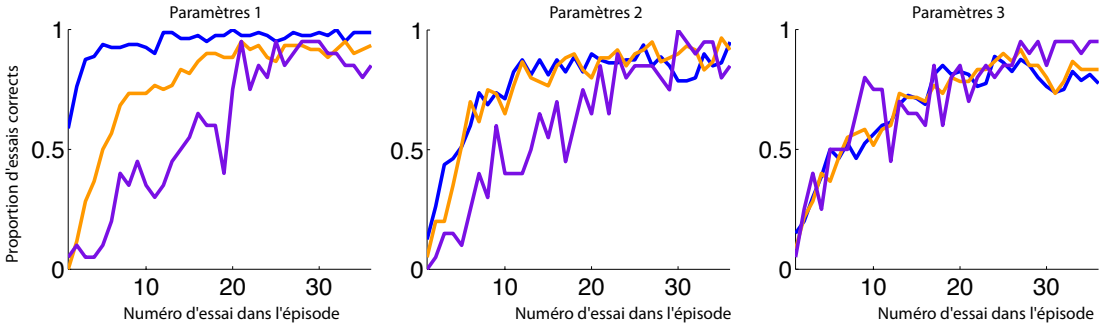


FIGURE 7.10 – Simulations du modèle pour les paramètres α_C ($[0,03; 0,001; 0,001]$) et p_{test} ($[0,4; 0,4; 0,01]$) seuls variables. Abscisse : numéro d'essai après le changement d'épisode. Ordonnée : proportion d'essais correctes pour les conditions TSaCa (bleu), TSaCn (jaune) et TSnCn (violet).

Notons à nouveau que les trois groupes isolés pour séparer les deux effets différents (effet contextes et effet task-sets, groupe 1 ; effet task-sets, pas d'effet contextes, groupe 2 ; pas d'effet task-sets ni contextes, groupe 3) ne sont, encore pas des groupes de niveau, bien que leur performance dans le cadre de cette expérience soit effectivement ordonnée. On peut montrer cela spécifiquement dans le cas du groupe 3, pour la condition TSnCn. Bien que l'effet ne soit pas significatif, dû à un nombre trop faible de données, on observe une tendance montrant que le groupe 3 est meilleur dans cette condition que les deux autres groupes. Cette observation n'est pas surprenante puisque ce groupe tend à apprendre quelque chose de nouveau, plutôt que d'essayer de réutiliser des schémas anciens comme les deux autres groupes. On voit bien ainsi que dans un environnement changeant, ce groupe serait favorisé, alors qu'il ne l'est pas dans notre expérience, qui favorise la réutilisation de task-sets.

Il est, par ailleurs, très probable que, plutôt que des groupes distincts, on ait un continuum de comportements entre ces trois extrêmes représentatifs de la présence ou de l'absence des deux effets recherchés.

Conclusion des expériences

Les deux expériences effectuées valident les prédictions effectuées par notre modèle et soutiennent ainsi, même si elles ne les démontrent pas, les hypothèses sur lesquelles il repose pour expliquer l'intégration du contrôle et de l'apprentissage. Elles soutiennent un principe d'économie : nous n'impliquons le contrôle cognitif que lorsque c'est vraiment nécessaire, n'activant autrement qu'un seul comportement en tâche par défaut et switchant uniquement quand nécessaire, c'est à dire lorsque le modèle actuel n'est plus prédictif de ce qui est observé. Elles soutiennent le principe d'une exploration indirecte et parallèle des différentes options, par un processus dans lequel on apprend a priori quelque chose de nouveau influencé initialement par ce qui a fonctionné précédemment, sauf si on se rend compte rapidement qu'il semble correspondre à quelque chose déjà appris. Elles soutiennent le principe d'un parallélisme hiérarchique entre la sélection d'une action face à un stimulus et la sélection d'une tâche face à un contexte, avec apprentissage de l'un par des renforcements externes et apprentissage de l'autre par un renforcement interne, la confiance dans le modèle actuel.

Quatrième partie

Discussion

Chapitre 8

Discussion

8.1 Résumé

Nous avons posé, dans cette thèse, la question de l'apprentissage de task-sets dans le but de leur utilisation pour le contrôle cognitif.

Nous avons proposé une théorie d'apprentissage, d'exploration et de contrôle, permettant d'apprendre ou de mettre à jour, par essais-erreur, un task-set ; de décider de la construction d'un nouveau task-set à ajouter au répertoire disponible quand nécessaire ; et de switcher quand nécessaire, pour réutiliser un task-set déjà appris.

La théorie proposée est le modèle minimal incluant ces caractéristiques, présentes séparément dans trois autres modèles auxquels elle peut être artificiellement réduite.

- En supprimant le rôle de *switch* joué par le signal de confiance, le modèle reste perpétuellement sur un seul task-set qui sélectionne les actions et s'ajuste, et donc s'adapte, ainsi que le modèle RL, aux changements d'environnements de renforcement.
- En supprimant le stockage des task-sets après un switch, le task-set test émerge toujours comme task-set par défaut. Notre modèle est ainsi réduit au modèle UU.
- En supprimant la singularisation d'un task-set par rapport aux autres, introduite par le rôle d'un task-set par défaut, on supprime la possibilité d'une période d'exploration, donc d'un agrandissement de l'espace des task-sets. On obtient donc un modèle de type

MMBRL.

Nous avons montré que la théorie proposée permettait effectivement de répondre à la question posée dans la thèse : elle permet d'apprendre un répertoire de task-sets, à partir de renforcements bruités, et d'apprendre leurs associations éventuelles aux contextes dans lesquels ils sont valides. Elle permet également d'apprendre à utiliser ces task-sets en fournissant intrinsèquement un mécanisme d'exploration efficace des task-sets déjà appris. Enfin, elle permet une grande flexibilité et une certaine économie, puisqu'elle assure la construction d'un répertoire ni plus, ni moins grand que nécessaire.

Nous avons testé, dans deux expériences comportementales, les prédictions effectuées par la théorie proposée. Nous avons montré que le comportement des sujets permettait de valider ces prédictions, contrairement à celles des autres modèles testés. La théorie permet par ailleurs de rendre compte de la grande variabilité comportementale observée à l'intérieur de l'ensemble de sujets.

Dans les prochaines sections, nous discutons les résultats présentés jusqu'ici. Tout d'abord, nous discutons les limitations de notre théorie et proposons de distinguer les limitations dues à des facilités d'implémentation, de celles intrinsèques à la théorie. Ensuite, nous discutons de la plausibilité biologique de cette théorie : les calculs proposés peuvent-ils effectivement être implémentés dans le cerveau humain ? Quelles régions du cerveau sont impliquées dans le maintien et l'utilisation des informations postulées par notre théorie ? Enfin, nous proposons des pistes quant à l'origine de la variabilité comportementale, observée chez les sujets, expliquée par des paramètres du modèle, mais non par des arguments biologiques.

8.2 Limitations

L'implémentation de la théorie d'intégration du contrôle et de l'apprentissage proposée dans cette thèse est nécessairement limitée, parfois par des choix de simplicité. Nous explicitons dans cette partie les limitations de notre modèle. Nous explicitons également la distinction entre celles qui relèvent d'un choix pour simplifier l'implémentation et ne sont donc pas des limitations théoriques, par rapport à celles qui, effectivement, sont des limitations théoriques.

8.2.1 Fini ou continu ? Task-sets, renforcements

Dans l'implémentation du modèle proposé, pour les simulations et les expériences, nous nous sommes limités à un renforcement binaire, indiquant seulement *gagné* (1) ou *perdu* (0). Pourtant, nous sommes souvent confrontés, dans la vie courante, à des retours de l'environnement plus complexes et notamment plus continus, avec une valeur quantitative portant une information (quantité d'argent, degré de satisfaction, etc.). On peut donc légitimement demander si ce choix d'implémentation constitue une limitation théorique de notre modèle, ou seulement une simplification.

L'estimation de la confiance dans les task-sets, essentielle à la théorie proposée, requiert que chaque task-set soit équipé d'un modèle direct interne, lui permettant de prédire (lorsque ce task-set est adapté aux contingences extérieures) les conséquences de la sélection d'une action face à un stimulus. C'est ce point clé qui permet l'estimation bayésienne de la confiance dans les task-sets, selon la formule $P(TS|observation)\ \alpha\ P(observation|TS)P(TS)$. En particulier, ce qui est requis par notre théorie est donc la capacité d'estimer, après l'observation des conséquences d'une action, la probabilité $P(observation|TS)$ dans le cadre d'un modèle direct lié au TS, ainsi que d'apprendre ce modèle direct.

Dans le cadre de renforcements binaires, ces deux exigences sont simplement remplies. Le modèle apprend les probabilités de renforcement ($\gamma(r, s, a)$) de manière fréquentielle, et peut donc estimer très simplement, après chaque essai, quelle était la probabilité de la conséquence observée.

Cependant, un modèle interne pourrait également être appris dans le cadre d'une fonction de renforcement plus complexe. Typiquement, dans Doya et al, 2002 [75], les auteurs utilisent le MMBRL et modélisent une erreur de prédiction gaussienne, permettant le calcul de la probabilité explicitement par $\exp(-\Delta_i^2/\sigma^2)$, où Δ est la distance entre l'état prédit et l'état observé. Bien que ce ne soit pas le cas dans Doya et al, 2003 [75], l'espérance (état prédit) et la variance (σ^2) de cette gaussienne pourraient être apprises fréquentiellement. On pourrait ainsi appliquer ce principe à l'observation de renforcements quantitatifs, et donc non plus simplement binaires.

Le fait qu'on ait choisi des renforcements binaires n'est donc pas un point crucial de la

théorie proposée. Par opposition, la présence d'un choix binaire quant à la sélection des task-sets est un point clé, qui est complètement indépendant de l'implémentation du modèle avec de renforcements binaires.

Rappelons en effet, que nous imposons, pour la sélection des task-sets, un seuil $\lambda > 0,5$ signifiant qu'on a plus confiance dans le fait qu'il soit approprié qu'inapproprié aux contingences extérieures. Ce seuil implique un choix binaire : être dans une période d'exploitation (un task-set est singularisé comme le task-set correct), ou ne pas y être. Au lieu d'utiliser la confiance ex-ante dans sa dimension continue (comme c'est effectué dans le MMBRL, par exemple, pour pondérer les choix et l'apprentissage), elle est utilisée pour effectuer un choix binaire. C'est cet aspect *oui ou non* qui permet d'introduire naturellement la notion de switch à la fin d'une période d'exploitation (ainsi qu'observé comportementalement) ; qui permet également d'introduire naturellement la période d'exploration et ainsi de proposer un mécanisme de construction de l'espace de task-sets, une méthode d'exploration implicite et parallèle des différentes possibilités, incluant la possibilité d'apprentissage d'un nouveau comportement.

La présence d'une décision discrète, au niveau hiérarchique plus élevé des task-sets semble donc être un point essentiel. Loin d'être une limitation, c'est au contraire un aspect crucial du modèle, qui permet l'intégration de l'apprentissage, de l'exploration et du contrôle.

8.2.2 Mémoire à long terme des TS

Nous construisons probablement, au cours de notre vie, un nombre très important de task-sets. Il semble peu crédible que l'ensemble de ces task-sets soient mis en concurrence et évalués à tout instant où du contrôle cognitif est nécessaire. Nous soutenons que le rôle du contexte permet de justifier le fait qu'à un moment donné, seul un petit nombre de task-sets est considéré par le cerveau et par notre modèle. En effet, la plupart des task-sets qui nous sont disponibles dans notre répertoire ne seront pas appropriés dans un contexte donné, impliquant que leur confiance a priori sera extrêmement faible et non dissociable du bruit inhérent au cerveau. On peut ainsi supposer que seul un nombre limité de task-sets émergent du bruit ambiant, sont stockés en mémoire de travail et se concurrencent effectivement pour savoir lequel est approprié dans l'environnement présent.

Cette hypothèse est soutenue par des travaux récents sur la mémoire de travail (McNab et Klingberg, 2008 [145], Bays et Husain, 2008 [14]). Ces études montrent que la mémoire de travail, plutôt que d'être limitée à une capacité finie, est limitée par une ressource globale finie, de telle sorte qu'un nombre non limité d'objets se partage dynamiquement cette ressource et peut être maintenu en mémoire de travail, chacun plus ou moins précisément en fonction de la proportion de ressource qui lui est allouée. Dans le cas des task-sets, le facteur de confiance dans les task-sets pourrait déterminer l'allocation de ressource, donc quels task-sets - existants dans la mémoire à long terme - émergent effectivement en mémoire de travail.

