HAL
open science

Codage hippocampique par transitions spatio-temporelles pour l'apprentissage autonome de comportements dans des tâches de navigation sensori-motrice et de planification en robotique

Julien Hirel

To cite this version:

Julien Hirel. Codage hippocampique par transitions spatio-temporelles pour l'apprentissage autonome de comportements dans des tâches de navigation sensori-motrice et de planification en robotique. Apprentissage [cs.LG]. Université de Cergy Pontoise, 2011. Français. NNT : . tel-00660862

HAL Id: tel-00660862

https://theses.hal.science/tel-00660862v1

Submitted on 17 Jan 2012

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.

Université de Cergy-Pontoise - Ecole doctorale Sciences et Ingénierie

THÈSE

présentée pour obtenir le titre de DOCTEUR en Sciences et Technologies de l'Information et de la Communication

---

CODAGE HIPPOCAMPIQUE PAR TRANSITIONS SPATIO-TEMPORELLES POUR L'APPRENTISSAGE AUTONOME DE COMPORTEMENTS DANS DES TÂCHES DE NAVIGATION SENSORI-MOTRICE ET DE PLANIFICATION EN ROBOTIQUE

---

par

Julien Hirel

ETIS - ENSEA / Université de Cergy-Pontoise / CNRS UMR 8051
6 avenue du Ponceau, 95014 Cergy-Pontoise Cedex, France

Soutenue le 06 décembre 2011 devant le jury composé de :

P. GAUSSIER,	ETIS, Université de Cergy-Pontoise	Co-directeur de thèse
M. QUOY,	ETIS, Université de Cergy-Pontoise	Co-directeur de thèse
R. CHATILA,	ISIR, CNRS/Université Paris 6	Rapporteur
F. ALEXANDRE,	LORIA, INRIA Nancy	Rapporteur
A. ARLEO,	NPA, CNRS/Université Paris 6	Examinateur
B. POUCET,	LNC, CNRS/Aix-Marseille Université	Examinateur
S. WIENER,	LPPA, CNRS/Collège de France	Examinateur
E. CRÜCK,	DGA/DS/MRIS	Examinateur

Remerciements

Je souhaiterais sincèrement remercier ici toutes les personnes qui ont contribué à la rédaction et à l'aboutissement de cette thèse. Elle est le fruit de trois années passées en compagnie de nombreuses personnes qui ont chacune apporté leur pierre à l'édifice. Il sera difficile de toutes les citer ici mais j'espère que ceux que j'aurais oublié se reconnaîtront.

Mes premiers remerciements iront à mes directeurs de thèse, Philippe Gaussier et Mathias Quoy, qui m'ont guidé tout au long de ce périple. J'ai pu partager leur curiosité et leur intérêt pour les neurosciences et la robotique et apercevoir grâce à eux cette vision globale des travaux passés, présents et futurs de l'équipe. Leurs conseils et les nombreuses discussions que nous avons pu avoir m'ont permis de prendre le recul nécessaire pour aborder une partie de la tâche herculéenne qu'est la modélisation du cerveau.

Je remercie également les membres de mon jury qui ont pris le temps de lire, revoir et commenter mes travaux ainsi que de se déplacer à Cergy, dans la lointaine banlieue parisienne, pour assister à ma soutenance. Merci, donc, à Frédéric Alexandre et Raja Chatila pour leur rôle de rapporteur et à Angelo Arleo, Bruno Poucet, Sidney Wiener et Eva Crück pour leur temps et l'évaluation de mes travaux, ainsi que les discussions sur la neurobiologie que nous avons pu avoir au travers de partenariat entre nos laboratoires.

De manière générale, je voudrais dire merci à l'équipe Neurocybernétique et au laboratoire ETIS de m'avoir accueilli chaleureusement durant cette thèse. L'excellente ambiance de travail sera probablement un des meilleurs souvenirs que je garderai de cette période de ma vie. Inbar Fijalkow a certainement joué un rôle, par sa direction du laboratoire, dans les excellentes conditions d'accueil des doctorants et du personnel en général dans le laboratoire. N'oublions pas non plus le CNRS et la DGA, pour le financement de ma thèse, sans qui je ne serais pas en train d'écrire ce manuscrit en cet instant.

De manière plus personnelle, je tiens à remercier individuellement les membres de l'équipe neurocyber avec qui j'ai eu le plaisir de travailler pendant ces trois années. Les nombreuses discussions, pas toujours scientifiques, qui ont occupé nos repas me manqueront. Merci à Matthieu Lagarde pour m'avoir transmis de nombreuses connaissances sur les outils informatiques de l'équipe, Christophe Giovannangeli pour m'avoir transmis ses travaux (touffus !), Nicolas Cuperlier pour son intérêt sans faille pour ses collègues et leurs travaux, Frédéric Demelo pour sa capacité à générer des débats incongrus, Cyril Hasson pour son côté provocateur et son ouverture d'esprit, Sofiane Boucenna pour sa bonne humeur et ses explications sur le Ramadan, Pierre Andry pour m'avoir fait découvrir qu'un prof pouvait aussi être un skater, Arnaud Blanchard pour sa bonne humeur et pour avoir pris soin de Promethe, Antoine de Rengervé pour sa gentillesse et sa volonté de rendre service, Adrien Jauffret pour ses blagues (pas toujours forcément drôles !), ainsi que tous les autres, qui me pardonneront, j'espère, de ne pas voir leur nom ici.

Je souhaiterais aussi remercier mes amis et colocataires qui ont parfois dû me supporter dans les périodes de stress d'avant-deadline. Enfin, comment ne pas remercier mes parents et ma soeur, dont la présence infaillible à mes côtés tout au long de ma vie a fait de moi la personne que je suis aujourd'hui.

Résumé

Cette thèse s'intéresse aux mécanismes permettant de faciliter l'acquisition autonome de comportements chez les êtres vivants et propose d'utiliser ces mécanismes dans le cadre de tâches robotiques. Des réseaux de neurones artificiels sont utilisés pour modéliser certaines structures cérébrales, à la fois afin de mieux comprendre le fonctionnement de ces structures dans le cerveau des mammifères et pour obtenir des algorithmes robustes et adaptatifs de contrôle en robotique.

Les travaux présentés se basent sur un modèle de l'hippocampe permettant d'apprendre des relations temporelles entre des événements perceptifs. Les neurones qui forment le substrat de cet apprentissage, appelés *cellules de transition*, permettent de faire des prédictions sur les événements futurs que le robot pourrait rencontrer. Ces transitions servent de support à la construction d'une carte cognitive, située dans le cortex préfrontal et/ou pariétal. Cette carte peut être apprise lors de l'exploration d'un environnement inconnu par un robot mobile et ensuite utilisée pour planifier des chemins lui permettant de rejoindre un ou plusieurs buts.

Outre leur utilisation pour la construction d'une carte cognitive, les cellules de transition servent de base à la conception d'un modèle d'apprentissage par renforcement. Une implémentation neuronale de l'algorithme de Q-learning, utilisant les transitions, est réalisée de manière biologiquement plausible en s'inspirant des ganglions de la base. Cette architecture fournit une stratégie de navigation alternative à la planification par carte cognitive, avec un apprentissage plus lent, et correspondant à une stratégie automatique de bas-niveau. Des expériences où les deux stratégies sont utilisées en coopération sont réalisées et des lésions du cortex préfrontal et des ganglions de la base permettent de reproduire des résultats expérimentaux obtenus chez les rats.

Les cellules de transition peuvent apprendre des relations temporelles précises permettant de prédire l'instant où devrait survenir un événement. Dans un modèle des interactions entre l'hippocampe et le cortex préfrontal, nous montrons comment ces prédictions peuvent expliquer certains enregistrements in-vivo dans ces structures cérébrales, notamment lorsqu'un rat réalise une tâche durant laquelle il doit rester immobile pendant 2 secondes sur un lieu but pour obtenir une récompense. L'apprentissage des informations temporelles provenant de l'environnement et du comportement permet de détecter des régularités. A l'opposé, l'absence d'un événement prédit peut signifier un échec du comportement du robot, qui peut être détecté et utilisé pour adapter son comportement en conséquence. Un système de détection de l'échec est alors développé, tirant parti des prédictions temporelles fournies par l'hippocampe et des interactions entre les aspects de modulation comportementale du cortex préfrontal et d'apprentissage par renforcement dans les ganglions de la base. Plusieurs expériences robotiques sont conduites dans lesquelles ce signal est utilisé pour moduler le comportement d'un robot, dans un premier temps de manière immédiate, afin de mettre fin aux actions du robot qui le mènent à un échec et envisager d'autres stratégies. Ce signal est ensuite utilisé de manière plus permanente pour moduler l'apprentissage des associations menant à la sélection d'une action, afin que les échecs répétés d'une action dans un contexte particulier fassent oublier cette association.

Finalement, après avoir utilisé le modèle dans le cadre de la navigation, nous montrons ses capacités de généralisation en l'utilisant pour le contrôle d'un bras robotique. Ces travaux constituent une étape importante pour l'obtention d'un modèle unifié et générique permettant le contrôle de plates-formes robotiques variés et pouvant apprendre à résoudre des tâches de natures différentes.

Abstract

This thesis takes interest in the mechanisms facilitating the autonomous acquisition of behaviors in animals and proposes to use these mechanisms in the frame of robotic tasks. Artificial neural networks are used to model cerebral structures, both to understand how these structures work and to design robust and adaptive algorithms for robot control.

The work presented here is based on a model of the hippocampus capable of learning the temporal relationship between perceptive events. The neurons performing this learning, called *transition cells*, can predict which future events the robot could encounter. These transitions support the building of a cognitive map, located in the prefrontal and/or parietal cortex. The map can be learned by a mobile robot exploring an unknown environment and then be used to plan paths in order to reach one or several goals.

Apart from their use in building a cognitive map, transition cells are also the basis for the design of a model of reinforcement learning. A biologically plausible neural implementation of the Q-learning algorithm, using transitions, is made by taking inspiration from the basal ganglia. This architecture provides an alternative strategy to the cognitive map planning strategy. The reinforcement learning strategy requires a longer learning period but corresponds more to an automatic low-level behavior. Experiments are carried out with both strategies used in cooperation and lesions of the prefrontal cortex and basal ganglia allow to reproduce experimental results obtained with rats.

Transition cells can learn temporally precise relations predicting the exact timing when an event should be perceived. In a model of interactions between the hippocampus and prefrontal cortex, we show how these predictions can explain in-vivo recordings in these cerebral structures, in particular when rat is carrying out a task during which it must remain stationary for 2 seconds on a goal location to obtain a reward. The learning of temporal information about the environment and the behavior of the robot allows the system to detect regularity. On the contrary, the absence of a predicted event can signal a failure in the behavior of the robot, which can be detected and acted upon in order to modulate the failing behavior. Consequently, a failure detection system is developed, taking advantage of the temporal predictions provided by the hippocampus and the interaction between behavior modulation functions in the prefrontal cortex and reinforcement learning in the basal ganglia. Several robotic experiments are conducted, in which the failure signal is used to modulate, immediately at first, the behavior of the robot in order to stop selecting actions which lead to failures and explore other strategies. The signal is then used in a more lasting way by modulating the learning of the associations leading to the selection of an action so that the repeted failures of an action in a particular context lead to the suppression of this association.

Finally, after having used the model in the frame of navigation, we demonstrate its generalization capabilities by using it to control a robotic arm in a trajectory planning task. This work constitutes an important step towards obtaining a generic and unified model allowing the control of various robotic setups and the learning of tasks of different natures.

Table des matières

Introduction **13**

1 Fondements neurobiologiques et éthologiques **19**
1.1 Région hippocampique . . . 19
1.1.1 Anatomie . . . 19
1.1.2 Corrélats spatiaux des activités neuronales . . . 22
1.1.3 Multi-modalité . . . 24
1.1.4 Conditionnement . . . 25
1.1.5 Codage temporel . . . 25
1.2 Cortex préfrontal . . . 26
1.2.1 Anatomie . . . 26
1.2.2 Contrôle inhibiteur du comportement . . . 27
1.2.3 Planification, processus cognitifs et adaptation . . . 27
1.3 Ganglions de la base . . . 28
1.3.1 Anatomie . . . 28
1.3.2 Circuit de récompense . . . 28
1.4 Tâche de navigation continue . . . 29
1.4.1 Protocole expérimental . . . 30
1.4.2 Corrélats spatiaux dans le cortex préfrontal . . . 31
1.4.3 Activité hors-champ des cellules de lieu . . . 32
1.4.4 Lien entre l'activité temporelle hippocampique et frontale . . . 34
1.5 Conclusion . . . 35

2 Modèles bio-inspirés de navigation et planification en robotique **37**
2.1 Etat de l'art des systèmes de navigation bio-inspirés . . . 37
2.1.1 Modélisation de cellules de lieu . . . 37
2.1.2 Cartes cognitives et navigation vers un but . . . 39
2.1.3 Apprentissage temporel . . . 42
2.1.4 Apprentissage par renforcement . . . 43
2.2 Fonctionnement de notre modèle . . . 44
2.2.1 Apprentissage de cellules de lieu . . . 44
2.2.2 Transitions et carte cognitive . . . 48
2.2.3 Apprentissage de séquences temporelles . . . 53
2.3 Discussion sur les modèles existants . . . 56

3 Apprentissage temporel de séquences d'événements perceptifs multi-modaux 59
3.1 Précision et adaptation des prédictions temporelles 60
3.1.1 Modèle de décomposition spectrale du temps 60
3.1.2 Apprentissage continu et généralisé 62
3.2 Utilisation du modèle 66
3.2.1 Expérience robotique d'apprentissage multi-modal 66
3.2.2 Compatibilité avec la carte cognitive 70
3.3 Discussion et applications 74

4 Codage de l'information temporelle et des buts : interaction hippocampo-corticale 77
4.1 Modélisation des interactions hippocampe-cortex préfrontal 77
4.1.1 Cadre expérimental : la tâche de navigation continue 77
4.1.2 Activités hors-champ de cellules de lieu 80
4.1.3 Sélection de l'action et arrêt du mouvement 83
4.2 Validation expérimentale 87
4.2.1 Expérience en simulation et activités hors-champ 87
4.2.2 Expérience sur robot réel et interactions Homme-Machine 90
4.3 Discussion et limitations 93

5 Ganglions de la base et apprentissage par renforcement 97
5.1 Etat de l'art des modèles neuronaux d'apprentissage par renforcement 98
5.1.1 Conditionnement et prédiction de récompense 98
5.1.2 TD-learning et modèles acteur-critique 99
5.1.3 Sélection de l'action et stratégies multiples 101
5.2 Modèle de Q-learning utilisant les neurones de transition 103
5.2.1 Implémentation neuronale 103
5.2.2 Modèle biologiquement plausible 106
5.2.3 Navigation multi-buts 109
5.2.4 Expériences en simulation 109
5.3 Apprentissage par renforcement et carte cognitive : complémentarité et coopération 111
5.3.1 Caractéristiques des deux systèmes 111
5.3.2 Expériences de coopération et données de lésion 112
5.4 Discussion 116

6 Utilisation de la détection de l'échec dans la modulation comportementale 119
6.1 Transitions temporelles et détection de l'échec 120
6.1.1 Modèle neuronal de détection de l'échec 120
6.1.2 Intégration dans le modèle des interactions hippocampe - cortex préfrontal - ganglions de la base 123
6.2 Modulation comportementale immédiate 126
6.2.1 Tâche de navigation continue et essais d'extinction 126
6.2.2 Apprentissage d'une ronde avec des points d'arrêt 129
6.2.3 Navigation multi-buts avec inhibition de chemins 133
6.3 Modulation comportementale à moyen terme 136
6.3.1 Modulation des associations transition-action 136
6.3.2 Apprentissage autonome de la tâche de navigation continue 139

6.4 Discussion 142

7 Vers un modèle générique d'apprentissage de tâches robotiques 145
7.1 Généralisation à des plates-formes robotiques variées 146
7.1.1 Planification avec un bras robotique 146
7.1.2 Intégration bras/robot mobile 155
7.2 Généralisation des signaux traités 159
7.3 Discussion 162

Conclusion 165

A Apprentissage temporel de signaux continus 187

B Outils de simulation de réseaux de neurones 189
B.1 Interface de conception : Coeos 189
B.2 Simulateur temps-réel distribué : Promethe 191

C Plates-formes robotiques 193
C.1 Robot mobile d'intérieur 193
C.2 Robot mobile d'extérieur 193
C.3 Bras robotique 195

La Nature n'utilise que les plus longs fils pour tisser ses motifs, de sorte que la plus petite pièce révèle la structure de la tapisserie toute entière.

– Richard Feynman

Introduction

La robotique est une discipline récente qui fait face à des problèmes complexes. Les actions qu'un humain accomplit sans même y réfléchir (prendre un livre dans une pile de livre, se déplacer dans une foule etc.) peuvent être difficiles à réaliser pour un système robotique. Par exemple, une grande communauté en robotique mobile travaille sur le "SLAM" (Simultaneous Localization And Mapping) qui pose la problématique de la construction d'une carte de l'environnement, tout en étant capable de se localiser dans cette carte à tout moment [Chatila and Laumond, 1985; Moutarlier and Chatila, 1990; Leonard and Durrant-Whyte, 1991]. Le terme d'*intelligence artificielle* (IA) a été introduit par Marvin Minsky et défini comme "la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique"[1]. Ainsi, la robotique constitue un excellent domaine d'application pour l'intelligence artificielle. Une grande partie des travaux effectués dans le domaine de l'IA adoptent une approche symbolique. Cette approche repose sur une représentation du monde sous forme de symboles et de relations entre ces symboles. Selon ce courant, les traitements cognitifs effectués par des systèmes intelligents se résumeraient à la manipulation de symboles. Cette approche fait cependant face au "Symbol Grounding Problem" [Harnad, 1990]. En d'autres termes, comment des symboles abstraits peuvent-il acquérir un sens ? Comment relier ces symboles à ce qu'ils représentent dans le monde réel ? Certains chercheurs affirment que ce problème est maintenant résolu, car les systèmes les plus récents peuvent apprendre seuls à former leur représentation symbolique du monde [Steels, 2008]. Ils s'accordent cependant sur le caractère indispensable de l'interaction entre l'agent qui apprend cette représentation et son environnement.

Cette interaction rejoint la notion d'enaction proposée par Varela [Varela et al., 1991]. Cette vue s'éloigne de l'approche symbolique de la cognition et met en avant l'importance de l'incarnation ("embodiment" en anglais). La cognition "incarnée" se ferait donc de manière subjective à travers le vécu de chaque être vivant et non à travers la manipulation de concepts abstraits. Notre démarche, dans cette thèse, s'inscrira dans cette approche de l'intelligence artificielle qui s'écarte des symboles. Ainsi, la sensori-motricité sera au cœur des systèmes développés. C'est la conjonction entre les perceptions et les actions du robot qui lui permettront de se construire une représentation de son environnement, des conséquences de ses actions sur l'environnement et in fine de se construire une représentation de lui même.

Les sciences du vivant sont une grande source d'inspiration pour les informaticiens. La nature a depuis longtemps trouvé des solutions à des problèmes sur lesquels les approches algorith-

[1] http://www.larousse.fr/encyclopedie/article/LIntelligence_Artificielle/11011577

miques classiques buttent encore. A travers l'étude du vivant et de ses mécanismes, on peut donc s'inspirer de ces solutions. L'observation des processus dynamiques existants dans le monde biologique a permis de mettre en place des modèles génériques et des principes algorithmiques utilisés dans une multitude de domaines (recherche opérationnelle, processus d'optimisation, etc.). En robotique, l'approche "animat" [Wilson, 1991], contraction d'animal artificiel, considère le robot comme un animal. Le robot/animat s'inspire du comportement de l'animal et de ses capacités d'adaptation. C'est dans ce cadre qu'ont été réalisés les travaux effectués pendant cette thèse.

Parmi les approches bio-inspirées utilisées en robotique, on peut citer les algorithmiques génétiques [Holland, 1975], très utilisés pour faire de l'optimisation. Ceux-ci reprennent le concept de l'évolution et de la sélection naturelle introduit par Darwin [Darwin, 1859]. Dans cette approche, le but est de proposer un ensemble de solutions, codées sous la forme d'un code "génétique". L'algorithme commence par sélectionner les meilleures solutions grâce à une fonction d'évaluation. En les faisant ensuite muter et se reproduire, l'algorithme conduit après de nombreuses itérations à une population donnant un ensemble de solutions tendant à maximiser le critère d'évaluation. Un autre exemple est celui des algorithmes de fourmis, qui s'inspirent du système de phéromones répandues par les fourmis pour marquer des chemins [Dorigo et al., 1996]. Ces algorithmes ont des applications dans l'exploration d'espaces et la recherche de chemins optimaux. Ils mettent en place des règles simples faisant émerger rapidement des solutions pas forcément optimales mais donnant de bonnes performances.

Les réseaux de neurones artificiels correspondent à un niveau différent de modélisation du vivant. Ils s'inspirent du fonctionnement même de l'être vivant, en interne, plutôt que d'une dynamique collective ou d'un mécanisme de sélection naturelle. Souvent, l'inspiration biologique donne naissance à des outils informatiques qui sont alors utilisés sans la contrainte de rester proche du système biologique d'origine. Ainsi les réseaux de neurones artificiels ont donné naissance à des outils de classification tels que les perceptrons [Rosenblatt, 1962]. Pour la résolution de problèmes non linéaires, des perceptrons multi-couches ont été utilisés [Rumelhart et al., 1986; Cun, 1985]. Plus récemment, les "deep networks" ont permis la résolution de problèmes de complexité croissante [Larochelle et al., 2007]. Ces réseaux de neurones utilisent un mécanisme de rétro-propagation du gradient d'erreur pour l'apprentissage. Ils sont couramment utilisés dans de multiples domaines impliquant de l'apprentissage supervisé et de la classification automatique.

Dans cette thèse, nous nous intéresserons plutôt aux modèles de réseaux de neurones artificiels qui ont vocation à rester proches de leurs équivalents biologiques et à reproduire les activités observées dans certaines structures cérébrales. Cela traduit une volonté double de transformer une inspiration biologique en un ensemble d'applications concrètes mais aussi d'améliorer la compréhension même de ce système biologique. La *neurocybernétique* désigne l'étude des interactions et mécanismes de communication de l'information dans les systèmes neuronaux. Le terme cybernétique a été inventé par Norbert Wiener en 1947 et peut être défini comme la science des systèmes auto-régulés. L'intérêt est porté sur les mécanismes d'interaction des différents composants du système qui peuvent faire émerger un comportement global. Ainsi, l'approche dite *bottom-up* en anglais vise à partir de la conception de ces composants simples pour laisser émerger des comportements complexes grâce à leur interaction. C'est dans cette approche que se situe la conception de réseaux de neurones. L'interaction entre une multitude de neurones modélisés par des équations d'activation et d'apprentissage simples peut mener un système à

traiter et catégoriser des formes complexes. On rejoint ainsi le concept de la subsomption en robotique, décrit par Rodney Brooks [Brooks, 1991]. Il suggère qu'un comportement intelligent peut être décomposé en une multitude de modules de comportements simples. Ces comportements seraient organisés hiérarchiquement en différents étages. Les informations fournies par les étages de bas-niveau (correspondant aux comportements les plus simples) remonteraient et participeraient à la prise de décision dans les étages supérieurs. Ces travaux ont été précurseurs d'une approche comportementale de la robotique. Des architectures de contrôle dans lesquelles de multiples comportements sont sommés ou arbitrés ont ainsi vu le jour [Arkin, 1998]. L'utilisation de réseaux de neurones artificiels permet de profiter de cette propriété d'émergence de comportements. Le calcul parallèle est aussi une propriété fondamentale des réseaux de neurones. Cela en fait un outil efficace dans des applications temps-réel.

L'approche neuromimétique de la robotique ne peut se faire que dans le cadre d'une étroite collaboration entre neurobiologistes, modélisateurs et roboticiens. Les neurobiologistes fournissent des données comportementales et électrophysiologiques lors de tâches réalisées par des animaux ou humains. Ces données, ainsi que l'anatomie des structures cérébrales étudiées tout au long de cette thèse, seront présentées dans le chapitre 1. Les modélisateurs, quant à eux, intègrent ces données pour créer des modèles computationnels du flot de l'information dans les différentes structures cérébrales impliquées. Les roboticiens implémentent ces modèles sur des substrats informatiques et matériels afin de produire des algorithmes de contrôle pour des robots, validant ou invalidant ainsi le fonctionnement du modèle. Cette thèse se focalise sur la modélisation biologiquement plausible de mécanismes cérébraux appliquée à la réalisation de tâches robotiques. Le paradigme de recherche est plutôt celui d'une modélisation de haut niveau des interactions entre structures cérébrales. Ainsi, c'est plutôt l'aspect fonctionnel de ces structures et comment leur interaction renforce ces fonctions qui nous intéressera. Nous verrons qu'une même structure cérébrale peut servir à des fonctions multiples au sein d'un modèle unique. Je n'utiliserai pas dans cette thèse de modèles de neurones à décharge, ou bien de modèles détaillés de colonnes corticales. Les implémentations neuronales réalisées seront faites avec des neurones à fréquence de décharge. L'apprentissage synaptique sera régi par des équations différentielles, plutôt que par des modélisations plus fines telles que la STDP (*Spike-Timing-Dependent-Plasticity*), prenant en compte les décharges individuelles des neurones [Song et al., 2000].

Une grande partie des travaux présentés dans cette thèse concernera la thématique de la navigation et de la planification chez les animaux et des application en robotique mobile. Le cadre de ces travaux ne se limite toutefois pas à la navigation et nous verrons comment les modèles développés peuvent être utilisés pour contrôler diverses plates-formes robotiques. Il sera question de savoir comment les animaux apprennent à former des représentations spatiales et/ou temporelles internes des environnements dans lesquels ils évoluent. La nature de ces représentations et le code neuronal utilisé sera un point crucial. De manière plus générale, cette thèse se place dans le cadre de *l'acquisition et de l'adaptation autonome de comportements*. L'objectif est d'obtenir un modèle générique d'apprentissage autonome de tâches robotiques. La question de l'autonomie dans l'acquisition par le robot de représentations concernant ses perceptions, comportements et buts est cruciale. Le robot doit être capable de construire cette représentation par lui-même. Cette construction peut se faire de manière isolée par l'utilisation de stratégies d'essai/erreur ou en interaction avec un partenaire. Ainsi lorsqu'un humain guide le robot pour lui faciliter l'apprentissage d'une tâche, celui-ci doit acquérir de manière autonome les représen-

tations lui permettant de reproduire cette tâche. Nous montrerons donc qu'il est possible pour un humain d'interagir avec un robot sans utiliser de solution spécifique pour l'interaction. L'intérêt de notre modèle sera de pouvoir déduire de manière autonome des comportements et contextes associés nécessaires à la réalisation d'une tâche, par ses propres moyens ou en étant guidé.

Pour cela, animaux et robots doivent posséder plusieurs capacités fondamentales. Ils doivent être capables de traiter des informations sensorielles de natures différentes (visuelles, proprioceptives, tactiles etc.) qui deviennent alors des perceptions. Il faut ensuite être capable de former un modèle cohérent des relations entre ces perceptions. Les animaux ou robots autonomes doivent également pouvoir distinguer les informations pertinentes des informations non pertinentes dans le cadre d'une tâche. Enfin, ces perceptions doivent être associées avec des comportements particuliers, qui peuvent mener le robot à la réalisation de sa tâche. Une importance toute particulière sera accordée à l'aspect temporel des apprentissages effectués dans le modèle. Ainsi, des codes neuronaux pourront apprendre des informations spatio-temporelles. Les tâches abordées dans cette thèse visent à utiliser toutes les capacités du modèle. Elles feront donc appel à des aspects de navigation motivée et de planification. La question du dressage sera étudiée, c'est-à-dire comment la tâche peut être progressivement modifiée pour permettre au robot de découvrir quelles conditions permettent de la réaliser, sans intervention directe de l'expérimentateur. Nous verrons comment, dans le cadre de la robotique, une interaction directe Homme-Machine peut faciliter cet apprentissage. Nous verrons aussi comment le robot peut adapter son comportement lorsque la tâche ou l'environnement changent. Toutes ces tâches seront motivées par l'envie de satisfaire un but. Dans ce cadre, le robot reçoit une récompense satisfaisant un besoin interne quand il accomplit la tâche. Il cherchera alors ensuite à réaliser la tâche pour satisfaire de nouveau ce besoin. Nous nous intéresserons particulièrement à la *tâche de navigation continue*, utilisée dans des expériences avec les rats. Cette tâche nécessite de naviguer vers un but, d'y rester immobile pendant un délai fixe et d'aller chercher une récompense délivrée à la fin du délai. Cette tâche a l'intérêt de requérir des comportements complexes de la part de l'animal : capacités de prédictions temporelles, de contrôle du mouvement, d'alterner entre plusieurs stratégies etc. Dans le cadre de la réalisation de cette tâche, nous tenterons d'expliquer et de reproduire certaines observations et enregistrements effectués chez les rats. Enfin, nous analyserons les répercutions de nos résultats sur la manière de penser les interfaces Homme/Robot.

Je commencerai à aborder les aspects de modélisation par une présentation de l'historique des modèles bio-inspirés de navigation et planification en robotique dans le chapitre 2. Le modèle sur lequel se basent mes travaux sera présenté dans ce chapitre. Ce modèle permet d'apprendre des séquences d'actions ou des représentations spatiales permettant de planifier des chemins. Une fois que ces représentations sont acquises, le robot peut alors émettre des prédictions sur la conséquence attendue de ses actions ou bien sur ses prochaines perceptions. Nous défendrons dans le chapitre 3 l'idée que ces prédictions peuvent être apprises sous forme de relations temporelles entre des événements perceptifs. Le siège de cet apprentissage serait l'hippocampe, recevant des informations multimodales du cortex entorhinal. Le robot pourra, dans ce cadre, construire une représentation temporelle précise de ses perceptions successives. Il pourra apprendre les actions ayant permis de passer d'une perception à une autre. Il sera alors aussi capable de planifier ses actions pour remonter la chaîne des événements le menant à la satisfaction de ses buts et donc à la réalisation de la tâche. Cette planification se fera dans le cadre d'interactions entre l'hippocampe et le cortex préfrontal. Dans le chapitre 4, les interactions

seront analysées dans le contexte d'observations neurobiologiques liées à ces structures. Nous montrerons comment notre modèle explique ces observations et quelles prédictions sur le fonctionnement des structures cérébrales étudiées en découlent.

Les prédictions faites par le robot lui donneront la capacité de prévoir précisément l'instant où un événement est attendu. Ainsi l'occurrence ou l'absence de cet événement représente une information importante sur le déroulement de la tâche. L'absence d'un événement attendu forme un signal de nouveauté qui est un indicateur de changements dans l'environnement ou dans les conditions de la tâche à accomplir. C'est dans cette détection de nouveauté que la nécessité d'avoir une modélisation temporelle précise est fondamentale. Cette information sur l'incohérence entre prédictions et réalité est nécessaire au robot pour pouvoir adapter rapidement son comportement lorsqu'il se trouve en situation d'échec. Cet échec peut survenir si les actions du robot n'ont plus les conséquences qu'elles avaient auparavant. Si le robot ne s'adapte pas rapidement, il peut persévérer dans un comportement désormais inadéquat pour résoudre la tâche. Nous verrons dans le chapitre 6 que les signaux de prédiction peuvent être utilisés dans ce sens et permettre l'adaptation du comportement du robot. Ce mécanisme de détection d'erreur permet à la fois d'éviter les effets de persévération et de faciliter l'apprentissage d'une représentation plus adaptée de la tâche.

D'un autre côté, la répétition régulière d'une tâche et des actions associées peut mener à l'adoption de comportements automatiques. Des habitudes se forment alors et les actions à effectuer relèvent plus du réflexe que du résultat d'un calcul cognitif intense pour planifier la satisfaction d'un but particulier. Nous défendrons donc, dans le chapitre 5, l'idée de boucles parallèles correspondant à des stratégies de navigation différentes. Ces stratégies coopéreraient pour l'apprentissage de tâches de navigation. Certaines stratégies pourraient alors impliquer des mécanismes de planification, par exemple par l'utilisation d'une carte cognitive. D'autres, inspirées de l'apprentissage par renforcement, acquerraient des comportements automatiques de manière plus lente. Nous discuterons des régions cérébrales impliquées dans les diverses stratégies, notamment le cortex préfrontal et les ganglions de la base. Leurs interactions dans le cadre de la modulation du comportement seront aussi discutées.

Enfin, nous verrons dans le chapitre 7 que le modèle développé dans cette thèse permet de contrôler aussi bien un robot mobile qu'un bras manipulateur. L'objectif est de s'orienter vers une architecture unifiée de contrôle de plates-formes robotiques, capable d'apprendre des tâches variées de manière isolée ou en interaction avec un humain. En utilisant la même architecture que pour le contrôle d'un robot mobile, nous montreront le contrôle d'un bras robotique ayant à manipuler des objets. Des travaux préliminaires étudierons la possibilité d'un contrôle fusionné dans une tâche nécessitant l'utilisation d'un bras monté sur un robot mobile. Nous verrons alors quelles limitations restent à surmonter afin de sortir du paradigme de modélisation consistant à associer un modèle neuronal à un type de tâche donné. Dans le cadre d'une approche réellement développementale de la robotique [Weng et al., 2001], un robot multi-fonctions serait capable de s'adapter à une multitude de tâches. Ces tâches pourraient lui être enseignées directement par un humain ou apprises à travers son exploration de l'environnement.

La souris est un animal qui, tué en quantité suffisante et dans des conditions contrôlées, produit une thèse de doctorat.

– Woody Allen

CHAPITRE 1

Fondements neurobiologiques et éthologiques

Les travaux réalisés dans le cadre de cette thèse se placent à la frontière entre la modélisation et la robotique. L'objectif est à la fois de produire des modèles prédictifs, pouvant rendre compte des activités observées par les neurobiologistes, et d'implémenter ces modèles pour produire des algorithmes robustes et adaptatifs. Tous ces travaux se basent sur de nombreuses informations concernant l'anatomie du cerveau, les structures cérébrales impliquées dans diverses tâches allant de la navigation au contrôle moteur, et les types d'activités enregistrées dans ces structures.

L'objectif de ce chapitre est donc d'introduire ces fondements neurobiologiques qui sont à la base des modèles utilisés et développés dans cette thèse. Je présenterai tout d'abord les structures étudiées ainsi que leur anatomie et discuterai leurs rôles en décrivant les enregistrements d'activités neuronales, données de lésions et observations comportementales effectuées lors de diverses expériences. Ensuite, je parlerai de résultats expérimentaux récents remettant en question les interprétations purement spatiales du rôle de l'hippocampe dans le cadre de tâches de navigation. Le lecteur est également invité à lire la thèse de Vincent Hok [Hok, 2007], qui fait un compte-rendu détaillé des régions hippocampique et préfrontale ainsi que de leur rôle dans la navigation.

1.1 Région hippocampique

1.1.1 Anatomie

Le terme région hippocampique inclura ici la formation hippocampique, composée du gyrus dentelé (DG), du subiculum et de l'hippocampe lui-même avec les différentes parties de la corne d'Amon (CA). Diverses régions parahippocampiques seront également étudiées, telles que le cortex entorhinal (EC), les cortex périrhinaux et postrhinaux, ainsi que le presubiculum et post subiculum, et enfin le septum. Dans le cerveau du rat, la formation hippocampique prend la forme d'un C allongé (fig. 1.1).

Cette section se contentera de présenter l'anatomie relative des différentes structures ainsi que les principales connexions excitatrices et inhibitrices les reliant entre elles. Pour plus de détails concernant la topologie précise des structures impliquées en 3D et la dissociation des différentes régions selon plusieurs axes, le lecteur pourra se référer à [Amaral and Witter, 1989], [Witter, 1993] ou encore [Insausti and Amaral, 2004]. Voir la figure 1.2 pour une représentation anatomique des différentes structures évoquées dans cette section.

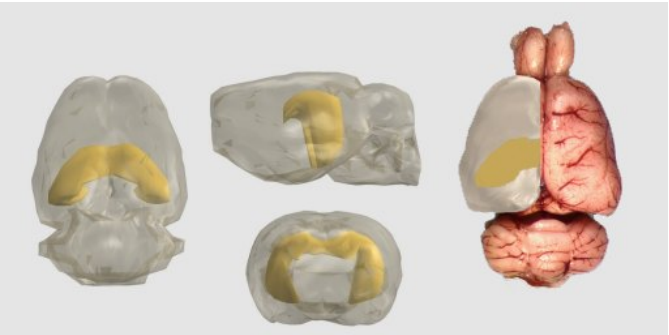


FIGURE 1.1 – Vue en 3D de l'hippocampe d'un rat[1].

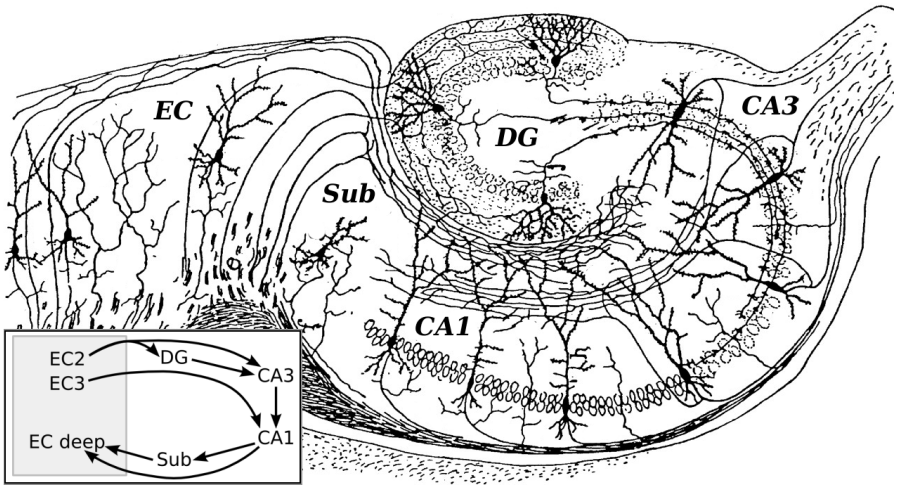


FIGURE 1.2 – Représentation d'une vue en coupe d'un hippocampe de rat. Version modifiée d'une image de [Cajal, 1911].

Cortex Entorhinal

Le cortex entorhinal est habituellement divisé en 6 couches séparées en deux catégories : les couches superficielles (de I à III) et les couches profondes (de IV à VI). Il est la source majeure d'information de l'hippocampe grâce à la voie perforante. Les couches superficielles II et III sont l'origine principale des connexions de la voie perforante. La couche II projette principalement vers le gyrus dentelé et la région CA3 de l'hippocampe. L'activation du cortex entorhinal a montré sa capacité à moduler l'activité neuronale et la plasticité dans le gyrus dentelé et l'amygdale [Yaniv et al., 2003]. Par ailleurs, la couche III fournit une connexion directe vers la région CA1 et vers le subiculum. D'autres connexions excitatrices en provenance des couches superficielles existent, partant vers les régions prélimbique et infralimbique du cortex préfrontal. Parmi les connexions efférentes des couches superficielles (principalement la couche V), on retrouve des projections vers l'amygdale, le septum, les cortex périrhinaux et postrhinaux.

Le cortex entorhinal reçoit des connexions afférentes substantielles des structures olfactives du télencéphale vers ses couches superficielles (I et II). La majorité des connexions provient ce-

[1]Source : http://synapses.clm.utexas.edu/anatomy/hippo/hippo.stm

pendant des cortex périrhinaux et postrhinaux, qui sont eux-mêmes fortement connectés avec des structures cérébrales très diverses. [Quirk and Vidal-Gonzalez, 2006] ont suggéré que l'amygdale puisse moduler les connexions entre les cortex péri-postrhinaux et entorhinaux. Le subiculum et la région CA1 ont des connexions réciproques avec le cortex entorhinal (couche V). Une source très importante d'activité pour les couches superficielles est le cortex préfrontal, notamment les régions prélimbiques et infralimbiques. D'autres connexions afférentes incluent notamment des entrées provenant du septum, de l'amygdale et du nucleus reuniens.

Formation hippocampique

L'hippocampe est au coeur des modèles présentés dans cette thèse et constitue une des structures les plus étudiées dans le cadre de la navigation spatiale. L'hippocampe lui-même est composé des différentes régions de la corne d'Amon tandis qu'on parle de formation hippocampique lorsqu'on veut inclure le gyrus dentelé et le subiculum.

Le gyrus dentelé reçoit de nombreuses projections de la couche II du cortex entorhinal via la voie perforante. Des connexions provenant des couches profondes existent également [Deller et al., 1996]. La majorité des corps cellulaires dans le gyrus dentelé sont des cellules granulaires. Celles-ci projettent à leur tour de manière exclusive sur la région CA3 de l'hippocampe par les fibres moussues. Outre la voie EC-DG-CA3, il existe aussi une projection directe des couches superficielles du cortex entorhinal vers CA3. Le rôle de l'amygdale dans la consolidation de mémoires à long terme liées à des mécanismes de récompense dans le gyrus dentelé a également été étudié [Almaguer-Melian et al., 2003].

La région CA3 projette à son tour vers la région CA1 en gardant une certaine topologie, par les connexions collatérales de Schaffer. CA3 possède également des collatérales récurrentes, avec des projections locales. Les régions CA3 et CA1 sont composées majoritairement de neurones pyramidaux. Des connexions directes du cortex entorhinal vers CA1 prennent leur origine dans la couche III. Le système complexe de connexions entre le cortex entorhinal, gyrus dentelé et les régions CA3 et CA1 est mieux connu sous le nom de boucle trisynaptique. En supplément de la connexion directe provenant du cortex entorhinal, CA1 reçoit des afférences d'autres structures corticales et sous-corticales, notamment le nucleus reuniens, dont les effets excitateurs ont été démontrés [Morales et al., 2007]. CA1 est la principale structure de sortie de l'hippocampe et possède donc de nombreuses connexions efférentes. Cependant, à l'inverse de CA3, CA1 ne possède pas de collatérales récurrentes. La projection principale se fait vers le subiculum. Parmi les autres connexions partant de CA1, on peut noter celles innervant le striatum (dans les ganglions de la base) et le cortex préfrontal median. Les connexions vers le préfrontal prennent leur origines dans l'hippocampe ventral et se focalisent majoritairement sur les cortex prélimbique et médial orbital [Jay and Witter, 1991].

Subiculum, septum

Les seules projections de l'hippocampe que reçoit le subiculum proviennent de la région CA1. Il reçoit accessoirement des projections des cortex entorhinaux, péri et postrhinaux. Une faible connectivité, réciproque, existe aussi vers les pré et parasubiculum. Les connexions principales du subiculum se font vers le présubiculum, le cortex entorhinal et dans une moindre mesure la région CA1. Le subiculum est une des régions principales de sorties de la formation hippocampique et projette vers de nombreuses structures corticales. Parmi celles-ci on peut compter le

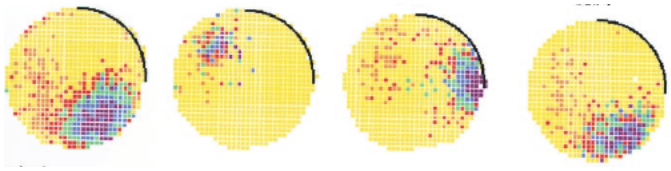


FIGURE 1.3 – Activité spatiale de 4 cellules de lieux enregistrées dans la région CA1 de l'hippocampe d'un rat navigant dans une arène circulaire. Une carte noire placée sur un mur fournit un repère visuel. Tiré de [Lenck-Santini et al., 2002].

cortex préfrontal médian, le septum, le nucleus accumbens et le nucleus reuniens.

Le septum, quant à lui, possède des connexions réciproques avec une bonne partie de la région hippocampique. Il projette notamment vers le gyrus dentelé, CA3 et CA1. Il reçoit de son côté des projections des régions CA3 et CA1 et du cortex entorhinal. Son rôle dans la régulation des niveaux d'acétylcholine [Oderfeld-Nowak et al., 1980] et du rythme thêta [Stewart and Fox, 1990] dans l'hippocampe a été démontré. Il participe donc de façon active à la modulation de l'apprentissage dans l'hippocampe, l'acétylcholine ayant un rôle d'inhibition sélective de la plasticité synaptique des connexions afférentes et efférentes dans l'hippocampe [Hasselmo and Schnell, 1994].

1.1.2 Corrélats spatiaux des activités neuronales

L'hippocampe est souvent considéré comme le siège des représentations spatiales de l'environnement dans le cerveau. Cela est du à la découverte des cellules de lieu par O'Keefe et Dostrovsky en 1971 [O'Keefe and Dostrovsky, 1971]. Ces cellules de lieu sont des neurones pyramidaux présents à la fois dans les régions CA3 et CA1. Elles ont de forts corrélats spatiaux et déchargent majoritairement à un endroit précis de l'environnement, leur activité diminuant au fur et à mesure que l'animal s'éloigne de cet endroit. Le gradient d'activité spatiale caractéristique de ces neurones est appelé "champ de lieu" (fig. 1.3). Cette découverte a amené [O'Keefe and Nadel, 1978] à attribuer à l'hippocampe un rôle dans la cartographie spatiale de l'environnement en créant une "carte cognitive". Cette carte pourrait être utilisée par la suite pour planifier un chemin vers des lieux connus par l'animal.

Les cellules de lieu hippocampiques ont des propriétés intéressantes. Lors de la répétition prolongée d'un trajet particulier dans un labyrinthe, les champs de cellules de lieux peuvent avoir tendance à s'étendre légèrement vers l'arrière (ils s'activent plus tôt dans la trajectoire du rat) [Mehta et al., 1997] tandis que, dans certains cas, ils se décalent plutôt vers des zones de récompense [Lee et al., 2006]. Ces données pourraient suggérer un aspect prospectif des activités de cellules de lieu. De plus, la forme des champs de lieux peut être asymétrique. Dans un trajet linéaire répété par un rat, la décharge des cellules de lieu a tendance à être plus faible lors de l'entrée sur le lieu et plus forte lors de la sortie du lieu [Mehta et al., 2000].

On peut distinguer les neurones des régions CA3 et CA1 qui ont des propriétés différentes. Le retour direct de EC sur CA1 joue un rôle dans la stabilisation des champs de lieu dans CA1. Des lésions de la couche III du cortex entorhinal qui projette vers CA1 ont montré un effet sur les champ de lieu dans CA1 (cellules de lieu plus bruitées, moins sélectives spatialement) mais pas dans CA3 [Brun et al., 2008]. De même des changements dans les amers visuels présents dans l'environnement peuvent entraîner des changements immédiat dans le codage spatial au ni-

veau de CA3, tandis que ces changements se répercutent bien plus tard (le lendemain) dans CA1 [Lee et al., 2004]. Enfin une étude récente a montré que les champs dans CA1 étaient moins susceptibles de changer avec l'apparition de nouveaux chemins (des raccourcis) dans un labyrinthe [Alvernhe et al., 2008]. Ces changements sont plus locaux dans CA1 que dans CA3, c'est-à-dire qu'ils touchent majoritairement les cellules de lieu ayant des champs de lieu à proximité des modifications introduites dans l'environnement. De plus, cette étude suggère que les cellules de lieux de CA3 et CA1 ne se contentent pas de donner une information sur la position de l'animal, mais également sur les chemins accessibles dans l'environnement. En effet l'ajout de raccourcis, fait en enlevant des parois transparentes et donc sans changer l'environnement visuel, entraîne de profond changements dans le codage spatial dans CA3 et CA1, même pour des cellules dont les champs sont éloignés de ce nouveau raccourci.

Des cellules de lieux ont été enregistrées dans d'autres structures. Les cellules de lieu du cortex entorhinal sont plus bruitées et possèdent des champs plus larges que les cellules hippocampiques [Quirk et al., 1992]. Dans le gyrus dentelé, des cellules répondant aussi bien à la position qu'à la direction de l'animal ont été enregistrées [Jung and McNaughton, 1993]. Leurs champs de lieu sont plus étroits que ceux trouvés dans CA3 et semblent suggérer un rôle du gyrus dentelé dans la transmission à l'hippocampe d'informations spatiales codées sous forme de populations de cellules spatialement très sélectives. La fusion des informations dans CA3 provenant de EC par des voies directes et indirectes, avec un rôle primordial du gyrus dentelé dans la séparation de formes, a été suggérée comme étant la base d'un mécanisme de séparation de formes permettant la facilitation de l'apprentissage dans l'hippocampe [Acsády and Káli, 2007]. Enfin on retrouve d'autres types de cellules avec des composantes spatiales. Le postsubiculum possède des "cellules de direction de la tête" qui réagissent lors que la tête de l'animal est tournée dans une direction particulière (dans un référentiel absolu) [Taube et al., 1990]. Récemment des "cellules de grille" ont été découvertes dans le cortex entorhinal [Hafting et al., 2005]. La particularité de chacune de ces cellules est de décharger non pas dans un seul lieu de l'environnement mais dans une multitude de points qui forment une grille hexagonale (en nid d'abeille) de l'espace. Finalement, des enregistrements ont montré la présence de neurones avec des corrélats spatiaux dans le nucleus accumbens et le noyau caudé [Mulder et al., 2004]. Ces neurones déchargent pendant de longues sections d'un labyrinthe et suggèrent l'implication de ces structures dans la création de "chunks" (ou amas) d'information, utiles pour diriger le comportement de l'animal.

Les cellules de lieu correspondant aux neurones pyramidaux de CA3 et CA1 et aux cellules granulaires du gyrus dentelé ont aussi la particularité de décharger en précession de phase par rapport au rythme thêta [Skaggs et al., 1996]. Cette précession de phase est une tendance des neurones à décharger plus tôt dans la phase thêta tandis que l'animal se déplace vers le centre du champ de lieu. Ces résultats sont souvent interprétés comme montrant un aspect *d'anticipation ou de prédiction* des cellules de lieu. La précession est plus marquée dans DG où les neurones déchargent en avance par rapport à CA1. De plus, les neurones de CA3 et CA1 déchargent dans les phases opposées du cycle thêta [Dragoi and Buzsáki, 2006]. La position actuelle est donc signalée un demi cycle thêta en avance dans CA3. Il semble également que la phase thêta proviennent de l'interaction entre des ensembles neuronaux, et non d'un système de pacemaker produit par des "neurones thêta".

1.1.3 Multi-modalité

Quelles informations sensorielles permettent la constitution d'un code spatial robuste dans le cortex entorhinal et l'hippocampe ? La vision joue certainement un rôle primordial dans la construction de ce code. [Muller and Kubie, 1987] montrent que la rotation des amers visuels présents dans un environnement entraîne une rotation identique des champs de lieu, de même pour les changements d'échelle. De plus, des barrières transparentes qui intersectent un champ de lieu détruisent complètement la dynamique de ce champ de lieu, ce qui discrédite l'hypothèse d'une représentation hippocampique de l'espace basée uniquement sur des perceptions visuelles instantanées. Dans des tâches où le rat doit se baser sur sa représentation interne de l'environnement, ces rotations induites par le déplacement des amers visuels peut fortement diminuer les performances du rat, tandis qu'il n'a aucun problème à naviguer vers un objectifs visible [Lenck-Santini et al., 2002]. La manipulation des amers visuels en remplaçant certains objets par des objets différents n'affecte pas la forme des cellules de lieux, ce qui impliquerait que l'aspect de la reconnaissance des objets est effectué en amont du cortex entorhinal (dans le cortex périrhinal par exemple) [Lenck-Santini et al., 2005].

En plus de ces informations visuelles, d'autres modalités peuvent venir compléter le code spatial. De nombreuses expériences ont montré le rôle du système vestibulaire et des informations d'intégration de chemin dans la construction du code spatial [Foster et al., 1989; Quirk et al., 1990; Fenton et al., 2010] et suggèrent que cette intégration a lieu avant l'hippocampe, dans le cortex entorhinal [Gothard et al., 2001]. D'autres informations sensorielles telles que les informations olfactives [Save et al., 2000] ou auditives [O'Keefe and Nadel, 1978] servent aussi à la construction du code spatial. Ce code peut donc être le résultat d'une intégration de ces différentes modalités et être affecté par les perturbations infligées à chacune de ces modalités [O'Keefe and Nadel, 1978].

Limiter le cortex entorhinal et l'hippocampe à la constitution d'un code purement spatial serait une erreur. De par ses nombreuses connexions afférentes, le cortex entorhinal reçoit des informations sensorielles de nombreuses aires cérébrales. Des cellules liées à la fois à des informations spatiales mais aussi comportementales (tourner à droite ou à gauche) ont été enregistrées à la fois dans EC et CA1 [Lipton et al., 2007]. Le code plus large et bruité des cellules entorhinales viendrait du fait que ces cellules représentent un contexte dans la tâche et permettent une sélection de sous-ensembles cellulaires dans l'hippocampe [Lipton et al., 2007; Lipton and Eichenbaum, 2008]. Ce type de contexte peut être motivationnel et entraîner des codes hippocampiques différents dans des situation de recherche de nourriture ou d'eau [Kennedy and Shapiro, 2009]. Des corrélats liés à des phases particulières d'une tâche ou à des comportements sont aussi présents dans l'hippocampe, et peuvent moduler l'activité spatiale des cellules de lieu [Hampson et al., 1993; Griffin et al., 2007]. D'autres modalités sensorielles sont aussi traitées par l'hippocampe et peuvent être fusionnées avec des informations spatiales, par exemple dans des tâches de discrimination olfactive [Eichenbaum et al., 1987; Wiener et al., 1989; Manns et al., 2007]. On voit donc une forte multi-modalité à la fois dans le cortex entorhinal et dans l'hippocampe, où la diversité des codages permet de représenter de nombreuses modalités sensorielles ainsi que de fusionner ces informations entre elles, menant ainsi à des codes hybrides incluant des informations spatiales.

On sait enfin que l'hippocampe a un rôle important dans la détection de la nouveauté et possède des mécanismes de neuromodulation facilitant l'acquisition de nouveaux motifs [Hasselmo and Fehlau, 2001]. L'acétylcholine permet notamment d'inhiber l'expression de motifs

anciennement appris pour faciliter l'acquisition des nouveaux [Hasselmo and Schnell, 1994].

1.1.4 Conditionnement

Outre les aspects spatiaux, je parlerai beaucoup d'aspects d'apprentissage de séquences et de prédiction dans cette thèse. Les conditionnement pavlovien et opérant sont des formes basiques de prédiction. Dans le conditionnement pavlovien, un stimulus inconditionnel est associé à une réponse réflexe (ex : la présentation de nourriture déclenche la salivation chez l'animal). Si de manière répétée, un premier stimulus (conditionnel) précède le stimulus inconditionnel déclencheur du réflexe, ce premier stimulus finira par être suffisant pour déclencher le réflexe (c'est-à-dire si un bruit particulier précède la présentation de nourriture, alors ce bruit finira par déclencher la salivation, même en l'absence de nourriture). Même si l'hippocampe n'est pas la structure primaire de l'apprentissage de conditionnements, il joue un rôle dans cet apprentissage pour le conditionnement de "trace" où il s'écoule un certain temps entre la disparition du stimulus conditionnel et l'apparition du stimulus inconditionnel. En effet l'hippocampe est nécessaire pour acquérir ce type de conditionnement mais pas pour le conditionnement de "délai" ou le stimulus conditionnel est toujours actif quand intervient l'inconditionnel [Clark and Squire, 1998]. Ce résultat serait lié à la nécessité d'acquérir une mémoire déclarative du lien entre premier et second stimulus. Quand l'attention du sujet est captée par des tâches cognitives contraignantes, cette acquisition est perturbée, ce qui laisse penser qu'elle implique des processus cognitifs de haut niveau [Carter et al., 2003]. Enfin ces associations stimulus-réponse ne peuvent subir une extinction que par des mécanismes d'inhibition active qui font intervenir l'hippocampe, le cortex préfrontal et aussi l'amygdale [Corcoran and Quirk, 2007; Milad et al., 2006; Sotres-Bayon et al., 2006]. D'autre part, le conditionnement opérant consiste à renforcer une action qui mène à un stimulus positif. C'est le point de départ de l'apprentissage de séquences, où une suite d'actions ordonnées doit être répétée pour résoudre une tâche.

1.1.5 Codage temporel

[Manns et al., 2007] souligne le rôle de l'hippocampe dans la représentation de séquences d'odeur. En plus du code spatial, il observe un code en population pour des événements olfactifs où les événements les plus rapprochés dans le temps ont le codage le plus proche. De même l'hippocampe semble nécessaire dans des tâches où un animal doit reproduire une trajectoire effectuée dans le sens inverse (même en l'absence d'informations visuelles ou olfactive) [Whishaw and Maaswinkel, 1998]. Ces résultats ont mené certains chercheurs à avoir une vue du cortex entorhinal et de l'hippocampe comme un système ayant un rôle principal dans la mémoire épisodique, plutôt que simplement dans la cartographie de l'environnement [Eichenbaum and Lipton, 2008]. Ainsi le cortex entorhinal porterait des informations sur le contexte temporel en supplément des informations spatiales. [Rondi-Reig et al., 2006] ont montré que la région CA1 jouait un rôle important dans une tâche de navigation dans un labyrinthe en étoile. Celle-ci est nécessaire aussi bien pour l'utilisation de stratégies spatiales allocentriques que pour des stratégies égocentriques faisant appel à des aspects de mémoire épisodique. Pour une revue exhaustive des différents mécanismes de traitement temporel dans le cerveau dans des échelles allant de plusieurs dizaines à plusieurs centaines de millisecondes, se référer à [Mauk and Buonomano, 2004].

1.2 Cortex préfrontal

1.2.1 Anatomie

Le cortex préfrontal peut être divisé entre plusieurs sous-régions. Dans le cortex préfrontal médian, la partie dorsale inclut les cortex précentraux et cingulaire antérieur tandis que la partie ventrale inclut les cortex prélimbique, infralimbique et médial orbital. Ce sont les régions sur lesquelles je focaliserai mon attention dans cette thèse. Les autres sous-régions incluent les cortex latéral et ventral orbital et la zone agranulaire du cortex insulaire. Ces régions peuvent également être découpées en termes fonctionnels. La partie dorso-médial est plutôt impliquée dans des aspect de mémoire liés à des réponses motrices et au traitement temporel de l'information. La partie ventrale médiale s'occuperait plutôt de la supervision de l'attention et la flexibilité du comportement. Enfin la partie orbito-frontale gère l'apprentissage inversif d'associations stimulus-récompense et les processus de choix dans des tâches avec renforcement retardé. Voir [Dalley et al., 2004] pour plus de détails sur le découpage anatomique et fonctionnel.

Le cortex préfrontal reçoit de nombreuses connexions très organisées des ganglions de la base. Il est également connecté de manière réciproque à plusieurs structures telles que le cortex pariétal, les aires corticales sensorielles, la substance noire, l'aire tegmentale ventrale, l'amygdale, l'hippothalamus latéral, le cortex entorhinal et le septum. L'interaction hippocampe-cortex préfrontal est très forte. CA1 et le subiculum projettent uniquement vers les aires infralimbique et prélimbique. Il ne semble pas exister de retour direct du PFC sur l'hippocampe, en revanche le nucleus reuniens (un noyau thalamique) semble jouer le rôle d'interface entre les 2 structures. Il a été montré que le nucleus reuniens avait un rôle excitateur sur CA1 [Bertram and Zhang, 1999] et sur le cortex préfrontal [Prisco and Vertes, 2006]. Le nucleus reuniens servirait également d'intermédiaire pour une influence indirecte du PFC sur CA1 [Vertes et al., 2007]. La figure 1.4 fournit un résumé des principales connexions excitatrices dans la région hippocampique et le cortex préfrontal.

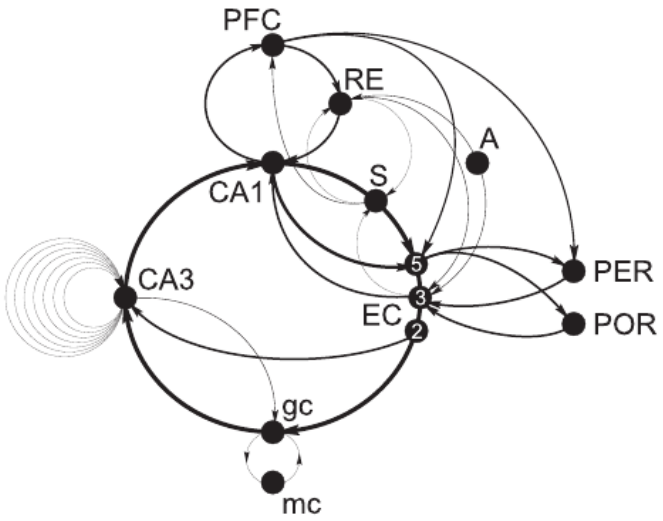


FIGURE 1.4 – Schéma des principales connexions excitatrices dans la boucle hippocampo-corticale. Tiré de [Hok, 2007].

1.2.2 Contrôle inhibiteur du comportement

Le cortex préfrontal peut être lié à des aspect de modulation du comportement pendant des phases d'attente. Des neurones enregistrés dans le PFC et dans le cortex moteur sont actifs pendant une phase d'attente avant le déclenchement d'une action (appuyer sur un levier) et participeraient au contrôle inhibiteur pro-actif de l'action [Narayanan and Laubach, 2009]. Une série d'expériences par Thorpe et al. montre l'implication du PFC dans la réalisation de tâches où des leviers, placés dans les 4 branches d'un labyrinthe, fournissent des récompenses tour à tour avec une séquence bien précise, la période de validité pour chaque levier pouvant être différente [Thorpe and Wilkie, 2002]. Les rats sont capables d'apprendre à prédire la fin de l'intervalle de temps où un levier est actif et de passer au suivant de manière préméditée. Les rats frontaux, quant à eux, montrent des signes de persévération sur des leviers dont la période de validité est terminée [Thorpe et al., 2002]. De plus les aspects spatiaux et temporels de ces tâches semblent pouvoir être appris de manière relativement dissociée [Thorpe and Wilkie, 2006; Thorpe et al., 2007].

1.2.3 Planification, processus cognitifs et adaptation

Outre le rôle du cortex préfrontal dans le contrôle pro-actif de l'action, il est essentiel dans nombre de tâches requérant un effort cognitif important ou des capacités de planification et d'adaptation [Dalley et al., 2004]. Des codes en population, signalant la stratégie motrice à utiliser, ont été enregistrés dans le PFC pendant une tâche d'alternance retardée [Baeg et al., 2003]. Ces codes changent lorsque les conditions de la tâche changent, montrant une capacité du PFC à adapter rapidement son codage à des nouvelles situations. Ces nouvelles expériences ont des codes neuronaux qui sont reproduits dans l'hippocampe et le cortex préfrontal en priorité pendant le sommeil de l'animal, facilitant ainsi une acquisition rapide des nouvelles règles à apprendre [Peyrache et al., 2009]. De même [Bast et al., 2009] suggèrent que la région intermédiaire (dans l'axe septotemporal) de l'hippocampe est une interface indispensable transmettant les informations nécessaires au cortex préfrontal pour encoder rapidement de nouvelles informations de but, indépendamment du code spatial utilisé dans l'hippocampe. Le nucleus reuniens faciliterait ainsi la communication entre les 2 structures afin de rendre possible l'acquisition de comportements orientés vers la satisfaction d'un but en faisant intervenir des processus cognitifs et émotionnels. Plusieurs études suggèrent que l'adaptation aux changements de conditions de la tâche requiert la participation du cortex préfrontal uniquement lorsque ces changements nécessitent de passer d'une stratégie comportementale à une autre (ex : stratégie de navigation allo-centrée ou égo-centrée) mais pas lorsque ces modifications changent juste une propriété de la stratégie courante (ex : stratégie de navigation égo-centrée, mais il faut maintenant tourner à droite au lieu de gauche) [Ragozzino et al., 1999; Rich and Shapiro, 2009]. Le passage d'une stratégie de lieu-action impliquant des circuit hippocampiques à une stratégie stimulus-réponse impliquant des circuits striataux se fait avec le cortex préfrontal comme médiateur. Ainsi [Rich and Shapiro, 2009] posent comme hypothèse que le PFC intègre de manière prédictive des informations sur les relations entre stimuli, actions et récompenses pour réaliser ces modulations.

Une série d'expériences par Granon et al. a permis de mettre en évidence le rôle du préfrontal dans la mémoire de travail, notamment dans des tâches de "Match-To-Sample" (MTS) et "Non-Match-To-Sample" (NMTS) [Granon et al., 1994]. Il semble que le PFC ne soit requis que pour des tâches de navigation d'une certaine complexité. La navigation dans un labyrinthe aquatique

vers une plate-forme cachée n'est perturbée chez les rats frontaux que lorsque le nombre de points de départ augmente (aucune différence avec le groupe contrôle pour 2 points de départ, perte de performances pour 4 points de départ) [Granon and Poucet, 1995]. Même lorsque la tâche n'inclut aucune complexité dans la sélection de l'action, les lésions du cortex préfrontal peuvent induire des pertes de performance si une attention soutenue est nécessaire [Granon et al., 1998]. Finalement, plutôt que de définir un rôle précis du PFC dans des mécanismes de mémoire de travail, d'attention ou de sélection de l'action, on peut émettre l'hypothèse que cette structure est nécessaire pour résoudre des tâches complexes intégrant de multiples facteurs [Granon and Poucet, 2000]. Ces résultats obtenus chez le rat sont très similaires aux résultats observés chez des patients humains atteints de lésions préfrontales. Ces derniers montrent des déficits de la mémoire de travail, de planification et d'attention dans une tâche où les patients doivent sélectionner des cartes en choisissant dans plusieurs tas dont les probabilités de perte et de gain d'argent sont différentes [Manes et al., 2002].

1.3 Ganglions de la base

1.3.1 Anatomie

Les ganglions de la base (BG) forment la dernière structure cérébrale présentée ici. Les structures cérébrales généralement incluses dans le terme "ganglions de la base" sont le striatum, le pallidum, le noyau sous-thalamique et la substance noire [Groenewegen, 2003]. Ces structures sont anatomiquement divisées en un certain nombre de sous-noyaux. Ainsi le striatum inclut le noyau caudé, le putamen et le nucleus accumbens tandis que le pallidum est formé de 2 parties interne et externe et enfin la substance noire se divise en 2 parties compacte et réticulée. L'aire tegmentale ventrale peut éventuellement aussi être incluse dans la famille des structures appartenant aux ganglions de la base. Ces structures peuvent généralement être distinguées en un groupe de structures d'entrée (noyau caudé, putamen et striatum ventral) qui reçoivent des projections directes du cortex cérébral et des structures de sortie (globus pallidus interne, substance noire réticulée, pallidum ventral) qui projettent en retour sur le cortex cérébral via le thalamus.

Les ganglions de la base ont donc la caractéristique de faire partie de nombreuses boucles au sein du cerveau. Des connexions réciproques importantes, comportant plusieurs boucles en parallèle, existent entre le cortex préfrontal et les ganglions de la base [Middleton and Strick, 2000]. Ces connexions sont parfois complexes, par exemple les régions prélimbique et médiale orbitale du PFC projettent vers la substance noire réticulée (SNr) via 3 voies différentes, ayant un effet d'excitation suivi d'une inhibition et enfin d'une excitation tardive éventuelle [Maurice et al., 1999]. Le striatum reçoit aussi d'importantes connexions excitatrices de l'hippocampe. Le striatum est très sensible aux effets de la dopamine et les ganglions de la base sont une partie importante du circuit dopaminergique. Les neurones dopaminergiques situés principalement dans la substance noire compacte et l'aire tegmentale ventrale projettent vers différentes régions du striatum et jouent un rôle probable dans le circuit de traitement des récompenses. Les circuits internes aux ganglions de la base sont détaillés dans la figure 1.5.

1.3.2 Circuit de récompense

Les informations sur les buts et objectifs de l'animal sont étroitement liées avec les informations sur les récompenses. Les ganglions de la base sont souvent associés aux circuits de récom-

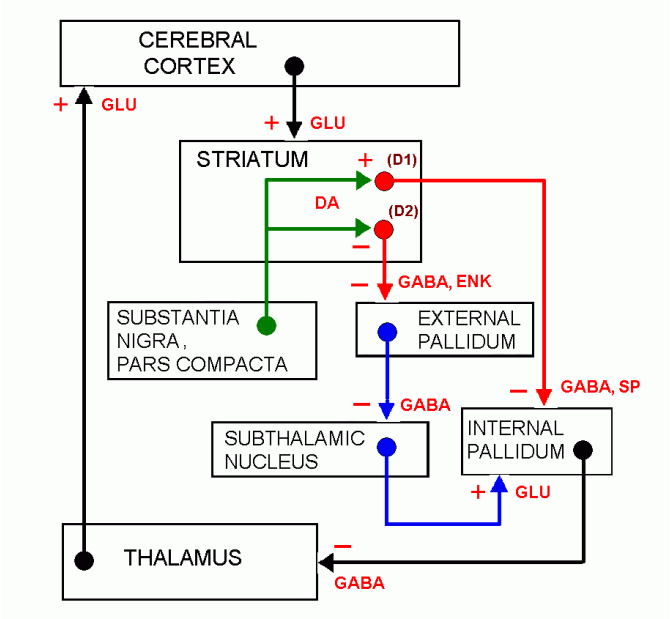


FIGURE 1.5 – Schéma représentant les principales voies synaptiques internes et externes des ganglions de la base.

pense [Groenewegen, 2003]. Cela vient de leurs neurones dopaminergiques qui semblent agir comme des prédicteurs de récompenses. Ces neurones déchargent selon ce qui semble être une erreur de prédiction sur l'arrivée d'une récompense [Schultz et al., 1993; Schultz, 1998]. Ils déchargent tout d'abord pendant l'arrivée de récompenses inattendues ou bien dans certains cas pour des aspects de nouveauté tels que l'ouverture d'une nouvelle boîte [Ljungberg et al., 1992]. Puis lorsque l'apprentissage a été effectué et la récompense peut être prédite avec certitude, on n'observe plus d'activité particulière lors de la réception de cette récompense. Finalement, si la récompense n'est pas donnée comme attendu, les mêmes neurones qui déchargeaient lors de la phase d'apprentissage montrent une période de dépression (activité réduite) pendant l'instant où la récompense aurait dû être reçue. Ces activités se retrouvent aussi dans le striatum et le cortex orbitofrontal du singe, et sont accompagnées d'activités liées à l'anticipation de la récompense et à une réponse après réception de la récompense [Schultz et al., 2000]. Le striatum intègre en plus des aspects liés aux mouvements prédicteurs de la réception de la récompense. De même que chez le singe, le cortex orbitofrontal de l'humain joue un rôle dans le traitement des récompenses négatives et positives [Rolls, 2000]. Plus généralement, les enregistrement montrent son implication dans le traitement de l'absence d'événements prédits, pertinents pour la tâche à effectuer [Schnider et al., 2007].