Cette hypothèse justifie deux approximations utilisées lorsque nous simulons le modèle.

La première consiste à débuter une simulation avec $n = 2$ task-set, l'un étant le TS_0 représentant le hasard, l'autre le premier task-set à apprendre. On considère en effet que, puisque les sujets n'ont a priori pas déjà appris de task-set correspondant à la situation expérimentale dans laquelle ils sont placés, aucun des task-sets de leur répertoire n'émerge et qu'ils commencent donc initialement seulement avec un task-set à apprendre.

La deuxième approximation consiste à limiter artificiellement la mémoire à long terme du modèle. En effet, si l'ordinateur a une mémoire parfaite et peut maintenir les calculs de λ même pour les valeurs les plus faibles, les sujets n'ont pas une mémoire aussi parfaite (Conway et al, 2003 [46]) et nous avons proposé que probablement, seules quelques valeurs les moins faibles émergeaient du bruit ambiant, correspondant à la capacité de la mémoire de travail. Nous avons donc, pour les simulations, imposé une limite au nombre de task-sets stockés par le modèle, sur lesquels il effectue les calculs de confiance.

Le valeur de cette limite (que nous avons testée entre 4 et 10) n'influence quasiment pas le comportement du modèle. Un effet notable de l'ajout de cette limite est une légère amélioration de sa performance dans l'expérience sans contextes, session contrôle. En effet, lorsque le modèle a une mémoire à long terme parfaite, il subit un effet d'interférence entre des task-sets qu'il a utilisés plusieurs épisodes auparavant, quasi identiques à ceux qu'il tente d'apprendre présentement. Cela ne pourrait pas se produire avec les sujets, qui n'ont pas la mémoire parfaite de l'ordinateur.

Notre théorie propose de modéliser l'interaction entre le contrôle cognitif et l'apprentissage, et non la mémoire à long terme des sujets. Comme aucun mécanisme d'oubli n'est modélisé, nous avons observé, paradoxalement, un effet négatif de la mémoire à long terme parfaite du modèle sur le contrôle cognitif. Limiter le nombre de task-sets stockés et utilisés par le modèle est donc une méthode simple de contourner ce problème.

Nous avons par ailleurs spécifiquement tenté de limiter la difficulté des problèmes de mémorisation des sujets dans le dessin expérimental. En effet, le déplacement d'un stimulus vers la case représentant l'action sélectionnée leur permet d'avoir une représentation visuelle du task-set, plus facile à mémoriser que la représentation purement motrice.

8.2.3 Généralisations

Un autre effet qui n'est pas modélisé, est un phénomène de généralisation que les sujets effectuent très facilement. Plus précisément, afin de tenter de limiter les biais vers certaines actions, nous avons sélectionné des task-sets pour lesquels chaque stimulus était associé à une réponse différente. La plupart des sujets se sont donc rendus compte que, si telle action était associée à tel stimulus, alors elle n'était associée à aucun autre stimulus. Heureusement, un temps de réponse suffisamment court tend à limiter (pas complètement) l'utilisation de cette heuristique. Malgré tout, cet effet pourrait justifier une performance légèrement meilleure des sujets par rapport au modèle, dans la condition contrôle de l'expérience sans contextes.

8.2.4 Contrôles séquentiel et épisodique, modifications potentielles

Contrôle séquentiel

Il n'est pas rare que nos actions aient des effets sur le monde non pas seulement par l'intermédiaire du renforcement qui nous est procuré, mais également par l'intermédiaire du prochain état, du prochain stimulus auquel on doit faire face. Nous n'avons pas modélisé cet aspect dans notre modèle, supposant que le stimulus à $t + 1$ était indépendant de l'action

à t, et que seules étaient observés par le modèle - et les sujets - les conséquences en terme de renforcement de leurs actions.

Notons à nouveau que, au coeur du modèle proposé, il y a l'estimation de la confiance qu'on a dans la validité d'un comportement, estimée grâce à un modèle interne prédisant ce qu'on doit observer comme conséquence de ses actions dans l'environnement dans lequel ce comportement est valide. Nous avons limité l'observation des conséquences à l'incertitude attendue sur le renforcement, cependant, nous aurions également pu modéliser la prédiction des conséquences séquentielles (portant sur l'état suivant) des actions. C'est d'ailleurs ce qui est fait dans le cadre du contrôle moteur pour MMBRL (Doya et al, 2002 [75], Imamizu et al, 2004 [112], 2007 [113], 2008 [111]).

On pourrait donc généraliser notre modèle à des comportements plus complexes que des task-sets, incluant une planification séquentielle des états suivants.

Contrôle épisodique

Nous nous sommes également limités au niveau du contrôle contextuel du modèle en cascade du contrôle cognitif (Koechlin et al, 2003[136]). Dans le cadre de cette théorie en cascade, on peut tirer un parallèle hiérarchique très fort entre les trois niveaux :

- au niveau sensoriel, en réponse à un stimulus, on sélectionne une action dont l'association au stimulus est renforcée par une récompense ou une punition ;
- au niveau contextuel, en réponse à un contexte, on sélectionne un task-set dont l'association au contexte est renforcée par la confiance dans le task-set ;
- au niveau épisodique, en réponse à un indice épisodique maintenu en mémoire de travail, on sélectionne un ensemble d'associations *episodic-set*, défini parallèlement à la notion de task-set comme un ensemble d'associations contexte - task-sets.

On pourrait donc imaginer, de manière exactement parallèle au niveau contextuel, la possibilité d'estimer la confiance dans l'*episodic-set* actuel, en fonction d'une part de ce qui est observé a posteriori, à savoir la récompense obtenue pour l'action et la récompense fictive (μ) obtenue pour le task-set ; et, d'autre part de l'information a priori, à savoir l'indice maintenu en mémoire de travail.

Cette extension a été testée dans un cadre mixte d'apprentissage des task-sets, des contextes, des episodic-sets et des indices épisodiques. Elle permet effectivement l'acquisition des ces différents niveaux et leur utilisation. Cependant, cela implique un niveau de complexité très élevé, demandant un nombre d'essais d'apprentissage très élevé, ce qui rend la possibilité d'un test expérimental du modèle généralisé au niveau épisodique très peu probable.

Notons cependant que, si on utilise cette extension au niveau épisodique, avec un seul contexte (conduisant à un type de task-switching fréquent où un indice indique la tâche à effectuer dans le prochain bloc, mais où l'indice n'est pas maintenu sous forme contextuelle pendant le bloc, et où, donc, le sujet doit maintenir de manière épisodique l'information), le modèle est isomorphe au modèle contextuel, avec l'ajout du maintien en mémoire de travail de l'indice épisodique, à la place de l'indice contextuel. Autrement dit, si on ajoute un étage *épisodique* au dessus de l'étage *contextuel* initial de notre théorie, mais si l'étage contextuel n'apporte pas d'information, le modèle à deux étages se réduit naturellement à un modèle à un seul étage. Notre modèle permet donc également de rendre compte simplement de ce type de paradigmes de task-switching épisodique.

8.2.5 Embranchements

Notre modèle peut sembler, à première vue, contradictoire avec la théorie des embranchements (*branching*) (Koechlin et al, 1999 [131], Koechlin et Hyafil, 2007 [132]). En effet, Koechlin et al proposent que le cortex fronto-polaire permet de mettre en attente une tâche dans le but de la terminer plus tard pendant qu'on en effectue une autre. Ils argumentent ainsi que notre capacité cognitive est limitée et qu'on ne peut maintenir qu'une seule tâche en attente pendant l'application d'un autre task-set.

Dans notre théorie, nous posons l'hypothèse que nous avons un task-set par défaut, qui est la tâche que nous effectuons à un moment donné, mais que nous continuons à surveiller la confiance que nous avons dans d'autres task-sets (un ou plus). On peut donc se demander si ce processus est équivalent à maintenir n task-sets pendant l'exécution d'une tâche, ce qui contredirait la limite de seulement une tâche maintenue en attente proposée par la théorie du *branching*.

Nous argumentons que la mise à jour de la confiance dans les task-sets autres que le task-set par défaut, ne correspond pas à une situation de mise en attente de tâches comme observé dans les tâches de branching. En effet, dans le branching, il y a une intention fixe, pendant l'exécution de la tâche secondaire, de revenir à la tâche primaire à la fin de l'épisode intermédiaire, donc le maintien d'un objectif lié à une tâche pendant l'exécution d'une autre tâche. Ce n'est pas le cas dans la situation qui nous intéresse : après un switch, les autres task-sets que le task-set par défaut sont abandonnés. L'intention d'y revenir ne se fait que lorsque leur confiance émerge à nouveau au-delà du seuil. Il n'y a donc pas de maintien d'un objectif lié aux autres task-sets pendant l'exécution du task-set par défaut, seulement une surveillance des autres options.

Si notre théorie ne contredit pas les limites du contrôle cognitif humain mises en valeurs par les problèmes d'embranchement, elle ne rend pas non plus compte de cet aspect du contrôle cognitif. Notons que la notion de mise en attente d'une tâche pendant qu'une autre est effectuée implique un stockage en mémoire de travail, ainsi qu'un aspect temporel ou séquentiel de la tâche, faute de quoi aucun maintien n'est nécessaire pour effectuer correctement la tâche initiale à la fin de l'embranchement. Les problèmes de mémoire et les problèmes séquentiels sont deux aspects principaux sur lesquels nous ne nous sommes pas attardés dans notre théorie, nous ne pouvons donc pas proposer simplement une généralisation permettant de rentre compte des phénomènes de *branching*.

8.2.6 Simultanéité

Notons enfin que nous avons imposé une contrainte particulière pour le type d'environnement dans lequel notre modèle est capable d'apprendre. En effet, nous avons spécifiquement imposé la nécessité qu'un task-set soit valide pendant une durée de temps suffisante pour l'apprendre (au moins à un certain degré), au moins initialement. Nous ne nous sommes donc pas placés dans la situation où la tâche valide et un contexte associé changent plus souvent, de telle sorte que plusieurs task-sets doivent être appris de manière simultanée. En l'absence d'un contexte, nous avons en effet argumenté que cela conduirait à une conjonction des tâches, de telle sorte qu'elles n'en formeraient plus qu'une. En présence de contexte, on pourrait cependant penser que les sujets pourraient parvenir à extraire la dimension

contextuelle puis les task-sets associés à chaque contexte, même si ces contextes changent régulièrement.

Notre théorie ne rend pas compte de ce phénomène potentiel. En effet, notre modèle extrait implicitement le contexte comme la dimension informative, et nécessite pour cela une certaine stabilité des task-sets. On peut supposer qu'un raisonnement plus explicite, permettant de tester des hypothèses quant à des dimensions naturellement candidates pour jouer le rôle de contexte, serait nécessaire pour parvenir à déterminer des contextes et des task-sets dans une expérience où ceux-ci changeraient fréquemment.

8.3 Implémentation neuronale des calculs

Nous avons proposé une théorie relativement algorithmique d'intégration entre l'apprentissage et le contrôle cognitif. Si nous supposons que cette théorie modélise le fonctionnement du cerveau humain, nous devons proposer une implémentation neuronale possible des calculs nécessaires au bon fonctionnement de la théorie. Nous avons déjà largement discuté dans la première partie bibliographique du rôle de la dopamine dans le codage d'une erreur de prédiction pour l'apprentissage par renforcement et d'un possible codage de fonctions de valeurs et de stratégies dans le striatum ventral et dorsal. Nous nous concentrons par la suite sur les calculs bayésiens.

8.3.1 Inférence bayésienne

Récemment, plusieurs études ont montré que des transferts d'informations semblables à des processus d'inférence bayésienne étaient représentés dans des réseaux de neurones biologiquement précis.