1.4 Tâche de navigation continue

Je vais maintenant détailler une série de résultats expérimentaux obtenus dans l'équipe du Laboratoire de Neurobiologie de la Cognition (LNC), à l'université d'Aix-Marseille. Les travaux effectués durant cette thèse sont le fruit d'une étroite collaboration entre nos deux équipes, et de nombreux aller-retour ont eu lieu pour tenter de modéliser les processus pouvant expliquer

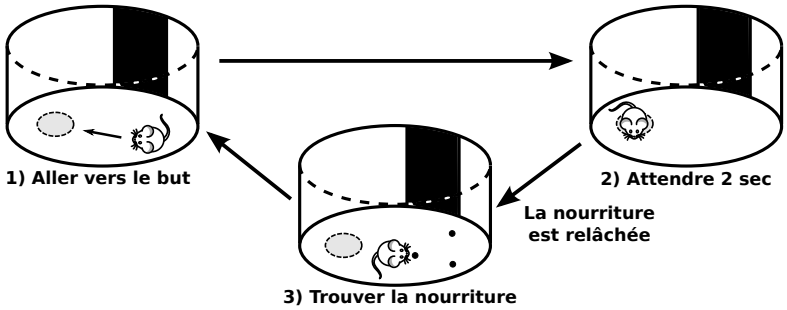


FIGURE 1.6 – Protocole expérimentale de la tâche de navigation continue.

les résultats obtenus chez le rat. Ces résultat proviennent de la mise en place de plusieurs expériences où des rats devaient réaliser une tâche de "navigation continue" (fig. 1.6).

1.4.1 Protocole expérimental

Cette expérience a été mise au point par [Rossier et al., 2000]. C'est une tâche de navigation continue, dans le sens où l'animal est constamment en train de résoudre la tâche de manière cyclique, sans interruption. La réalisation de la tâche se fait en trois étapes :

1. Le rat doit se déplacer vers une zone but marquée (tâche indicée) ou non (tâche de navigation continue).

2. Le rat doit attendre de manière immobile pendant un délai de 2 secondes sans quitter la zone but, pour déclencher le lâcher d'une boulette de nourriture.

3. La boulette de nourriture atterrit au centre du dispositif et peut rebondir n'importe où dans l'enceinte. La troisième phase consiste donc à rechercher cette boulette. Il doit s'écouler au moins 3 secondes après le lâcher de la nourriture avant que le rat ne puisse aller retenter d'obtenir une autre récompense.

On peut différencier deux versions de cette tâche, l'une où le but est marqué par une zone de couleur au sol et l'autre où il n'est pas signalé. Dans le premier cas, la navigation vers le but ne fait pas intervenir de mécanismes complexes de navigation, un simple "homing" (navigation guidée visuellement vers un amer) suffit. Dans le second cas, il est nécessaire d'avoir construit une représentation spatiale de l'environnement et d'avoir mémorisé l'emplacement du but. Les amers visuels sont fournis par une carte noire recouvrant une partie du mur circulaire et qui sert de point de repère. L'arène est régulièrement nettoyée pour éviter la présence d'informations olfactives et un fond sonore uniforme empêche le rat de se servir de repères auditifs.

La tâche de navigation continue présente un protocole très intéressant à étudier car elle demande l'utilisation d'une grande variété de stratégies comportementales. Elle requiert tout d'abord des capacités de navigation vers un but non marqué. Ensuite un mécanisme d'inhibition du mouvement doit être mis en place et lié à des facultés d'estimation temporelle. En effet par défaut le rat se déplace constamment et déteste rester immobile, particulièrement quand il est éloigné des bords de l'arène. Le temps de 2 secondes utilisé dans l'expérience correspond à la période maximale pendant laquelle le rat tolère de rester immobile. Finalement, lorsque la nourriture a été relâchée par un dispositif situé au dessus de l'arène et laissant tomber la nourriture

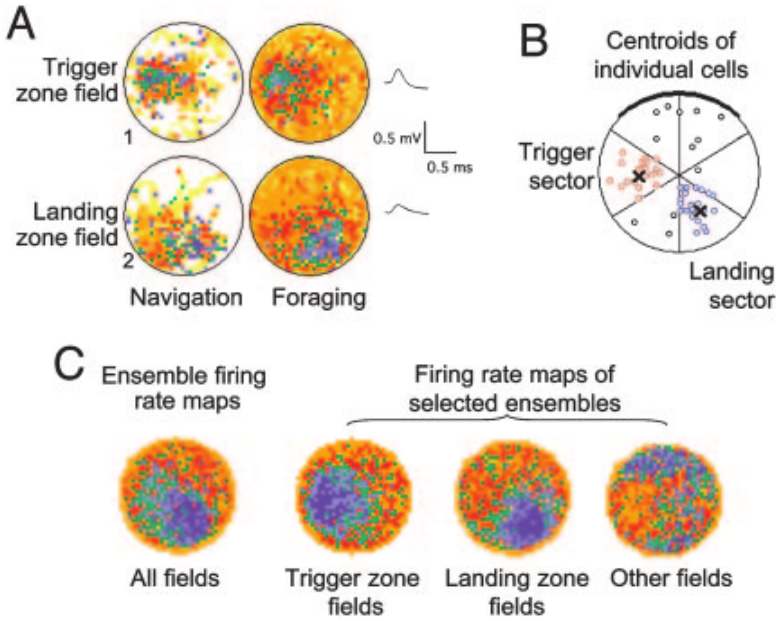


FIGURE 1.7 – A) Différentiation de l'activité entre les épisodes de navigation et d'exploration, montrant qu'on retrouve les mêmes champs. B) Répartition des centroïdes des cellules spatiales dans l'environnement. C) Cartes d'activité cumulées pour les sous-ensembles de cellules enregistrées. Tiré de [Hok et al., 2005].

au sol, où elle rebondit, le rat doit commencer une phase de recherche dans l'environnement. De plus, cette tâche nécessite des capacités d'apprentissage sur des composantes spatiales (pour la navigation) et temporelles (pour la phase d'attente) qui peuvent être dissociées relativement aisément. Etant donné que le rat est immobile pendant la phase d'attente, les activités observées ne peuvent pas être liées à des variations d'informations spatiales. De même une dissociation claire peut être établie entre le but (que l'on peut considérer comme un sous-but permettant au final de recevoir la nourriture) et la consommation de la récompense elle-même, qui se situent dans des endroits différents. Le codage spatial de la zone du but ne peut donc pas être confondu avec une réponse immédiate à la réception d'une récompense sous forme de nourriture, puisque celle-ci intervient a posteriori.

1.4.2 Corrélats spatiaux dans le cortex préfrontal

Des électrodes ont été implantées dans le cortex préfrontal de rats réalisant la tâche de navigation continue. De précédents résultats montraient une absence de corrélats spatiaux dans le PFC lors d'une exploration non motivée de l'arène [Poucet, 1997]. Cependant, durant la réalisation de la tâche, près de 25% des neurones observés dans les aires prélimbique et infralimbique du cortex préfrontal avaient des corrélats spatiaux [Hok et al., 2005]. Les champs de lieux observés dans le PFC sont plus bruités et plus large que ceux de l'hippocampe. De manière très intéressante, ces champs ne sont pas uniformément répartis dans l'arène. Une majorité de champs répondent à l'emplacement du lieu but et à l'endroit où les boulettes de nourriture viennent tomber sur le sol (fig. 1.7).

Quelques cellules répondent également près de la carte servant de repère visuel au mur et

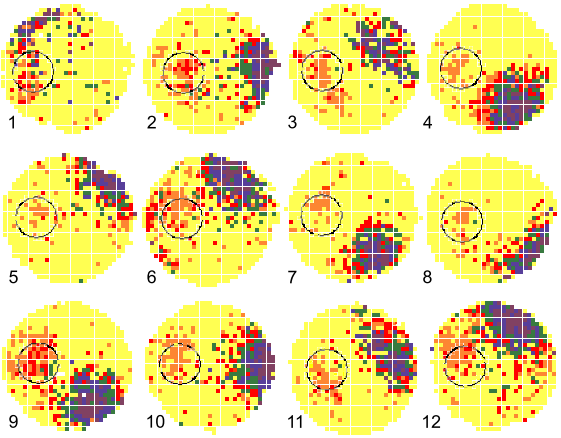


FIGURE 1.8 – Champs de lieu secondaires pour des cellules de lieu dans CA1. Outre le champ de lieu principal, on peut voir une hausse d'activité au niveau de la zone but. Tiré de [Hok et al., 2007a].

une petite minorité couvre d'autres zones de l'environnement. L'importance de la zone but dans la tâche n'est pas à démontrer. De plus malgré le fait que la nourriture s'éparpille partout dans l'environnement, les rats ont tendance à se diriger en premier vers la zone où la boulette tombe au sol après avoir déclenché le lâcher (en se dirigeant probablement grâce au son émis par la chute), puis partent à la recherche de leur récompense. Ces 2 lieux revêtent donc une importance particulière dans la tâche et il semble donc ressortir de ces résultats que le cortex préfrontal mémorise des lieux qui ont un contexte motivationnel fort et qui sont particulièrement pertinents pour la réalisation de la tâche. Ces activités sont observées à l'identique pendant les épisodes de navigation vers le but ou de recherche de nourriture.

1.4.3 Activité hors-champ des cellules de lieu

L'hippocampe est une source d'information majeure des aires infralimbique et prélimbique. De plus, avec ses cellules de lieu, il fournit un codage spatial très complet. Il serait donc fortement plausible que les informations spatiales retrouvées au niveau préfrontal aient une origine au niveau hippocampique. Des enregistrement ont donc été effectués dans CA1 pendant la tâche de navigation continue avec l'objectif de vérifier si les activités hippocampiques elles-mêmes possédaient des informations concernant les buts [Hok et al., 2007b,a]. *Les résultats de cette étude ont montré que la majorité des cellules de lieu dans CA1,* ***en plus*** *d'avoir un champ de lieu principal couvrant en endroit précis de l'arène, avaient une activité accrue au niveau de la zone du but*. Spatialement, cette activité apparaît comme un champ de lieu secondaire, de moindre importance, situé au niveau du but (fig. 1.8). Il faut noter que le lieu but n'est pas sur-représenté dans la population des cellules de lieu. L'environnement est couvert de manière relativement uniforme (avec une plus grande proportion de neurones possédant des champs au bord de l'arène). Il n'y a donc pas de sur-représentation observée des lieux présentant une forte valeur motivationnelle au niveau hippocampique.

Les propriétés les plus intéressantes de l'activité "hors-champ" sont révélées lorsqu'on l'observe d'un point de vue temporel. En effet ce regain d'activité n'a lieu que pendant la phase d'attente, qui mène à la réception de la nourriture. *Ces champs de lieu secondaires ne sont observés que pendant les phases de navigation motivée vers le but et pas pendant les phases d'exploration*

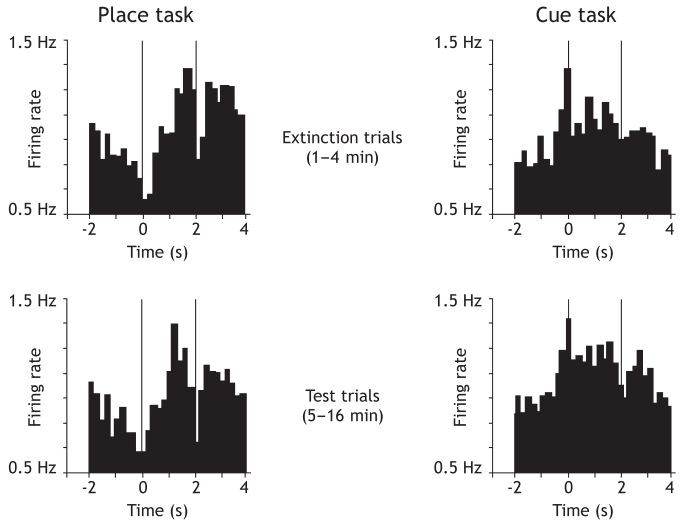


FIGURE 1.9 – Forme temporelle des activités hors-champ des cellules de lieu dans CA1 pour les tâches de navigation continue (à gauche) et indicée (à droite). Les traits verticaux signalent le début et la fin de la période d'attente de 2 secondes. Tiré de [Hok et al., 2007b].

lorsque le rat est à la recherche de la nourriture. La forme temporelle de l'activité hors-champ des cellules de CA1 est celle d'un pic d'activité atteignant un sommet peu avant la fin du délai de 2 secondes (fig. 1.9). Pour les rats réalisant la tâche indicée, un premier pic apparaît aussi juste avant l'arrivée sur le but marqué. Etant donnée l'immobilité du rat pendant la phase d'attente, l'évolution de l'activité observée ne peut être liée à des aspect spatiaux ou comportementaux, et semble donc suggérer un rapport avec un mécanisme d'estimation de l'écoulement du temps. Un tel mécanisme pourrait servir à prédire de manière précise la fin du délai d'attente. Afin de tester l'hypothèse de l'existence d'un système d'estimation et de prédiction temporelle, des essais ont été conduits, pendant lesquels la récompense n'était pas fournie à la fin du délai d'attente. Dans les essais normaux, le son émis par la chute de la nourriture fournissait un signal marquant la disponibilité de la récompense et donc la fin de la période d'attente. En l'absence de récompense, aucun stimulus ne marque la fin de la période de 2s. Ces essais d'extinction ont montré que les rats avaient bien appris la durée de la période d'attente et étaient capables d'estimer l'écoulement du temps de manière suffisamment précise pour pouvoir déterminer quand les 2 secondes étaient passées [Hok et al., 2007b]. En effet les rats reprennent leur mouvement, passé le délai de 2 secondes, qu'ils aient reçu leur récompense ou non. On pourrait imaginer que ce mécanisme ne soit nécessaire que dans la tâche où le but n'est pas signalé, le rat ayant besoin de détecter l'absence de la récompense qui pourrait lui indiquer qu'il n'est pas bien positionné sur le but, afin d'affiner sa représentation spatial de la zone but. Cependant, tous les rats montrent une capacité à reprendre leur mouvement de manière précise à la fin des 2 secondes, que le but soit marqué ou non. Il semblerait donc que des relations temporelles entre des événements sensoriels et/ou motivationnels soit constamment apprises, que la tâche requiert leur utilisation ou non. Ces relations peuvent mener à un système de prédictions qui peut servir à détecter des irrégularités dans le déroulement d'une tâche ou d'un comportement et adapter les actions en fonction de ces changements.

1.4.4 Lien entre l'activité temporelle hippocampique et frontale

Les résultats précédents suggèrent une implication du cortex préfrontal dans la gestion des buts de la tâche tandis que l'hippocampe serait impliqué dans le codage des aspects spatio-temporels. Cependant il est difficile de savoir l'importance de ces structures dans la réalisation de la tâche de préférence spatiale. Une expérience a alors été réalisée, dans laquelle le PFC de rats ayant appris à faire la tâche a été inactivé [Hok, 2007]. Ces lésions du préfrontal n'ont eu aucun effet significatif sur les performances post-opération des rats. De plus, les activités hors-champs des cellules de lieu sont maintenues après l'inactivation. Le seul effet notable est une légère augmentation de l'activité générale dans l'hippocampe, montrant un certain défaut d'inhibition global. Ces résultats laissent penser que le cortex préfrontal n'est pas la source directe des activités hors-champs enregistrées dans l'hippocampe et n'est pas indispensable pour résoudre la tâche correctement. Il faut cependant noter que les lésions ont été faites après apprentissage de la tâche, quand les rats avaient atteint des niveaux de performance asymptotiques. A ce niveau, il est probable que des mécanismes de navigation "automatiques" de plus bas niveau aient pris le relais sur des processus plus cognitifs impliquant le PFC. De même, si son utilité est limitée dans la réalisation de la tâche après apprentissage, il n'est pas dit que le PFC ne joue pas un rôle important dans l'acquisition des comportements nécessaires à cette réalisation.

Finalement, la propagation des activités temporelles de l'hippocampe a été testée dans une variante de la tâche de navigation continue [Burton et al., 2009]. Dans cette expérience, les rats ont été entraînés à résoudre la tâche avec un lieu but changeant chaque jour. Au début de la journée, une session de 15 min se déroulait avec un disque de couleur marquant la zone du but, puis celui-ci était retiré pour le reste des sessions. Des électrodes ont été implantées dans l'aire prélimbique du cortex préfrontal. Les résultats montrent une sélectivité spatiale des neurones préfrontaux beaucoup moins grande que dans les expériences précédentes, ce qui est probablement lié au changement régulier de la position du but. On peut aussi noter que les rats n'ont plus de préférence particulière pour la zone de chute de la nourriture dans leur recherche de l'environnement, et que celle-ci n'est plus particulièrement représentée dans les activité spatiales préfrontales. D'un autre côté, une grande partie des neurones montre un lien avec les aspects temporels de la tâche. La forme des activités hors-champ présentes dans l'hippocampe se retrouve dans les neurones préfrontaux (à savoir un pic d'activité qui culmine juste avant la fin de la période de 2 secondes). De plus, une grande partie des neurones présentant des corrélats spatiaux présente aussi cette composante temporelle, peu de cellules purement spatiales ont été enregistrées. On peut donc parler d'un code spatio-temporel. Il semblerait aussi que l'aspect temporel soit bien plus marqué dans le cortex préfrontal que dans l'hippocampe. La supériorité du codage temporel dans cette expérience où la phase d'attente est le principal invariant de la tâche, par opposition à la tâche originale où le but est fixe et le code spatial bien plus présent, montre une capacité du cortex préfrontal à mémoriser les invariants pertinents d'une tâche.

Afin de déterminer l'origine du codage temporel présent dans le PFC, des lésions de l'hippocampe ventral et intermédiaire, qui sont les aires projetant vers le PFC, ont été réalisées. Ces lésions laissent intact l'hippocampe dorsal qui est plus impliqué dans la navigation que les parties ventrales, et affectent donc peu les capacités de navigation des rats. Après les lésions de l'hippocampe, les quelques activités spatiales présentes au niveau préfrontal sont peu affectées. La liaison direct provenant du cortex entorhinal pourrait expliquer cette capacité du préfrontal à maintenir un code spatial malgré l'absence d'entrées hippocampiques. En revanche, les lésions ont un effet significatif sur les caractéristiques temporelles de l'activité dans le PFC qui sont ma-

joritairement supprimées. On peut donc supposer que ces activités temporelles ont une origine hippocampique. De plus les performances des rats se trouvent diminuées. Les capacités de navigation sont inchangées mais les rats montrent une incapacité à attendre pendant les 2 secondes et ont tendance à quitter le but précocement. Cette incapacité pourrait être liée à un défaut de prédiction temporelle, provoqué par la perte du code temporel au niveau préfrontal, ou bien à une impulsivité accrue.

1.5 Conclusion

Dans cette partie nous avons pu voir que le circuit neuronal incluant l'hippocampe, le cortex préfrontal et les ganglions de la base était impliqué dans de nombreux processus d'apprentissage, adaptation et modulation du comportement dans des tâches très diverses. Deux vues principales de l'hippocampe s'opposent à l'heure actuelle : l'aspect spatial avec la création de cartes cognitives et l'aspect de mémoire épisodique avec la formation de séquences [Eichenbaum et al., 1999]. Les travaux présentés dans cette thèse visent à intégrer ces deux aspects complémentaires dans une approche unique en proposant un codage temporel de suites d'événements multimodaux, intégrant une forte composante spatiale. Ces codes peuvent ensuite être réutilisés dans un système complexe d'interaction entre le cortex préfrontal, dont j'ai montré ici le rôle dans l'adaptation et la modulation du comportement en intégrant des informations temporelles, et les ganglions de la base, qui fournissent des signaux d'évaluation des comportements utilisés grâce à leur système de traitement et prédiction de récompenses. Il est donc primordial de comprendre comment les informations fournies par l'hippocampe sont utilisées par ces structures et comment l'interaction entre toutes ces aires cérébrales peut mener à un système capable de s'auto-réguler en adaptant ses actions et ses comportement en fonction de ses attentes et des résultats obtenus.

L'intelligence artificielle se définit comme le contraire de la bêtise naturelle.

– Woody Allen

CHAPITRE 2

Modèles bio-inspirés de navigation et planification en robotique

Dans le chapitre précédent, j'ai présenté les observations et enregistrements neuronaux effectués chez des animaux devant réaliser certaines tâches (de navigation ou autre). Ces résultats donnent des pistes sérieuses pour la compréhension des mécanismes cérébraux permettant la réalisation de ces tâches. Ainsi, des modèles de plus en plus poussés de ces mécanismes ont vu le jour durant ces dernières décennies. Certains ont mené à des implémentations à l'aide de réseaux de neurones artificiels, et ont parfois été utilisés pour le contrôle de robots devant réaliser des tâches similaires à celles données aux animaux.

Les nombreuses observations liées à l'espace faites dans l'hippocampe ont mené au développement de nombreux modèles de l'hippocampe pour la navigation spatiale. Les capacités du cortex préfrontal, dans le contrôle et l'inhibition du comportement dans des situations faisant appel à des fonctions cognitives de haut niveau, mènent à son implication dans les modèles de planification. Je présenterai donc tout d'abord dans ce chapitre un état de l'art des modèles neuronaux bio-inspirés liés à la navigation et la planification. Je décrirai ensuite plus en détail les modèles utilisés dans l'équipe Neurocybernétique avant de discuter des propriétés et des limitations des divers modèles existants.

2.1 Etat de l'art des systèmes de navigation bio-inspirés

2.1.1 Modélisation de cellules de lieu

Depuis la découverte des cellules de lieu en 1978 [O'Keefe and Nadel, 1978], l'hippocampe est devenu la structure centrale de la majorité des modèles cérébraux de la navigation spatiale. D'un point de vue computationnel, les cellules de lieu fournissent à la fois une information quant au positionnement de l'animal et un moyen de discrétiser l'environnement continu en une série d'états. De nombreux modèles probabilistes et systèmes de sélection de l'action (tels que les processus de décision markoviens, par exemple), nécessitent une description du système en termes d'états distincts. Les cellules de lieu permettent cette catégorisation de l'espace qui peut être utilisée dans des processus décisionnels.

Dès le début des années 1990, les premiers modèles computationnels de cellules de lieux ont commencé à être implémentés. [Burgess et al., 1994] ont créé un modèle de cellules de lieu

se basant sur la perception de deux amers visuels. Des courbes de réglage sont utilisées pour simuler la réponse de la cellule de lieu en fonction de la distance à chaque amer, le cadre est donc plutôt celui d'un modèle mathématique de la cellule de lieu que d'un modèle biologiquement plausible. Des couches successives avec des inhibitions locales permettent de stabiliser les champs de lieu et d'avoir un mécanisme de compétition. Le modèle est ensuite implémenté en simulation [Burgess et al., 1994] puis sur un robot Khepera capable de détecter les murs environnants [Burgess et al., 1997]. Dans le même temps, [Gaussier and Zrehen, 1995] présentent l'architecture PerAc qui deviendra la base de notre modèle (présenté plus en détails dans la section 2.2). L'idée est de réaliser des associations de bas niveau entre perceptions et actions, construisant ainsi une architecture sensori-motrice. Des configurations d'amers visuels, correspondant à des cellules de lieu, sont alors apprises et associées avec une direction. Le modèle est d'abord utilisé en simulation, avant une implémentation sur robot réel permettant l'apprentissage de cellules de lieu visuelles [Banquet et al., 1997; Gaussier et al., 2000]. Par ailleurs, [Redish and Touretzky, 1997] développent un modèle théorique des processus cognitifs liés à la navigation dans la région hippocampique. Dans leur modèle, le code spatial des cellules de lieu est créé à partir d'informations visuelles, correspondant à des vues locales, et d'intégration de chemin, provenant d'informations idiothétiques par le biais du subiculum. Le code spatial serait formé dans l'hippocampe, et des référentiels différents seraient activés dans le cortex entorhinal afin de sélectionner des sous-ensembles spatiaux dans l'hippocampe. Enfin la sélection de l'action se ferait dans les ganglions de la base, plus précisément dans le nucleus accumbens, sur la base des informations spatiales fournies par l'hippocampe.

Plus tard, [Arleo and Gerstner, 2000] puis [Arleo and Rondi-Reig, 2007] ont développé un modèle de la création du code spatial des cellules de lieu intégrant des informations visuelles et proprioceptives. Le système allothétique (visuel) projette dans EC superficiel, tandis que le système idiothétique (intégration de chemin) projette dans EC profond. Des filtres de Walsh sont utilisés pour détecter différentes fréquences dans les informations visuelles afin de créer des cellules de vues, à leur tour utilisés pour le code spatial des cellules de lieu. L'apprentissage de cellules de lieu visuelles se fait dans les couches superficielles du cortex entorhinal grâce à un apprentissage hebbien non supervisé. Aucun modèle biologiquement plausible d'intégration de chemin n'est donné mais il est modélisé par des cellules déchargeant comme une fonction gaussienne de la position et de la direction. Finalement le recrutement des cellules de lieux dans l'hippocampe, à l'aide des informations visuelles et idiothétiques, est fait de manière autonome en se basant sur un seuil d'activité minimal. Le modèle est testé dans un environnement réel entouré de murs striés (pour faciliter la détection de fréquences) et mène à la création de plus de 800 cellules de lieu pour un environnement relativement petit.

Les découvertes récentes des cellules de grilles dans le cortex entorhinal [Hafting et al., 2005] ont entraîné la conception de modèles mettant en jeu ces cellules dans la construction du code des cellules de lieu. Ainsi [Gorchetchnikov and Grossberg, 2007] proposent un modèle de code spatial dans l'hippocampe dépendant des activités de cellules de grilles. Leur hypothèse est que l'hippocampe possède plusieurs voies parallèles et que les différences de granularité spatiale observées pour les cellules de grilles (des cellules avec des grilles de tailles très variables sont présentes dans le cortex entorhinal) se transmettent dans le code hippocampique. Ainsi l'hippocampe posséderait des codes spatiaux et temporels à de multiple échelles. Un autre modèle de cellules de lieu élaborées à partir de cellules de grilles suggère le rôle prépondérant du gyrus dentelé dans l'encodage de formes [Rolls et al., 2006]. DG serait la clé d'un processus d'ortho-

gonalisation et de codage en population. Cette vision du trio EC-DG-CA servant à l'encodage et la récupération de formes est souvent mise en opposition à d'autres modèles de l'hippocampe où celui-ci est utilisé comme une mémoire auto-associative. Ainsi les connexions récurrentes de CA3 seraient la clé du codage des formes. [Káli and Dayan, 2000], par exemple, proposent un modèle où des attracteurs sont formés grâce aux connexions récurrentes de CA3 (les voies directes EC-CA3 et indirectes EC-DG-CA3 ne sont pas différenciées) et permettent d'avoir des représentations spatiales, directionnelles ou non, dans l'hippocampe à partir d'informations de position et de direction dans le cortex entorhinal. [Samsonovich and McNaughton, 1997] fournissent également un modèle de construction de cellules de lieu sous formes d'attracteurs continus se basant sur des informations de mouvement. Ainsi une même population de neurones peut encoder des contextes spatiaux très variés avec un codage d'attracteurs en population.

Parmi les autres travaux de modélisation spatiale dans l'hippocampe, on peut noter les "Brain-Based Device" [Fleischer et al., 2007] permettant la reconstruction de cellules de lieux à partir de multiples capteurs. Ces cellules présentent des aspect d'activité prospective et rétrospective. [Milford and Wyeth, 2008], quant à eux, se placent dans le contexte d'un système faiblement bio-inspiré et mettent en avance les performances de leur système de cellules de lieu par rapport aux approches de SLAM traditionnelles. Leur modèle permet la création de cellules de pose (lieu et direction) à partir d'informations purement visuelles. Le système est testé en passif lors d'un trajet en voiture sur 66 km où plus de 12000 lieux sont catégorisés. Enfin, si d'autres systèmes de navigation se basent sur des activités de cellules de lieu pour la sélection de l'action, il se contentent parfois de modéliser cette activité en simulation comme une fonction gaussienne de la position (connue) du robot [Foster et al., 2000; Dollé et al., 2010].

2.1.2 Cartes cognitives et navigation vers un but

L'idée de la carte cognitive a été formée par [Tolman, 1948]. La carte cognitive correspondrait à une représentation interne de l'espace permettant d'inférer des chemins. Ainsi elle va de paire avec la notion d'un apprentissage latent où l'animal pourrait, dès l'apparition d'un raccourci ou d'un but particulier, sélectionner des chemins optimaux pour rejoindre ce but à partir de ses expériences passées. La première théorie de l'hippocampe comme étant le siège d'une carte cognitive cérébrale a été formulée par [O'Keefe and Nadel, 1978], en se basant sur sa découverte des cellules de lieu.

La navigation vers un but donné n'implique pas forcément la construction d'une représentation interne complexe de l'environnement, nécessitant de raisonner sur les chemins disponibles. Ainsi, dans un environnement ouvert avec un but unique, il peut suffire d'associer une action à un lieu. Dans [Burgess et al., 1994, 1997], les cellules de lieu répondent de manière différenciée (par rapport à la phase theta) si leur centre se trouve devant ou derrière le rat. Une fois le but découvert, le rat regarde dans toutes les directions et son orientation est associée avec les cellules de lieu se trouvant devant lui. Ainsi à chaque cellule de lieu est associée son orientation par rapport au but (Nord, Sud, Est, Ouest). Pendant la navigation, il suffit de prendre la direction inverse de celle indiquée par le lieu où l'on se trouve pour rejoindre le but. Par ailleurs, l'architecture PerAc présentée dans [Gaussier and Zrehen, 1995] nous permettait d'associer des cellules de lieux avec des directions. Ces associations sensori-motrices créaient alors un bassin d'attraction pouvant être utilisé par un robot pour rejoindre un lieu but. Des expériences de navigation ont d'abord été réalisées en simulation, puis sur robot réel dans un environnement de bureaux [Ban-

quet et al., 1997; Gaussier et al., 2000]. Dans un autre cadre, [Koene et al., 2009] réalisent une expérience à l'aide d'un simulateur en 3D, où un robot apprend à naviguer vers un but et peut reconnaître des lieux distants. Parmi les lieux reconnus à distance, ceux qui ont été associés à une valeur de récompense sont sélectionnés et le robot se dirige dans leur direction grâce à un mécanismes de soustraction de coordonnées. Ces systèmes montrent cependant rapidement leurs limites, dans des environnement de grande taille où la récompense peut avoir été donnée loin de l'endroit où le robot se trouve actuellement.

[Redish and Touretzky, 1998] proposent un modèle où les connexions récurrentes de CA3 servent uniquement à mémoriser les chemins utiles pour rejoindre un but. Ainsi les connexions entre lieux successifs sont apprises grâce à la superposition des champs de lieux adjacents. Durant les simulations, l'apprentissage des trajectoires vers le but est cependant fait de manière supervisée, le robot étant guidé vers le but afin que les chemins appris soit uniquement ceux menant au but de manière optimale. De plus, lors des phases de reproduction post-apprentissage, la sélection de l'action nécessaire pour passer au prochain lieu prédit par le modèle est faite de manière algorithmique, en comparant les positions des lieux respectifs. Ce modèle représente donc un début de carte cognitive, incomplète et apprise sous supervision, et ne fait pas le lien avec un mécanisme de sélection de l'action neuronal.

Dans des environnements plus complexes, où il devient indispensable de planifier son parcours pour rejoindre un but, l'utilisation d'une carte cognitive devient nécessaire (voir section 2.2.2 pour plus de détails). [Mallot et al., 1995] puis [Franz et al., 1997] proposent un réseau neuronal, sans correspondance biologique, qui crée une carte cognitive reliant des cellules de vue. Ces cellules sont catégorisées en fonction d'entrées visuelles et reliées entre elles lorsqu'elles sont perçues successivement grâce à une règle de Kohonen [Kohonen, 1989] modifiée. Un comportement de *homing* basé sur les différences visuelles entre les lieux permet ensuite de rejoindre un lieu voisin, sélectionné après avoir mentalement vérifié tous les lieux accessibles et choisi le meilleur. Le modèle a été implémenté sur robot réel. Dans un registre plus biologiquement inspiré, [Muller et al., 1996] forment un modèle de la carte cognitive dans l'hippocampe représentant l'environnement complet et pas uniquement les chemins les plus courts vers un but donné. Il supposent eux aussi que les lieux adjacents sont liés par les collatérales récurrentes de CA3 grâce à la superposition des champs de lieux lors de l'exploration de l'environnement. Dans leur modèle, le chemin le plus court vers un but peut être trouvé en explorant la carte pour trouver le chemin offrant le moins de résistance synaptique (la résistance synaptique diminuant en fonction de la proximité de deux lieux). Cependant aucun modèle neuronal de sélection de l'action lié à ce fonctionnement n'est donné. Peu après, [Banquet et al., 1997] introduisent l'idée des cellules de transition et d'une carte cognitive utilisant ces transitions, modèle développé dans cette thèse. Dans ce modèle, des transitions entre lieux seraient apprises au niveau hippocampique et une carte cognitive reliant ces transitions serait stockée au niveau préfrontal (voir section 2.2.2). Le modèle est implémenté en simulation avant d'être utilisé sur robot réel en créant une carte de transitions entre cellules de lieu visuelles [Gaussier et al., 2002; Banquet et al., 2005; Cuperlier et al., 2007].

[Voicu and Schmajuk, 2000] proposent également une version améliorée du système de carte cognitive développé dans [Schmajuk and Thieme, 1992]. Le système est vaguement bio-inspiré, il est implémenté en simulation dans un monde "grille" (comme un échiquier où le robot se déplace de case en case et chaque case correspond à un lieu). Ce type de monde simplifie grandement le problème car il supprime le besoin d'une intégration de chemin pour connaître l'action

nécessaire pour passer d'un lieu à l'autre et assure qu'une action effectuée mènera toujours dans le lieu voulu (pas de problème de décalage). Les cases successives visitées sont liées entre elles si elles sont adjacentes, dans une carte hétéro-associative. Des buts sont aussi liés aux cases sur lesquelles ils sont situés. Dans le modèle initial, l'activité du lieu courant se propageait dans le graphe jusqu'à ce qu'elle atteigne un but. Un lieu était ensuite choisi parmi les lieux voisins pour rejoindre ce but. Dans la version de [Voicu and Schmajuk, 2000], l'activité se propage depuis le but, évitant le recalcul de la propagation à chaque changement de lieu. Dans une première phase, un mécanisme ad-hoc de sélection de l'action choisit la direction en fonction du lieu à atteindre. A travers la répétition de la tâche, l'action associée à chaque lieu finit par être apprise et l'utilisation de la carte cognitive devient superflue.