Koechlin et al, 1999 [130], proposent un réseau de neurones permettant de modéliser la perception du mouvement. Ils proposent une connectivité équilibrant les connexions *feed-forward* et latérales ainsi que les connexions excitatrices et inhibitrices de telle sorte que, de manière continue, un schéma d'entrée est comparé avec l'information a priori encodée dans l'ensemble de la population. Cela permet exactement d'encoder une inférence bayésienne,

de telle sorte que l'activité d'un neurone représente la probabilité a posteriori que l'entrée correspondante soit observée, sachant les réponses précédentes de la population de neurones. Les auteurs montrent que ce modèle reposant sur un principe d'inférence bayésienne permet d'expliquer des propriétés connues des neurones de l'aire visuelle MT, responsable de la perception du mouvement. Ils en déduisent qu'il est possible que ces neurones effectuent effectivement ce type d'inférence bayésienne.

Bien que ce modèle soit basé sur un réseau de neurones, il reste assez loin d'une implémentation biologiquement plausible d'un calcul bayésien puisqu'il est basé sur des neurones ponctuels, dont l'activité est modélisée par un simple taux de décharge. Une méthode plus proche de la réalité des neurones demande un codage de leur activité par la distribution temporelle de leurs pics d'activation (*spike*), plutôt que simplement par leur fréquence moyenne, par exemple avec un modèle de neurone *integrate and fire* (Lapicque, 1907 [139], Abbott et al, 1999 [2], Hodgkin et Huxley, 1952 [109]).

D'autres études (Ma et al, 2006 [141], Beck et Pouget, 2007 [15]) montrent que, comme nécessaire pour une inférence bayésienne, les populations de neurones représentent des distributions de probabilités, dans un *encodage probabiliste par population*. Ils observent que la structure statistique de la variabilité observée pour l'activité des neurones du cortex (de type *Poisson*) permet de réduire des calculs d'inférence bayésienne à des combinaisons linéaires d'activités de populations de neurones. Ils montrent ainsi qu'on peut modéliser, avec des neurones respectant des propriétés biologiques connues (statistiques temporelles de la variabilité des *spikes*), des calculs d'inférence bayésienne.

Dans un cadre différent, Deneve (2008, [69] [70]) montre également que la dynamique de *spiking neurons* peut être considérée comme une forme d'inférence bayésienne dans le temps : dans ce modèle, chaque spike porte une information qui ne pouvait pas être prédite par l'activité passée. On a ainsi une représentation déterministe de l'information à l'échelle temporelle et spatiale d'un spike, plutôt qu'au niveau d'un encodage probabilistique par population. Malgré cet aspect déterministe, les statistiques temporelles des spikes restent biologiquement plausibles (Poisson) et une inférence Bayésienne est effectuée.

Nous voyons donc que, bien que la plupart de ces études soient limitées à des zones sensorielles du cortex, et non associatives ou préfrontales, plusieurs modèles semblent montrer

que des calculs d'inférence bayésienne peuvent être représentés par des neurones.

8.3.2 Données neuronales

Nous avons montré que des modèles de réseaux de neurones biologiquement plausibles permettaient de faire des calculs d'inférence bayésienne et permettaient de répliquer des effets observés empiriquement. Dans cette partie, nous montrons que des données d'électrophysiologie ont montré que les calculs de base sont effectivement effectués par des neurones dans le cerveau

Tout d'abord, il est largement admis (Koch et Segev, 2000 [129]) que les neurones peuvent effectuer une intégration (spatiale et temporelle) de ses entrées. Cela correspond à une opération de type pondération linéaire. Koch et Segev (2000, [129]) montrent que de nombreuses autres opérations peuvent être effectuées par des neurones individuels, notamment des opérations multiplicatives. Carandini et Heeger (1994, [38]) montrent que, par l'effet d'un certain type d'inhibition latérale de l'ensemble de la population (dit *shunting inhibition*), on peut observer une normalisation de l'activité d'un neurone, soit l'implémentation d'une opération divisive.

8.3.3 Apprentissage

Nous avons déjà très largement proposé des mécanismes biologiquement plausibles d'apprentissage des task-sets, reposant notamment sur le rôle essentiel de la dopamine pour l'encodage d'une erreur de prédiction. Celle-ci permet alors la mise en place de potentiation (ou dépotentiation) à long terme de synapses. Les modèles acteurs-critiques (ou acteur-directeur) ont permis de montrer que les calculs liés à l'apprentissage, notamment la mise à jour des valeurs de prédiction à l'aide de l'erreur de prédiction et l'utilisation de ces valeurs pour déterminer une stratégie, étaient probablement éffectués dans le striatum ventral et dorsal.

Pour les modèles internes directs, ce qui est appris est directement dépendant du feedback observé. On peut donc supposer le même mécanisme, lié au rôle de la dopamine, pour leur

apprentissage, même si celui-ci a probablement lieu dans d'autres zones du cortex, comme nous l'argumenterons plus bas.

Par contre, l'apprentissage des associations contexte - task-set ne dépend pas d'un signal de renforcement extérieur, mais d'un signal de renforcement interne. On peut donc légitimement se demander comment cet apprentissage pourrait être implémenté dans le cortex.

Nous argumenterons plus bas que les modèles internes, γ, sont probablement représentés dans le cortex préfrontal ventro-médial (vmPFC), tandis que le cortex cingulaire antérieur représente les valeurs de confiance a priori, λ. Ces deux quantités sont les quantités nécessaires au calcul de μ, la confiance ex-post, qui est le signal de renforcement interne utilisé pour apprendre les associations contexte-task-set. Or, le vmPFC et l'ACC, sont les deux entrées principales du locus coeruleus (LC), noyau produisant la norépinephrine (voir section 3.2.2, Aston-Jones et al, 2005 [6]). On peut ainsi émettre l'hypothèse, comme Yu et Dayan, 2005 [217], que la NE signale la confiance ex-post μ, ou plutôt l'incertitude inattendue ex-post, $1 - \mu$. Par ailleurs, Harley, 2004 [103], montre que la norépinephrine (à l'instar de la dopamine) promeut la plasticité à long terme et est donc impliquée dans l'apprentissage.

On déduit de ces données qu'il est possible que l'apprentissage des associations contexte-task-set soit implémenté par un mécanisme similaire à celui pour l'apprentissage des associations stimulus-actions, mais avec la norépinéphrine (représentant la confiance ex-post μ) comme signal d'apprentissage améliorant la plasticité à long terme.

8.4 Implémentation fonctionnelle

Nous avons indiqué des pistes, au niveau neuronal, pour l'implémentation des calculs requis par notre modèle. A un niveau plus fonctionnel, nous proposons ici d'étudier l'implication de différentes régions du cerveau, préfrontales ou sous-corticales, dans la représentation des différentes quantités et des différents objets manipulés par notre modèle.

8.4.1 Associations stimulus-actions : prémoteur, striatum

Nous argumentons dans cette section que les associations stimulus-actions sont représentées dans le cortex frontal dorsolatéral, plus spécifiquement dans le cortex prémoteur, et que leurs valeurs sont représentées dans le striatum.

Il est largement accepté que le cortex prémoteur dorsal (dlPM) est nécessaire à l'apprentissage et au maintien de mappings visuo-moteurs arbitraires. En particulier, les études sur les singes montrent que l'ablation du dlPM entraîne une incapacité des singes à apprendre des nouveaux task-sets, ou à les maintenir (Murray et al, 2000 [151], Petrides, 1982 [167], Halsband et Passingham, 1982 [100]). L'étude de patients humains présentant des lésions dans le cortex frontal latéral a également montré que ces patients présentaient un déficit pour apprendre des task-sets, quel que soit le mode d'apprentissage, dirigé ou renforcé (Petrides, 1997 [168]).

Si le rôle du cortex préfrontal dorso-latéral et du cortex prémoteur dorsal sont peu dissociés plus haut, Koechlin et al, 2003 [136], ont spécifiquement dissocié leur rôle, montrant que si le dlPFC maintenait les task-sets, c'était bien le cortex prémoteur dorso-latéral qui était critique pour les associations stimulus-actions.

On sait également que le striatum joue un rôle essentiel dans l'encodage des valeurs des couples stimulus-actions (voir la section 1.1.2, Samejima et al, 2005 [187], 2007 [186], Pessiglione et al, 2008 [165]).

On peut donc poser l'hypothèse que la boucle cortico-basale incluant le cortex prémoteur dorso-latéral et le striatum est essentielle dans la représentation des associations stimulus-actions, ainsi que de leurs valeurs.

8.4.2 Associations contexte-TS : dlPFC, pariétal, hippocampe ?

Nous argumentons dans cette section que les associations contexte-task-sets sont représentées dans le cortex préfrontal dorso-latéral postérieur, en lien avec le cortex pariétal et peut-être l'hippocampe.

L'étude de Koechlin et al, 2003 [136], validant le modèle en cascade du cortex préfrontal

latéral a mis en valeur le fait que la sélection d'un task-set à partir d'un contexte s'effectuait dans le cortex préfrontal latéral postérieur. L'implications du cortex préfrontal latéral postérieur dans la sélection des tâches est par ailleurs confirmé dans de nombreuses études portant sur le task-switching et mettant en avant son rôle dans la configuration des tâches, le maintien d'associations, ce dont nous avons déjà parlé en détail dans la section 3.1.4 (Hyafil et al, 2009 [110], Mitchell et al, 2008 [149], Murray et al, 2007 [152]).

Notons par ailleurs que des activations liées au feedback, plus spécifiquement aux erreurs, sont observées dans le cortex préfrontal dorso-latéral (pour une revue, voir notamment Dosenbach, 2006 [72]). Il est donc possible que les valeurs d'association entre contextes et task-sets soient stockées dans le cortex préfrontal latéral.

Le cortex pariétal, plus précisément le sulcus pariétal inférieur, est connu pour être impliqué dans le réseau de régions essentielles au task-switching (Dosenbach et al, 2006 [72], Shafritz et al, 2005 [197], Wu et al, 2004 [215], Jubault et al, 2007 [121], Koechlin et al, 2003 [136]), en particulier pour le stockage de tâches ou de séquences d'actions et pour le maintien d'information. Dosenbach propose en particulier que le pariétal joue un rôle pour loader, transmettre et implémenter des associations stimulus actions en fonction d'un but, soit un task-set adapté à un contexte.

L'hippocampe a également été impliqué dans la mémoire d'associations, le task-switching et le contrôle dépendant de contextes (Graham et al, 2009[98], Frank et al, 2006 [87], Turnock et al, 2008 [206], Strange et al, 2001 [201]). En particulier, [206] propose que cette région joue un rôle de *gating* entre le cortex préfrontal et les ganglions de la base permettant de contrôler, en fonction du contexte, les associations stimulus-réponses utilisées. C'est donc un candidat possible pour l'implémentation des associations contexte-task-set et le contrôle contextuel.

On peut donc poser l'hypothèse qu'un réseau impliquant le cortex préfrontal latéral postérieur, le cortex pariétal inférieur et l'hippocampe sont impliqués dans les associations entre contextes avec les task-sets, les task-sets étant plus vraisemblablement représentés dans les régions préfrontales, les contextes dans les régions pariétales ou hippocampales.

8.4.3 Modèles internes prédictifs : vmPFC

Notre théorie suppose que chaque task-set est muni d'un modèle direct, soit une représentation interne permettant de prédire les conséquences des actions dans l'environnement dans lequel ce task-set est valide. Ce modèle permet d'évaluer la vraisemblance d'une observation. Dans cette section, nous défendons que le cortex préfrontal ventro-médial (vmPFC) essentiellement, et le cortex orbito-frontal (OFC) secondairement, sont impliqués dans la représentation, la mise à jour et l'utilisation pour l'évaluation de la vraisemblance d'une conséquence de ce modèle interne.

Le cortex orbito-frontal est depuis longtemps connu pour être impliqué dans l'estimation de valeurs abstraites (O'Doherty et al, 2001 [157], Hampshire et al, 2005 [101]). Plus précisément, de nombreuses études et modèles montrent que l'OFC est essentiel pour apprendre à prendre des décisions qui dépendent d'une estimation correcte de la valeur attendue des décisions : par exemple, prise en compte de la magnitude et de la probabilité de la récompense (Frank et al, 2006 [84]), prise en compte de l'objectif (Valentin et al, 2007 [208]). Kepecs et al, 2008 [127], mesurent l'activité de neurones de l'OFC de rats pendant une tâche de décision sous incertitude et montrent que des signaux de confiance sur la conséquence de la décision sont observés dans l'OFC. Plusieurs études (Cools et al, 2004 [50], Constantinidis et al, 2004 [45]) rappellent d'ailleurs que des lésions de l'OFC portent atteinte, entre autres, à la capacité d'effectuer du *reversal learning*, paradigme qui demande d'interpréter les erreurs observées comme incompatibles avec les prédictions.