En parallèle, [Frezza-Buet and Alexandre, 2002] utilisent, pour la planification et la navigation d'un robot, un modèle de colonnes corticales présenté dans [Frezza-Buet et al., 2001] pour l'apprentissage de séquences. Un modèle biologique, implémenté sous forme de modèle computationnel, est présenté. Il reprend le principe de maxi-colonnes corticales caractérisant des états du système et pouvant être reliées entre elles pour former une carte cognitive. Ces maxi-colonnes sont composées d'un ensemble de mini-colonnes, la séparation des mini-colonnes étant réalisée pour apprendre différentes séquences incluant un même état. Une rétro-propagation d'activités de motivation, correspondant aux besoin de trouver de la nourriture ou de l'eau, permet une forme d'activation "appelante" des colonnes. Une seconde forme d'activation provenant des entrées sensorielles du système permet la co-activation de colonnes correspondant à des chemins permettant de rejoindre un but de manière optimale. L'action associée peut alors être exécutée. Le modèle permet d'apprendre des séquences d'événements ou de conjonctions d'événements (une mémoire associative permet de caractériser un état correspondant à la perception simultanée de deux événements). La catégorisation des événements est cependant réalisée de manière ad-hoc. Des expériences en simulation sont réalisées dans un monde grille.

[Koene et al., 2003; Hasselmo and Eichenbaum, 2005] proposent un modèle où 2 voies séparées existent dans l'hippocampe. La voie EC2-CA3-CA1 servirait à encoder et récupérer des épisodes de navigation en utilisant les collatérales récurrentes de CA3. Cela permettrait donc de connaître des lieux successifs et de prédire en chaque endroit quels lieux sont accessibles. D'un autre côté la carte topologique de l'environnement serait apprise dans EC3 (via des connexions récurrentes) et recevrait les positions des différents but mémorisés par le cortex préfrontal. Le rôle de EC3 serait alors d'encoder une sorte de contexte temporel permettant une activation sélective dans l'hippocampe. La fusion de ces deux sources d'information dans CA1 permettrait alors de sélectionner le prochain lieu le plus pertinent pour rejoindre le but.

Ces modèles proposent souvent des solutions en termes de chemins et de lieux à atteindre et la conversion de cette information en termes d'actions à effectuer est souvent faite de façon algorithmique (en calculant grâce aux positions respectives des lieux, connues de manière ad-hoc, la direction à prendre) ou bien par essai-erreur (par un mécanisme d'exploration et de remontée de gradient sur l'activité du lieu à atteindre). Nous verrons dans la section 2.2.2 la solution proposée par [Banquet et al., 1997] à ce problème de sélection de l'action dans la carte cognitive. Le modèle proposé par [Hasselmo, 2005] fournit une architecture neuronale complète du processus d'apprentissage et d'utilisation d'une carte cognitive et de l'association avec un mécanisme de sélection de l'action, tout en s'affranchissant des parties algorithmiques présentes dans de nombreux modèles. La carte cognitive, située dans le cortex préfrontal, est constituée d'une chaîne de colonnes corticales. Différentes sous-populations de neurones dans ces colonnes fournissent

des voies parallèles de transfert d'information. Ainsi une voie permet de rétro-propager l'activité liée à un but tandis qu'une autre voie propage l'activité du lieu courant vers les lieux adjacents. Les colonnes corticales représentent à la fois des états du système, des actions et des buts et sont intercalées de la manière suivante : Etat-Action-Etat-Action-Etat-But. Ainsi les états de la carte cognitive sont reliés entre eux par une colonne corticale représentant l'action à effectuer pour passer du premier état au deuxième. Le mécanisme de sélection de l'action se fait donc par la prédiction par le système des actions disponibles et états atteignables depuis la position actuelle, tandis que la rétro-propagation du but vient sélectionner la meilleure de ces actions. La nécessité d'un codage unique de chacune des actions nécessaires pour passer entre les états en fait une architecture gourmande en nombre de neurones. Le modèle est tout d'abord testé en simulation avec un monde grille [Hasselmo, 2005], et une implémentation utilisant des neurones impulsionnels sert à résoudre une tâche de GO-NOGO [Koene and Hasselmo, 2005]. Cependant la compatibilité avec un système utilisant des cellules de lieu n'est pas démontrée.

Enfin, [Martinet et al., 2011] proposent, eux aussi, un modèle de carte cognitive dans le cortex préfrontal qui se base sur des colonnes corticales. Leur modèle ressemble à celui de [Hasselmo, 2005] et fait appel à des propagations d'activités à la fois depuis le but et le lieu actuel, ainsi qu'à une alternance de colonnes représentant des états et des actions. Le modèle apporte cependant plusieurs améliorations : 1) Il utilise une représentation hippocampique de l'espace sous forme de cellules de lieu. 2) Il propose un système de carte cognitive multi-échelles. Ici deux niveaux de mini-colonnes (composant une même colonne corticale) fournissent deux représentations spatiales de granularités différentes. Tandis que la couche inférieure prend des entrées directes de l'hippocampe et fournit un code spatial plus compact, la couche supérieure prend ses entrées de la couche inférieure et intègre des informations de changement de direction, ce qui permet à une mini-colonne de répondre parfois pour un bras entier d'un labyrinthe. Des expériences en simulation sont conduites dans des labyrinthes de Tolman, de petite et grande taille, où l'intérêt du second niveau de représentation dans le choix du chemin le plus court est démontré pour le labyrinthe de grande taille.

2.1.3 Apprentissage temporel

La composante temporelle dans l'apprentissage prend une place très importante dans tous les mécanismes de conditionnement. En effet, dans ce cadre, un délai précis entre deux stimuli permet d'associer une réponse réflexe, liée au second stimulus, à la présentation du premier. Après apprentissage, le premier stimulus prédit temporellement de manière précise l'apparition du second. Un modèle de conditionnement pavlovien et d'apprentissage par renforcement utilisant l'architecture ART (Adaptive Resonance Theory) a été longuement étudié dans [Grossberg and Schmajuk, 1987; Grossberg et al., 1987]. L'aspect de la prédiction temporelle biologiquement plausible est abordé dans [Grossberg and Schmajuk, 1989] avec le système de décomposition spectrale du temps. Ce modèle suppose l'existence dans DG de cellules granulaires réagissant avec des dynamiques temporelles différentes et donnant, à travers l'activité de la population de neurones, une décomposition spectrale du temps écoulé depuis un événement particulier. Le modèle utilise cette décomposition spectrale pour prédire l'arrivée d'une récompense et détecter son absence. Le modèle de décomposition spectrale a été largement réutilisé dans des modèles de prédiction temporelle de récompense et de conditionnement [Grossberg and Merrill, 1992; Brown et al., 1999; Contreras-Vidal and Schultz, 1999]. Un modèle de timing basé sur

un codage en population du temps par des cellules granulaires a également été conçu par [Buonomano and Mauk, 1994]. Ce modèle concerne le cervelet et, comme les autres modèles de ce type, fait intervenir des mécanismes adaptés à l'apprentissage de timings de l'ordre de dizaines de millisecondes à plusieurs secondes. Les modèles cités se focalisent sur l'aspect temporel de la prédiction et font des associations entre un stimulus et une réponse, poussant parfois le modèle jusqu'à l'apprentissage de conditionnements secondaires mais pas de séquences de stimuli.

Le modèle présenté par [Banquet et al., 1997] pour l'apprentissage de transitions intègre un apprentissage temporel faisant appel à une modélisation des cellules granulaires du gyrus dentelé inspirée de celle de [Grossberg and Schmajuk, 1989]. Les transitions permettent donc l'apprentissage de séquences temporelles pouvant être associées à des actions motrices (voir section 2.2.3 pour plus de détails). Ainsi, il a été utilisé pour la reproduction par imitation de trajectoires en navigation [Gaussier et al., 1998], ou plus tard pour la reproduction de gestes par un bras robotique [Andry et al., 2001; Rengervé et al., 2010]. Avec la même volonté de produire un modèle générique d'apprentissage de séquences temporelles entre des événements perceptifs, [Frezza-Buet et al., 2001] présentent un modèle utilisant des colonnes corticales. Leur modèle peut apprendre la relation temporelle entre des événements successifs encodés par des colonnes corticales différentes. Une mémoire temporelle est assurée par des neurones dont l'activité diminue linéairement après leur activation, gardant donc une trace des événements perçus. Ainsi, des colonnes corticales activées de manière rapprochée dans le temps sont liées entre elles. L'association de certaines de ces colonnes à des buts permet par la suite de retrouver les séquences permettant d'atteindre ces buts. Ce modèle a plus tard été utilisé dans le cadre de la navigation d'un robot mobile [Frezza-Buet and Alexandre, 2002].

D'autres modèles se concentrent sur l'apprentissage de séquences (d'événements perceptifs, d'actions etc.). Certains font abstraction d'une modélisation fine de l'écoulement du temps dans ces séquences et mémorisent principalement l'ordre des événements sensoriels. [Levy, 1996] utilise les connexions récurrentes de CA3 pour mémoriser des séquences de stimuli et émettre des prédictions. De même les modèles de navigation présentés dans la section 2.1.2 et basés sur le même principe d'encodage de séquences dans CA3 [Hasselmo and Eichenbaum, 2005; Redish and Touretzky, 1998; Voicu and Schmajuk, 2000] peuvent être vus comme des modèles d'apprentissage de séquences ayant la particularité de travailler avec des informations spatiales (voir section 2.3 pour une discussion sur ce sujet). Les particularités des cellules de lieux concernant la précession de décharge par rapport à la phase theta peuvent également fournir des informations sur la relation temporelle entre des informations spatiales [Dragoi and Buzsáki, 2006]. Des séquences temporelles complexes (où un état est répété dans une unique séquence, ABCDBD par exemple) peuvent être apprises avec une modélisation temporelle fine. Ceci a été réalisé en utilisant un modèle de neurones impulsionnels et un apprentissage hebbien bruité, l'aspect temporel étant donnée par des neurones intégrateurs à fuite [Tijsseling and Berthouze, 2003]. Des CTRNN (oscillateurs neuronaux) ont également été utilisés pour fournir des échelles temporelles variables dans un système chargé de décomposer des actions sous forme de séquences de primitives motrices d'échelles variables [Yamashita and Tani, 2008].

2.1.4 Apprentissage par renforcement

De nombreux travaux de modélisation de l'hippocampe et des ganglions de la base ont été réalisés dans le cadre de l'apprentissage par renforcement dans des tâches de navigation. Ces travaux

sont détaillés dans le chapitre 5 consacré à l'apprentissage par renforcement. Une partie de ce chapitre se consacrera également à l'utilisation de multiples stratégies de navigation en parallèle. Un exemple sera donnée avec une planification par carte cognitive et un apprentissage par renforcement.

2.2 Fonctionnement de notre modèle

2.2.1 Apprentissage de cellules de lieu

La pierre angulaire du modèle de l'hippocampe est la construction des cellules de lieu. A vrai dire, ces cellules sont présentes en amont de l'hippocampe lui-même, dans le cortex entorhinal. Nous présentons ici un modèle d'apprentissage visuel de cellules de lieu basé sur la reconnaissance d'amers dans un panorama de l'environnement [Gaussier and Zrehen, 1995; Gaussier et al., 2000] (fig. 2.1). Une description détaillée et discussion du système est disponible dans [Giovannangeli, 2007].

A chaque instant, le robot prend un panorama visuel de l'environnement qui l'entoure. Une caméra panoramique ou une caméra mono-directionnelle montée sur un système pan-tilt permet de faire un panorama en une ou plusieurs prises de vue. Les images composant le panorama passent ensuite par un traitement visuel. Le gradient de ces images, qui sont en noir et blanc, est calculé. Une convolution avec un filtre correspondant à une différence de gaussiennes est réalisée sur le gradient obtenu. Cette opération permet d'isoler des zones de fort contraste, en effet le calcul du gradient, puis les inhibition latérales créées par la convolution permettent de faire ressortir des zones contrastées qui sont entourées par une zone de faible contraste, autrement dit des points saillants. Dans l'image résultante, les points les plus saillants sont sélectionnés par un mécanisme de compétition locale. Ce processus permet donc de sélectionner un certain nombre de points saillants répartis dans tout le panorama visuel, qui correspondront aux amers visuels utilisés pour la navigation. Un système d'identification et de mémorisation de ces amers entre alors en jeu. Une "imagette" (petite zone de l'image située autour du point saillant) subit une transformation log-polaire afin de rendre la représentation robuste aux rotations et effets d'échelle. Un réseau neuronal a pour but d'apprendre et de mémoriser cette représentation log-polaire. Un système de recrutement permet d'encoder de nouveaux amers si ceux-ci n'avaient jamais été appris. L'activité de cette population de neurones permet donc d'identifier un amer perçu. L'équation pour l'apprentissage des amers en fonction des vues locales est la suivante :

$$\frac{dW_{ij}^{VA}(t)}{dt} = R_i^A(t) \cdot X_j^V(t) \tag{2.1}$$

W_{ij}^{VA} représente le poids synaptique entre un neurone de vue locale j et un neurone d'amers i, $R_i^A(t)$ est un signal marquant le recrutement d'un neurone (en passant à 1 puis en repassant à 0) déclenché lorsque l'activité maximale de la population de neurones passe sous un seuil de vigilance ν^A et X^V est l'activité des neurones de vue locale, après la transformation log-polaire (un neurone par pixel).
L'équation pour le calcul de l'activité X_i^A d'un neurone i correspond à un amer visuel est la suivante :

$$X_i^A(t) = 1 - \frac{\sum_j |W_{ij}^{VA}(t) - X_j^V(t)| \cdot H_\epsilon(W_{ij}^{VA}(t))}{\sum_j H_\epsilon(W_{ij}^{VA}(t))} \text{ si le neurone i est recruté} \tag{2.2}$$

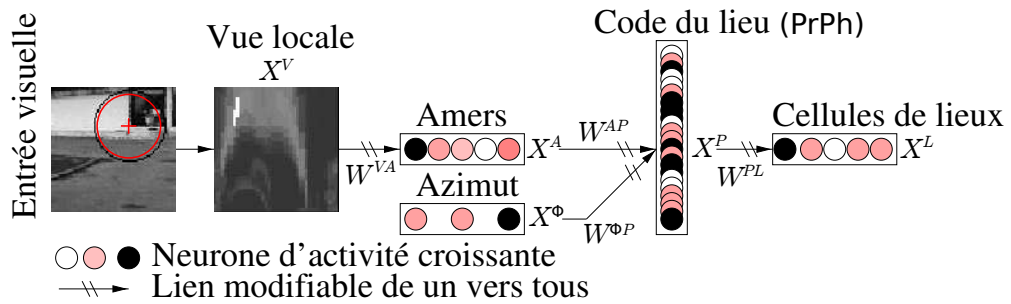


FIGURE 2.1 – Construction du code visuel des cellules de lieu. Les vues sont traitées pour en extraire des amers visuels. Après apprentissage et identification des amers, un code visuel est créé en associant tous les amers d'un panorama avec leur azimut. Cette configuration sert de code pour les cellules de lieu. Tiré de [Giovannangeli, 2007].

avec H la fonction de Heavyside telle que $H_\epsilon(x) = \begin{cases} 0 & \text{si } x < \epsilon \\ 1 & \text{si } x \geq \epsilon \end{cases}$

L'activité $X_i^A(t)$ d'un neurone non recruté est une valeur aléatoire entre 0 et B, B représentant le niveau de bruit synaptique maximal. La fonction H_ϵ permet de détecter les poids synaptiques ayant appris (ϵ est une valeur positive proche de 0).

D'un autre coté, le système de cellules de direction de la tête fournit des informations sur l'orientation du robot. Ce système est alimenté par des données provenant d'une boussole magnétique, mais peut également se baser uniquement sur une boussole visuelle [Giovannangeli and Gaussier, 2007] et être complété par des informations proprioceptive. En effet, chez l'animal, les cellules de direction de la tête font intervenir les systèmes visuels et vestibulaires. La boussole magnétique permet de s'affranchir des coûts computationnels liés à la boussole visuelle, et simule son utilisation. En utilisant l'orientation du robot ainsi que la position du point saillant dans l'image, on peut en déduire l'azimut de l'amer visuel. Ainsi le système possède une voie d'identification des amers ("What") et une voie de localisation ("Where"). Physiologiquement parlant le *What* serait transmis par le cortex perirhinal (Pr), ce qui correspond aux données récoltées par les neurobiologistes [Lenck-Santini et al., 2005], tandis que le *Where* serait transmis par le parahippocampe (Ph). En fusionnant ces informations dans une population de neurones (appelée PrPh) faisant un produit sur les signaux transmis par les deux voies, on peut encoder une multitude d'amers visuels avec l'angle auquel ils sont perçus. Un panorama visuel est ainsi représenté par une population d'amers visuels et de leur azimuts. L'apprentissage dans PrPh est contrôlé par les équations suivantes :

$$\frac{dW_{ij}^{\phi P}(t)}{dt} = R_i^P(t) \cdot H_\gamma(X_j^\phi(t)) \tag{2.3}$$

$$\frac{dW_{ij}^{AP}(t)}{dt} = R_i^P(t) \cdot H_\gamma(X_j^A(t)) \tag{2.4}$$

$R_i^P(t)$ est un signal de recrutement (passant à 1 puis revenant à 0) déclenché lorsque l'activité maximale de la population de neurones passe sous un seuil de vigilance ν^P. La fonction H_γ détecte les neurones fortement activés en entrée du PrPh (γ étant proche de 1).

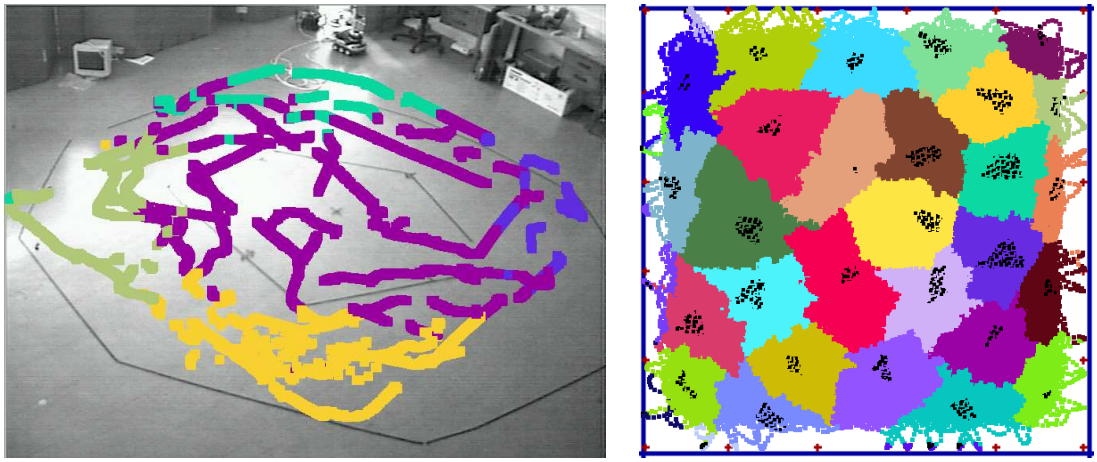
L'équation pour le calcule de l'activité X_i^P du neurone i du PrPh est la suivante :

$$X_i^P(t) = \max(\lambda_i(t) \cdot X_i^P(t - dt), f(\max_j W_{ij}^{\phi P} \cdot X_i^{\phi} \cdot \max_k W_{ik}^{AP} \cdot X_k^A - \theta)) \tag{2.5}$$

avec f une fonction seuil telle que $f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$

X^A et X^{ϕ} sont les activités des neurones codant respectivement pour l'identité et l'azimut des amers reconnus. Un seul amer est traité à chaque pas de simulation d'où la nécessité pour le PrPh de fonctionner comme une mémoire. $\lambda_i(t)$ est un facteur d'oubli dépendant du temps depuis que l'amer a été perçu, du nombre de panoramas effectués etc. Il permet de conserver les activités liées à la perception d'un amer pendant les différentes prises de vue du panorama, voire sur plusieurs panoramas, avant qu'elles soient remises à jour par une meilleure reconnaissance ou oubliées (voir [Giovannangeli, 2007]). θ est le seuil d'activation des neurones du PrPh.

Une seconde population de neurones, située dans le cortex entorhinal, reçoit cette information et opère un processus d'apprentissage similaire à (2.1). Ces neurones constituent les cellules de lieu entorhinales, qui répondent à une configuration d'amers donnée (fig 2.2). L'apprentissage de nouvelles cellules de lieu est contrôlé par un signal de neuromodulation, qui peut être donné de manière supervisée, ou par un système de meta-contrôle, ou bien encore par un simple seuil d'activité minimale sur la population de cellules de lieu. L'activité des cellules de lieu est calculée avec l'équation (2.2) mais un paramètre supplémentaire permet de définir une fraction ρ des meilleurs amers reconnus, permettant de reconnaître une cellule de lieu avec une partie seulement des amers appris. Par exemple, avec un paramètre ρ à 0.25, la reconnaissance d'un quart des amers appris suffira à reconnaître la cellule de lieu de façon robuste. L'association d'une configuration d'amers avec leurs azimuts produit une cellule de lieu dont l'activité spatiale est similaire aux champs de lieu observés chez les animaux.



(a) Cellules de lieu apprises dans un environnement réel en utilisant une caméra panoramique pour la vision. Les couleurs représentent la cellule gagnante à chaque endroit visité.

(b) Cellules de lieu apprises dans un environnement simulé avec 20 amers placés régulièrement au niveau des murs. Les points noirs représentent une activité maximale de la cellule.

FIGURE 2.2 – Champs d'activité de cellules de lieu visuelles.

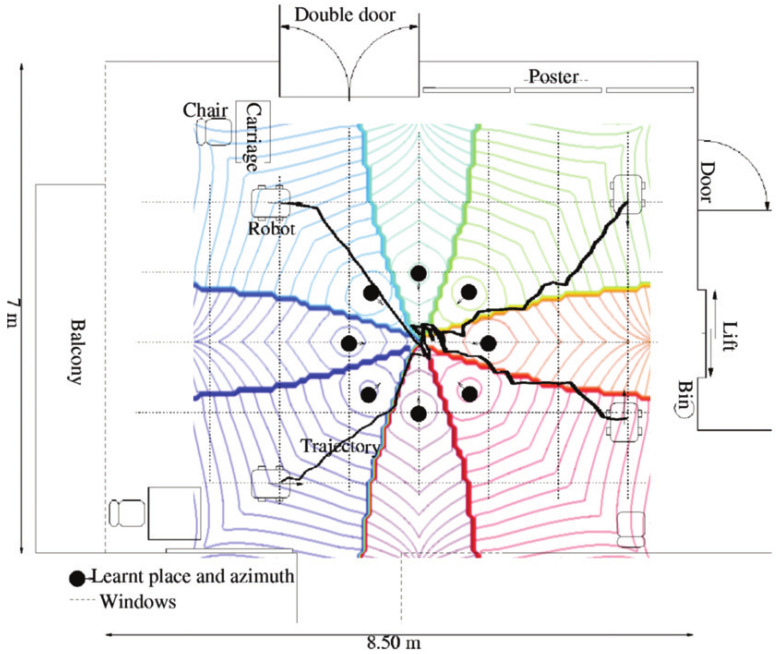


FIGURE 2.3 – Bassin d'attraction construit avec 8 cellules de lieu, permettant de rejoindre un but grâce à des associations sensori-motrices. Tiré de [Giovannangeli, 2007].

En effet quand le robot se déplace, les changements d'angle perçu des amers, ou même la disparition de ces amers suffisent à faire baisser l'activité de la cellule. Le modèle détaillé ici se base sur des informations uniquement visuelles pour construire un code spatial. Il est cependant possible de rendre ce codage plus robuste en intégrant également des données proprioceptives. C'est ce que vise à faire le modèle développé dans [Gaussier et al., 2007], qui utilise l'intégration de chemin pour créer des cellules de grille entorhinales. Cette information est ensuite fusionnée avec les informations visuelles pour rendre les cellules de lieu plus robustes aux ambiguïtés visuelles.

Les cellules de lieu entorhinales, possédant des champs larges et bruités, peuvent servir de base à des stratégies de navigation simples, ne nécessitant donc pas la participation de l'hippocampe. Dans ces stratégies de type "sensori-motrices" (où une action est associé à une sensation), un lieu est associé à une direction à prendre. La direction peut être donnée par l'humain supervisant le robot lors de l'apprentissage. Avec 4 cellules de lieu ou plus, on peut alors construire un bassin d'attraction autour d'un but (fig. 2.3), chacune des cellules donnant la direction vers le but [Gaussier et al., 2000]. Ce genre de stratégie peut également être utilisée pour apprendre une ronde, les corrections données par l'expérimentateur servant à affiner le bassin d'attraction correspondant à la ronde voulue [Giovannangeli and Gaussier, 2010]. Ainsi l'autonomie du robot lui permet de tenter de reproduire une tâche par lui même tout en sachant intégrer les directives qui lui sont fournies par un humain le guidant en cas d'erreur.

Ces expériences ont cependant des limitations et ne peuvent fonctionner que dans un environnement ouvert simple. Si des obstacles sont présents et que plusieurs chemins se présentent pour rejoindre le lieu, ce modèle est incapable de refléter le choix du robot de parcourir l'un ou l'autre, chaque lieu étant associé à une unique action. Enfin le robot est également incapable

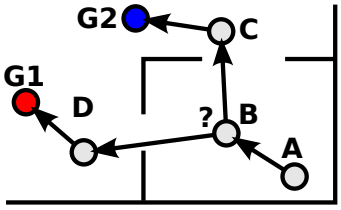


FIGURE 2.4 – Cas où une bifurcation se présente, menant à deux buts différents. Selon le contexte motivationnel, l'action à effectuer en B peut être différente, d'où la nécessité d'une planification de chemin allant au delà d'associations lieu-action.

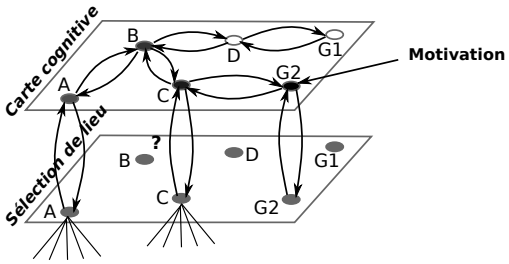
d'adapter rapidement, et de façon autonome, ses trajectoire en cas de changement de l'environnement (passage bloqué, apparition d'un raccourci). Pour résoudre ce genre de problèmes, une solution peut être de mettre en place des mécanismes de frustration, qui détectent une stagnation dans les progrès du robot à réaliser sa tâche et entraînent un changement de stratégie quand la frustration devient trop grande [Hasson and Gaussier, 2010]. On peut également faire appel à des mécanismes de planification, pouvant adapter plus rapidement les actions du robot en fonction de multiples plans.

2.2.2 Transitions et carte cognitive

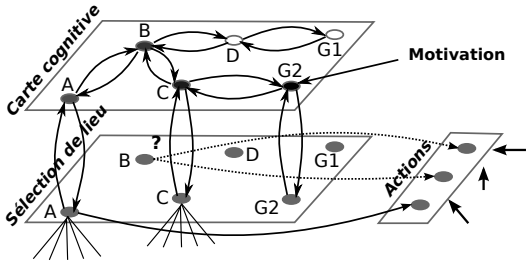
Les approches sensori-motrices utilisant un bassin d'attraction atteignent leurs limites lors de tâches avec des buts distants, qui peuvent être loin de l'endroit où se trouve le robot. Les champs de lieu qui permettent de sélectionner une action ne peuvent alors être discriminant à cause de la distance. On pourrait alors répéter la création d'associations lieu-action, mais cette tâche peut devenir fastidieuse pour de grands environnements. De plus cette représentation statique de l'environnement ne permet pas d'adapter les chemins pris aux changements de topologie, ou bien de gérer la volonté de satisfaire plusieurs buts (à part en créant des bassins d'attraction différents pour chaque but [Gaussier et al., 2000]). Dans une situation où plusieurs chemins peuvent être empruntés pour rejoindre un ou plusieurs buts, le choix d'un chemin ou d'un autre peut être différent en fonction du contexte motivationnel (voir fig. 2.4), ce qui rend inutilisable une simple association lieu-action. Ce sont ces constatations qui ont motivé la conception d'une architecture de planification utilisant une carte cognitive.

Le principe de la carte cognitive est assez simple, les lieux appris dans l'environnement sont liés entre eux par des connexions synaptiques et forment une carte topologique de l'environnement. Lorsqu'un but est découvert, la connexion entre le neurone qui code la motivation pour rejoindre ce but et le neurone représentant le lieu actuel est renforcée. Ainsi l'activité de la motivation se propage à travers la carte cognitive en diminuant à chaque connexion synaptique (dont les poids sont inférieurs à 1). L'activité d'un neurone est calculée comme la valeur maximale des activités post-synaptiques qu'il reçoit. Ainsi plus le lieu est proche du but plus son activité dans la carte cognitive est grande, et on peut alors trouver le chemin optimal vers le but par une simple remontée de gradient. Ce système est similaire à l'algorithme de Bellman-Ford, calculant le chemin le plus court dans un graphe.

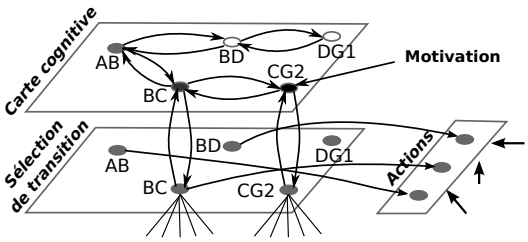
Un problème fondamental de modélisation neuronale intervient cependant lorsque l'on crée une carte cognitive composée de lieux (fig. 2.5). En effet, même si la carte permet de sélectionner le prochain lieu à atteindre, une association directe avec l'action à effectuer ne peut être faite,



(a) Construction d'une "carte cognitive". Pendant l'apprentissage, l'information spatiale passe de la couche de reconnaissance des lieux à la couche "but" (apprentissage de la topologie entre A, B, C ...). Pendant la planification l'activité de motivation est rétro-propagée dans la carte.



(b) La planification est impossible en utilisant des lieux uniquement. En effet une situation peut-être liée à 2 mouvements différents, il est donc impossible de choisir l'action à effectuer.



(c) Utilisation des transitions. En chaque lieu plusieurs transitions peuvent être apprises. Grâce à la rétro-propagation de l'activité de motivation, la meilleure transition peut être sélectionnée et l'action correspondante peut être effectuée.

FIGURE 2.5 – Schémas de carte cognitive montrant la nécessité de l'apprentissage de transitions.

car un même lieu peut correspondre à des actions différentes selon les situations. Une association directe lieu-action ne permet donc pas de résoudre le problème de sélection de l'action. De plus, un système additionnel serait nécessaire pour connaître les lieux accessibles depuis le lieu actuel, comparer leurs activités dans la carte cognitive et sélectionner le lieu approprié. La solution proposée par [Banquet et al., 1997] consiste à utiliser une carte liant des transitions entre lieux, plutôt que les lieux eux-mêmes. Ainsi une transition peut être directement associée à une action (le passage d'un lieu A à un lieu B se fait en allant toujours dans la même direction). La transition sélectionnée par la carte cognitive peut directement mener à l'exécution de l'action correspondante.

Le schéma théorique (fig. 2.6) du rôle de l'hippocampe dans le modèle de [Banquet et al., 1997] nous montre les bases de l'architecture utilisant les transitions entre lieux. Le cortex entorhinal reçoit des signaux des aires corticales associatives puis il filtre et intègre ces informations multimodales pour les transmettre d'une part aux cellules pyramidales de CA3 et d'autres part au gyrus dentelé. DG opère alors une discrimination des signaux et organise une hiérarchie temporelle qui est ensuite retransmise sur CA3. Cette hiérarchie temporelle permet à CA3 d'avoir connaissance des événements passés et de les mettre en correspondance avec l'événement présent, agissant alors comme une mémoire associative en stockant les transitions possibles entre

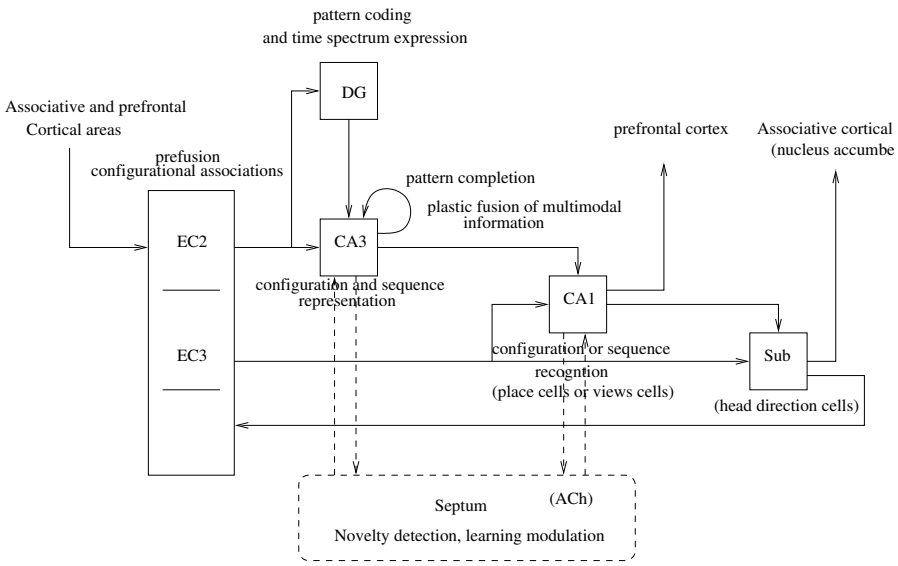


FIGURE 2.6 – Schéma du modèle fonctionnel de l'hippocampe. Tiré de [Banquet et al., 1997].

ces événements. CA3 opère donc une fusion des différentes entrées et permet de faire une "complétion de formes" temporelle. La reconnaissance de la séquence en cours se fait au niveau de CA1 à l'aide des informations de EC et de CA3. Cette information est ensuite envoyée vers le Subiculum, par exemple pour être traitée de manière indépendante de l'orientation du robot grâce aux cellules de direction de la tête. Elle part aussi vers le PFC pour servir aux processus cognitifs de plus haut niveau.

Le modèle simulé a été implémenté et utilisé dans des tâches de navigation robotique [Gaussier et al., 2002; Cuperlier et al., 2007]. L'architecture résultante (fig. 2.7) était capable d'apprendre des transitions entre lieux, de construire une carte cognitive reliant ces transitions et de naviguer vers un but en utilisant la carte. Dans ce cadre, les signaux fournis par EC sont uniquement spatiaux et correspondent à des activités de cellules de lieu. Les activités des cellules de lieu sont calculées à partir d'informations visuelles puis soumises à une compétition de type *Winner-Take-All* (WTA) afin de sélectionner uniquement la cellule ayant la réponse la plus forte en un point précis. On parlera donc par la suite de *lieu actuel* en désignant la cellule de lieu ayant l'activité la plus haute en un point donné. La fonction temporelle au niveau de DG est simplifiée à la simple mémorisation du lieu passé. L'association apprise au niveau de CA3 est donc la transition d'un lieu à un autre en dehors de toute information concernant le temps mis à effectuer cette transition. Les équations gouvernant l'apprentissage sont les suivantes :

$$\frac{dW_{ij}^{EC-CA3}(t)}{dt} = R_i^{CA3}(t) \cdot \theta \cdot X_j^{EC}(t) \tag{2.6}$$

$$\frac{dW_{ij}^{DG-CA3}(t)}{dt} = R_i^{CA3}(t) \cdot X_j^{DG}(t) \tag{2.7}$$

$R_i^{CA3}(t)$ est le signal de recrutement déclenché lorsque l'activité maximale de la population tombe en dessous de $1 - \theta$. θ est le seuil d'activation des neurones de CA3. Les activités neuronales étant entre 0 et 1, son utilisation dans l'apprentissage des poids W^{EC-CA3} assure que la

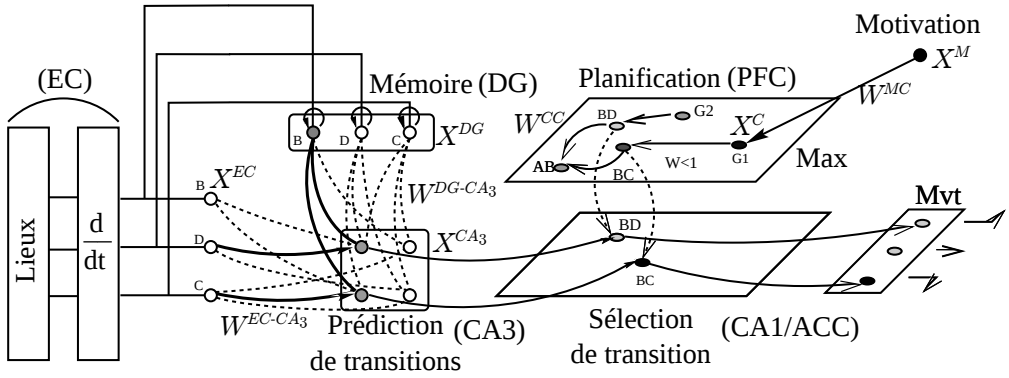


FIGURE 2.7 – Schéma du modèle de boucle hippocampo-corticale pour la planification.