L'ensemble de ces arguments (évaluation nécessaire à une décision, signal de confiance donc probabilité de la conséquence observée) soutient l'idée d'une représentation du modèle interne prédictif dans l'OFC. Cependant, la région voisine (et souvent confondue) du vmPFC pourrait également être une bonne candidate à la représentation du modèle interne prédictif et au calcul de la vraisemblance $\gamma(r, s, a)$ indispensable au modèle. En particulier, Hampton et al, 2006 [102] montrent que l'activité du vmPFC est fortement corrélée à la probabilité a priori que le choix du modèle soit correct, ce qui correspond strictement à $\gamma(r, s, a)$ dans notre modèle. De même, des études menées par l'équipe de Behrens et Rushworth (Boorman et al, 2009 [23], Behrens et al, 2008 [16], Rushworth et al, 2008 [181]) montrent que l'activité du vmPFC corrèle avec la valeur estimée par un modèle, a priori, d'un choix d'une

action, ou à la probabilité de la conséquence observée pour l'action.

On peut donc poser l'hypothèse que les zones voisines de l'OFC et du vmPFC encodent un modèle interne lié au task-set, ce qui leur permet d'estimer la vraisemblance des conséquences observées.

8.4.4 Signaux de confiance dans le TS : cortex préfrontal médial

Notre modèle s'intéresse à deux types de signaux de confiance dans les task-sets, chacun essentiel, mais très distincts dans la valeur qu'ils représentent ainsi que dans leur rôle.

- λ mesure a priori la confiance dans la validité d'un task-set pour la décision à effectuer. λ sert à la décision, et nécessite pour sa mise à jour, des informations contextuelles, des informations de volatilité et la dernière mesure de μ.
- μ mesure a posteriori la confiance dans la validité d'un task-set pour la décision qui a été effectuée. μ sert à l'évaluation, et nécessite pour sa mise à jour, la dernière mesure de λ et la prédiction du modèle interne $\gamma(r, s, a)$.

λ : confiance ex-ante : ACC

Nous proposons ici que λ est implémenté dans le cortex préfrontal médian, plus précisément dans l'ACC.

Nous avons déjà montré de manière très générale que l'ACC était impliqué de manière essentielle dans le contrôle cognitif, notamment pour réagir efficacement aux signaux (conflit, erreurs, risque, etc.) indiquant la nécessité d'augmenter le niveau de contrôle et éventuellement de switcher (Aarts et al, 2008 [1], Hyafil et al, 2009 [110], Brown et al, [33], Dosenbach et al, 2006[72], Wager et al, 2005 [209]). Ces données argumentent en faveur de la présence du calcul et de la représentation d'un signal servant aux décisions sur le contrôle des task-sets dans l'ACC, ce qui est exactement le rôle de λ. D'autres études permettent d'argumenter plus précisément en faveur de la représentation de λ dans l'ACC.

Tout d'abord, de nombreuses études, notamment de l'équipe de Rushworth et Behrens (Kennerley et al, 2006 [126], Behrens et al, 2007 [17], Rushworth et al, 2006 [183], 2008

[181]) montrent que l'ACC intégre l'information obtenue, en tenant compte de l'historique des actions et des réultats, pour guider les choix volontaires. Cela coïncide avec le rôle de λ, qui n'interprète pas identiquement une erreur dans différents environnements historiques de renforcement. Hayden et al, 2009 [105], ont également montré que toute l'information - pas seulement celle liée aux récompenses effectivement obtenues, mais également les renforcements fictifs - est prise en compte dans l'ACC. Cela correspond bien au rôle de λ, qui est d'intégrer toute l'information présente afin de prendre une décision au niveau hiérarchique des task-sets.

Des études montrent d'ailleurs que l'ACC réagit spécifiquement aux renforcements, qu'ils soient négatifs ou positifs, s'ils sont pertinents pour l'adaptation du comportement (Quilodran et al, 2008 [171]). O'Doherty et al ont d'ailleurs observé (2001 [157]), dans une tâche de *probabilistic reversal learning*, une activation de l'ACC pour des punitions, ou pour des récompenses exclusivement pendant la période de *reversal*, lorsque celles-ci apportent de l'information sur la tâche.

Hampton et al, 2006 [102], montrent que l'ACC s'active plus lors des essais switch, et que son activité est corrélée à la probabilité a priori que la décision soit incorrecte, ce qui signale une nécessité de plus de contrôle. Enfin, Sallet et al, 2007 [185], montre que ce qui est encodé dans l'ACC n'est pas une valeur absolue, mais relative, dépendante du contexte et de l'historique. Ils proposent que l'ACC « évalue et signale quand les attentes en terme de récompense sont violées par des stimuli inattendus ».

De toutes ces observations, on déduit que l'ACC intègre tout type d'information lorsqu'elle est utile pour évaluer la pertinence du comportement actuel, en tenant compte de l'historique et du contexte, afin de signaler la nécessité de plus de contrôle ou d'un switch comportemental. On peut donc postuler que l'ACC encode une quantité inversement corrélée à λ, par exemple $1 - \lambda$, représentant non pas la confiance dans le comportement actuel, mais plutôt le doute, et qui signalerait la nécessité de plus de contrôle et de switcher par son augmentation ($\lambda < 0,5$).

Nous avons vu dans le paragraphe 3.1.4 que la plupart des auteurs s'accordaient à dire que l'ACC émettait un signal lorsqu'était perçue la nécessité de plus de contrôle, ou de switcher, à l'attention des régions latérales du cortex préfrontal. On peut donc supposer que

le switch entre différents task-sets est implémenté ainsi, par influence des régions médiales qui implémentent λ, sur les régions latérales du cortex préfrontal, qui maintiennent les task-sets et leurs associations aux contextes.

μ : confiance ex-post

Le calcul de la confiance ex-post, μ, requiert crucialement l'utilisation du modèle interne pour l'inférence bayésienne. μ est ensuite utilisé pour le calcul de λ. En l'absence d'information contextuelle importante, μ et λ sont très proches, on peut donc supposer qu'ils sont encodés dans des régions proches, les régions du cortex préfrontal médian. C'est pourquoi nous émettons l'hypothèse que les valeurs de confiance ex-post sont encodées de manière intermédiaire entre l'ACC et le vmPFC. Notons que ces régions sont connectées avec les régions latérales du cortex préfrontal, ce qui est nécessaire puisque μ sert à renforcer les associations contextes-TS, et λ à décider du TS par défaut maintenu

8.4.5 TS par défaut, autres task-sets, TS test : rôles des boucles fronto-basales ?

Dans notre théorie, plusieurs task-sets sont considérés en parallèle, mais un task-set particulier est singularisé pour son rôle dans l'apprentissage et la sélection de l'action (le task-set par défaut pendant une période d'exploitation, le task-set test pendant une période d'exploration). Nous suggérons ici que ce seul task-set acteur et apprenant est représenté dans les boucles entre le cortex frontal et les ganglions de la base.

Nous avons déjà vu que les ganglions de la base, notamment par l'intermédiaire de la boucle qu'ils forment avec le cortex prémoteur, jouent un rôle essentiel dans l'apprentissage par essai-erreur et dans la sélection de l'action. En effet, on a vu que le striatum représentait la valeur des paires stimulus actions (Pessiglione et al, 2006 [166], Samejima et al, 2005 [187], 2007 [186]), qu'il utilisait des signaux d'erreur de prédictions pour mettre à jour ces valeurs (Pessiglione et al, 2006 [166], Seymour et al, 2004 [196], Schonberg et al, 2007 [190], Kahnt et al, 2009 [122]). Crucialement, on a vu que le striatum ventral était indispensable

à l'apprentissage et le striatum dorsal à l'utilisation de l'apprentissage pour la sélection des actions (Atallah et al, 2007 [7]).

Par ailleurs, le modèle en cascade du contrôle cognitif (Koechlin et al, 2003 [136], 2007 [134]) propose que la partie postérieure du dlPFC, qui effectue la sélection du task-set en fonction du contexte, influence le cortex pré-moteur, responsable de la sélection des actions en réponse aux stimuli, sous le contrôle du task-set.

Il semble donc pertinent de supposer que la représentation des associations stimulus-actions qui composent le task-set par défaut est implémentée dans la boucle entre les ganglions de la base et le cortex, incluant le cortex prémoteur et le striatum, puisque ce premier est effectivement influencé par la représentation du task-set dans le dlPFC, et que ce deuxième permet l'apprentissage et la sélection d'actions effectivement réservés uniquement au task-set par défaut.

Lors d'une période d'exploration, seul le task-set test est acteur et apprenant. C'est un task-set transient, destiné à disparaître, sauf s'il devient un nouveau task-set test. On pourrait supposer alors que ses associations stimulus-actions sont représentées dans la boucle prémotrice des ganglions de la base, sans que ce task-set soit représenté dans le dlPFC, contrairement au task-set par défaut qui représente effectivement un comportement appris et validé.

8.5 Différences interindividuelles

Nous avons montré que les sujets se comportaient de manières très diverses face aux problèmes d'apprentissage et de contrôle. Aucun facteur explicatif simple (âge, sexe, niveau d'étude) n'a pu rendre compte des différences observées. Nous proposons dans cette section des pistes de réflexion, sur quels facteurs pourraient impliquer cette variabilité.

8.5.1 Différences stables ou circonstancielles ?

A partir des debriefings de chaque sujet, nous avons pu les classer dans deux ou trois groupes, pour l'expérience sans contexte et avec contexte, respectivement. Nous avons montré la

pertinence de ces groupes de deux manières :

– par les données comportementales, qui confirment la présence ou l'absence de certains effets, montrant un comportement qualitativement différents entre les différents groupes de sujets ;
– par le fitting des modèles sur les données comportementales, qui confirment que certains paramètres expliquent les différences comportementales observées.

Quelle est la cause de ces différences qualitatives de comportement ? On peut, en particulier, se demander si elles sont robustes dans le temps : un sujet serait-il classé dans le même groupe s'il participait à la même expérience un autre jour ? On sait en effet que de nombreux facteurs variables, peuvent influencer le comportement. Par exemple, Pessiglione et al, 2007 [164], montrent l'importance de la motivation (monétaire par exemple) sur la performance. Dreher et al, 2007 [78], montrent que, selon la période du cycle menstruel, on observe des réactions différents face aux récompenses, chez les sujets féminins.

On pourrait donc penser que les différences observées entre les sujets, sont uniquement dues à des facteurs environnementaux, différents pour chaque sujet : leur niveau de fatigue, leur niveau de motivation à effectuer la tâche au mieux, ou d'autres facteurs.

Bien que nous n'ayons aucun moyen de conclure sur ce point, nous défendons ici, que, ces différences circonstancielles ne sont pas le facteur explicatif principal des différences interindividuelles observées, mais que celles-ci sont liées à des facteurs robustement liés à chaque sujet.

En effet, de nombreuses études psychologiques ont démontré des différences comportementales stables, indépendantes des facteurs environnementaux, dans le domaine du contrôle exécutif. Plutôt que de s'intéresser à un des facteurs psychométriques les plus connus et utilisés, le quotient intellectuel (QI), les études psychométriques portant sur le contrôle exécutif tendent à utiliser le facteur psychométrique *Gf*, nommé *G-factor* pour *General Fluid Intelligence factor*, soit facteur d'intelligence générale fluide (Kane et al, 2002 [124]). Il est reconnu (Engle et al, 1999 [80]) que ce facteur fournit une mesure stable et pertinente d'aspects moins sociaux et plus flexibles de l'intelligence, que le QI.

Les études psychométriques, par exemple Kane et al, 2002 [124], Engle et al, 1999 [80] montrent que ce facteur psychométrique stable, Gf, est très relié à la capacité de mémoire

de travail. Les auteurs définissent cette capacité de mémoire de travail, comme la capacité d'exercer un contrôle attentionel exécutif, soit de stocker et maintenir en mémoire de travail toutes (et uniquement) les informations pertinentes, en résistant aux interférences possibles appliquées par les distracteurs. Bien que le recouvrement entre le facteur Gf et cette capacité liée au contrôle cognitif ne soit pas parfait (Conway et al, 2003 [46]), tous les auteurs s'accordent sur le très fort lien existant entre les deux.

Puisque la performance dans des tâches semble très corrélée à la mesure d'un indice psychométrique stable, on peut déduire qu'il semble exister une composante stable, indépendante des facteurs environnementaux, pour nos performances dans des tâches de contrôle cognitif. Contrairement à d'autres domaines de la cognition (tâches sémantiques, mémoire à long terme, mémoire visuelle), les performances dans le domaine des fonctions exécutives semblent largement indépendantes du gradient socio-économique (Noble et al, 2007 [155]). Par contre, plusieurs études tenant compte des différences individuelles dans le domaine du contrôle exécutif relient ces performances à l'activité préfrontale (McNab et Klingberg, 2008 [145], Kane et al, 2002 [124]).

Il semble donc nécessaire de chercher la cause des différences individuelles fortes observées dans nos données comportementales, dans un facteur biologique, lié au préfrontal, stable plutôt qu'environnemental.

Dans la suite de cette section, nous évoquons les pistes de ce type explorées pour étudier la variabilité interindividuelle dans la littérature.

8.5.2 Neuromodulateurs

Nous avons évoqué plusieurs études impliquant une manipulation pharmacologique, provoquant un changement de comportement lors d'une tâche cognitive (par exemple Pessiglione et al, 2006 [166], Frank et al, 2006 [87], 2004 [88], Chamberlain et al, 2006 [39]). Ces études manipulent le niveau tonique de présence d'un neuromodulateur, par exemple la dopamine, la noradrénaline ou la sérotonine. Puisque le changement de ce niveau tonique provoque des changements de comportement, il est naturel de considérer la possibilité que les différences comportementales observées entre les sujets pourraient être dues à des différences dans ce

niveau tonique individuel (Cools et al, [50]).

Une série d'études menée par Cools et al (2004 [50], [47], 2009 [48]) confirme cette possiblité dans le cas particulier de la dopamine. En particulier, Cools et al, 2009 [48], soumettent des sujets à une tache de *reversal learning*. Ils mesurent également, par tomographie par émission de positrons, leur niveau individuel de production de dopamine dans le striatum. Ils montrent une corrélation forte entre le niveau de dopamine striatale et la performance des sujets, dans le sens prédit par les études pharmacologiques : les sujets ayant un fort niveau de dopamine striatale apprennent mieux à partir des récompenses, tandis que ceux ayant un faible niveau de dopamine striatale apprennent mieux à partir des punitions.

On voit donc que les niveaux toniques individuels des neuromodulateurs pourraient être des causes de variabilité comportementale entre les sujets. Ces différences de niveaux toniques de neuromodulateurs peuvent être provoquées par des différences génétiques entre les sujets, ce que nous montrons dans le paragraphe suivant.

8.5.3 Génétique

Plusieurs techniques sont utilisées pour évaluer le rôle d'un gène cible dans une fonction cognitive (Green et al, 2008 [99]). Après qu'un (ou plusieurs) polymorphismes d'un gène cible sont identifiés, on peut classer les sujets selon la forme de polymorphisme qu'ils présentent. On peut alors utiliser ces groupes pour les corréler avec des effets comportementaux, avec des paramètres de modèles fittés sur des données comportementales, avec le niveau d'activité d'une région du cerveau ou avec le niveau de connectivité entre différentes régions du cerveau. Notons que beaucoup de prudence et de circonspection sont en général nécessaires pour tirer des conclusions de ces études génétiques (Goldberg et Weinberger, 2004 [94]) : en effet, entre le gène et le comportement, se trouvent plusieurs maillons logiques, chacun complexe : codage par le gène de la structure de la protéine, qui elle même a un effet sur les neurones du cerveau, qui sont les corrélats neuraux d'une fonction cognitive. Cette chaîne n'est pas simplement linéaire, un gène ayant des effets multiples sur de multiples processus. Par ailleurs des effets de corrélation non causaux peuvent également être observés.

Malgré cela, certains résultats robustes existent, notamment dans le domaine d'étude de

l'apprentissage par renforcement et celui du contrôle cognitif. Frank et al (2007 [86], 2009 [85]) ont par exemple montré que les gènes DARPP32 et DRD2, liés aux récepteurs D1 et D2 de la dopamine, respectivement, présentaient des polymorphismes qui pouvaient être corrélés avec des paramètres d'apprentissage par récompense et par punition, respectivement, fittés sur le comportement des sujets.

Un gène, COMT (Catechol-O-Methyltransférase), est particulièrement étudié, ainsi que son polymorphisme $val^{158}met$ (Blasi et al, 2005 [22], Frank et al, 2007 [86], Frank et al, 2009 [85], Dreher et al, 2009 [77]). En effet, son effet est particulièrement bien connu : ce gène produit une enzyme dégradant les mono-amines, en particulier la dopamine. La version *met* de cette enzyme est moins stable que la version *val*, conduisant à une quantité de dopamine corticale plus importante. Plusieurs études montrent ainsi un effet comportemental paramétrique de la présence de l'allèle *met* du gène COMT, les sujets présentant deux allèles *met* ayant les meilleures performances dans des tâches de contrôle ou d'apprentissage (Blasi et al, 2005 [22], contrôle attentionnel ; Frank et al, 2007 [86], 2009 [85], *probabilist learning*).

Malgré la robustesse de certains résultats impliquant ces polymorphismes dans la variabilité interindividuelle, certaines interprétations restent complexes. En effet, Fan et al, 2003 [81], étudiant les gènes DRD4 et MAOA (liés à la dopamine corticale) dans une tâche de conflit en IRMf, montrent que le polymorphisme impliquant la plus grande quantité de dopamine corticale implique une meilleure performance et une plus grande activité de l'ACC. Blasi et al, 2005 [22], par contre, étudiant le gène COMT dans une tâche de contrôle attentionnel, montrent que le polymorphisme impliquant la plus grande quantité de dopamine corticale, s'il implique bien également la meilleure performance comportementale, implique une activité plus faible, et non plus grande, de l'ACC. Les uns concluent que la dopamine aide à l'activation de l'ACC donc à une meilleure performance par plus d'activation, les autres que la dopamine aide à une plus grande efficacité de l'ACC donc à une meilleur performance par moins d'activation.

Pour conclure, il semble donc que, si les effets restent peu clairs quant aux conséquences observables en IRMf, on observe des résultats réplicables liant différents gènes de la dopamine à des différences de comportements entre les sujets. Notamment, le polymorphisme du gène COMT semble lié aux différences interindividuelles de performances des tâches

préfrontales. On peut donc émettre l'hypothèse que ce gène COMT explique les différences interindividuelles observées dans notre tâche, puisque celle-ci implique l'acquisition de task-sets pour le contrôle cognitif, dont nous avons montré qu'il implique le préfrontal ; on peut alors espérer observer un lien entre le polymorphisme de COMT et le paramètre p_{test}, qui règle la stratégie exploratoire et permet, d'après notre modèle, d'expliquer les différences individuelles. C'est donc une piste à explorer pour chercher l'origine de la variabilité comportementale observée dans les expériences de cette thèse et pouravancer dans la compréhension des mécanismes d'intégration de l'apprentissage et du contrôle cognitif dans le cortex préfrontal..

Chapitre 9

Conclusion

Nous avions posé, dans cette thèse, la question de l'apprentissage de task-sets pour le contrôle. Nous avons proposé une théorie computationnelle permettant d'intégrer des notions d'apprentissage et de contrôle. Cette théorie repose sur deux principes simples : 1) la surveillance implicite, parallèle et constante, des différentes options stratégiques disponibles, pour comparer les conséquences observées de la stratégie utilisée, avec les conséquences qu'on est en droit d'attendre dans le cadre des autres options stratégiques, 2) le choix de singulariser le rôle d'une stratégie particulière par rapport aux autres possibles. Ces deux principes induisent naturellement des mécanismes de switch (donc de contrôle des stratégies à utiliser), d'exploration des options disponibles, et d'apprentissage de nouvelles options. Ils permettent donc bien à notre théorie de faire intéragir l'apprentissage, à bas niveau, de stratégies ; l'apprentissage, à haut niveau, d'un répertoire de stratégies et des contextes dans lesquels elles sont valides ; l'exploration, à haut niveau, du répertoire de stratégies, ainsi que des stratégies non encore apprises ; et le contrôle, à haut niveau, de la sélection de ces stratégies pour la sélection des actions.

Nous avons également montré, dans deux expériences comportementales, que cette théorie computationnelle avait une validité pour expliquer l'intégration de l'apprentissage et du contrôle cognitif dans le cortex préfrontal humain : en effet, les effets comportementaux qu'elle prédit, contrairement à d'autres modèles, sont effectivement observés chez des sujets humains adultes sains. Elle permet, par ailleurs, de rendre compte du fait qu'on peut

observer des comportements stratégiques qualitativement différents dans différents groupes de sujets.

Cependant, deux aspects différents de la validité de cette théorie doivent encore être étudiés, et font l'objet de directions de recherches futures potentielles.

Tout d'abord, les ramifications mathématiques de la théorie proposée doivent encore faire l'objet d'une étude spécifique : comme tout modèle algorithmique, il faudrait déterminer spécifiquement les conditions dans lesquelles il converge, ou, tout du moins, permet une performance correcte. Par ailleurs, la théorie a spécifiquement été construite afin de rendre compte du comportement humain, plutôt qu'afin de tenter d'optimiser l'apprentissage et la prise de décision. On peut, malgré tout, se poser la question de l'optimalité : en effet, on peut supposer que la manière dont sont intégrées les fonctions d'apprentissage et de contrôle cognitif a contribué à l'évolution et la survie de l'espèce humaine. Cela implique, si la théorie représente bien le fonctionnement du cerveau humain pour l'apprentissage et le contrôle, qu'on pourrait définir un cadre pour lequel elle est optimale.

Le deuxième axe de validation de notre théorie qui doit encore été étudié, est la validation de son implémentation dans le cerveau. Cet axe offre une piste de continuation de ce projet de recherche, pour des études en *model-based* IRMf, permettant de tester les prédictions effectuées dans la partie discussion, quant à une possible représentation, dans différentes zones du cortex préfrontal et des ganglions de la base, des quantités prédites par notre théorie. Une autre piste pour la continuation de ce projet de recherche dans le but de valider biologiquement notre théorie, est de profiter de la variabilité inter-indivuelle observée comportementalement et expliquée par les paramètres du modèles, pour effectuer une étude génétique portant sur des gènes cibles de l'apprentissage et du contrôle cognitif préfrontal, tels que COMT. On pourrait exhiber potentiellement une cause biologique des différences observées, ainsi qu'un lien fort entre la représentation des quantités responsables des différences dans notre modèle, et les substrats neuraux impliqués par les gènes ciblés.

Rappelons cependant, comme le montre une étude menée par Rueda et al, 2005 [179], sur des enfants entre quatre et six ans, que si le développement du contrôle exécutif semble être sous fort contrôle génétique, il est également très sensible à l'éducation pendant le développement. En conséquence, même si nous montrons que des différences comportemen-

tales sont causées par une prédétermination génétique, nous ne pouvons pas oublier le rôle essentiel joué par l'éducation pour le contrôle exécutif.

Notre théorie tente, justement, d'expliquer comment l'apprentissage s'intègre au contrôle cognitif. Bien que les situations auxquelles elle a été appliquée dans cette thèse soient très peu écologiques, elle pourrait s'appliquer à de nombreuses situations quotidiennes : en effet, nous sommes confrontés en permanence à la nécessité d'apprendre et d'exercer du contrôle cognitif de manière simultanée. Un des exemples les plus frappants est celui de l'évolution des technologies à laquelle nous sommes confrontés, nous obligeant en permanence à apprendre de nouveaux task-sets, tout en restant capables d'utiliser ceux déjà appris ainsi que d'identifier le contexte correct d'utilisation.