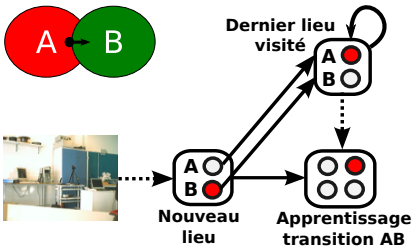


FIGURE 2.8 – Apprentissage

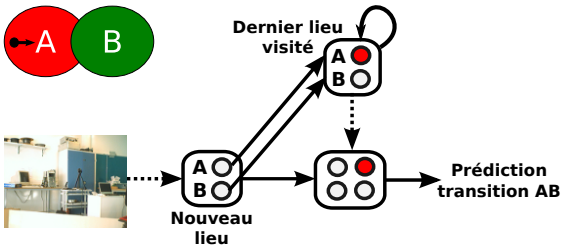


FIGURE 2.9 – Prédiction de transition.

voie synaptique provenant de EC seule n'est pas suffisante pour activer un neurone de CA3. L'équation pour le calcul de l'activité X^{CA3} des neurones dans CA3 est celle-ci :

$$X_i^{CA3}(t) = f(\sum_j W_{ij}(t) \cdot X_j(t) - \theta) \tag{2.8}$$

Les X_j et W_{ij} représentent indifféremment les connexions synaptiques venant de EC et DG.

Une fois l'association entre le lieu passé (A) et le nouveau lieu (B) apprise, toute nouvelle entrée dans A va réactiver la mémoire correspondante dans DG et donc la transition AB précédemment apprise (fig. 2.8 et 2.9). Ce système permet donc de prédire toutes les transitions existantes (celles dont le robot a déjà fait l'expérience) depuis le lieu actuel. Après exploration de l'environnement, on est donc en mesure de prédire en chaque point les lieux directement atteignables mais sans précision par rapport au temps nécessaire. En termes de charge computationnelle, le nombre de transitions dans un environnement ne dépasse pas en moyenne 5 ou 6 fois le nombre de lieux (à cause de la forme des champs de lieu) [Cuperlier et al., 2006].

Lors d'un changement de lieu gagnant (A→B), un signal de neuromodulation est émis pour signaler que l'on vient de réaliser une transition (AB). L'activité de la transition réalisée remonte alors vers le PFC. Le modèle suppose que la carte cognitive est stockée dans le préfrontal ou bien dans le pariétal (avec le préfrontal comme intermédiaire avec l'hippocampe) et rejoint donc la famille de modèles [Hasselmo and Eichenbaum, 2005; Martinet et al., 2011] supposant une carte cognitive corticale, à l'encontre de ceux proposant une carte hippocampique [Muller et al., 1996; Redish and Touretzky, 1998].

Des observation in-vivo semblent cependant suggérer le rôle important de l'hippocampe intermédiaire qui projette fortement vers le PFC dans la transmission d'information nécessaires

à l'encodage rapide de nouveaux buts [Bast et al., 2009]. Dans notre modèle, il existe au niveau préfrontal une mémoire de travail des deux dernières transitions. Les connexions récurrentes sur la carte cognitive sont renforcées entre les neurones codant pour ces transitions (2.9). Lors de l'exploration de l'environnement, la carte cognitive est donc progressivement créée au fur et à mesure que le robot passe d'un lieu à un autre. Les équations régissant l'apprentissage dans la carte cognitive sont les suivantes :

$$\frac{dW_{ij}^{CC}(t)}{dt} = T(t) \cdot ((\gamma - W_{ij}^{CC}) \cdot X_i^C(t) \cdot X_j^C(t) - W_{ij}^{CC}(t) \cdot (\lambda_1 \cdot X_j^C(t) - \lambda_2)) \quad (2.9)$$

$$\frac{dW_{ij}^{MC}(t)}{dt} = S(t) \text{ pour } i, j = \text{argmax}_{k,l}(X_l^C(t) \cdot X_k^M(t)) \quad (2.10)$$

$T(t)$ est un signal binaire (0 ou 1) actif lorsqu'une transition est effectuée (passage d'un lieu à un autre). Ce signal contrôle l'apprentissage sur les connexions récurrentes W^{CC}. γ est un paramètre inférieur à 1 réglant la diffusion de l'activité de motivation dans la carte. λ_1 et λ_2 sont des paramètres d'oubli actif et passif respectivement sur les connexions récurrentes. $S(t)$ est un signal marquant la satisfaction d'un but (découverte d'une ressource etc.). Ce signal contrôle l'apprentissage des connexions synaptiques W^{MC} entre les neurones de motivations d'activité X^M et les neurones de la carte cognitive d'activité X^C.

$$X_i^C(t) = \begin{cases} f(\max_j W_{ij}(t) \cdot X_j(t)) & \text{si } T(t) = 0, S(t) = 0 \\ X_i^{MEM}(t) & \text{sinon} \end{cases} \quad (2.11)$$

X^{MEM} est l'activité provenant de la mémoire stockant les deux dernières transitions effectuées. L'activité X^C des neurones de la carte cognitive correspond à la propagation des activités de motivation en phase d'utilisation, mais correspond uniquement aux activités des dernières transitions réalisées en phase d'apprentissage (lorsqu'une transition est réalisée ou un but est satisfait).

Lors de la découverte d'une ressource (nourriture, eau, etc.), la motivation associée à cette ressource (faim, soif, etc.) est associée, dans la carte cognitive, au lieu où la ressource a été trouvée. Cette motivation se propage ensuite dans le graphe, indiquant le chemin le plus court pour rejoindre la ressource depuis n'importe quel lieu connu. On peut alors faire la fusion entre ces activités de motivation et les activités de prédiction de transition venant de l'hippocampe au niveau du nucleus accumbens. Une compétition par WTA a lieu entre les transitions dans ACC. Le niveau de prédiction des différentes transitions par l'hippocampe est sensiblement le même. L'activité provenant de la carte cognitive vient donc biaiser le choix d'une transition parmi celles qui sont prédites, en sélectionnant ainsi la transition menant le plus rapidement vers le but.

Chaque transition apprise, lorsque la neuromodulation signale un changement de lieu, est associée au niveau du nucleus accumbens avec le mouvement réalisé pour l'effectuer. Ce mouvement provient du système d'intégration de chemin qui est remis à zéro à chaque changement de lieu. Un champ de neurone accumule les informations sur la direction du robot pendant toute sa traversée d'un lieu. Il fournit donc, à la sortie du lieu, la direction et la longueur de la trajectoire effectuée dans le lieu. Cette direction est considérée comme étant celle à prendre pour effectuer la transition. Lors de la sélection d'une transition particulière pour rejoindre un but on peut reproduire en termes de direction le mouvement associé, appris précédemment lors de l'exploration de l'environnement. Ainsi le robot peut remonter de transition en transition en empruntant le chemin le plus court pour rejoindre son but. On peut toutefois noter une limitation

au codage du mouvement par une direction. En effet la direction apprise après exploration totale de l'environnement tend vers le vecteur moyen reliant les 2 champs de lieu. Si le robot part du centre du premier champ de lieu il arrivera alors probablement à destination. Cependant quand le robot pénètre dans les champs de lieu des cellules par leur bord, il peut arriver que le mouvement appris nous mène sur une autre cellule de lieu que celle espérée. L'utilisation des activités de toutes les cellules de lieu voisines et la fusion des mouvements proposés pour rejoindre le prochain lieu peut constituer une solution à ce problème. Une version plus complexe de la planification [Cuperlier, 2006] utilise une compétition souple sur les activités des cellules de lieu, en gardant un nombre k de gagnant et en affinant ainsi la localisation du robot. Ce modèle fait ensuite appel au connexions récurrentes de CA3, non utilisées dans le modèle original, pour sélectionner uniquement les prédictions pertinentes pour la position actuelle du robot.

D'un point de vue neurobiologique, la plausibilité des cellules de transition est difficile à prouver. En effet, observées d'un point de vue spatial, elles ressemblent à des cellules de lieux. [Dragoi and Buzsáki, 2006] suggèrent que le rythme thêta est utilisé pour organiser l'activation séquentielle des cellules de lieux dans l'hippocampe afin d'apprendre les relations temporelles entre lieux successifs. [Manns et al., 2007] montrent des corrélations aussi bien temporelles que spatiales entre des événements olfactifs et spatiaux. La similarité des codes hippocampiques est plus importante entre deux événements proches dans l'espace, mais aussi dans le temps. Certains résultats montrent que les champs de lieu dans l'hippocampe sont perturbés par l'ajout ou le retrait de raccourcis à l'aide de barrières transparentes qui modifient la topologie de l'environnement mais pas les informations visuelles [Alvernhe et al., 2008]. Cela suggère que les neurones de l'hippocampe portent des informations sur les chemins existants plutôt que simplement sur une position de l'animal dans l'espace. Les propriétés asymétriques des champs de lieu pourraient être expliquées par un codage en transitions. En effet, les champs s'étendent dans le sens opposé du mouvement du rat avec la répétition d'une trajectoire [Mehta et al., 1997]. Ainsi, la transition serait apprise, permettant une prédiction du prochain lieu de plus en plus précoce. De plus l'activité neuronale de ces cellules de lieu semble être plus forte lors de la sortie du lieu que lors de l'entrée sur le lieu [Mehta et al., 2000]. Cette asymétrie pourrait s'expliquer par la forme des activités de prédiction de transitions, avec une activité neuronale plus forte lorsqu'on s'approche du moment où la transition va être effectuée.

Enfin, des enregistrements concernant l'activité prospective dans l'hippocampe supportent l'hypothèse de cellules de transition. [Kennedy and Shapiro, 2009] ont montré une activité prospective des cellules hippocampiques en fonction du but visé mais pas des futurs actions. D'autres études montrent que l'activité des cellules hippocampiques peut être dépendante des futurs choix de l'animal (action de tourner à gauche/droite dans un labyrinthe) [Lipton and Eichenbaum, 2008]. Ces résultats pourraient correspondre aux capacités de prédiction des événements futurs que fournissent les cellules de transition.

2.2.3 Apprentissage de séquences temporelles

En se basant sur le même modèle de [Banquet et al., 1997], une implémentation légèrement différente de celle utilisée en navigation a été conçue dans une optique d'imitation, d'apprentissage et de reproduction de séquences d'actions [Gaussier et al., 1998; Moga et al., 2003]. Ces actions peuvent par exemple être des commandes motrices d'un robot, permettant ainsi à ce robot de reproduire des mouvements qu'on lui a appris, notamment pour suivre une trajectoire définie.

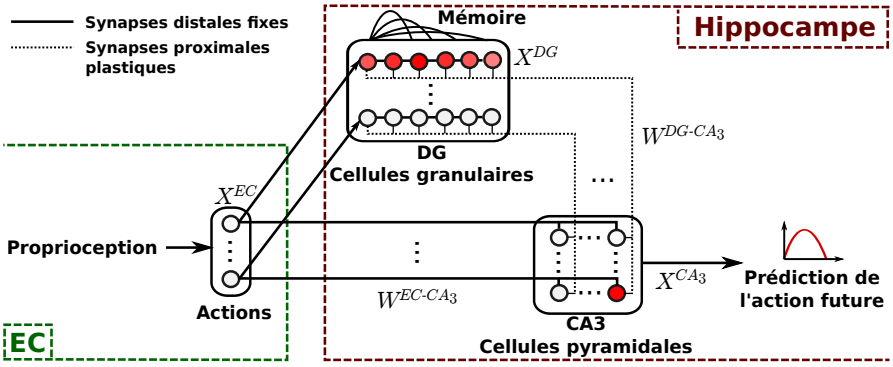


FIGURE 2.10 – Modèle hippocampique d'apprentissage de séquences d'actions. CA3 apprend des associations entre une mémoire temporelle de l'action passée et la nouvelle action.

Le timing jouant une part très importante dans ce type de tâche, le modèle se base sur une architecture de décomposition temporelle biologiquement plausible, assez précise, au niveau de DG. Cette décomposition temporelle est inspirée du système de décomposition spectrale du temps proposé par [Grossberg and Schmajuk, 1989]. Les cellules granulaires de tailles variées qui le composent fournissent un mécanisme de décomposition spectrale dans le temps de l'activité liée à l'apparition d'un événement. En donnant à chacune de ces cellules un temps, une intensité et une durée d'activation différents, on peut alors apprendre de façon précise des transitions entre événements sensoriels ou proprioceptifs. Bien sûr le timing de ces transitions doit rester dans les échelles de temps qui correspondent à l'utilisation de l'hippocampe, à savoir la mémoire à court/moyen terme avec des temps se comptant en secondes au maximum. D'autres structures cérébrales interviennent pour l'apprentissage de transitions sur des délais plus long (minutes et plus).

La figure 2.10 montre le modèle hippocampique d'apprentissage de séquences temporelles. On voit que ces cellules d'une batterie de DG, codant pour la mémoire temporelle d'un événement, réagissent avec des timing différents. L'apprentissage au niveau des poids des connexions synaptiques entre ces cellules et celles de CA3, qui reçoivent également les informations concernant les événements perçus, permet d'apprendre le timing entre deux événements (2.12). Une discussion plus détaillée du modèle de décomposition spectrale utilisé dans l'apprentissage de séquences temporelles se situe dans la section 3.1.1. L'équation d'apprentissage pour CA3 est la suivante :

$$W_{ij}^{DG-CA3} = \sum_k X_k^{EC} \cdot W_{ik}^{EC-CA3} \cdot X_j^{DG} / \sqrt{\sum_k (X_k^{DG})^2} \tag{2.12}$$

L'équation pour le calcul des activités neuronales est :

$$X_i^{CA3} = \sum_j W_{ij}^{DG} \cdot X_j^{DG} \tag{2.13}$$

A la différence du modèle de CA3 utilisé dans la navigation, il n'y a ici pas de recrutement. Les connexions entre EC, DG et CA3 sont topologiques, CA3 formant une matrice de toutes les combinaisons possibles entre les neurones de EC et les batteries de DG. Chaque neurone de CA3 correspond donc à une transition bien précise entre événements, et seul le timing est appris

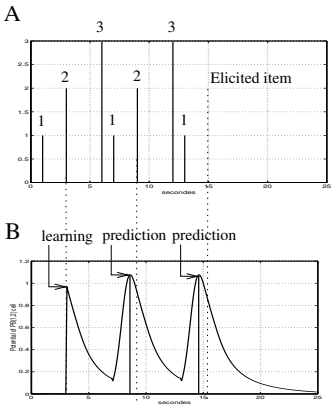


FIGURE 2.11 – (A) Simulation d'apprentissage d'une séquence de 3 événements. (B) Enregistrement de l'activité de CA3 pour le neurone prédisant le passage de l'événement 1 à 2. L'activité de prédiction prend une forme de cloche. Tiré de [Gaussier et al., 1998].

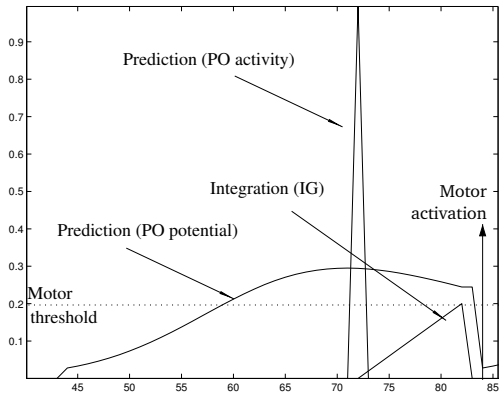


FIGURE 2.12 – Exemple d'activation d'une action prédite par le système de reproduction de séquences. Lorsque la dérivée de la prédiction devient nulle, le système déclenche un pic d'activité. Cette activité est intégrée jusqu'à atteindre un seuil d'activation de la commande motrice prédite. Le système enchaîne ainsi les actions de la séquence. Tiré de [Andry et al., 2001].

par la modification des poids synaptiques entre DG et CA3, les connexions entre EC et CA3 étant fixées à 1.

Comme pour le modèle utilisé pour la navigation, la perception d'un événement déjà rencontré va initier la prédiction de tous les événements futurs possibles par CA3. En effet la perception d'un événement particulier entraîne l'activation de la batterie de cellules correspondante au niveau de DG et permet via les connexions synaptiques de DG vers CA3 d'activer les cellules codant pour des transitions entre événements déjà apprises. De plus, le timing de ces transitions est également prédit. Les activités de prédiction ont la forme d'une cloche, le pic de celle-ci marquant le moment où l'événement prédit est attendu. Cette prédiction s'effectue toujours en avance de phase par rapport au timing initial appris, de manière à réellement prédire l'événement peu avant son arrivée (fig. 2.11). La précision de la prédiction est d'autant meilleure que le timing est court (avec toutefois une réserve sur les timing très court, les neurones ayant un temps de réaction se comptant en millisecondes). Plus la durée d'une transition est grande, plus la prédiction arrivera en avance par rapport au timing réel et perdra sa capacité d'estimation du délai appris initialement [Gaussier et al., 1998]. Ces propriétés coïncident avec les observations faites chez les animaux sur les mécanismes de prédiction temporelle, qui ont inspiré la loi de Weber [Grossberg and Schmajuk, 1989].

Les événements appris par le système peuvent être des ordres moteurs. Lors de l'apprentissage d'une séquence, on apprend donc les transitions entre toutes les actions motrices qui la composent. La présentation au système du premier mouvement de cette séquence lui permet ensuite de prédire le second. Si on conçoit le système pour qu'il réalise l'action prédite avec le timing appris pendant la démonstration, on peut alors reproduire toute la séquence (fig. 2.12). Ce modèle a été utilisé dans de multiples expériences. Une étude des prédictions et de la capacité à reproduire des séquences de signaux artificiels est faite dans [Moga et al., 2003]. Dans [Gaussier et al., 1998], le système est utilisé pour reproduire une trajectoire donnée avec un robot mobile. Dans [Andry et al., 2001; Rengervé et al., 2010], il est utilisé pour apprendre par imitation et

reproduire des séquences de gestes à l'aide d'un bras robotique. Une des difficultés du système se situe dans la reproduction de séquences complexes. En effet la mémoire dans DG ne contient que le dernier événement, et le système n'est pas capable de discriminer plusieurs répétitions distinctes (dans une même séquence) d'un même événement. Les prédictions de CA3 donneront à chaque fois l'éventail de toutes les événements futurs possibles. Il est donc nécessaire de pouvoir séparer les différentes occurrences d'un même événement au cours d'une séquence. Cela peut être fait en utilisant des états cachés [Andry et al., 2005], ou en construisant des états différents en combinant les événements moteurs avec des informations provenant d'oscillateurs neuronaux, fournissant un contexte temporel [Lagarde et al., 2007].

2.3 Discussion sur les modèles existants

Le modèle développé dans l'équipe propose une vue très intéressante et innovante du code spatial hippocampique. L'existence de cellules de transitions, qui coderaient pour une transition entre deux lieux plutôt que pour les lieux eux-mêmes, est une hypothèse forte. Le modèle générique ne limite pas l'apprentissage de transitions aux seules informations spatiales. Comme je l'ai présenté dans la section 2.2.3, il peut aussi s'appliquer à l'apprentissage de transitions entre commandes motrices. Cet aspect d'apprentissage de séquence est au coeur du modèle, plus encore que l'aspect spatial. [Eichenbaum et al., 1999] discutent deux vues qui s'opposent généralement dans les modèles de l'hippocampe : le rôle joué dans la mémoire épisodique donnant une importance particulière à l'aspect temporel, et le rôle joué dans la localisation et la navigation donnant une importance particulière à l'aspect spatial. Ils concluent en proposant, de leur point de vue, que l'aspect de mémoire épisodique est central dans le fonctionnement de l'hippocampe et que les aspects spatiaux représentent un sous-ensemble de ces capacités de mémoire. Le modèle présenté ici partage cette vision de l'hippocampe réunifiant les composantes temporelles et spatiales. Ainsi le rôle principal serait l'apprentissage de séquences et la hiérarchisation temporelle d'événements perceptifs multi-modaux, dont les informations spatiales formeraient un sous-ensemble.

Ce modèle générique a été décliné en deux implémentations. L'une tire plutôt parti des capacités d'apprentissage spatial du modèle et permet d'utiliser des transitions entre lieux pour planifier un parcours. Toutefois l'aspect temporel précis est mis de côté. L'autre implémentation tire parti de la modélisation fine des séquences temporelles, et est utilisée pour reproduire des séquences d'actions. La modalité spatiale des signaux entorhinaux est cependant écartée. [Lagarde et al., 2008b] montrent que ces architectures d'apprentissage temporel ou spatial ne sont pas opposées et peuvent être complémentaires. Dans l'expérience, une ronde est apprise à la fois comme un bassin sensori-moteur avec des associations de type lieu-action, et comme une séquence d'orientations du robot qui peut être reproduite avec un timing précis. Bien que la trajectoire apprise par les 2 systèmes soit la même, la reproduction n'est pas tout à fait identique. En effet, le système temporel reproduit uniquement la séquence d'actions et va donc décaler la trajectoire si le point de départ du robot est décalé. Le système spatial va, quant à lui, rejoindre la trajectoire initiale dans l'environnement. Dans cette expérience, les 2 systèmes d'apprentissage sont utilisés en parallèle et ne coopèrent pas. Aucun mécanisme de sélection de l'action ne permet de favoriser une stratégie plutôt qu'une autre.

Les travaux présentés dans cette thèse visent à proposer un modèle unifié de l'apprentissage spatio-temporel. Une unique architecture prendrait alors avantage des capacités de modélisa-

tion temporelle fine du modèle et travaillerait avec des informations multi-modales pour inférer les relations temporelles entre différents événements sensoriels ou proprioceptifs. Ainsi les systèmes d'apprentissage de séquences et de planification spatiale seraient réunis. Ce modèle pourrait alors être utilisé pour résoudre des tâches nécessitant de faire appel à des capacités d'apprentissage spatial et temporel. La tâche de navigation continue présentée dans la section 1.4 inclut ces différentes composantes. Pour la reproduire, le système robotique doit être capable de naviguer vers un but non marqué dans un environnement ouvert, apprendre la relation temporelle entre un signal spatial (présence sur le lieu but) et l'arrivée d'un signal sensoriel (le son) marquant la disponibilité d'une récompense. De plus, cette prédiction temporelle doit être liée à un comportement moteur particulier : l'absence de mouvement. Nous verrons donc dans les chapitres suivants comment les composantes spatiales, temporelles et de sélection de l'action peuvent être intégrées dans un modèle qui pourra être utilisé pour résoudre une grande variété de tâches. En effet il existe en réel besoin de converger vers des modèles capables de s'adapter à des tâches variées plutôt que d'adapter le modèle de manière ad-hoc à chaque tâche particulière.

On n'a pas la même perception du temps selon les species, c'est ce qui fait que je peux passer ma main entre toi et moi comme ça, parce que pour l'oxygène, une seconde, c'est peut-être dix secondes, et pour le béton, une seconde, c'est peut-être un millième de seconde.

– J.C. VanDamme

CHAPITRE 3

Apprentissage temporel de séquences d'événements perceptifs multi-modaux

Ce chapitre sera consacré à un modèle de l'hippocampe permettant l'apprentissage de séquences d'événements perceptifs. Le modèle présenté dans la section 2.2.3 permet de traiter des informations sensorielles (visuelles etc.) et de construire des perceptions (spatiales etc.). Il permet ensuite d'apprendre la relation temporelle entre des perceptions successifs. Grâce à une décomposition spectrale du temps, il permet d'émettre des prédictions temporelles fines sur l'enchaînement de ces événements. Le modèle a été utilisé dans diverses applications : apprentissage de séquences d'actions motrices [Andry et al., 2001; Lagarde et al., 2007] et apprentissage du rythme d'une interaction impliquant une reconnaissance de gestes [Andry et al., 2011] ou d'expressions faciales [Boucenna et al., 2008]. Dans chacun de ces cas, l'architecture est adaptée pour la tâche et fonctionne pour un seul type de signal. La question de savoir si une architecture unique pouvait être utilisée pour traiter tous ces différents types de signaux s'est donc posée. De plus, il fallait s'assurer de la compatibilité du modèle de décomposition spectrale avec le traitement de signaux spatiaux en navigation, qui étaient traités jusqu'ici avec un système simplifié de mémoire temporelle.

Je présenterai donc dans ce chapitre les améliorations apportées au système de décomposition spectrale afin de le rendre plus précis et plus adaptatif. Les prédictions peuvent être comparées aux signaux reçus et permettent de repérer des différences. Cette comparaison forme la base d'un système de détection de la nouveauté, ou de l'échec et dépend fortement de la précision des prédictions. Je montrerai ensuite comment le système peut apprendre de manière parallèle à prédire plusieurs types de signaux, et comment les équations d'apprentissage peuvent être modifiées pour utiliser une architecture d'apprentissage unique à ces fins. Une expérience robotique effectuée en simulation et avec un robot réel en environnement intérieur sera mise en place pour vérifier les capacités de prédiction du système. Enfin, nous verrons comment ce système d'apprentissage de séquences d'événements multi-modaux peut s'intégrer dans une architecture de planification utilisant une carte cognitive dans des tâches de navigation.

3.1 Précision et adaptation des prédictions temporelles

3.1.1 Modèle de décomposition spectrale du temps

L'architecture neuronale de décomposition spectrale, donnant une mémoire temporelle du délai écoulé depuis un événements récent, se base sur une série de cellules granulaires avec des propriétés différentes de décharge dans le temps. L'activité de ces cellules est représentée par une fonction gaussienne du temps écoulé depuis la présentation du stimulus associé à l'événement. Chaque événement codé de manière indépendante dans EC possède des connexions topologiques vers une batterie de cellules granulaire dans DG, spécifique à ce code particulier. L'activité de la cellule $i(0 < i \leq n_C)$ d'une batterie de DG est définie par les équations suivantes :

$$X_i^{DG}(t) = A_i(t) \cdot f_i \cdot e^{-(t-t_A-f_i)^2/d_i} \tag{3.1}$$

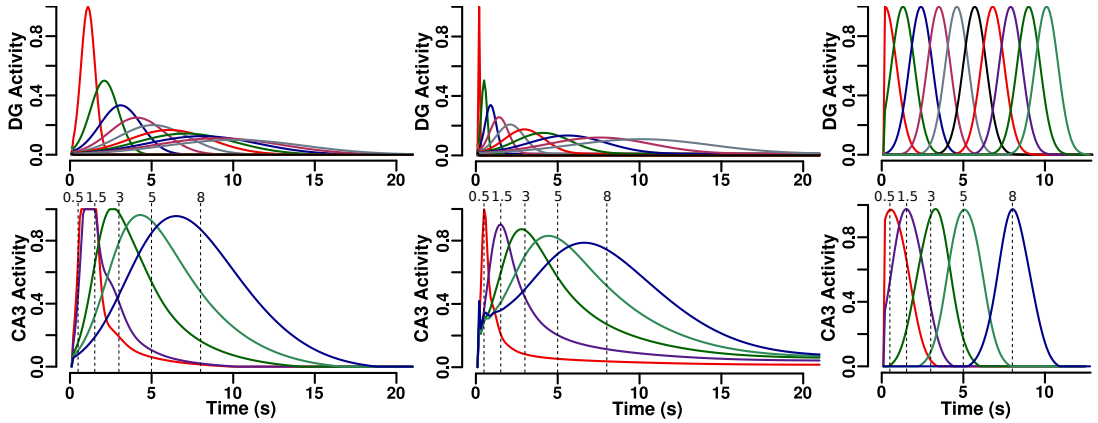
$$\text{avec } _i = (i \cdot T_{max})/n_C \qquad \text{moyenne de la gaussienne} \tag{3.2}$$

$$f_i = 1/i \qquad \text{amplitude de la gaussienne} \tag{3.3}$$

$$d_i = ((i \cdot T_{max})/C)^2 \qquad \text{variance de la gaussienne} \tag{3.4}$$

T_{max} est la période (en secondes) couverte par l'activité d'une batterie, n_C le nombre de cellules dans chaque batterie, C un paramètre gérant la variance des gaussiennes , $A_i(t)$ est à 1 si la batterie à laquelle appartient le neurone i est activée (une seule batterie peut être activée à la fois), 0 sinon et t_A est l'instant d'activation de la batterie.

Cependant ces équations montrent leurs limites dans certains cas particuliers. La répartition de la moyenne des gaussiennes est uniforme sur l'intervalle de temps couvert par la batterie. Selon le nombre de cellules dans la batterie, la première cellule peut mettre quelques centaines de millisecondes, voire quelques secondes avant de s'activer. Cela empêche l'apprentissage de timings très courts et ne correspond pas aux données biologiques montrant des délais d'activation de quelques centaines de millisecondes. De plus cette répartition uniforme répartit autant de cellules au début de la période (où les activités sont très différenciées et permettent d'apprendre des timings précis) qu'à la fin (où toutes les activités sont presque confondues). Il serait plus intéressant d'avoir plus de cellules réagissant avec des timings courts afin de pouvoir prédire plus précisément sans augmenter le nombre de cellules dans une batterie. Enfin, les activités de prédiction engendrées par le modèle original ont la forme d'un pic d'activité plus ou moins large. Ce pic est plus large pour les timings plus long. De plus, il est plus en avance de phase par rapport au délai appris pour les timing plus long. Ces propriétés sont en accord avec la loi de Weber [Grossberg and Schmajuk, 1989], qui formalise des observations identiques concernant les prédictions temporelles chez l'animal. Or, lorsque ce pic d'activité est étroit, l'activité de la prédiction est nulle pendant une longue période, avant et après la prédiction. Certains systèmes de planification ont besoin de connaître à l'avance les différentes transitions pouvant survenir dans le futur. Il peut donc être nécessaire d'avoir des prédictions ayant un minimum d'activité en dehors de leur pic principal. Afin de prendre en compte ces besoins, j'ai modifié les équations



(a) Modèle original **(b)** Modèle modifié. **(c)** Modèle régulier.

FIGURE 3.1 – Activités des cellules d'une batterie de DG et activités de prédiction correspondantes dans CA3. Cinq transitions ont été apprises avec des timings de 0.5s, 1.5s, 3s, 5s et 8s. Le modèle original a du mal à apprendre les timings courts car la première cellule de DG réagit 1s après activation de la batterie. Le modèle modifié s'en sort beaucoup mieux et prédit correctement les événements à 0.5 et 1.5s. Le modèle de base d'ondelettes fournit des prédictions régulières centrées sur le timing appris. Paramètres : $T_{max} = 10, n_C = 10, C = 0.06$ **(b)** $M = 0.01, T_0 = 0.1$ **(c)** $d_i = 1, f_i = 1, T_0 = 0.1, M = 0$

du modèles de la manière suivante :

$$X_i^{DG}(t) = A_i(t)(a + f_i(1-a) \cdot e^{-(t-t_A-m_i)^2/d_i}) \tag{3.5}$$

$$m_i = (T_{max} - T_0 + 1)^{(i-1)/(n_C-1)} + T_0 - 1 \tag{3.6}$$

$$f_i = 1/i \tag{3.7}$$

$$d_i = (m_i \cdot T_{max} \cdot C)^2 \tag{3.8}$$

a est un niveau minimum d'activité pour chaque cellule quand la batterie est activée et T_0 est le timing d'activation (en secondes) de la première cellule. $(i-1)/(n_C-1)$ est compris entre 0 et 1, donc tous les m_i sont compris entre T_0 et T_{max} inclus.

L'utilisation d'un calcul exponentiel des moyennes des gaussiennes permet d'avoir plus de cellules granulaires répondant pour des timings courts. Cela permet d'avoir une plus grande précision de prédiction pour ces timings courts, en accord avec la loi de Weber, que pour les timings longs. Le niveau d'activité minimale a assure que des transitions puissent être apprises pour des timings très longs, dépassant largement le timing d'activation de la dernière cellule. Dans ce cas, le timing de la transition ne peut être appris, le système apprend uniquement que la transition est possible.

La figure 3.1 montre l'activité des cellules d'une batterie dans le temps et les prédictions temporelles correspondantes dans CA3 pour le modèle original et modifié. Une figure supplémentaire montre les activités de prédiction lorsque le spectre temporel utilisé dans DG est une base de gaussiennes régulièrement espacées avec des amplitudes et variances constantes. Pour peu que l'activité de ces cellules soit relativement orthogonale, elles forment une base de signaux temporels disjoints permettant d'apprendre un signal de manière très précise (voir section 3.2.1). La caractéristique de leurs prédictions est qu'elles gardent la même forme (un pic centré

exactement sur le délai appris) quelque soit la longueur du délai appris. Cependant ce genre de modèle s'éloigne d'une modélisation biologiquement plausible.

3.1.2 Apprentissage continu et généralisé

Le modèle d'apprentissage temporel présenté dans la section précédente fonctionne en apprenant la relation temporelle entre deux événements. Le terme d'événement est utilisé pour marquer le caractère transitoire des signaux traités par le système. Ces signaux peuvent correspondre à un changement d'état du système (changement de cellule de lieu gagnante etc.) ou bien à la perception d'un stimulus (un son, un flash de lumière etc.) ou encore à une commande motrice (tourner de 90°). Ils ont cependant tous la particularité d'être codés par une activité neuronale courte et intense. Le système apprend alors à lier ces pics d'activité dans le temps.

L'équation originalement utilisée dans l'apprentissage de séquences (2.12) fait un apprentissage en un coup. La modification des poids synaptiques est modulée par l'activité dans EC, et n'est donc effectuée que pendant le court laps de temps où l'événement se produit. Lors d'une nouvelle occurrence de l'événement, un nouveau timing est appris, écrasant totalement l'apprentissage précédant. Ce système ne permet pas de construire des prédictions temporelles qui s'affinent dans le temps, avec la répétition d'une séquence, ou qui moyennent un timing qui peut être variable. De plus, en cas de changement dans l'environnement, le système peut s'adapter aux nouvelles conditions mais uniquement de manière brutale et instantanée, ce qui le rend très susceptible au bruit. Je vais donc m'intéresser ici aux mécanismes qui permettent de mettre en place un apprentissage itératif, qui adaptera ses prédictions temporelles constamment en gardant une trace des apprentissages précédents.