La complexité de l'interaction entre le contrôle et l'apprentissage se ressent très bien dans l'étude, à l'âge adulte, de comportements complexes, tels que la conduite d'une voiture ou l'apprentissage d'un instrument de musique. Beaucoup de techniques pédagogiques, efficaces, sont adoptées pour leur efficacité empirique : un professeur de chant, par exemple, sait qu'il ne doit faire porter son attention à son étudiant que sur un aspect technique précis, jusqu'à ce que celui-ci soit relativement automatisé, avant de passer à un autre niveau d'apprentissage. Il sait également qu'il est important, à certains instants, d'imposer à l'étudiant de contrôler quelque chose de non pertinent pour le chant, afin qu'il ne puisse pas tenter de contrôler un certain aspect technique du chant qui doit être, de préférence, appris d'une manière automatisée. Le professeur utilise donc, intuitivement et par expérience, le fait que nous ne pouvons pas contrôler plusieurs aspects de notre comportement à la fois, soit pour limiter l'objet de l'apprentissage, soit pour empêcher le contrôle dans les situations où celui-ci est néfaste à la performance.

On peut donc penser qu'une meilleure compréhension de l'interaction entre l'apprentissage et le contrôle cognitif, tenant compte de la structure hiérarchique du contrôle cognitif et de l'apprentissage, pourrait permettre, à long terme, d'optimiser certaines méthodes de pédagogie, dont le but est justement de faciliter l'apprentissage des étudiants, et la construction de compétences robustes, qu'ils peuvent utiliser de manière flexible. Cela reste cependant une application lointaine, mais enthousiasmante, de la théorie d'intégration du contrôle cognitif et de l'apprentissage proposée dans cette thèse.

Bibliographie

[1] E Aarts, A Roelofs, and M Van Turennout. Anticipatory activity in anterior cingulate cortex can be independent of conflict and error likelihood. *Journal of Neuroscience*, 28(18) :4671–4678, Apr 2008.

[2] L F Abbott. Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain Res Bull*, 50(5-6) :303–4, Jan 1999.

[3] G E Alexander and M D Crutcher. Functional architecture of basal ganglia circuits : neural substrates of parallel processing. *Trends Neurosci*, 13(7) :266–71, Jul 1990.

[4] G E Alexander, M R DeLong, and P L Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci*, 9 :357–81, Jan 1986.

[5] W F Asaad, G Rainer, and E K Miller. Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6) :1399–407, Dec 1998.

[6] Gary Aston-Jones and Jonathan D Cohen. Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J. Comp. Neurol.*, 493(1) :99–110, Dec 2005.

[7] Hisham E Atallah, Dan Lopez-Paniagua, Jerry W Rudy, and Randall C O'reilly. Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat Neurosci*, 10(1) :126–31, Jan 2007.

[8] C Azuar, P Reyes, E Volle, S Kinkingnehun, E Bravo, R Kouneiher, B Dubois, E Koechlin, and RA Levy. Architecture of cognitive control in the human prefrontal cortex : A lesion behavior mapping study in patients with prefrontal lesions. *Society for Neuroscience*, Oct 2009.

[9] A Baddeley. Working memory. *Science*, 255(5044) :556–9, Jan 1992.

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

[11] David Badre and Mark D'Esposito. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of cognitive neuroscience*, 19(12) :2082–99, Dec 2007.

[12] David Badre and Anthony D Wagner. Computational and neurobiological mechanisms underlying cognitive flexibility. *Proc Natl Acad Sci USA*, 103(18) :7186–91, May 2006.

[13] Hannah M Bayer and Paul W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1) :129–41, Jul 2005.

[14] Paul M Bays and Masud Husain. Dynamic shifts of limited working memory resources in human vision. *Science*, 321(5890) :851–4, Aug 2008.

[15] Jeffrey M Beck and Alexandre Pouget. Exact inferences in a neural implementation of a hidden markov model. *Neural computation*, 19(5) :1344–61, May 2007.

[16] Timothy E. J Behrens, Laurence T Hunt, Mark W Woolrich, and Matthew F. S Rushworth. Associative learning of social value. *Nature*, 456(7219) :245–249, Nov 2008.

[17] Timothy E J Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nat Neurosci*, 10(9) :1214–1221, Sep 2007.

[18] R Bellman. The theory of dynamic programming. *Proceedings of the National Academy of Sciences of . . .*, Jan 1952.

[19] R Ernest Bellman. Introduction to the mathematical theory of control processes,Äé - page 137. page 245, Jan 1967.

[20] M Bertin, N Schweighofer, and K Doya. Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Networks*, 20(6) :668–675, Aug 2007.

[21] D Bertsekas. Dynamic programming : deterministic and stochastic models. 1987.

[22] G Blasi. Effect of catechol-o-methyltransferase val158met genotype on attentional control. *Journal of Neuroscience*, 25(20) :5038–5045, May 2005.

[23] Erie D Boorman, Timothy E J Behrens, Mark W Woolrich, and Matthew F S Rushworth. How green is the grass on the other side ? frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, 62(5) :733–43, Jun 2009.

[24] M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12(5) :201–208, May 2008.

[25] M Botvinick, J Cohen, and C Carter. Conflict monitoring and anterior cingulate cortex : an update. *Trends in Cognitive Sciences*, 8(12) :539–546, Dec 2004.

[26] M Botvinick, Y Niv, and A Barto. Hierarchically organized behavior and its neural foundations : A reinforcement learning perspective. *Cognition*, page 19, Oct 2008.

[27] Matthew Botvinick and David C Plaut. Doing without schema hierarchies : a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, 111(2) :395–429, Apr 2004.

[28] Matthew M Botvinick. Multilevel structure in behaviour and in the brain : a model of fuster's hierarchy. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 362(1485) :1615–1626, Apr 2007.

[29] Matthew M Botvinick and David C Plaut. Such stuff as habits are made on : A reply to cooper and shallice (2006). *Psychological Review*, 113(4) :917–928, Sep 2006.

[30] S Bouret and S Sara. Network reset : a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, 28(11) :574–582, Nov 2005.

[31] K Brodmann. Vergleichende lokalisationslehre der großhirnrinde in ihren prinzipien dargestellt auf grund des zellenbaues. *Leipzig : Barth*, 1909.

[32] A Brovelli, N Laksiri, B Nazarian, M Meunier, and D Boussaoud. Understanding the neural computations of arbitrary visuomotor learning through fmri and associative learning theory. *Cerebral Cortex*, 18(7) :1485–1495, Oct 2007.

[33] J Brown and T Braver. A computational model of risk, conflict, and individual difference effects in the anterior cingulate cortex. *Brain Research*, 1202 :99–108, Apr 2008.

[34] J Brown, J Reynolds, and T Braver. A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, 55(1) :37–85, Aug 2007.

[35] Joshua W Brown and Todd S Braver. Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712) :1118–21, Feb 2005.

[36] S Budhani, A.A Marsh, D.S Pine, and R.J.R Blair. Neural correlates of response reversal : Considering acquisition. *NeuroImage*, 34(4) :1754–1765, Feb 2007.

[37] P Calabresi, P Gubellini, D Centonze, B Picconi, G Bernardi, K Chergui, P Svenningsson, A A Fienberg, and P Greengard. Dopamine and camp-regulated phosphoprotein 32 kda controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. *Journal of Neuroscience*, 20(22) :8443–51, Nov 2000.

[38] M Carandini and D J Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264(5163) :1333–6, May 1994.

[39] S. R Chamberlain. Neurochemical modulation of response inhibition and probabilistic learning in humans. *Science*, 311(5762) :861–863, Feb 2006.

[40] J Chein and W Schneider. Neuroimaging studies of practice-related change : fmri and meta-analytic evidence of a domain-general control network for learning. *Cognitive Brain Research*, 25(3) :607–623, Dec 2005.

[41] J D Cohen and D Servan-Schreiber. Context, cortex, and dopamine : a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1) :45–77, Jan 1992.

[42] Jonathan D Cohen, Todd S Braver, and Joshua W Brown. Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, pages 1–7, 2002. 21-26 : reviews DA-PFC, models WM.

[43] Jonathan D Cohen, Samuel M Mcclure, and Angela J Yu. Should i stay or should i go ? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 362(1481) :933–942, Mar 2007.

[44] Michael X Cohen and Michael J Frank. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research*, 199(1) :141–156, Apr 2009.

[45] Christos Constantinidis and Emmanuel Procyk. The primate working memory networks. *Cognitive, affective & behavioral neuroscience*, 4(4) :444–65, Dec 2004.

[46] A Conway. Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, 7(12) :547–552, Dec 2003.

[47] R Cools. Differential responses in human striatum and prefrontal cortex to changes in object and rule relevance. *Journal of Neuroscience*, 24(5) :1129–1135, Feb 2004.

[48] R Cools, M. J Frank, S. E Gibbs, A Miyakawa, W Jagust, and M D'esposito. Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. *Journal of Neuroscience*, 29(5) :1538–1543, Feb 2009.

[49] Roshan Cools, Luke Clark, Adrian M Owen, and Trevor W Robbins. Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *Journal of Neuroscience*, 22(11) :4563–7, Jun 2002.

[50] Roshan Cools and Trevor W Robbins. Chemistry of the adaptive mind. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 362(1825) :2871–2888, Oct 2004.

[51] Richard P Cooper and Tim Shallice. Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4) :297–338, Sep 2000.

[52] Richard P Cooper and Tim Shallice. Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, 113(4) :887–916, Jan 2006.

[53] Richard P Cooper and Tim Shallice. Structured representations in the control of behavior cannot be so easily dismissed : A reply to botvinick and plaut (2006). *Psychological Review*, 113(4) :929–931, Oct 2006.

[54] G Corrado and K Doya. Understanding neural coding through the model-based analysis of decision making. *Journal of Neuroscience*, 27(31) :8178–8180, Aug 2007.

[55] R Crites and A Barto. An actor/critic algorithm that is equivalent to q-learning. *Advances in Neural Inf. Proc. Systems 7*, Jan 1995.

[56] E. A Crone. Neural evidence for dissociable components of task-switching. *Cerebral Cortex*, 16(4) :475–486, Jun 2005.

[57] J L Cummings. Anatomic and behavioral aspects of frontal-subcortical circuits. *Ann N Y Acad Sci*, 769 :1–13, Dec 1995.

[58] Nathaniel D Daw. Dopamine : at the intersection of reward and action. *Nat Neurosci*, 10(12) :1505–7, Dec 2007.

[59] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12) :1704–1711, Dec 2005.

[60] Nathaniel D Daw, John P O'doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095) :876–879, Jun 2006.

[61] P Dayan and Y Niv. Reinforcement learning : The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2) :185–196, Apr 2008.

[62] Peter Dayan and Bernard W Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36(2) :285–98, Oct 2002.

[63] Peter Dayan and Terrence J Sejnowski. Exploration bonuses and dual control. Jan 1996.

[64] Peter Dayan and Angela Yu. Phasic norepinephrine : A neural interrupt signal for unexpected events. *Network : Computation in Neural Systems*, 17(4) :335–350, Dec 2006.

[65] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian q-learning. Apr 1998.

[66] S Dehaene and J P Changeux. The wisconsin card sorting test : theoretical analysis and modeling in a neuronal network. *Cereb Cortex*, 1(1) :62–79, Jan 1991.

[67] S Dehaene and J P Changeux. A hierarchical neuronal network for planning behavior. *Proc Natl Acad Sci USA*, 94(24) :13293–8, Nov 1997.

[68] S Dehaene and J P Changeux. Reward-dependent learning in neuronal networks for planning and decision making. *Prog Brain Res*, 126 :217–29, Jan 2000.

[69] Sophie Deneve. Bayesian spiking neurons i : inference. *Neural computation*, 20(1) :91–117, Jan 2008.

[70] Sophie Deneve. Bayesian spiking neurons ii : learning. *Neural computation*, 20(1) :118–45, Jan 2008.

[71] B Doll, W Jacobs, A Sanfey, and M Frank. Instructional control of reinforcement learning : A behavioral and neurocomputational investigation. *Brain Research*, Aug 2009.

[72] Nico U.F Dosenbach, Kristina M Visscher, Erica D Palmer, Francis M Miezin, Kristin K Wenger, Hyunseon C Kang, E. Darcy Burgund, Ansley L Grimes, Bradley L Schlaggar, and Steven E Petersen. A core system for the implementation of task sets. *Neuron*, 50(5) :799–812, Jun 2006.

[73] Kenji Doya. Metalearning and neuromodulation. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6) :495–506, Jan 2002.

[74] Kenji Doya. Modulators of decision making. *Nat Neurosci*, 11(4) :410–416, Apr 2008.

[75] Kenji Doya, Kazuyuki Samejima, Ken ichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6) :1347–69, Jun 2002.