Les capacités du modèle ne le limitent pas à la prédiction et l'anticipation d'événements ponctuels. Il peut aussi être utilisé pour prédire la forme d'un signal continu. Un tel signal pourrait être le niveau d'activité d'une cellule de lieu au cours de la trajectoire d'un robot se déplaçant, par exemple. Un événement particulier doit cependant toujours activer la batterie de cellules granulaires correspondante, mais celle-ci peut alors être utilisée pour prédire la forme du signal continu, en se basant sur le délai écoulé et l'identité du dernier événement. Ce type de prédiction a été testé avec un apprentissage de type Least-Mean-Square (LMS) [Moga, 2000]. Pour cela, on utilise la règle de Widrow-Hoff (3.9) [Widrow and Hoff, 1960]. Cette règle permet à l'activité de prédiction de converger vers le signal prédit, en minimisant l'erreur moyenne. Un apprentissage continu peut entraîner un problème de convergence, c'est-à-dire que la partie la plus récente du signal appris va être favorisée au détriment de la partie la plus ancienne. Il peut donc être nécessaire de déclencher l'apprentissage de manière aléatoire (avec une fréquence moyenne) pour supprimer cet effet. Cependant, les études en simulation que j'ai réalisées montrent que cet effet de récence est négligeable lorsque la vitesse d'apprentissage α est suffisamment faible et les variances des gaussiennes dans DG suffisamment étroites (voir annexe A). L'équation d'apprentissage du LMS utilisée est la suivante :

$$\frac{dW_{ij}^{DG-CA3}(t)}{dt} = \alpha \cdot (\sum_k X_k^{EC}(t) \cdot W_{ik}^{EC-CA3}(t) - X_i^{CA3}(t)) \cdot X_j^{DG}(t) \tag{3.9}$$

α est un paramètre de vitesse d'apprentissage. X^{EC} est l'activité provenant des neurones de EC et servant de signal inconditionnel (le signal "cible" à apprendre du LMS). X^{DG} est l'activité des neurones de DG servant de signal conditionnel.

Le système d'apprentissage aléatoire n'est de toute façon pas adapté à l'apprentissage de timings pour des événements ponctuels, dont l'activité est présente pendant des laps de temps très courts. Afin de rendre le système capable de travailler avec des signaux de toutes natures, il serait intéressant de parvenir à une équation pouvant traiter à la fois les signaux continus et transitoires. La règle d'apprentissage du LMS fournit un bon point de départ dans le sens où elle permet d'avoir un apprentissage continu, qui adapte les poids synaptiques en fonction du signal désiré (activité venant de EC). Dans l'équation d'apprentissage en un coup (2.12), le terme de modification synaptique est normalisé et prend en compte l'énergie globale du signal présent dans DG. Cela permet de ne pas favoriser la vitesse d'apprentissage des timings plus courts, par rapport aux timings longs pour lesquels les activités dans DG sont plus faibles. Une version du LMS appelée NLMS (pour Normalized Least Mean Square) intègre cette normalisation [Nagumo, 1967]. L'équation correspondante est la suivante :

$$\frac{dW_{ij}^{DG-CA}(t)}{dt} = \alpha \cdot (\sum_k X_k^{EC}(t) \cdot W_{ik}^{EC-CA}(t) - X_i^{CA}(t)) \frac{X_j^{DG}(t)}{\sum_l X_l^{DG}(t)^2 + \sigma_1} \tag{3.10}$$

σ_1 est une valeur positive proche de 0 empêchant la divergence des poids pour des activités très faibles dans DG.

Le NLMS permet d'apprendre les signaux continus de manière adéquate (car il fonctionne de manière similaire au LMS). Cependant, la question de la modulation de l'apprentissage des signaux ponctuels se pose : quand l'apprentissage doit-il être fait ? Comme auparavant, on peut moduler l'apprentissage par l'activité sur EC. On apprend donc uniquement pendant les phases transitoires d'activité marquant un événement. Cette modulation permet effectivement d'adapter le timing de prédiction et d'apprendre un délai moyen entre deux événements. En revanche, aucun oubli n'est possible. Si un événement déclenche la prédiction d'un second événement qui ne se produira plus, cette prédiction ne sera jamais oubliée. En effet, aucune activité au niveau de EC ne viendra déclencher la modification des poids synaptiques, qui pourrait mener à un oubli. A l'opposé, on pourrait laisser l'apprentissage se faire en permanence. Ainsi les poids synaptiques seraient adaptés constamment, adaptant la prédiction à chaque fois que l'événement prédit se produit. Dans ce cas, l'activité de prédiction anticipant l'arrivée de l'événement est perçue comme une erreur à corriger par le LMS qui va tendre à diminuer les poids synaptiques responsables de cette activité. Une prédiction jamais accomplie sera donc oubliée. Par contre, aucune modulation de cette erreur n'intervient dans l'équation du LMS, le système apprend aussi vite qu'il oublie. Or la période d'anticipation pendant laquelle la prédiction est active mais pas le signal prédit (dans EC), correspondant à l'oubli, est beaucoup plus longue que la période pendant laquelle la coactivation de CA3 par EC et DG entraîne l'apprentissage (fig. 3.2).

Il faut donc moduler la vitesse d'apprentissage afin que le système soit particulièrement sensible à la courte période d'apprentissage, tandis qu'il oublie lentement les prédictions apprises. Une solution est de rendre le système sensible aux brusques changements d'activité du signal désiré. Il s'adapte alors lentement pendant la phase de prédiction où le signal désiré reste stable avec une activité nulle, et il s'adapte rapidement pendant la phase de co-activation où l'événement est signalé par une activité transitoire. Pour ce faire, j'ai donc ajouté un terme de modulation de la vitesse d'apprentissage, calculé pour chacun des neurones et se basant sur la différence entre l'activité actuelle du signal désiré et une moyenne de ce signal dans le temps, calculée avec une fenêtre glissante [Hirel et al., 2010a]. Les équations correspondantes sont les

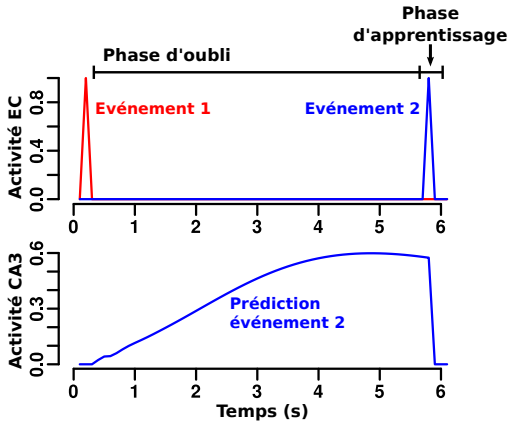


FIGURE 3.2 – Activité dans EC et CA3 pour la prédiction d'une transition entre deux événements ponctuels. L'activité anticipatrice de la prédiction entraîne une phase d'oubli beaucoup plus longue que la phase d'apprentissage qui se fait pendant l'activation dans EC.

suivantes :

$$\frac{dW_{ij}^{DG-CA3}(t)}{dt} = \alpha \cdot \eta_i(t) \frac{(\sum_k X_k^{EC}(t) \cdot W_{ik}^{EC-CA3}(t) - X_i^{CA3}(t)) \cdot X_j^{DG}(t)}{\sum_l X_l^{DG}(t)^2 + \sigma_1} \tag{3.11}$$

$$\eta_i(t) = |X_i^{EC}(t) - m_i^{EC}(t)| + \sigma_2 \tag{3.12}$$

$$m_i^{EC}(t) = \gamma \cdot m_i^{EC}(t - dt) + (1 - \gamma) \cdot X_i^{EC}(t) \tag{3.13}$$

η_i est une neuromodulation de l'apprentissage, m_i^{EC} est une moyenne glissante de X_i^{EC}, σ_2 est une valeur faible assurant une vitesse d'apprentissage minimale et γ un paramètre contrôlant la proportion entre activité passée et actuelle dans le calcul de la moyenne glissante.

En outre, cette susceptibilité de l'apprentissage à des changements d'activité correspond bien au rôle de l'hippocampe dans la détection de nouveauté, évoquée précédemment (sec. 1.2.3). Le septum, en particulier pourrait intervenir dans la régulation de l'apprentissage en fonction de la nouveauté et donc être ma source de la modulation η (fig. 3.3). Avec cette équation, le modèle peut apprendre à adapter ses prédictions temporelles lors de changements dans la temporalité des séquences, oublier certaines prédictions devenues obsolètes et affiner ses prédiction pour des transitions régulières (fig. 3.4). L'affinage de ces prédictions se fait en deux temps : une première phase où les prédictions se forment jusqu'à atteindre un niveau d'activité asymptotique, représentant un certain niveau de confiance dans la prédiction, puis une deuxième phase où le pic de prédiction devient plus étroit et plus précis. En plus, il peut aussi bien apprendre à prédire des événements ponctuels que l'évolution de signaux continus. Ces capacités de généralisation sont primordiales si l'on veut pouvoir traiter des cas variés avec une même architecture, et être capable de réaliser des inférences sur les relations temporelles entre des signaux de natures très différentes.

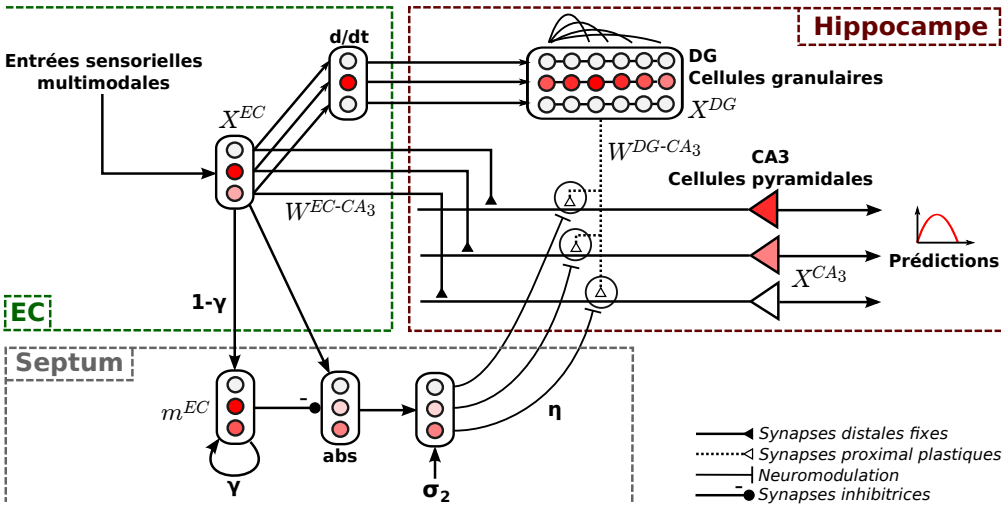
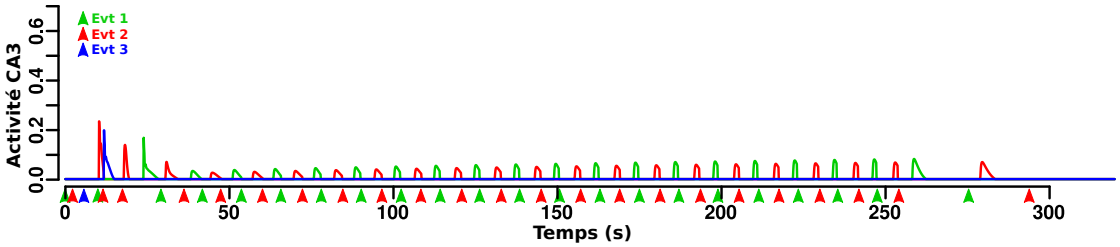
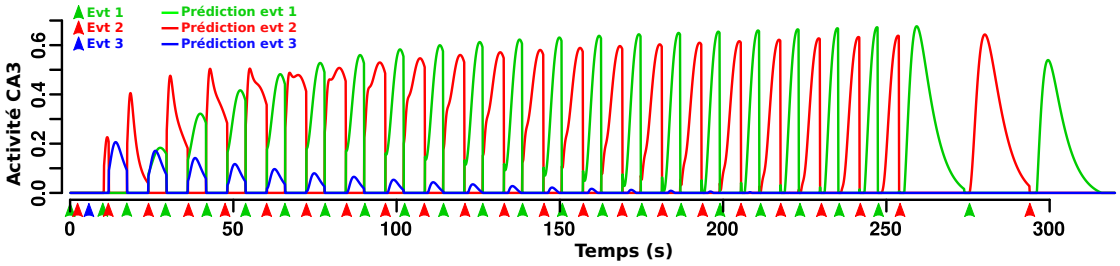


FIGURE 3.3 – Modèle d'apprentissage dans CA3 avec modulation par le septum. Le septum joue le rôle de détecteur de nouveauté en détectant des variations dans le signal provenant de EC. Plus ces variations sont grandes, plus l'apprentissage sera rapide (voir équation 3.11).

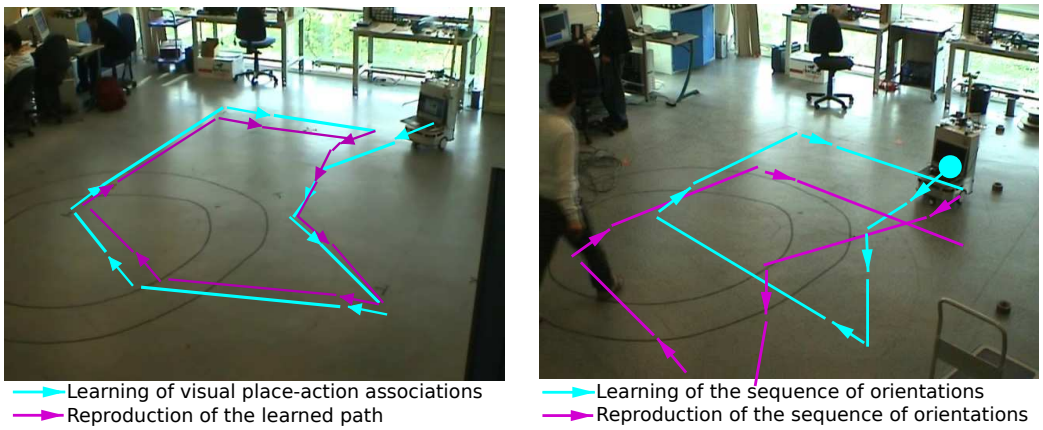


(a) NMLS, équation originale (3.10).



(b) NMLS, équation avec modulation de l'apprentissage (3.11).

FIGURE 3.4 – Activités de prédiction dans CA3 pendant l'apprentissage d'une séquence avec 3 événements. Le NLMS original ne parvient pas à construire les prédictions, qui sont trop rapidement oubliées. Avec la modulation, les prédictions sont apprises correctement. On voit plusieurs propriétés du système : 1) l'événement 3 qui n'est présenté qu'une fois est prédit à la suite du 2 et cette transition est progressivement oubliée car l'événement 3 n'est plus présenté (courbe bleue). 2) L'événement 2 est d'abord présenté très rapidement après le 1, ce qui mène à une prédiction avec un pic très rapide. Puis le délai est augmenté. On peut alors voir le pic de prédiction se déplacer progressivement vers un timing plus long (courbe rouge). 3) L'événement 1 suit le 2 toujours avec le même délai, la prédiction s'affine, présentant un pic d'activité plus étroit et plus fort marquant une plus grande précision (courbe verte).



(a) Stratégie lieu-action

(b) Stratégie de séquence temporelle d'actions

FIGURE 3.5 – Apprentissage et reproduction d'une même trajectoire par 2 stratégies différentes. Lors de la reproduction, le point de départ est déplacé. Alors que la stratégie de navigation par lieux utilise les informations visuelles pour rejoindre la trajectoire originale, la stratégie de séquence d'actions décale l'intégralité de la trajectoire apprise. Tiré de [Lagarde et al., 2008b].

3.2 Utilisation du modèle

3.2.1 Expérience robotique d'apprentissage multi-modal

Est-il nécessaire d'utiliser un système unique pour prédire des signaux de types différents ? Quels sont ces signaux de natures variées pouvant être traités en parallèle ? Dans la section 2.3, je discutais les travaux de [Lagarde et al., 2008b] montrant l'apprentissage d'une trajectoire aussi bien en termes de séquence temporelle d'actions que d'associations lieu/action (fig. 3.5). Ces travaux traitaient les deux types d'apprentissage comme des systèmes parallèles, utilisant deux architectures différentes. Une voie traitait les informations visuelles et spatiales tandis que l'autre traite les informations temporelles et proprioceptives. Aucune hypothèse n'était faite quant à la fusion des actions proposées par chacune des voies. Mais que se passe-t-il dans le cas où une tâche nécessite de lancer une séquence d'actions après l'arrivée sur un lieu donné, où bien quand la perception d'un stimulus sensoriel doit s'intercaler au milieu d'une séquence d'actions ? Des interactions transversales entre les systèmes de séquences temporelles et de traitement de l'information spatiale peuvent être nécessaire pour résoudre certaines tâches. L'utilisation d'un modèle d'apprentissage commun, qui peut apprendre les relations temporelles entre toutes ces modalités, prend alors tout son sens.

De plus, des expériences ont montré que de nombreuses modalités intervenaient au niveau de la construction du code des cellules de lieu dans l'hippocampe des animaux [Eichenbaum et al., 1987; O'Keefe and Nadel, 1978]. Les modalités visuelles et d'intégration de chemin sont particulièrement importantes [Arleo and Rondi-Reig, 2007] et leur intégration serait faite dès le cortex entorhinal [Gothard et al., 2001]. Ces résultats correspondrait bien à un modèle où l'intégration d'un code multi-modal est faite dès EC et transmise à l'hippocampe qui pourrait alors inférer des relations entre toutes ces modalités.

Nous avons donc décidé de tester les capacités du modèle à émettre des prédictions en tra-

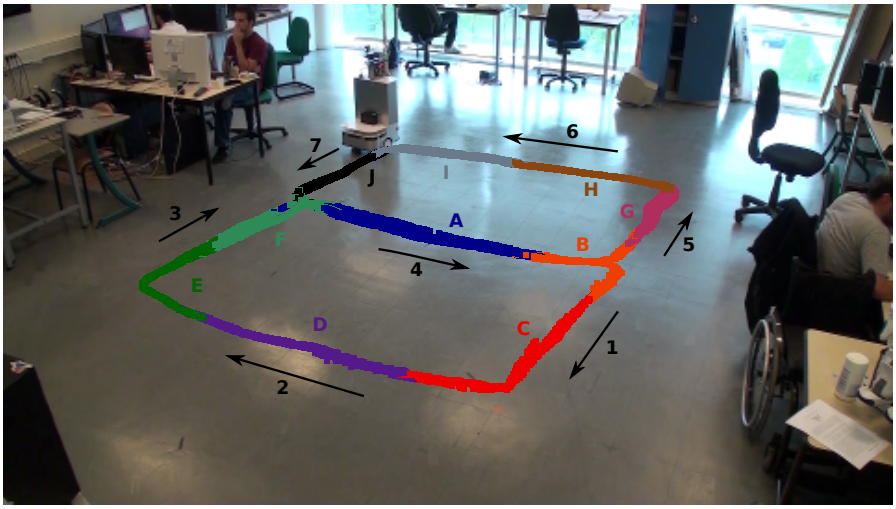


FIGURE 3.6 – Trajectoire en forme de 8 prise par le robot pendant l'expérience. Les couleurs correspondent à la cellule de lieu la plus active à chaque endroit.

vaillant à la fois avec des signaux spatiaux et proprioceptifs. Une expérience a été réalisée à la fois en simulation [Hirel et al., 2010a] et sur robot réel. Les résultats étant relativement similaires, je ne détaillerai ici que ceux récupérés sur robot réel, qui présentent l'avantage de montrer la viabilité du système dans de vraies conditions de navigation, en environnement intérieur. Dans cette expérience, le robot (robulab 10, voir annexe C) suit une trajectoire définie avec une vitesse constante, dans un environnement de bureaux (fig. 3.6). Le robot est guidé par un système externe afin de rester sur cette trajectoire, l'apprentissage est donc fait de manière passive, sans modifier le comportement du robot.

On veut ici tester les capacités du système à apprendre à détecter des régularités dans le comportement. Les chapitres suivants aborderont la problématique du contrôle de l'action à l'aide de ce qui a été appris. Un système pan-tilt avec une caméra permet une prise de vue de panorama. Une boussole électronique fournit l'orientation du robot. Celle-ci pourrait être remplacée par une boussole neuronale utilisant des informations visuelles [Giovannangeli and Gaussier, 2007]. Ces informations sont utilisées pour construire des cellules de lieux selon la procédure indiquée dans la section 2.2.1. La trajectoire peut donc être caractérisée spatialement comme une série de cellules de lieu. De plus, l'orientation du robot est discrétisée sur un champ de 8 neurones représentant des directions différentes (chacun couvrant 45°). Cette discrétisation en 8 directions permet de pouvoir apprendre des séquences d'orientation sans être trop sensible aux variations légères d'orientation du robot qui pourraient perturber le système avec une discrétisation trop fine. La trajectoire peut ainsi être vue également comme une suite de directions différentes, liées dans le temps. Le type de mémoire utilisée dans DG est une base régulière de gaussiennes (voir fig. 3.1). L'implémentation du modèle est réalisée en utilisant le simulateur de réseaux de neurones artificiels temps-réel et distribués *Promethe* (voir annexe B), développé dans l'équipe [Lagarde et al., 2008a; Quoy et al., 2000]. Trois types de signaux différents sont présents dans EC et utilisés par le système de prédiction (fig. 3.7) :

Signal transitoire d'entrée dans un nouveau lieu : Lors d'un changement de la cellule de lieu gagnante (suite à une compétition de type WTA), un pic d'activité identifiant le nouveau

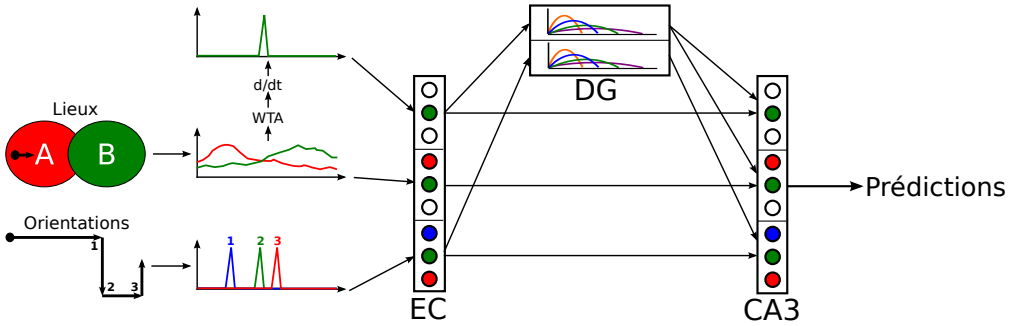


FIGURE 3.7 – Types de signaux présentés au système pour l'apprentissage des prédictions

lieu est émis. Ce pic remet à zéro la mémoire dans DG et active la batterie correspondante au nouveau lieu. Le système prédit, à partir du temps passé dans le lieu actuel, le moment d'entrée dans les prochains lieux possibles.

Activités de cellules de lieu : L'activité de toutes les cellules de lieu, avant compétition, est prédite. A chaque cellule de lieu correspond un neurone prédicteur dans CA3, qui va prédire l'évolution de l'activité de la cellule en fonction du temps passé dans le lieu actuel. La mémoire temporelle utilisée dans DG est la même que pour le signal transitoire.

Changement d'orientation du robot : A chaque changement de direction du robot, un pic d'activité correspondant à la nouvelle direction est émis. Une mémoire parallèle dans DG est remise à zéro, et garde une trace temporelle de la nouvelle direction. Le système prédit les prochains changement de direction possibles en fonction du temps écoulé depuis le dernier changement et son orientation.

Ces 3 voies sont pour le moment séparées, les événements proprioceptifs sont donc prédits uniquement en fonction des événements proprioceptifs passés et les activités de cellules de lieu uniquement en fonction des changements de lieu. Une discussion sur les possibilités de fusionner l'apprentissage en utilisant toutes les modalités possibles pour prédire un événement et sélectionner de manière statistique les événements pertinents sera faite dans le chapitre 7.

Une période d'apprentissage est laissée au robot pendant 8 tours. Après cette période, on arrête l'apprentissage afin de pouvoir analyser la différence entre les prédictions et le signal réel, sans interférence de nouveaux apprentissages. En situation réelle, le robot apprend cependant constamment. Les résultats montrent une bonne capacité du système à prédire à la fois les événements futurs en les anticipant et l'évolution des signaux continus d'activités de cellules de lieu (fig. 3.8). La RMSE (*Root Mean Square Error*) pour la prédiction des activités de cellules de lieu est de 0.077, pour des activités variant entre 0 et 0.7, sur les 2 tours de prédiction réalisés. L'équation utilisée pour le calcul est la suivante :

$$RMSE = \sqrt{\frac{\sum_{t=t_0}^{t_{max}} \sum_{i=1}^{N} (X_i^{EC}(t) - X_i^{CA3}(t))^2}{N \cdot t_{max}}} \tag{3.14}$$

t_0 est l'instant auquel les 2 tours de prédiction, sans apprentissage, commencent et t_{max} l'instant où ils se terminent. Chaque valeur de t correspond à une itération de simulation. X^{EC} et X^{CA3}

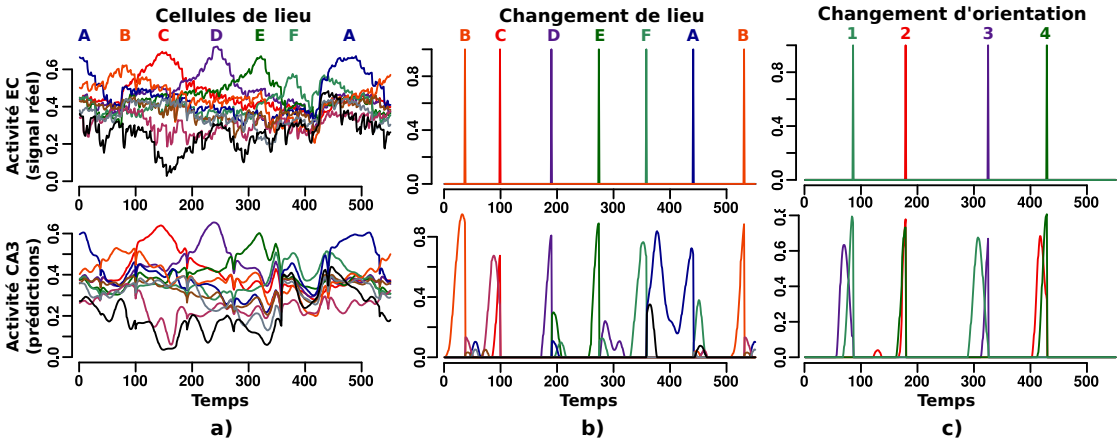


FIGURE 3.8 – a) Prédiction continue de l'activité des cellules de lieu. De faibles chutes d'activité peuvent être observées lors de la remise à zéro de la mémoire sur DG. **b)** Prédiction de l'instant d'arrivée dans les prochains lieux. De petits pics de prédiction secondaires apparaissent et sont dus au bruit sur l'activité des cellules de lieu, qui fait osciller entre 2 gagnants dans le WTA. Ces pics ont des activités faibles dues à la rareté de ces cas. Les grands pics secondaires lors du passage entre F et A témoignent d'une zone de forte compétition entre plusieurs cellules de lieu. **c)** Prédiction de l'instant du prochain changement de direction et de la prochaine orientation attendue.

correspondent respectivement aux activités des signaux originaux et des prédictions. N est le nombre de neurones codant pour les différents signaux à prédire.

La faible erreur de prédiction est due majoritairement aux périodes de remise à zéro de la mémoire dans DG, qui entraînent une chute temporaire de l'activité de prédiction, ainsi qu'aux points de bifurcation de la trajectoire où le système doit prédire deux possibilités de chemins différentes. On peut voir aussi que les événements ponctuels sont prédits avec précision, un pic d'activité se déclarant et culminant juste au moment de l'arrivée de l'événement prédit. Il est intéressant de noter que les pics de prédiction sont très étroits et n'ont presque pas de précession par rapport au timing réel de l'événement prédit. Cela est dû principalement à l'utilisation d'une base sur DG avec des variances faibles, et à l'apprentissage répété de la trajectoire qui a permis d'affiner les prédictions. Enfin, le modèle ne fait pas de différence entre les différentes répétitions d'un même événement au sein d'une séquence. La trajectoire en 8 implique un point de bifurcation où le robot part à gauche ou à droite une fois sur deux. Le système se contente de prédire les événements futurs possibles en fonction du dernier événement. Sa mémoire ne remontant pas aux événements précédents, il ne peut différencier les passages où le robot tourne à droite de ceux où il tourne à gauche, et il se contente de signaler la possibilité des deux cas. [Lagarde et al., 2007] proposent de remédier à ce problème, afin de pouvoir reproduire une séquence complexe d'actions, en rajoutant des oscillateurs neuronaux fournissant un contexte temporel et permettant de retirer l'ambiguïté dans les différentes états du système. Enfin, l'architecture produit des prédictions basées sur des régularités temporelles dans la reproduction de la trajectoire, la vitesse du robot peut donc avoir une grande influence sur ces régularités. La vitesse du robot étant ici constante, il n'y a pas d'interférences liées à des ralentissement ou accélérations. Une solution pour rendre les prédictions indépendantes de la vitesse pourrait être de moduler la perception de l'écoulement du temps dans la mémoire de DG en fonction de la proprioception concernant la vitesse du robot.

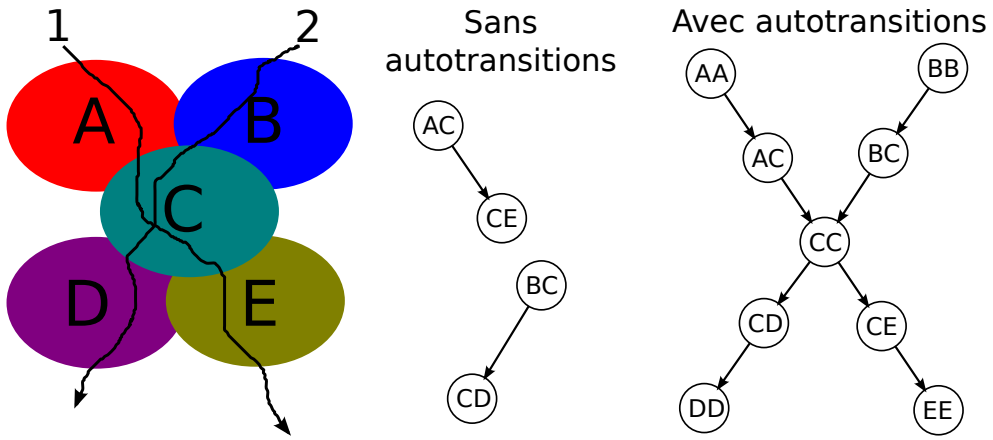
On peut voir dans cette expérience qu'un même signal (les activités de cellules de lieu) peut être traité de manières différentes et fournir des prédictions qui ne se focalisent pas sur les mêmes aspects. Ici, en utilisant les signaux bruts, on peut obtenir une prédiction de l'évolution de ces signaux tandis qu'en analysant les changements de vainqueur dans une compétition de type WTA, on peut prédire les timings de changement de lieu. En présentant au système le résultat direct de la compétition (sans utiliser la dérivée du signal), nous obtiendrions des prédictions hybrides (voir expérience du chapitre 4). Les neurones de prédiction prédiraient alors les transitions vers le lieu auxquels ils correspondent, avec des pics de prédiction, de manière similaire au prédictions utilisant le signal spatial transitoire dans l'expérience présentée ici. Mais, en supplément, ils prédiraient également l'évolution de l'activité de la cellule de lieu, uniquement quand le robot se déplace dans ce lieu. Ils utiliseraient donc l'information du temps passé dans un lieu pour prédire l'évolution de son activité. Nous appellerons les neurones codant pour ce genre de transitions, d'un lieu vers lui-même, des neurones d'*autotransition*. Ces autotransitions peuvent être utiles dans la construction de la carte cognitive. En effet, la faculté du système à créer un code pour des autotransitions, en plus des transitions habituelles, peut accélérer l'apprentissage de la carte cognitive et permettre d'inférer des chemins plus rapidement lors d'une exploration incomplète de l'environnement (fig 3.9). Bien sûr, lors des phases de planification, ces transitions n'ont pas d'action associée et ne sont jamais sélectionnées, mais elles permettent de connecter plusieurs chemins explorés entre eux pour mieux sélectionner le chemin le plus court vers un but. Finalement, notre modèle peut traiter des signaux de différentes modalités, mais peut aussi fournir plusieurs types d'informations sur un même signal.

Finalement, cette expérience montre les capacités de notre modèle à travailler aussi bien avec des signaux spatiaux et proprioceptifs que temporels. Ainsi, un code temporel utilisant des informations spatiales et proprioceptives pourrait expliquer le rôle de l'hippocampe dans la construction de stratégies de navigation allocentriques purement spatiales et égocentriques utilisant une mémoire épisodique [Rondi-Reig et al., 2006].

3.2.2 Compatibilité avec la carte cognitive

L'expérience précédente a permis de vérifier les capacités de prédiction du modèle dans des conditions où les signaux appris pouvaient être indépendants et de natures diverses. Cependant, le modèle a été utilisé de manière passive, la navigation étant contrôlée par un système externe algorithmique qui se contentait de maintenir le robot sur une trajectoire donnée. Je vais donc maintenant discuter de l'intégration de ce modèle unifié permettant de fournir des prédictions temporelles fines dans les architectures de navigation existantes, notamment celle de la carte cognitive (les stratégies de type lieu-action peuvent se passer de l'hippocampe).

Les neurones de prédiction utilisés dans l'expérience précédente ne sont pas réellement des cellules de transitions telles qu'elles ont été présentées dans la section 2.2.2. En effet ces neurones prédisent l'arrivée d'un nouvel événement (A) après n'importe quel autre événement (B, C, D) et ne font pas de distinction entre les différentes transitions (BA, CA, DA). Pour obtenir des transitions distinctes, on pourrait modifier la topologie des connexions entre les populations neuronales de EC, DG et CA3. Une solution serait de créer une matrice de NxN neurones dans CA3 (N étant le nombre d'événements distincts codés par EC). Cette matrice serait une représentation de toutes les combinaisons possibles de transitions entre états. Ce modèle était originalement utilisé dans les architectures d'apprentissage de séquences d'actions. Il a cependant



(a) Deux trajectoires successives traversant un lieu en commun.

(b) Construction de la carte cognitive associée aux 2 trajectoires, avec et sans autotransitions.

FIGURE 3.9 – Exemple montrant l'intérêt des autotransitions. Sans les autotransitions, deux trajectoires passant par un lieu commun mais sans aucune transition commune ne peuvent être connectées. Grâce aux autotransitions, celles-ci sont connectées et des chemins tels que A-C-D peut être inférés à partir de bouts de trajectoires précédentes. Sans les autotransitions, ce chemin devrait d'abord être exploré pour pouvoir être utilisé.

quelques limitations :

1. Une matrice de NxN neurones est coûteuse en terme de mémoire. En outre, une grande majorité de ces neurones sont inutiles car une grande partie des combinaisons de transitions entre événements ne se produira jamais. En navigation, par exemple, le nombre de transitions apprises est en moyenne de 5xN [Cuperlier, 2006]. C'est pour cette raison que l'architecture d'apprentissage de transitions utilisée en navigation fait appel à un mécanisme de recrutement dans CA3, qui recrute uniquement les neurones correspondant à des transitions existantes.