[76] Jean-Claude Dreher and Karen Faith Berman. Fractionating the neural substrate of cognitive control processes. *Proc Natl Acad Sci USA*, 99(22) :14595–600, Oct 2002.

[77] Jean-Claude Dreher, Philip Kohn, Bhaskar Kolachana, Daniel R Weinberger, and Karen Faith Berman. Variation in dopamine genes influences responsivity of the human reward system. *Proc Natl Acad Sci USA*, 106(2) :617–22, Jan 2009. Notes : N'apporte presque rien.

[78] Jean-Claude Dreher, Peter J Schmidt, Philip Kohn, Daniella Furman, David Rubinow, and Karen Faith Berman. Menstrual cycle phase modulates reward-related neural function in women. *Proc Natl Acad Sci USA*, 104(7) :2465–70, Feb 2007.

[79] T Egner. Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, 12(10) :374–380, Oct 2008.

[80] R W Engle, S W Tuholski, J E Laughlin, and A R Conway. Working memory, short-term memory, and general fluid intelligence : a latent-variable approach. *Journal of experimental psychology General*, 128(3) :309–31, Sep 1999.

[81] Jin Fan, John Fossella, Tobias Sommer, Yanghong Wu, and Michael I Posner. Mapping the genetic variation of executive attention onto brain activity. *Proc Natl Acad Sci USA*, 100(12) :7406–11, Jun 2003.

[82] M J Frank, B Loughry, and R C O'Reilly. Interactions between frontal cortex and basal ganglia in working memory : a computational model. *Cognitive, affective & behavioral neuroscience*, 1(2) :137–60, Jun 2001.

[83] Michael J Frank. Dynamic dopamine modulation in the basal ganglia : a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of cognitive neuroscience*, 17(1) :51–72, Jan 2005.

[84] Michael J Frank. Hold your horses : a dynamic computational role for the subthalamic nucleus in decision making. *Neural networks : the official journal of the International Neural Network Society*, 19(8) :1120–36, Oct 2006.

[85] Michael J Frank, Bradley B Doll, Jen Oas-Terpstra, and Francisco Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat Neurosci*, 12(8) :1062–8, Aug 2009.

[86] Michael J Frank, Ahmed A Moustafa, Heather M Haughey, Tim Curran, and Kent E Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc Natl Acad Sci USA*, 104(41) :16311–6, Oct 2007.

[87] Michael J Frank, Randall C O'reilly, and Tim Curran. When memory fails, intuition reigns : midazolam enhances implicit inference in humans. *Psychological science : a journal of the American Psychological Society / APS*, 17(8) :700–7, Aug 2006.

[88] Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. By carrot or by stick : cognitive reinforcement learning in parkinsonism. *Science*, 306(5703) :1940–3, Dec 2004.

[89] S Fusi, W Asaad, E Miller, and X Wang. A neural circuit model of flexible sensorimotor mapping : Learning and forgetting on multiple timescales. *Neuron*, 54(2) :319–333, Apr 2007.

[90] J M Fuster and G E Alexander. Neuron activity related to short-term memory. *Science*, 173(997) :652–4, Aug 1971.

[91] Joaquín M Fuster. Frontal lobe and cognitive development. *J Neurocytol*, 31(3-5) :373–85, Jan 2002.

[92] Charles R Gallistel, Stephen Fairhurst, and Peter Balsam. The learning curve : implications of a quantitative analysis. *Proc Natl Acad Sci USA*, 101(36) :13124–31, Sep 2004.

[93] J. C Gittins and D. M Jones. A dynamic allocation index for the sequential design of experiments. pages 241–266. Colloq. Math. Soc. János Bolyai, Vol. 9, 1974.

[94] T Goldberg and D Weinberger. Genes and the parsing of cognitive processes. *Trends in Cognitive Sciences*, 8(7) :325–335, Jul 2004.

[95] P S Goldman-Rakic. The prefrontal landscape : implications of functional architecture for understanding human mentation and the central executive. *Philos Trans R Soc Lond, B, Biol Sci*, 351(1346) :1445–53, Oct 1996.

[96] Jeffrey S Landau Eugene S Gollin. Successive reversal performance in young children as a function of the delay interval between reversls. *Child development*, page 11, March 1966.

[97] J Grafman. The structured event complex and the human prefrontal cortex. *Principles of frontal lobe function*, Jan 2002.

[98] Steven Graham, Elaine Phua, Chun Siong Soon, Tomasina Oh, Chris Au, Borys Shuter, Shih-Chang Wang, and Ing Berne Yeh. Role of medial cortical, hippocampal and striatal interactions during cognitive set-shifting. *NeuroImage*, 45(4) :1359–1367, May 2009.

[99] A.E Green, M.R Munafo, C.G DeYoug, John Fossella, Jin Fan, and J.R Gray. Using genetic data in cognitive neuroscience : from growing pains to genuine insights. *Nat Rev Neurosci*, 9 :710–720, Sep 2008.

[100] U Halsband and R Passingham. The role of premotor and parietal cortex in the direction of action. *Brain Research*, 240(2) :368–72, May 1982.

[101] A Hampshire and A. M Owen. Fractionating attentional control using event-related fmri. *Cerebral Cortex*, 16(12) :1679–1689, Dec 2005.

[102] A. N Hampton, P Bossaerts, and John P O'doherty. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32) :8360–8367, Aug 2006.

[103] Carolyn W Harley. Norepinephrine and dopamine as learning signals. *Neural Plast*, 11(3-4) :191–204, Jan 2004.

[104] H F Harlow. The formation of learning sets. *Psychological Review*, 56(1) :51–65, Jan 1949.

[105] Benjamin Y Hayden, John M Pearson, and Michael L Platt. Fictive reward signals in the anterior cingulate cortex. *Science*, 324(5929) :948–50, May 2009.

[106] T Hazy, M Frank, and R Oreilly. Banishing the homunculus : Making working memory work. *Neuroscience*, 139(1) :105–118, Apr 2006.

[107] Thomas E Hazy, Michael J Frank, and Randall C O'reilly. Towards an executive without a homunculus : computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 362(1485) :1601–1613, Apr 2007.

[108] Seth A Herd, Marie T Banich, and Randall C O'reilly. Neural mechanisms of cognitive control : an integrative model of stroop task performance and fmri data. *Journal of cognitive neuroscience*, 18(1) :22–32, Jan 2006.

[109] A L Hodgkin and A F Huxley. Propagation of electrical signals along giant nerve fibers. *Proc R Soc Lond, B, Biol Sci*, 140(899) :177–83, Oct 1952.

[110] A Hyafil, C Summerfield, and E Koechlin. Two mechanisms for task switching in the prefrontal cortex. *Journal of Neuroscience*, 29(16) :5135–5142, Apr 2009.

[111] H Imamizu and M Kawato. Neural correlates of predictive and postdictive switching mechanisms for internal models. *Journal of Neuroscience*, 28(42) :10751–10765, Oct 2008.

[112] Hiroshi Imamizu, Tomoe Kuroda, Toshinori Yoshioka, and Mitsuo Kawato. Functional magnetic resonance imaging examination of two modular architectures for switching multiple internal models. *J Neurosci*, 24(5) :1173–81, Feb 2004.

[113] Hiroshi Imamizu, Norikazu Sugimoto, Rieko Osu, Kiyoka Tsutsui, Kouichi Sugiyama, Yasuhiro Wada, and Mitsuo Kawato. Explicit contextual information selectively contributes to predictive switching of internal models. *Exp Brain Res*, 181(3) :395–408, Jul 2007.

[114] T Jaakkola, M Jordan, and S Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, Jan 1994.

[115] R Jacobs, M Jordan, and A Barto. Task decomposition through competition in a modular connectionist architecture : The what .... *Machine learning : from theory to applications : ...*, Jan 1993.

[116] R Jacobs, M Jordan, S Nowlan, and G Hinton. Adaptive mixtures of local experts. *Neural computation*, Jan 1991.

[117] Harry J Jerison and Dahlia W Zaidel. Evolution of the brain. neuropsychology. *Neuropsychology. Handbook of perception and cognition (2nd ed.). San Diego, CA, US : Academic Press.*, pages 53–82, 1994.

[118] Daphna Joel, Yael Niv, and Eytan Ruppin. Actor-critic models of the basal ganglia : new anatomical and computational perspectives. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6) :535–47, Jan 2002.

[119] Kevin Johnston, Helen M Levin, Michael J Koval, and Stefan Everling. Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. *Neuron*, 53(3) :453–462, Feb 2007. singes pro, anti saccades.

[120] M Jordan and R Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, Jan 1994.

[121] T Jubault, C Ody, and E Koechlin. Serial organization of human behavior in the inferior parietal cortex. *Journal of Neuroscience*, 27(41) :11028–11036, Oct 2007.

[122] Thorsten Kahnt, Soyoung Q Park, Michael X Cohen, Anne Beck, Andreas Heinz, and Jana Wrase. Dorsal striatal-midbrain connectivity in humans predicts how reinforcements are used to guide decisions. *Journal of cognitive neuroscience*, 21(7) :1332–45, Jul 2009.

[123] S Kakade and P Dayan. Dopamine : Generalization and bonuses. *Neural Networks*, Jan 2002.

[124] Michael J Kane and Randall W Engle. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence : an individual-differences perspective. *Psychonomic bulletin & review*, 9(4) :637–71, Dec 2002.

[125] M Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6) :718–27, Dec 1999.

[126] Steven W Kennerley, Mark E Walton, Timothy E J Behrens, Mark J Buckley, and Matthew F S Rushworth. Optimal decision making and the anterior cingulate cortex. *Nat Neurosci*, 9(7) :940–947, Jul 2006.

[127] Adam Kepecs, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210) :227–231, Sep 2008.

[128] Hackjin Kim, Shinsuke Shimojo, and John P O'doherty. Is avoiding an aversive outcome rewarding ? neural substrates of avoidance learning in the human brain. *PLoS Biol*, 4(8) :e233, Jul 2006.

[129] C Koch and I Segev. The role of single neurons in information processing. *Nat Neurosci*, 3 Suppl :1171–7, Nov 2000.

[130] E Koechlin, J L Anton, and Y Burnod. Bayesian inference in populations of cortical neurons : a model of motion integration and segmentation in area mt. *Biological cybernetics*, 80(1) :25–44, Jan 1999.

[131] E Koechlin, G Basso, P Pietrini, S Panzer, and J Grafman. The role of the anterior prefrontal cortex in human cognition. *Nature*, 399(6732) :148–51, May 1999.

[132] E Koechlin and A Hyafil. Anterior prefrontal function and the limits of human decision-making. *Science*, 318(5850) :594–598, Oct 2007.

[133] E Koechlin and T Jubault. Broca's area and the hierarchical organization of human behavior. *Neuron*, 50(6) :963–974, Jun 2006.

[134] E Koechlin and C Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6) :229–235, Jun 2007.

[135] Etienne Koechlin, Adrian Danek, Yves Burnod, and Jordan Grafman. Medial prefrontal and subcortical mechanisms underlying the acquisition of motor and cognitive action sequences in humans. *Neuron*, 35(2) :371–81, Jul 2002.

[136] Etienne Koechlin, Chrystèle Ody, and Frédérique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648) :1181–5, Nov 2003.

[137] Frédérique Kouneiher, Sylvain Charron, and Etienne Koechlin. Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci*, 12(7) :939–45, Jul 2009.

[138] Kai A Krueger and Peter Dayan. Flexible shaping : How learning in small steps helps. *Cognition*, 110(3) :380–394, Mar 2009.

[139] L Lapique. Recherches quantitatives sur l'excitation électrique de nerfs traitée comme une polarisation. *J. Physiol. Pathol. Gen.*, 1907.

[140] Conor Liston, Shanna Matalon, Todd A Hare, Matthew C Davidson, and B.J Casey. Anterior cingulate and posterior parietal cortices are sensitive to dissociable forms of conflict in a task-switching paradigm. *Neuron*, 50(4) :643–653, May 2006.

[141] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9(11) :1432–1438, Nov 2006.

[142] Farshad A Mansouri, Keiji Tanaka, and Mark J Buckley. Conflict-induced behavioural adjustment : a clue to the executive functions of the prefrontal cortex. *Nat Rev Neurosci*, 10(2) :141–152, Feb 2009.