2. Avec la matrice de neurones, chaque neurone code pour une seule transition. Ainsi, on perdrait les capacités du système à posséder des neurones pouvant prédire l'évolution d'un signal, quelque soit l'événement précédent, ou ayant des codes hybrides comme les cellules prédisant à la fois des transitions et autotransitions.

L'idéal serait alors d'utiliser un système de recrutement, qui permettrait d'utiliser uniquement le nombre de neurones nécessaire, et de ne pas fixer une topologie fixe dans le modèle. Le recrutement dans le système utilisé pour la navigation était basé sur un seuil minimal d'activité. Lors d'un changement de lieu, un signal de neuromodulation était émis, déclenchant l'apprentissage dans l'hippocampe. Si l'activité présente dans CA3 était insuffisante, cela signifiait qu'aucun neurone n'était co-activé par EC et DG et ne codait pour la transition réalisée. Un nouveau neurone était donc recruté. Ce mécanisme de recrutement ne peut être utilisé ici directement au niveau de CA3. On veut laisser à CA3 la possibilité d'apprendre plusieurs signaux continus en parallèle, comme cela était le cas pour la prédiction des activités de cellules de lieu dans la section précédente. Dans ce cadre là, en utilisant un recrutement dans CA3, un neurone serait

recruté pour la prédiction de la première cellule. Puis, ce neurone émettant des prédictions en permanence, son activité serait suffisante pour empêcher le recrutement des prochains neurones. Un codage en population des lieux dans EC rendrait cette propriété de prédiction de multiple signaux en parallèle indispensable pour l'apprentissage de prédictions spatiales, puisque que le code d'un unique lieu pourrait correspondre à une multitude de signaux à apprendre en parallèle.

J'ai donc pris le parti de différencier les régions CA3 et CA1 du modèle de l'hippocampe (fig. 3.10). La région CA3 posséderait des neurones prédicteurs d'événements, ne différenciant pas les transitions. Les connexions entre EC et CA3 seraient topologiques tandis que les connexions entre DG et CA3 seraient de type *un-vers-tous*, permettant ainsi à chaque neurone de CA3 de prédire un unique événement en se basant sur toutes les informations de mémoire temporelle possibles. Afin de différencier les multiples transitions correspondant à l'activité d'un de ces neurones de prédiction d'événement, un recrutement serait effectué dans CA1. Les neurones de CA1 possèdent des connexions afférentes de la région CA3 et de EC. On peut donc y fusionner les informations de prédiction provenant de CA3 avec des informations sur le dernier événement provenant de EC (on suppose l'existence d'une mémoire identifiant le dernier événement dans EC3, qui projette vers CA1). Les informations de cellules de grille pourraient aussi participer à la formation de cette mémoire pour les signaux spatiaux. Le recrutement dans CA1 serait alors déclenché par la neuromodulation correspondant à l'arrivée d'un nouvel événement. Les neurones recrutés seraient alors des neurones de transition utilisables par l'architecture de navigation. L'apprentissage, déclenché pour le neurone recruté ou le neurone ayant la plus forte activité si celle-ci est au dessus du seuil de recrutement, est régi par l'équation suivante :

$$\frac{dW_{ij}(t)}{dt} = f(ACh(t) \cdot (\alpha \cdot X_j(t) - \gamma \cdot W_{ij}(t))) \tag{3.15}$$

$ACh(t)$ est la neuromodulation acétylcholine (1 si un événement a lieu, 0 sinon), α la vitesse d'apprentissage et γ un facteur d'oubli.

Cette équation permet de renforcer les connexions synaptiques provenant de neurones présynaptiques activés lors de l'apprentissage. Le terme d'oubli permet d'effacer des associations apprises mais devenues obsolètes car elles ne sont plus rencontrées. La fonction seuil f permet d'éviter la divergence des poids synaptiques.

L'équation pour le calcul de l'activité dans CA1 est :

$$X_i^{CA1}(t) = f(\sum_j W_{ij} \cdot X_j - \theta) \tag{3.16}$$

θ est le seuil d'activation utilisé pour inhiber les neurones non co-activés par EC et CA3. Les X_j et W_{ij} représentent indifféremment les connexions synaptiques provenant de EC et CA3.

Cette distinction entre CA3 et CA1 est cohérente avec de multiples observations faites chez le rongeur. Des lésions effectuées dans EC3 ont entraîné une dégradation des corrélats spatiaux dans CA1 (champs de lieu plus larges, moins cohérents) tandis que les activités dans CA3 étaient inchangées [Brun et al., 2008]. Les modifications d'amers visuels entraînant des déplacements de champs de lieu se répercutent plus vite dans CA3 que dans CA1 [Lee et al., 2004]. Enfin, l'ajout de raccourcis dans un labyrinthe modifie les champs de lieu dans CA3 et CA1, mais touche plus les champs de lieu éloignés du raccourci dans CA3 que CA1 [Alvernhe et al., 2008]. Ces observations suggèrent effectivement que la voie EC-CA1 sert à consolider dans CA1 les représentations spatiales provenant de CA3, en y fusionnant les informations spatiales provenant

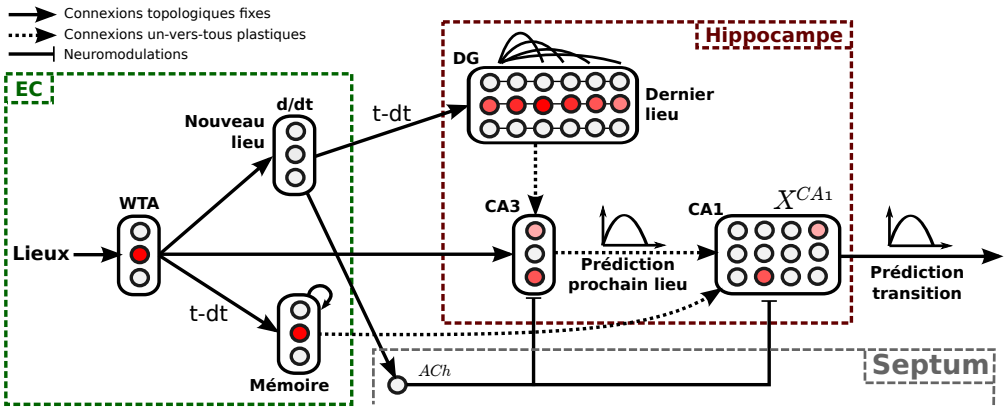


FIGURE 3.10 – Modèle avec différenciation de CA3 et CA1 dans le cas de l'apprentissage de transitions entre lieux. Les neurones de CA3 sont des prédicteurs du prochain lieu, tandis que ceux de CA1 codent spécifiquement pour chaque transition d'un lieu à un autre.

de EC. Ce modèle, où des prédictions de séquences provenant de CA3 sont fusionnées avec un contexte temporel provenant de EC, rejoint celui de [Hasselmo and Eichenbaum, 2005]. Néanmoins, dans leur modèle, ces deux voies sont inversées. EC3 fournit des prédictions en termes de séquences tandis que CA3 encode un contexte temporel qui permet de faire une sélection au niveau de CA1 parmi les séquences prédites.

Dans notre modèle, le recrutement est donc effectué dans CA1 et contrôlé par la neuromodulation déclenchée par tout événement. Dans les travaux précédents de l'équipe, le recrutement était déclenché par un seuil d'activité spécifiquement choisi pour être au dessus d'une activation provenant uniquement de DG et en dessous d'une co-activation par EC et DG (attente d'une conjonction d'activation EC-DG). Ce mécanisme permettait de séparer les transitions prédites des transitions réellement effectuées. Par exemple, lors du passage du lieu A au lieu B, seule la transition AB est co-activée, alors que les autres transitions prédites (AC, AD etc.) ont une activité sous le seuil de recrutement. Le signal de neuromodulation était aussi celui utilisé par la carte cognitive pour apprendre le lien entre la transition effectuée et la précédente. Etant donné que la transition effectuée a une activité plus forte, grâce à la co-activation, un simple WTA permettait de l'identifier pour faire remonter l'information vers la carte cognitive. Or, ce système ne peut marcher que dans un cas simple où toutes les prédictions ont des activités identiques et constantes (pas d'informations temporelles de prédiction). En effet, avec le système temporel les activités de prédiction sont très variables et, lors d'un changement de lieu, il peut être difficile d'identifier la transition effectuée. Le simple ajout d'une activité supplémentaire par la voie EC-CA3 peut ne pas suffire à compenser la différence d'activité avec d'autres prédiction temporelles plus fortes, mais non réalisées. J'ai donc modifié le calcul de l'activité des neurones dans CA3 afin que le signal de neuromodulation inhibe l'activité des neurones n'étant pas co-activés. Ainsi lors d'une transition entre 2 lieux, seuls les neurones co-activés correspondant au code de cette transition déchargent. Cela permet de mettre en place un mécanisme de recrutement dans CA1 se basant sur un seuil d'activité minimale, identifiant le fait qu'un neurone codant pour la transition a déjà été recruté. Le fait que les prédictions puissent retomber à zéro en dehors de leur pic d'activité principal pourrait entraîner le recrutement de plusieurs neurones pour la même transition avec des timings très différents. La modification du modèle de DG permettant de dé-

finir un niveau d'activité de prédiction minimal (sec. 3.1.1) est donc indipensable pour éviter ce recodage d'une même transition. Enfin, grâce à la modulation, la transition effectuée peut être remontée vers la carte cognitive pour être liée aux autres transitions.

Ce mécanisme de neuromodulation est très similaire au fonctionnement de l'acétylcholine (ACh) dans l'hippocampe. Une modulation par ACh peut être déclenchée par le septum, qui joue le rôle de détecteur de nouveauté [Meeter et al., 2004], ce qui correspond au changement de lieu dans notre modèle. Par ailleurs, l'acétylcholine agit sur les synapses en inhibant la transmission des synapses proximales mais pas celle des synapses distales [Hasselmo and Schnell, 1994]. Cette modulation favorise un apprentissage de nouvelles formes en empêchant les formes anciennement apprises de perturber cet apprentissage. Or, les connexions DG-CA3 sont proximales et les connexions EC-CA3 sont distales. Les temps d'actions plutôt longs de l'acétylcholine [Hasselmo and Fehlau, 2001] ont conduit certains chercheurs à baser leurs modèles d'encodage et utilisation de formes dans l'hippocampe sur les différentes phases du rythme thêta [Hasselmo et al., 2002; Koene et al., 2003]. Ce type de modulation pourrait être une alternative dans notre modèle à la modulation par ACh.

3.3 Discussion et applications

Le modèle présenté dans ce chapitre permet d'améliorer les capacités d'autonomie du robot en facilitant l'adaptation de ses représentations temporelles sous forme de chaînes d'événements. Cela rend le comportement du robot plus flexible et adaptable aux changements pouvant intervenir dans son environnement, tout en gardant une certaine stabilité pour ne pas être trop sensible au bruit. La capacité à moyenner l'apprentissage temporel pourra apporter des progrès significatifs dans les diverses applications utilisant le système de prédiction. Plusieurs travaux se concentrent sur l'apprentissage d'un rythme d'interaction, que ce soit dans le cadre d'une imitation passant par la détection de gestes [Andry et al., 2011] ou dans le cadre d'une interaction émotionnelle détectant des expressions faciales [Boucenna et al., 2008]. Ces systèmes apprennent le délai temporel entre deux actions de l'humain interagissant avec le robot. Or ce délai peut présenter une certaine variabilité, et un apprentissage moyen constituerait un bien meilleur outil de travail qu'un apprentissage en un coup qui n'est pas à même de bien caractériser la dynamique d'interaction d'une personne.

Lorsque l'apprentissage d'une tâche commence, il est difficile de déterminer quelles informations vont être pertinentes pour sa réalisation. Notre hypothèse est que les différents signaux sont traités à tout moment par des boucles parallèles. Différents mécanismes peuvent être appris par des structures dédiées (cervelet pour des conditionnements moteurs, ganglions de la base pour les attentes de récompense, cortex préfrontal pour le contrôle des stratégies de sélection de l'action etc.). De la même manière, l'hippocampe apprend des relations temporelles entre des signaux de natures différentes. Au fur et à mesure de la répétition de la tâche, ces signaux peuvent être sélectionnés pour leur pertinence dans l'accomplissement d'un but particulier. L'expérience présentée dans ce chapitre utilise une boucle de prédiction distincte pour chaque type d'information : les changements d'orientation sont prédits en fonction du délai écoulé depuis le dernier changement d'orientation, de même pour les lieux. Cependant, rien n'empêcherait, en modifiant la topologie des connexions, de prédire le prochain changement d'orientation en fonction du dernier changement de lieu, et vice-versa. Dans ce cas, afin de pouvoir arriver à extraire les événements ou conjonctions d'événements multi-modaux réellement pertinents dans la sé-

quence, il faudrait ajouter un système de catégorisation faisant appel à des notions de *positive* et *negative patterning* tel que celui présenté dans [Schmajuk and DiCarlo, 1992]. Une discussion plus détaillée sur ces aspects de patterning sera faite dans le chapitre 7.

Pour conclure, la construction de prédictions temporelles fines permet d'obtenir une représentation fiable des événements attendus. En comparant ces attentes avec ce qui se passe réellement, nous obtenons un système de détection de nouveauté. On peut aussi bien parler de système de détection de l'échec si l'on considère que les prédictions sont liées à des actions effectuées par le robot. Chaque transition prédite par le système correspond en effet à une action que le robot doit effectuer si il veut réaliser cette transition. On peut alors considérer que l'architecture apprend en fait les conséquences, décalées dans le temps, des actions qu'il effectue (par exemple : si je suis dans le lieu A et que je vais au nord, je devrais arriver dans le lieu B d'ici peu). Dès lors, si le robot effectue une action liée à une transition et que cette transition n'arrive pas au moment attendu, la situation peut être considérée comme de la nouveauté ou comme un échec. Nous verrons dans le chapitre 6 comment ces situations d'échec peuvent être détectées et comment cette information peut être utilisée pour moduler le comportement du robot en conséquence.

Publications personnelles

Hirel, J., Gaussier, P., and Quoy, M. (2010a). Model of the hippocampal learning of spatio-temporal sequences. In *Artificial Neural Networks – ICANN 2010*, volume 6354 of *Lecture Notes in Computer Science*, pages 345–351. Springer Berlin / Heidelberg

Placez votre main sur un poêle une minute et ça vous semble durer une heure. Asseyez vous auprès d'une jolie fille une heure et ça vous semble durer une minute. C'est ça la relativité.

– Albert Einstein

CHAPITRE 4

Codage de l'information temporelle et des buts : interaction hippocampo-corticale

Dans le chapitre 3, j'ai décrit un modèle d'apprentissage dans l'hippocampe de séquences temporelles liant des événements multi-modaux. Ce modèle a été utilisé pour apprendre à prédire, de manière passive, des signaux de natures variées. Cependant, la question de l'utilisation de ces prédictions n'a pas encore été abordée. Au travers des chapitres suivants, nous verrons comment cette information peut être transmise dans un réseau incluant le cortex préfrontal et les ganglions de la base, et intégrée à des mécanismes de planification et de sélection de l'action.

Dans ce chapitre, je me concentrerai sur les interactions entre l'hippocampe et le cortex préfrontal. Je m'intéresserai particulièrement à la façon dont les informations concernant les aspects temporels de la tâche, provenant de l'hippocampe, peuvent interagir avec les informations de buts, provenant du PFC. Nous verrons alors comment le modèle s'intègre dans un cadre expérimental venant de la neurobiologie : l'expérience de navigation continue. Je discuterai de la manière dont les résultats expérimentaux obtenus chez les rongeurs peuvent être expliqués par un modèle des interactions hippocampe-PFC. Nous verrons aussi comment cette tâche peut être reproduite par un robot en utilisant les neurones de transition et une extension des modèles de navigation existants. Enfin, des expériences en simulation et sur robot réel nous fourniront des résultats permettant de mettre en avant certaines caractéristiques particulières du modèle. Nous pourrons alors émettre des prédictions quant au fonctionnement des régions préfrontales et hippocampiques lors de la réalisation de la tâche.

4.1 Modélisation des interactions hippocampe-cortex préfrontal

4.1.1 Cadre expérimental : la tâche de navigation continue

En quoi la tâche de navigation continue est-elle intéressante dans le cadre de la modélisation d'un système d'apprentissage de séquences temporelles ? En fait, le protocole expérimental se révèle aussi intéressant du point de vue des neurobiologistes que des roboticiens. En effet, les enregistrements effectués chez les rats au niveau du lieu but, pendant la phase d'attente, permettent d'étudier les activités dans l'hippocampe sans qu'il y ait changement de lieu (animal immobile) ni consommation de nourriture représentant une récompense. D'un point de vue robotique, cette tâche implique l'utilisation de multiples stratégies motrices (navigation vers un but, attente,

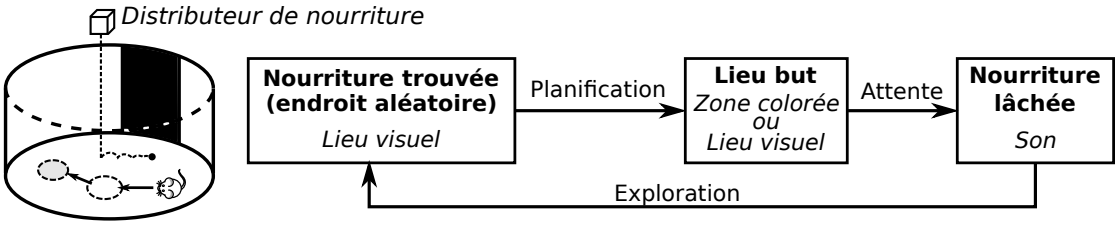


FIGURE 4.1 – La tâche de navigation continue (indicée ou non indicée) vue comme une séquence de comportements moteurs déclenchés par des événements perceptifs multi-modaux.

exploration aléatoire). De plus, ces stratégies motrices interviennent de manière séquentielle et sont liées à des événements perceptifs précis. On peut donc décomposer la réalisation de la tâche en tant que séquence de comportements moteurs contrôlés par des stimuli perceptifs (fig 4.1).

Comment l'aspect temporel intervient-il dans cette séquence ? En fait, la régularité temporelle principale se trouve dans le délai d'attente avant le lâcher de la nourriture qui est fixé à 2 secondes. Pour le rat, apprendre la durée du délai pourrait ne pas être nécessaire, puisqu'il lui suffit de rejoindre le lieu but et d'attendre jusqu'à ce que la nourriture tombe. Cependant, dans la tâche non indicée (c'est-à-dire pour laquelle le lieu d'arrêt n'est pas indiqué par une zone colorée), il est nécessaire pour le rongeur de savoir s'il a bien atteint la bonne zone, afin qu'il puisse éventuellement corriger sa position. Il ne peut se rendre compte de son erreur de positionnement que s'il détecte l'absence de récompense. Cela montre qu'il est capable de prédire l'arrivée de celle-ci avec une certaine précision temporelle. Dans le cas de la tâche indicée, un tel mécanisme de prédiction temporelle ne devrait pas être indispensable, étant donné que le lieu but est visible. Cependant, lors d'essais où la récompense n'était pas donnée après la fin du délai de 2 secondes, les rats reprenaient leurs mouvements après 2 secondes sur le lieu but, que ce soit dans le cas indicé ou non [Hok et al., 2007b]. Il semble donc que :

- Les rats ont appris le timing lors des essais où la récompense est donnée à chaque fois, dans les cas indicé et non indicé, alors que cet apprentissage n'est pas indispensable pour la réalisation de la tâche.
- Les rats sont capables d'utiliser leur connaissance du délai pour détecter l'absence de récompense et reprendre leur mouvement par la suite.

C'est finalement dans le cas des essais d'*extinction* où aucune récompense n'est donnée que l'on comprend la nécessité d'un système d'estimation temporelle. En effet, sans celui-ci, le rat serait incapable de savoir combien de temps il doit attendre sa récompense. Il pourrait donc attendre indéfiniment une récompense qui ne viendra jamais, ou bien quitter le lieu but une fois un certain niveau de frustration atteint (mais probablement bien après la fin du délai). L'apprentissage du délai permet donc ici de détecter précisément le moment où une récompense attendue aurait du être reçue, et d'agir en conséquence. Ces résultats viennent renforcer ceux obtenus lors d'autres tâches de conditionnement avec un délai et correspondent bien à l'hypothèse de modèle fixée dans les chapitres précédents, supposant un apprentissage des relations temporelles entre les divers événements perçus lors de la tâche, même lorsque cet aspect temporel n'est pas crucial pour la réalisation.

Dans notre modèle, cette prédiction du délai entre l'entrée sur le lieu but et le lâcher de la nourriture pourrait facilement être apprise. En effet, si l'arrivée sur le but et le son déclenchent

des événements perceptifs, alors la transition *Lieu but* → *Son* sera apprise avec le délai correspondant. Lors de la reproduction, le pic de prédiction donnera alors le timing auquel l'arrivée du son est attendue. Le son étant une modalité à part entière, nous pouvons catégoriser simplement au niveau entorhinal un événement correspondant à la perception d'un son. La catégorisation de l'entrée sur le lieu but peut néanmoins être différente selon le type de tâche, indicée ou non. Dans le cadre de la tâche indicée, l'arrivée sur le lieu but est détectée directement par un retour visuel marquant l'entrée sur la zone de couleur. Dans le cas non indicé, un premier traitement spatial doit être effectué. Dans l'expérience chez les rongeurs, le lieu but est appris par *shaping*, en réduisant progressivement la taille pour laisser au rat le temps de construire un code spatial suffisamment précis pour caractériser ce lieu de petite taille (environ 20cm tandis que les rats font dans les 12cm). On n'observe cependant pas de sur-représentation du lieu but dans la répartition des cellules de lieu. Ce code spatial est donc probablement construit en intégrant les informations provenant des cellules de lieu avoisinantes (et possiblement des informations directes sur les amers visuels). L'implémentation du système pour la catégorisation spatiale d'un lieu de taille précise devrait donc faire appel à des aspects de *patterning*. Cette catégorisation s'effectue probablement dans DG. En effet, des observations montrent que les aspects émotionnels et motivationnels, gérés par l'amygdale, influent sur les apprentissages à long terme ayant lieu dans DG [Almaguer-Melian et al., 2003]. Ces aspects de patterning sont discutés plus en détail dans le chapitre 7. En attendant la mise en place d'un tel système, les expériences menées dans ce chapitre utiliseront un lieu but indicé, ou bien feront l'association entre la cellule de lieu gagnante et le son (limitant alors la précision du code spatial du lieu but à la taille du champ de lieu de cette cellule après compétition).

Outre cette simplification concernant le code spatial du lieu but dans la tâche non indicée, une autre simplification concernera la modélisation de l'objectif de la tâche lui-même. Dans le cadre de la modélisation de tâches orientées vers un but, il est souvent question de récompenses, motivations, buts, besoins etc. Nous nous référerons au terme de *récompenses* pour la réception d'une récompense matérielle (nourriture, eau, etc.) satisfaisant un *besoin* de bas niveau (faim, soif, etc.). Le terme de *but* sera utilisé pour des objectifs plus abstraits intervenant dans la réalisation de la tâche, et la *motivation* représente la volonté de satisfaire ce but. Pour l'expérience de navigation continue, la récompense réelle est la consommation de la nourriture. Le lâcher de la nourriture, déclenché par la phase d'attente, représente un sous-objectif permettant d'atteindre cette récompense. Cependant la phase de recherche menant à la nourriture correspond à un comportement d'exploration aléatoire de l'environnement, la nourriture pouvant être tombée n'importe où. Nous ferons donc la simplification de considérer le but de la tâche comme étant le lâcher de la nourriture (donc la perception du son correspondant). Chez le rat, la valeur motivationnelle de ce son doit être apprise par conditionnement à travers les nombreuses répétitions du protocole. Le son est alors associé à la consommation future de la nourriture et représente un objectif en soi, menant à cette nourriture. Nous considérerons ainsi que cet apprentissage a déjà eu lieu. La tâche sera alors modélisée en définissant le son comme un but à satisfaire. La phase de recherche de la nourriture correspondra à une période d'exploration aléatoire de durée fixe. La fin de cette période sera marquée par le moment où la motivation à satisfaire le but, dont le niveau augmente constamment, atteindra un seuil suffisant pour déclencher l'utilisation d'une stratégie de planification visant à atteindre le but (fig 4.2). L'utilisation d'une carte cognitive peut ne pas être nécessaire pour la navigation. Une répétition de la tâche pourrait entraîner l'apprentissage par renforcement d'une stratégie de navigation plus "automatique". De même,

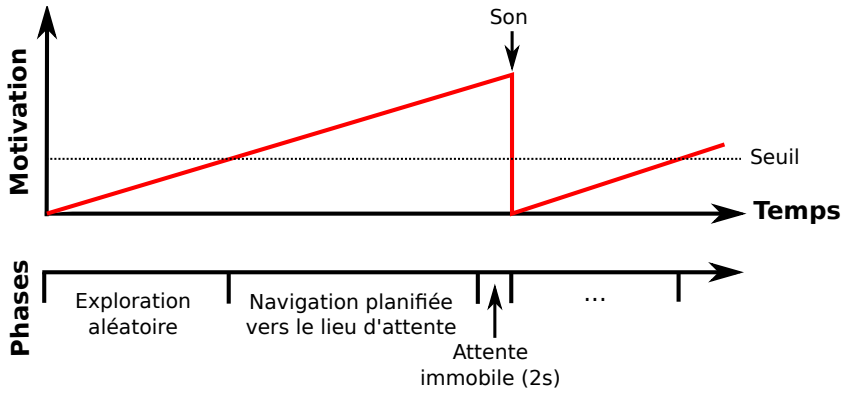


FIGURE 4.2 – Niveau de motivation à satisfaire le but, représenté par le son, au cours des différents épisodes de la tâche de navigation continue. L'utilisation de la navigation par planification pour rejoindre le but est déclenchée lorsque la motivation dépasse un seuil fixé. La durée de la période d'exploration est fonction de ce seuil.

les nombreuses visites du lieu but et son importance dans la tâche pourraient mener à l'utilisation de stratégies de "retour au nid" se basant sur des informations visuelles ou proprioceptives. Ces questions concernant l'existence de plusieurs stratégies de navigation en parallèle seront abordées dans le chapitre 5.

4.1.2 Activités hors-champ de cellules de lieu

Le modèle de planification en navigation utilisant la carte cognitive, présenté dans le chapitre 2, est cohérent avec plusieurs observations neurophysiologiques :

- Des neurones avec des corrélats spatiaux liés aux buts sont présents dans le cortex préfrontal [Hok et al., 2005].
- Des neurones avec des champs de lieux larges et bruités sont présents au niveau entorhinal (les cellules de lieu de notre modèle) tandis que des neurones avec des champs plus étroits et dépendants du contexte sont situés dans l'hippocampe (les cellules de transition). Voir la fin de la section 2.2.2 pour une discussion sur la plausibilité neurobiologique des cellules de transition.

Par ailleurs, la version spatio-temporelle du modèle, présentée dans le chapitre 3, prédit que l'activité des transitions, observées d'un point de vue temporel et non purement spatial, prend la forme d'une cloche avec un maximum d'activité précédant l'arrivée attendue de l'événement prédit. La forme de cette activité est similaire à celle observée pour les cellules de lieu durant la phase d'attente de la tâche de navigation continue (voir fig. 1.9 et 3.1). La majorité des cellules de lieu présentent un pic secondaire d'activité durant cette période, avec un maximum d'activité précédant la fin du délai de 2s. Il est difficile de savoir si ce pic d'activité est relatif au début ou à la fin du délai. Pour cela, il faudrait conduire des expériences avec des délais variables, mais apprendre à un rat à se tenir immobile pendant plus de 2 secondes s'avère difficile. Néanmoins, la similarité des activités laisse penser que ces activités hors-champ de cellules de lieu pourraient en fait être liées à des activités de prédiction de cellules de transition.

Même si le modèle peut expliquer la forme des activités hors-champ, il n'explique pas pourquoi celles-ci sont observées pour la majorité des neurones de CA1, même loin de leur champ

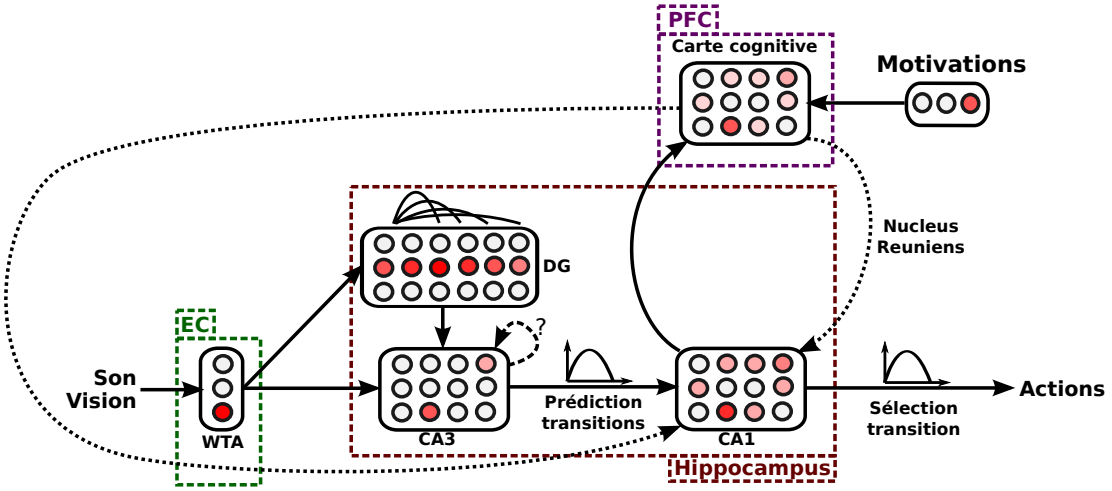


FIGURE 4.3 – Hypothèses rejetées concernant les activités hors-champ dans l'hippocampe. Les pointillés représentent les rebouclages possibles du PFC vers CA1 pouvant expliquer la diffusion des activités de prédiction. Les résultats montrant la persistance des activités hors-champ après inactivation du PFC invalident cependant ces modèles.

principal. Cette activité est clairement liée à des informations de buts, car elle n'apparaît spatialement pour aucun autre lieu que le lieu but. Temporellement, elle semble être prédictrice du lâcher de nourriture, correspondant au but de la tâche, qui a lieu à la fin du délai d'attente. Les informations de but étant codées au niveau préfrontal dans notre modèle, il existe alors forcément un retour du PFC sur l'hippocampe expliquant cette activité liée au but dans CA1. Sans ce retour, aucune inférence ne pourrait être faite au niveau hippocampique quant à la valeur motivationnelle d'une transition plutôt qu'une autre. Plusieurs hypothèses de modèle ont été envisagées par le passé concernant le retour de la carte cognitive préfrontale sur la sélection des transitions hippocampiques. Nous allons voir ici que les nouvelles données nous permettent de choisir entre une possibilité de retour direct au niveau de CA1 [Gaussier et al., 2002] ou bien une fusion des informations hippocampiques et préfrontal dans le nucleus accumbens [Banquet et al., 2005].

L'hypothèse principale de notre modèle étant le fait que certains neurones de l'hippocampe sont en fait des cellules de transition implique que les champs de lieu secondaires observés in-vivo sont en réalité des *transitions secondaires*. L'idée selon laquelle les buts sont codés au niveau préfrontal, de pair avec le fait que ces transitions secondaires n'arrivent qu'au lieu but, suggèrent l'implication du PFC dans la propagation de la prédiction du son aux différentes transitions. Une première hypothèse serait que l'activité de prédiction de la transition *But* $\rightarrow$ *Son* se propage vers le PFC et reboucle sur CA1. En considérant le PFC comme le support de la carte cognitive, cette activité pourrait se diffuser dans la carte et revenir exciter de façon diffuse l'hippocampe. Ce rebouclage du cortex préfrontal sur l'hippocampe pourrait se faire via le cortex entorhinal, ou bien plus probablement par le nucleus reuniens. Ce dernier semble en effet être une interface majeure de communication bi-directionnelle entre le cortex préfrontal et CA1 [Bertram and Zhang, 1999; Vertes et al., 2007]. Une autre explication plausible serait que l'activité de prédiction, une fois fusionnée avec des informations de but au niveau du PFC, soit projetée de manière diffuse vers CA1 via des connexions très étendues du PFC vers CA1. Les hypothèses de modèles proposées sont résumées dans la figure 4.3).

Pour vérifier ces hypothèses, une expérience a été conduite durant laquelle une inactivation temporaire du cortex préfrontal a été pratiquée sur des rats devant effectuer la tâche de navigation continue [Hok, 2007]. L'inactivation a eu peu d'effets sur les performances des rats. La seule différence notable a été une augmentation du niveau moyen d'activité dans l'hippocampe. Mis à part cette observation, les activités spatiales dans l'hippocampe, et notamment les activités hors-champ, sont restées inchangées, de même que les performances des rats pour résoudre la tâche. Ces résultats suggèrent que le préfontal n'est pas indispensable pour la résolution de la tâche, et invalident les hypothèses le considérant comme la source des activités hors-champ. Même s'il joue peut-être un rôle dans la formation de cette activité, il ne constitue en tout cas pas la seule voie de propagation.

D'un autre côté, les rats dont le PFC a été inactivé avaient atteint des niveaux asymptotiques de performances et maîtrisaient parfaitement la tâche. On sait aussi que le PFC reçoit bien des informations spatiales et temporelles de l'hippocampe [Burton et al., 2009] et est capable de caractériser les buts de la tâche. Il est probable que le rôle du PFC soit de travailler de concert avec l'hippocampe pour la facilitation de l'acquisition des comportements nécessaires à la réalisation de cette tâche. Il traiterait donc des informations spatio-temporelles, en les intégrant avec des signaux de satisfaction de but, et fournirait un feedback à l'hippocampe. Ceci permettrait l'acquisition de mécanismes automatiques de bas niveau contrôlant probablement le comportement de l'animal lorsque la tâche est bien connue. Une fois l'acquisition de ces mécanismes de bas niveau effectuée, les processus cognitifs plus complexes ne sont plus nécessaires.

Je vais donc maintenant présenter un modèle de la boucle hippocampe-PFC cohérent avec les activités hors-champ observées et robuste aux lésions du PFC [Hirel et al., 2010c]. Dans ce modèle (fig. 4.4), le PFC est impliqué dans l'apprentissage de l'association entre un événement et la satisfaction d'un but. Le cortex entorhinal intègre ici des informations visuelles (sous la forme d'activités spatiales de lieux) et sonores. Chaque changement de lieu ou perception d'un son déclenche un événement et la relation temporelle entre ces événements est apprise. L'événement correspondant au son intervient de façon simultanée avec la satisfaction du but (le son avait été défini comme un but à atteindre dans la section 4.1.1).

Ce signal de satisfaction est transmis de manière diffuse aux neurones d'événements dans EC. Cela entraîne un nouvel apprentissage dans CA3 : en plus du neurone de prédiction codant pour le son, tous les autres neurones apprennent à prédire un événement arrivant 2 secondes après l'entrée sur le lieu but (qui est le dernier événement). Outre l'événement assigné à chacun des neurones de CA3 par les connexions topologiques venant de EC, l'apprentissage leur fait ici prédire en plus la satisfaction du but. Ainsi ces activités hors-champ seraient des prédictions secondaires de la satisfaction d'un but, et résulteraient uniquement du circuit d'apprentissage temporel de l'hippocampe, ce qui expliquerait leur persistance après l'inactivation du PFC. Le PFC serait donc nécessaire pour l'acquisition de ces prédictions secondaires mais, après l'apprentissage, la prédiction de la satisfaction du but est indépendante du préfrontal.