[143] A McGovern, D Precup, B Ravindran, and S Singh. Hierarchical optimal control of mdps. *Proceedings of the Tenth Yale Workshop on Adaptive ...*, Jan 1998.

[144] F Mcnab, A Varrone, L Farde, A Jucaite, P Bystritsky, H Forssberg, and T Klingberg. Changes in cortical dopamine d1 receptor binding associated with cognitive training. *Science*, 323(5915) :800–802, Feb 2009.

[145] Fiona McNab and Torkel Klingberg. Prefrontal cortex and basal ganglia control access to working memory. *Nat Neurosci*, 11(1) :103–7, Jan 2008.

[146] Vincent Meininger. Neuro-anatomie. page 135, Jan 1983.

[147] E K Miller. The prefrontal cortex and cognitive control. *Nat Rev Neurosci*, 1(1) :59–65, Oct 2000.

[148] E K Miller and J D Cohen. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, 24 :167–202, Jan 2001.

[149] D Mitchell, R Rhodes, D Pine, and R Blair. The contribution of ventrolateral and dorsolateral prefrontal cortex to response reversal. *Behavioural Brain Research*, 187(1) :80–87, Feb 2008.

[150] S Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3) :134–140, Mar 2003.

[151] E A Murray, T J Bussey, and S P Wise. Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Experimental brain research Experimentelle Hirnforschung Expérimentation cérébrale*, 133(1) :114–29, Jul 2000.

[152] L. J Murray and C Ranganath. The dorsolateral prefrontal cortex contributes to successful relational memory encoding. *Journal of Neuroscience*, 27(20) :5515–5522, May 2007.

[153] A Nagano-Saito, M Leyton, O Monchi, Y. K Goldberg, Y He, and A Dagher. Dopamine depletion impairs frontostriatal functional connectivity during a set-shifting task. *Journal of Neuroscience*, 28(14) :3697–3706, Apr 2008.

[154] Randall O'Reilly Nicolas P Rougier. Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, page 18, Jul 2002.

[155] Kimberly G Noble, Bruce D Mccandliss, and Martha J Farah. Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Sci*, 10(4) :464–480, Jul 2007.

[156] J O'doherty. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669) :452–454, Apr 2004.

[157] J O'doherty, M L Kringelbach, E T Rolls, J Hornak, and C Andrews. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci*, 4(1) :95–102, Jan 2001.

[158] John P O'doherty, Alan Hampton, and Hackjin Kim. Model-based fmri and its application to reward learning and decision making. *Ann N Y Acad Sci*, 1104 :35–53, May 2007.

[159] R. C O'reilly. Biologically based computational models of high-level cognition. *Science*, 314(5796) :91–94, Oct 2006.

[160] Randall C O'reilly and Michael J Frank. Making working memory work : a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2) :283–328, Feb 2006.

[161] Randall C O'reilly, Michael J Frank, Thomas E Hazy, and Brandon Watz. Pvlv : the primary value and learned value pavlovian learning algorithm. *Behav Neurosci*, 121(1) :31–49, Feb 2007.

[162] A M Owen. Cognitive planning in humans : neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog Neurobiol*, 53(4) :431–50, Nov 1997.

[163] Anitha Pasupathy and Earl K Miller. Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433(7028) :873–6, Feb 2005.

[164] M Pessiglione, L Schmidt, B Draganski, R Kalisch, H Lau, R. J Dolan, and C. D Frith. How the brain translates money into force : A neuroimaging study of subliminal motivation. *Science*, 316(5826) :904–906, May 2007.

[165] Mathias Pessiglione, Predrag Petrovic, Jean Daunizeau, Stefano Palminteri, Raymond J Dolan, and Chris D Frith. Subliminal instrumental conditioning demonstrated in the human brain. *Neuron*, 59(4) :561–567, Aug 2008.

[166] Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J Dolan, and Chris D Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106) :1042–1045, Aug 2006.

[167] M Petrides. Motor conditional associative-learning after selective prefrontal lesions in the monkey. *Behavioural Brain Research*, 5(4) :407–13, Aug 1982.

[168] M Petrides. Visuo-motor conditional associative learning after frontal and temporal lesions in the human brain. *Neuropsychologia*, 35(7) :989–97, Jul 1997.

[169] M Petrides and D Pandya. Association pathways of the prefrontal cortex and functional observations. *Principles of frontal lobe function*, Jan 2002.

[170] Michael L Platt and Scott A Huettel. Risky business : the neuroeconomics of decision making under uncertainty. *Nat Neurosci*, 11(4) :398–403, Apr 2008.

[171] R Quilodran, M Rothe, and E Procyk. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*, 57(2) :314–325, Jan 2008.

[172] Antonio Rangel, Colin Camerer, and P. Read Montague. A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci*, 9(7) :545–556, Jul 2008.

[173] P Redgrave. What is reinforced by phasic dopamine signals ? *Brain Research Reviews*, 58(2) :322–339, Aug 2008.

[174] Peter Redgrave and Kevin Gurney. The short-latency dopamine signal : a role in discovering novel actions ? *Nat Rev Neurosci*, 7(12) :967–75, Dec 2006.

[175] R Rescorla. Informational variables in pavlovian conditioning. *The psychology of learning and motivation : Advances ...*, Jan 1972.

[176] J Reynolds and R O'Reilly. Developing pfc representations using reinforcement learning. *Cognition*, Jul 2009.

[177] K Richard Ridderinkhof, Markus Ullsperger, Eveline A Crone, and Sander Nieuwenhuis. The role of the medial frontal cortex in cognitive control. *Science*, 306(5695) :443–7, Oct 2004.

[178] Nicolas P Rougier, David C Noelle, Todd S Braver, Jonathan D Cohen, and Randall C O'reilly. Prefrontal cortex and flexible cognitive control : rules without symbols. *Proc Natl Acad Sci USA*, 102(20) :7338–43, May 2005.

[179] M Rosario Rueda, Mary K Rothbart, Bruce D Mccandliss, Lisa Saccomanno, and Michael I Posner. Training, maturation, and genetic influences on the development of executive attention. *Proc Natl Acad Sci USA*, 102(41) :14931–6, Oct 2005.

[180] M F S Rushworth, M E Walton, S W Kennerley, and D M Bannerman. Action sets and decisions in the medial frontal cortex. *Trends Cogn Sci (Regul Ed)*, 8(9) :410–7, Sep 2004.

[181] Matthew F S Rushworth and Timothy E J Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci*, 11(4) :389–397, Apr 2008.

[182] Matthew F S Rushworth, Mark J Buckley, Timothy E J Behrens, Mark E Walton, and David M Bannerman. Functional organization of the medial frontal cortex. *Curr Opin Neurobiol*, 17(2) :220–7, Apr 2007.

[183] MFS Rushworth and ME Walton. The anterior cingulate cortex : rewrd-guided action selection and the value of actions. page 39, Jul 2006.

[184] Katsuyuki Sakai. Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, page 29, Apr 2008. Voir summary points list.

[185] Jérôme Sallet, René Quilodran, Marie Rothé, Julien Vezoli, Jean-Paul Joseph, and Emmanuel Procyk. Expectations, gains, and losses in the anterior cingulate cortex. *Cognitive, affective & behavioral neuroscience*, 7(4) :327–36, Dec 2007.

[186] Kazuyuki Samejima and Kenji Doya. Multiple representations of belief states and action values in corticobasal ganglia loops. *Ann N Y Acad Sci*, 1104 :213–28, May 2007.

[187] Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. Representation of action-specific reward values in the striatum. *Science*, 310(5752) :1337–40, Nov 2005.

[188] A. G Sanfey. Social decision-making : Insights from game theory and neuroscience. *Science*, 318(5850) :598–602, Oct 2007.

[189] Susan J Sara. The locus coeruleus and noradrenergic modulation of cognition. *Nat Rev Neurosci*, 10(3) :211–223, Mar 2009.

[190] T Schonberg, N. D Daw, D Joel, and J. P O'doherty. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, 27(47) :12860–12867, Nov 2007.

[191] Franz Schubert. Quintette en ut majeur, d956. 1889.

[192] W Schultz. A neural substrate of prediction and reward. *Science*, 275(5306) :1593–1599, Mar 1997.

[193] Wolfram Schultz. Getting formal with dopamine and reward. *Neuron*, 36(2) :241–63, Oct 2002.

[194] Nicolas Schweighofer, Saori C Tanaka, and Kenji Doya. Serotonin and the evaluation of future rewards : theory, experiments, and possible neural mechanisms. *Ann N Y Acad Sci*, 1104 :289–300, May 2007.

[195] K Semendeferi, A Lu, N Schenker, and H Damasio. Humans and great apes share a large frontal cortex. *Nat Neurosci*, 5(3) :272–6, Mar 2002.

[196] Ben Seymour, John P O'doherty, Peter Dayan, Martin Koltzenburg, Anthony K Jones, Raymond J Dolan, Karl J Friston, and Richard S Frackowiak. Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992) :664–7, Jun 2004.

[197] Keith M Shafritz, Paul Kartheiser, and Aysenil Belger. Dissociation of neural systems mediating shifts in behavioral response and cognitive set. *NeuroImage*, 25(2) :600–6, Apr 2005.

[198] E R Sowell, P M Thompson, C J Holmes, T L Jernigan, and A W Toga. In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nat Neurosci*, 2(10) :859–61, Oct 1999.

[199] Anja Stemme, Gustavo Deco, and Astrid Busch. The neurodynamics underlying attentional control in set shifting tasks. *Cognitive neurodynamics*, 1(3) :249–59, Sep 2007.

[200] Anja Stemme, Gustavo Deco, and Astrid Busch. The neuronal dynamics underlying cognitive flexibility in set shifting tasks. *Journal of computational neuroscience*, 23(3) :313–31, Dec 2007.

[201] B A Strange, R N Henson, K J Friston, and R J Dolan. Anterior prefrontal cortex mediates rule learning in humans. *Cereb Cortex*, 11(11) :1040–6, Nov 2001.

[202] Par Richard S Sutton and Andrew G Barto. Reinforcement learning. 1998.

[203] R Sutton, D Precup, and S Singh. Between mdps and semi-mdps : A framework for temporal abstraction in reinforcement .... *Artificial intelligence*, Jan 1999.

[204] S Tanaka, K Samejima, G Okada, K Ueda, Y Okamoto, S Yamawaki, and K Doya. Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Networks*, 19(8) :1233–1241, Oct 2006.

[205] Saori C Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, and Shigeto Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci*, 7(8) :887–93, Aug 2004.

[206] M Turnock and S Becker. A neural network model of hippocampal–striatal–prefrontal interactions in contextual conditioning. *Brain Research*, 1202 :87–98, Apr 2008.

[207] M Usher, J D Cohen, D Servan-Schreiber, J Rajkowski, and G Aston-Jones. The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401) :549–54, Jan 1999.

[208] V. V Valentin, A Dickinson, and J. P O'doherty. Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, 27(15) :4019–4026, Apr 2007.

[209] Tor D Wager, John Jonides, Edward E Smith, and Thomas E Nichols. Toward a taxonomy of attention shifting : individual differences in fmri during multiple shift types. *Cognitive, affective & behavioral neuroscience*, 5(2) :127–43, Jun 2005.

[210] Jonathan D Wallis. Orbitofrontal cortex and its contribution to decision-making. *Annu Rev Neurosci*, 30 :31–56, Jan 2007.

[211] C Watkins and P Dayan. Q-learning. *Machine learning*, Jan 1992.

[212] R A Wise and P P Rompre. Brain dopamine and reward. *Annual review of psychology*, 40 :191–225, Jan 1989.

[213] Bianca C Wittmann, Nathaniel D Daw, Ben Seymour, and Raymond J Dolan. Striatal activity underlies novelty-based choice in humans. *Neuron*, 58(6) :967–73, Jun 2008.

[214] D Wolpert and M Kawato. Multiple paired forward and inverse models for motor control. *Neural networks : the official journal of the International Neural Network Society*, 11(7-8) :1317–1329, Oct 1998.

[215] Tao Wu, Kenji Kansaku, and Mark Hallett. How self-initiated memorized movements become automatic : a functional mri study. *J Neurophysiol*, 91(4) :1690–8, Apr 2004.

[216] Wako Yoshida and Shin Ishii. Resolution of uncertainty in prefrontal cortex. *Neuron*, 50(5) :781–789, Jun 2006.

[217] A Yu and P Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4) :681–692, May 2005.