Finalement, je n'ai pas évoqué ici le rôle des connexions récurrentes sur CA3, qui pourraient aussi expliquer la propagation de l'activité de prédiction du son aux autres neurones prédicteurs dans l'hippocampe. Cette hypothèse ne supprimerait toutefois pas le besoin d'un retour du PFC, qui peut expliquer pourquoi cette propagation n'a lieu que sur le lieu but. Dans ce cas, le signal de satisfaction du but provoquerait un apprentissage synaptique dans les connexions récurrentes afin que le neurone de prédiction du son projette de manière diffuse sur les autres neurones de CA3. Il faudrait cependant garder un modèle de contexte dans EC pour expliquer l'absence

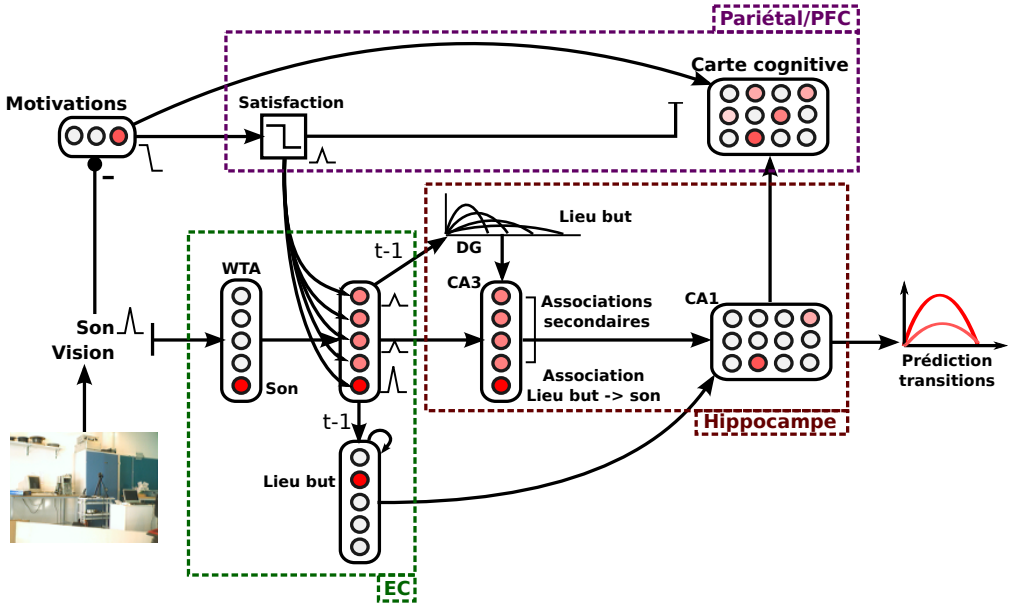


FIGURE 4.4 – Modèle d'apprentissage secondaire dans l'hippocampe, expliquant les activités hors-champ. Des prédictions secondaires sont apprises par les neurones de prédiction d'événements lors de la satisfaction d'un but, via le retour du PFC sur EC. Tous les neurones de prédictions dans CA3 prédisent deux transitions différentes : leur propre transition originale et une prédiction secondaire liant le lieu but avec la satisfaction du but. Cette activité de prédiction apparaît comme un champ de lieu secondaire quand on la représente spatialement.

des activités hors-champ lors d'épisodes de navigation non motivée. Le modèle se rapprocherait donc très fortement de celui présenté ici.

4.1.3 Sélection de l'action et arrêt du mouvement

Je vais maintenant aborder le problème du contrôle du mouvement du robot. Le système actuel permet d'apprendre les transitions entre les différents événements survenant lors de la tâche et peut prédire précisément dans le temps l'arrivée de la satisfaction du but lorsqu'il attend sur le lieu but. Cependant ce modèle est, jusqu'ici, passif et ne permet pas d'expliquer comment l'animal contrôle son mouvement et choisit de s'arrêter sur le lieu but. Dans ce cadre, il convient de s'intéresser aux mécanismes de sélection de l'action faisant intervenir la planification et la carte cognitive. Ces mécanismes impliquent une compréhension des interactions entre l'hippocampe, le cortex préfrontal et le nucleus accumbens (voir fig. 4.6). Le modèle de carte cognitive suggère que celle-ci est stockée soit dans le cortex préfrontal, soit dans le cortex pariétal, auquel cas le cortex préfrontal servirait d'intermédiaire et permettrait de récupérer des informations sur des parties de la carte qui sont pertinentes pour la tâche en cours. L'absence de corrélats spatiaux clairs dans l'expérience de [Burton et al., 2009] où le but change de place chaque jour pourrait être expliqué par la propagation dans la carte cognitive de l'activité de ces buts, entraînant alors une faible sélectivité spatiale des activités enregistrées mais laissant l'aspect temporel intact. Une question ouverte concerne la structure cérébrale accueillant le mécanisme de sélection de transition par la carte cognitive. La sélection se fait par l'intégration des activités de transitions

prédites par l'hippocampe avec l'activité de la carte cognitive qui vient biaiser le choix d'une transition à l'aide d'une compétition de type WTA. La fusion de ces 2 voies pourrait se faire au niveau de CA1 directement, sur lequel le PFC projette par l'intermédiaire du nucleus reuniens, ou bien au niveau du nucleus accumbens qui reçoit des projections de CA1 et du PFC.

On sait, par le biais des enregistrements effectués dans CA1, que les activités de prédictions temporelles des transitions parviennent dans CA1. Ces activités sont fortement variables : faibles lorsqu'on s'éloigne du timing de la transition prédite, fortes lorsqu'on est proche de ce timing. Si le biais apporté par la carte cognitive se faisait au niveau de CA1 directement, celui-ci devrait pouvoir biaiser une compétition entre des prédictions de transitions d'activités très variable. En pratique, il est très difficile de pouvoir effectuer ce biais. Il est beaucoup plus aisé de biaiser le choix d'une transition lorsque les différentes prédictions ont sensiblement le même niveau d'activité, comme cela est le cas dans le modèle original de navigation qui n'utilisait pas de description temporelle des transitions. Devant l'impossibilité de travailler directement avec les prédictions temporelles fines de transitions dans la compétition de sélection d'une transition, le modèle a été adapté pour introduire une étape de binarisation des activités de prédiction. Ainsi, toute transition projette vers un neurone de prédiction au niveau de ACC qui aura un état excité ou non. L'état excité sera déclenché par une activité de prédiction dépassant un certain seuil (ce qui correspond à une transition prédite donc réalisable, quelque soit le timing appris). Cette binarisation, supposée avoir lieu dans ACC afin de laisser intactes les prédictions temporelles dans l'hippocampe pour être conforme aux observations neurophysiologiques, permet la mise en place d'une compétition intégrant les activités de la carte cognitive.

On suppose donc que la sélection d'une transition particulière à effectuer est réalisée dans le nucleus accumbens en intégrant les activités de prédictions de l'hippocampe et de carte cognitive du cortex préfrontal. Il reste alors à faire le lien entre ces activités de prédiction et sélection de transitions et le mécanisme de sélection de l'action permettant au robot de bouger ou de s'arrêter. Dans une première étape de modélisation, nous avons supposé l'existence de neurones inhibiteurs du mouvement présents dans le PFC (fig. 4.5). Ainsi l'activation des ces neurones, directement liés aux systèmes de commandes motrices régulant la vitesse du robot, permettrait de commander l'arrêt du déplacement. Dans cette version du modèle, une compétition aurait lieu dans le PFC entre les transitions prédites par l'hippocampe et le gagnant de cette compétition contrôlerait l'activation ou non de l'inhibition du mouvement. Lors de l'arrivée sur le lieu but, l'autotransition associée à ce lieu commencerait à être prédite et déclencherait l'arrêt du déplacement. La prédiction de l'arrivée du son augmenterait peu avant la fin du délai et finirait par gagner la compétition, déclenchant la reprise du mouvement. Cette compétition serait modulée par la motivation pour limiter ces épisodes d'arrêt du mouvement aux phases de navigation vers le but, et non aux phases d'exploration aléatoire. Ce modèle avait cependant des limitations :

- La compétition entre deux activités de prédiction temporelle de transitions est très sensible à la forme de ces activités. Or, cette compétition contrôle l'arrêt et la reprise du mouvement. Un réglage très précis du modèle et un long apprentissage est alors nécessaire pour obtenir le comportement adéquat.

- Aucune hypothèse n'était émise quant à la manière dont chacune des transitions était associée à l'inhibition du mouvement ou non.

- Les résultats ultérieurs de lésion du PFC pendant la tâche de navigation continue étaient en désaccord avec le modèle. En effet, dans le cas d'un contrôle actif du mouvement

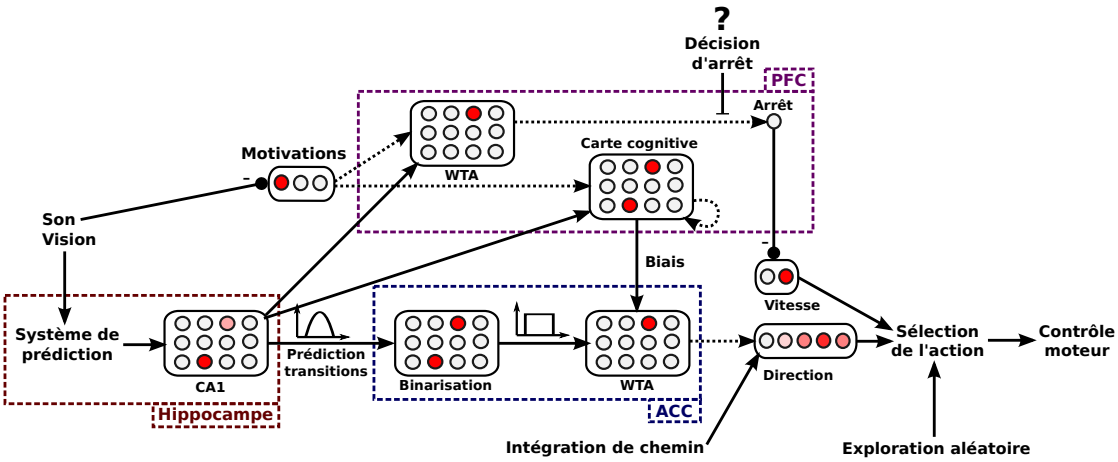


FIGURE 4.5 – Hypothèse de modèle rejetée pour la sélection de l'action menant à l'arrêt sur le lieu but. Le modèle se base sur un contrôle actif du mouvement par le PFC, dépendant d'une compétition sur les activités de prédiction de transitions modulée par la motivation. Il n'est toutefois pas compatible avec les données d'inactivation du PFC montrant une conservation des performances dans la tâche.

par le préfrontal, l'inactivation de celui-ci abolirait le contrôle du mouvement lors de la phase d'attente. Les résultats expérimentaux montrent au contraire une conservation des performances des rats après inactivation du PFC.

Ces observations nous ont donc amené à revoir notre modèle (fig. 4.6). En suivant les mêmes principes de raisonnement que dans la section 4.1.2, nous avons supposé que l'apprentissage approfondi de la tâche avait du mener à l'acquisition de mécanismes sensori-moteurs de bas niveau permettant sa réalisation. Ainsi le PFC serait impliqué dans l'acquisition de la stratégie de contrôle du mouvement mais, une fois celle-ci apprise, ne serait plus indispensable. Nous sommes donc revenus sur la base de notre modèle d'associations transition-action où une correspondance entre une transition et son action en termes de direction pouvait être apprise. Le modèle a alors été étendu pour inclure un contrôle de l'action en termes de vitesse. Cette vitesse serait, au même titre que la direction, codée sur un champ de neurones, représentant différentes vitesses allant de l'immobilité à une certaine vitesse de déplacement maximale. Une transition d'un événement à un autre serait alors associée à une action (vitesse et direction) à effectuer pour atteindre l'événement cible : une transition purement spatiale d'un lieu A à un lieu B serait associée au fait de se déplacer dans une direction donnée, tandis que la transition *Lieu but* $\rightarrow$ *Son* pourrait être associée à une vitesse nulle, puisque le son ne peut être perçu que si le robot s'arrête. Le contrôle du mouvement dépendrait donc directement du résultat de la compétition entre les transitions biaisée par la carte cognitive, et de l'action associée à la transition gagnante.

Supposons que le système ait appris à associer la transition *Lieu but* $\rightarrow$ *Son* à l'action de s'arrêter et que les autres transitions (purement spatiales) soient associées à un mouvement dans leurs directions respectives. Lors des phases où le robot est motivé pour atteindre le but (le son), l'activité est propagée dans la carte cognitive à partir de la transition *Lieu but* $\rightarrow$ *Son* (car elle mène à la satisfaction du but). La remontée des transitions spatiales amène le robot au lieu but, puis le biais de la carte cognitive fait gagner la transition *Lieu but* $\rightarrow$ *Son* qui déclenche l'arrêt du mouvement. Après 2 secondes, le son est entendu et la motivation correspondante est inhibée puisque le but a été satisfait. Il n'y a alors plus d'activité dans la carte cognitive, qui ne biaise plus

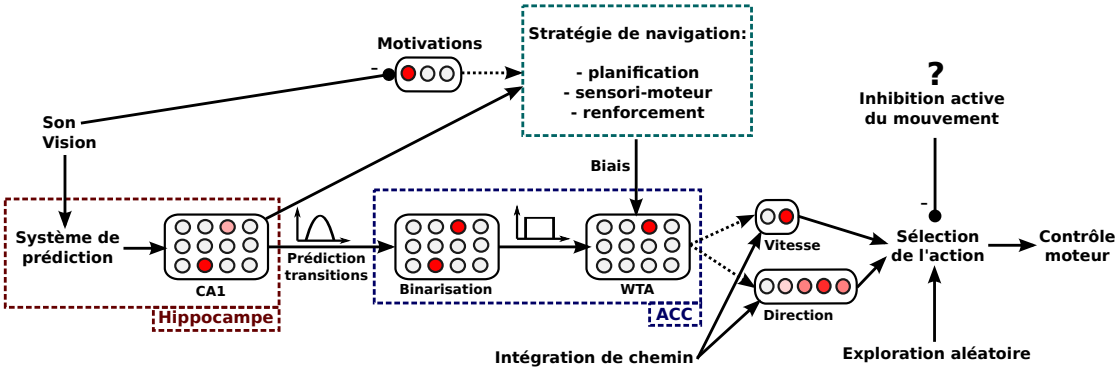


FIGURE 4.6 – Modèle de sélection de l'action liée aux prédictions de transitions fournies par l'hippocampe. La sélection d'une transition est faite par le biais d'une stratégie de navigation utilisant les transitions, et l'action correspondante est effectuée. En l'absence d'action proposée par le système, l'exploration aléatoire sélectionne une action. Un mécanisme, restant à modéliser, contrôlant le choix de l'arrêt peut inhiber le mouvement et mener à l'association entre la prédiction du son et l'immobilité.

la compétition entre les transitions. Aucune transition ne gagne la compétition et plusieurs actions sont proposées. Les activités neuronales correspondant aux actions proposées sont faibles à cause des inhibitions latérales liées à la compétition et à l'absence de vainqueur. Un bruit dans la sélection de l'action peut alors entraîner un comportement d'exploration aléatoire. Ce comportement prend alors le relais, jusqu'à ce que la motivation réapparaisse après un certain temps et que des transitions soient de nouveau sélectionnées. Si le robot avait d'autres buts à satisfaire dans l'environnement, d'autres stratégies de navigation vers ces buts pourraient prendre le relais plutôt qu'une exploration aléatoire. Ce système permet donc de reproduire le comportement du rat lors de la réalisation de la tâche de navigation continue. Une propriété intéressante du système est que le choix de la transition à effectuer se fait par un biais sur la compétition entre les prédictions de transitions. Ce biais peut provenir de la carte cognitive mais aussi de n'importe quel autre architecture de navigation utilisant des transitions (stratégie sensori-motrices, apprentissage par renforcement etc.). *Le modèle est donc indépendant du type de stratégie de navigation utilisée*, du moment que celle-ci sélectionne adéquatement les transitions pour atteindre le but.

Dans cette description du fonctionnement du système, on ne parle que des essais récompensés. En effet, le système apprend ici les caractéristiques temporelles de la tâche, en apprenant les relations temporelles entre les divers événements, mais n'utilise pas cette composante temporelle. C'est l'arrivée du son qui met fin à la phase d'attente, et non sa prédiction. Pour que ce système puisse déclencher la reprise du mouvement même en l'absence du son, il lui faut un mécanisme de détection de l'absence de ses événements prédits. Cette limitation sera discuté dans la section 4.3 et le système correspondant sera détaillé dans le chapitre 6. Nous nous contenterons pour l'instant de réaliser la tâche sans les essais d'extinction.

Enfin, ce fonctionnement suppose que les associations entre transitions et actions ont déjà été apprises a priori. L'association entre une transition et son action en termes de direction, en utilisant une intégration de chemin, a été détaillée dans le chapitre 2. Un mécanisme similaire pourrait être utilisé pour mémoriser les informations proprioceptives concernant la vitesse du robot pendant que la transition était effectuée. Ainsi une transition est associée à la direction et la vitesse qui ont permis de l'effectuer. Cet apprentissage est effectué avec une règle de type

LMS (3.9), où le signal désiré est celui fourni par l'intégration de chemin. Si la stratégie par défaut du robot est de se déplacer, cela laisse en suspens la question de savoir comment la décision d'arrêter le mouvement sur le lieu but a été prise (pour pouvoir être ensuite associée avec la transition *Lieu but* $\rightarrow$ *Son*). Une solution simple et implémentable d'un point de vue robotique, mais peu plausible biologiquement, est de superviser le mouvement du robot et de lui envoyer un signal explicite d'arrêt lorsqu'il se trouve sur le lieu but. Ainsi il entendra le son et sera capable de faire l'association entre sa perception et ses actions. Une méthode un peu moins ad-hoc consisterait à profiter des mécanismes d'évitement d'obstacle de bas niveau du robot et à le faire s'arrêter en se plaçant devant lui. Dans ce cas, l'apprentissage serait fait en interaction avec un humain et éviterait une prise de contrôle directe du robot par l'humain. La question de l'apprentissage autonome de l'arrêt du mouvement, comme cela se passe pour les animaux, est plus complexe. Ce point sera discuté dans la section 4.3 et nous verrons comment le système développé dans le chapitre 6 peut apporter une solution à ce problème.

4.2 Validation expérimentale

4.2.1 Expérience en simulation et activités hors-champ

Le modèle décrit dans la section 4.1 a été implémenté en simulation en utilisant le simulateur de réseaux de neurones *Promethe* (voir annexe B). Cette implémentation reprend donc le modèle du retour du cortex préfrontal sur le cortex entorhinal, permettant l'apprentissage d'associations secondaires dans l'hippocampe et menant à la production d'activités hors-champ. La partie sélection de l'action est également présente, la décision de l'arrêt du mouvement étant apprise par supervision directe de l'expérimentateur. Le robot simulé apprend à reproduire la tâche de navigation continue et les activités dans différentes parties du modèle sont enregistrées.

Les expériences sont conduites dans un environnement ouvert, sans obstacles. 20 amers visuels simulés sont répartis le long des murs et fournissent les informations visuelles utilisées par le système. Durant une phase préliminaire, le robot explore son environnement de façon aléatoire et construit sa carte cognitive. Les cellules de lieu sont apprises de manière autonome avec un recrutement sous un seuil minimal d'activité. L'activité des cellules de lieu après compétition est directement transmise à CA3 (sans le système de détection de changement du gagnant), les neurones de CA3 sont donc des prédicteurs de transitions mais aussi de l'autotransition du lieu vers lui-même (voir chapitre 3). L'hippocampe apprend les transitions entre ces différents lieux et fournit l'information à la carte cognitive qui construit une représentation des chemins possibles dans l'environnement. L'exploration dure suffisamment longtemps pour que le robot acquiert une carte exhaustive des lieux et transitions, ainsi que les actions (c'est-à-dire les directions) qui sont associées à ces transitions.

Un lieu but non indicé est placé dans le coin nord-ouest de l'arène. Un système automatique produit un son simulé lorsqu'il détecte la présence du robot sur le lieu but pendant plus de 2 secondes, de manière similaire au montage expérimental utilisé avec les rats. Lors de l'exploration, la vitesse de déplacement par défaut du robot ne lui permet pas de rester assez longtemps sur le but pour entendre le son. Il est donc incapable d'apprendre la présence d'un lieu but dans l'environnement. Pendant une seconde phase d'apprentissage, l'expérimentateur dirige le robot et le fait s'arrêter sur le lieu but pendant 2 secondes. Après avoir entendu le son, le système apprend à relier temporellement l'entrée sur la cellule de lieu la plus active (et normalement la plus

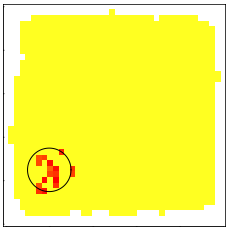


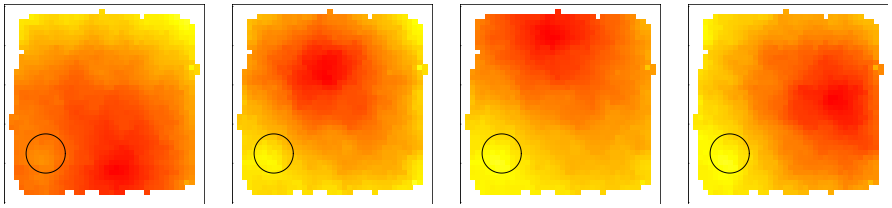
FIGURE 4.7 – Corrélats spatiaux de l'activité d'un neurone entorhinal codant pour la modalité sonore.

proche) et l'arrivée du son. Il apprend également la durée du délai qui sépare ces événements. La transition apprise est associée avec l'action donnée par l'expérimentateur, à savoir rester immobile (ce qui correspond à une commande de vitesse nulle pour le robot). La perception du son ayant été définie comme un but, la satisfaction de ce but déclenche une activation diffuse dans le cortex entorhinal, entraînant alors l'apprentissage de prédictions secondaires dans l'hippocampe. Les diverses cellules de prédiction dans les régions CA3 et CA1 se voient alors activées lors de l'entrée sur le lieu but (en fait lors de l'entrée sur la cellule de lieu la plus proche). Cette activité apparaît comme une activité hors-champ dans l'hippocampe. Les corrélats spatiaux des activités de différentes parties du modèle qui montrent ces champs secondaires (fig. 4.8).

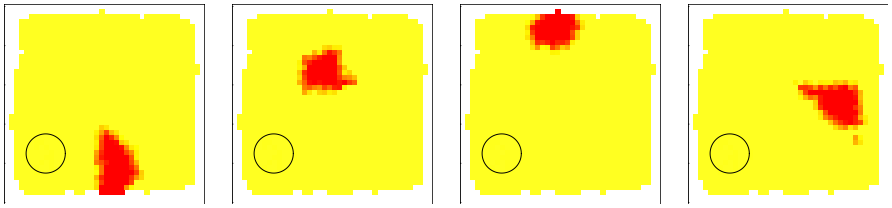
Les cellules de lieu entorhinales, avant compétition, présentent des champs de lieu larges et bruités. L'activité qui résulte de leur compétition est beaucoup plus restreinte. La largeur des champs de lieu après compétition est très dépendante de la densité de répartition des cellules de lieu. Moins il y a de cellules, plus les champs après compétition seront larges. Cette densité est réglée par le seuil de recrutement utilisé lors de l'apprentissage. Le neurone codant pour l'événement représentant la perception du son possède aussi des corrélats spatiaux puisque ce son n'est produit que sur le lieu but (fig. 4.7). Bien qu'il ne traite que la modalité sonore, la visualisation spatiale de son activité pourrait mener à la conclusion erronée que le neurone est une cellule de lieu !

Les cellules granulaires dans DG possèdent des champs de lieu de tailles variables. Les cellules ayant des temps de réponse très courts (qui sont donc excitées peu après l'activation d'une batterie et sur une période de temps courte) possèdent des champs très réduits, avec une activité plutôt concentrée sur le bord des lieux pour lesquels elles gardent une trace temporelle. Les cellules avec des temps de réponse plus longs ont des champs de lieu plus diffus (de la même taille que les cellules de lieu entorhinales après compétition).

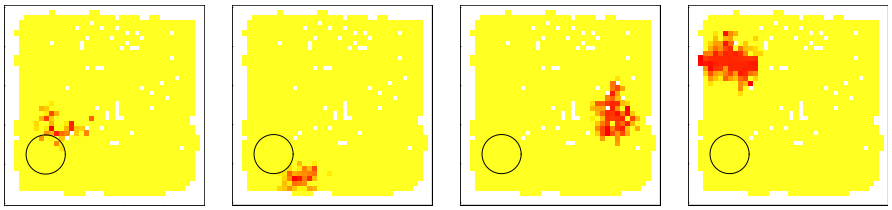
Les cellules pyramidales de CA3 possèdent des champs plus larges que les cellules entorhinales après compétition. Ces neurones prédisent en effet l'arrivée dans un lieu depuis tous les lieux avoisinants, en plus de l'autotransition du lieu vers lui-même. La représentation spatiale de l'activité montre clairement une activité secondaire au niveau du lieu but, même pour les cellules ayant un champ principal loin du but. Cette activité vient de la prédiction de la satisfaction du but lors de la phase d'attente, apprise par les associations secondaires. D'un autre côté, les cellules de CA1 possèdent un champ secondaire moins marqué, à cause de l'activité venant de EC qui remet les prédictions dans leur contexte spatial, et tend à réduire les prédictions secondaires qui sont loin du lieu pour lequel le neurone de transition de CA1 est censé coder. L'activité provenant de CA3 peut suffire à augmenter le potentiel d'un neurone dans CA1, mais la co-activation avec EC est nécessaire pour que ce potentiel dépasse le seuil d'activation. Ainsi l'activité observée pour le champ secondaire est sous le seuil d'activation des neurones et n'est donc pas transmise



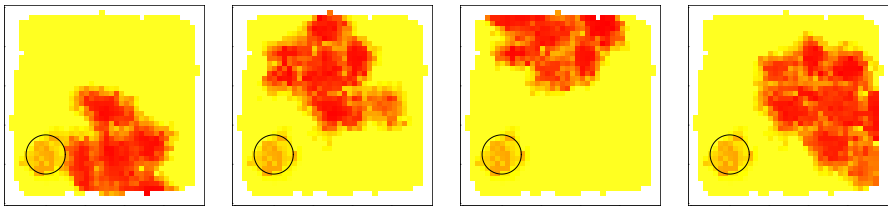
(a) EC (avant compétition) - Cellules de lieu



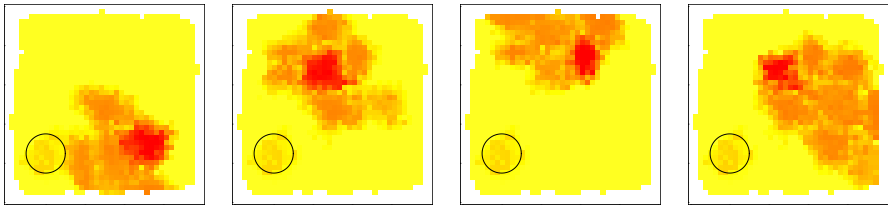
(b) EC (après compétition) - Cellules de lieu



(c) DG - Cellules de trace temporelle



(d) CA3 - Cellules de prédiction d'événement



(e) CA1 - Cellules de transition

FIGURE 4.8 – Corrélats spatiaux des activités de multiples neurones pendant l'expérience, après apprentissage de la tâche. Tous les neurones de CA3 et CA1 possèdent un champ d'activité secondaire, en plus de leur champ principal. Les champs de lieu sont larges par rapport à l'environnement mais cette taille relative dépend de l'environnement simulé et des paramètres du modèle tels que la tolérance aux écarts d'azimut pour les amers. Des champs plus petits pourraient être obtenus en changeant ces paramètres. Le cercle indique la position du lieu but.

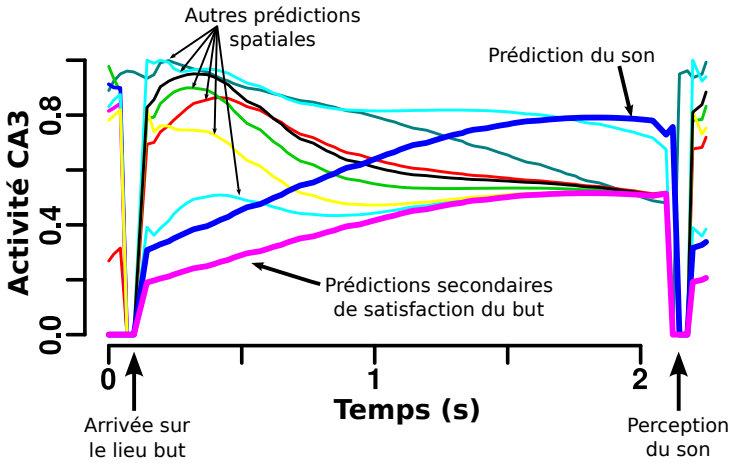


FIGURE 4.9 – Activité de différentes cellules pyramidales de CA3 au niveau du lieu but. L'activité primaire est la prédiction de la perception du son en fonction du temps passé sur le lieu but. Le pic d'activité précède le timing attendu de la perception du son. Les prédictions secondaires de satisfaction du but ont la même forme de prédiction mais avec un niveau d'activité plus bas. D'autres transitions spatiales vers les lieux voisins sont aussi prédites.

à la carte cognitive. On peut aussi observer que leurs champs sont plus étroits que dans CA3. Les transitions prédites par les neurones de CA3 étant sélectionnées dans CA1, le champs d'activité d'un neurones de CA1 est en fait une sous-partie du champ d'activité d'un neurone de CA3.

Enfin, la forme temporelle des activités de prédiction dans CA3 lors de l'attente sur le lieu but ressemble aux observations chez le rat (fig. 4.9). L'activité a une forme de cloche dont le maximum d'activité se situe peu avant la perception du son. L'activité est plus forte pour le neurone codant explicitement pour la prédiction du son, tandis que les autres neurones ayant appris secondairement à prédire la satisfaction du but ont une activité moindre. Leur prédiction principale correspond à d'autres transitions dans l'environnement. On peut aussi observer d'autres prédictions déclenchées par l'entrée sur le lieu but, qui prédisent la possibilité de rejoindre des lieux adjacents. Dans un système biologique, ces prédictions seraient probablement codées sur des populations de neurones.

4.2.2 Expérience sur robot réel et interactions Homme-Machine

L'expérience de navigation continue implémentée en simulation a par la suite été réalisée avec un robot réel. La simulation permet, via une exploration exhaustive de l'environnement et la récupération de nombreuses données, de valider les principes du modèle. L'affichage des corrélats spatiaux des activités, montré dans la section précédente, ne pourrait être aussi détaillé sur robot réel. Les vitesses de simulation permettent en effet d'obtenir des résultats dans lesquels le robot parcourt tout l'environnement de nombreuses fois et permettent d'afficher une activité moyenne en chaque endroit. Les expériences sur robot réel permettent, quant à elles, de valider le fonctionnement dans des conditions réelles avec des signaux bruités. Elles permettent également de vérifier les applications du modèle en robotique mobile.

Dans cette expérience, un robot robulab 10 (voir annexe C) est utilisé, et peut se déplacer dans un environnement de 5x5m. La base carrée du robot fait 45x45cm au sol. Le robot

se déplace à une vitesse maximale de 15cm/s qui peut être réduite par le système de détection d'obstacles. L'environnement visuel du robot correspond à une salle de bureau normale. Des planches de 30cm de hauteur limitent les déplacement du robot à la partie centrale de la salle mais ne gênent pas la vision, la caméra voyant par dessus ces planches. Cette arène est nécessaire car les capteurs ultrason utilisés pour l'évitement d'obstacle ne permettent pas au robot d'éviter les plans de tables qui sont en hauteur et les pieds de bureaux très étroits, non détectés par les capteurs ultrason peu précis situés à 10cm du sol. Plus récemment, l'utilisation d'un télémètre laser, plus précis, a permis de régler le problème de la détection des pieds de bureaux mais pas des plans de tables. Grâce à l'arène, le robot peut explorer son environnement de manière entièrement autonome, sans intervention de l'expérimentateur et sans risque de collision. Deux obstacles sont placés au milieu de l'arène. Un lieu but est marqué par une zone de couleur de 60x80cm sur le sol. Cette zone peut être identifiée par un détecteur de couleur placé sous le robot. Par contre, le système de vision panoramique, situé sur le dessus du robot, ne peut pas repérer la zone de couleur. Ce système de vision capture une nouvelle image toutes les 300ms environ, effectuant donc un panorama à 180°composé de 7 prises de vues en 2s. Lors de l'apprentissage d'une cellule de lieu, des vues sont apprises sur 360°, mais, lors de la navigation, la reconnaissance des cellules de lieu est effectuée uniquement avec des prises de vue sur 180°à l'avant du robot. Les informations visuelles de l'architecture sont donc mises à jour toutes les 300ms. La limitation est d'ordre mécanique et liée à la vitesse du système pan-tilt ainsi qu'au temps de stabilisation nécessaire pour obtenir une image nette. Le temps d'une boucle de simulation est de 100ms pour le simulateur neuronal. La navigation se fait en utilisant la carte cognitive et les informations des amers visuels environnants. La présence sur le lieu but est confirmée par la détection de la couleur une fois que le robot est sur la zone colorée. Au niveau entorhinal, la détection de la couleur est représentée par un événement sensoriel particulier. Dans le cadre de cette tâche avec un but indicé, la détection de la couleur signale donc l'entrée sur le lieu but.

Comme dans l'expérience en simulation, le robot commence par une phase d'exploration qui lui permet d'apprendre des cellules de lieu visuelles, les transitions qui les relient et la carte cognitive formant la topologie de l'environnement (fig. 4.10). Les changements de lieu ainsi que la détection de la zone de couleur déclenchent des événements dans EC. Les transitions apprises dans l'environnement incluent donc des transitions spatiales d'un lieu vers un autre, mais également les transitions des lieux voisins vers la zone de couleur. Les actions associées en termes de direction sont également apprises. Le robot connaît donc après apprentissage tous les chemins menant à n'importe quel lieu, ou à la zone de couleur. Le système automatique de production du son déclenche celui-ci lorsque le robot reste plus de 7s dans la zone de couleur. Ce délai est choisi en fonction de la vitesse par défaut du robot, pour que le comportement de déplacement ne lui permette pas de rester assez longtemps pour entendre ce son. Lors de l'exploration initiale, le son n'est donc jamais entendu, et la zone de couleur n'est pas encore associée à la satisfaction d'un but.

Dans un second temps, on apprend au robot à réaliser la tâche. A la différence de la simulation, on utilise ici l'interaction Homme-Machine pour faire apprendre la phase d'attente. Une laisse attachée à un cou artificiel permet à l'humain de guider le robot vers la zone colorée, ce qui évite d'attendre qu'il passe dessus de manière aléatoire. Une fois sur la zone, l'humain se place devant le robot et lui bloque le passage (le robot cherche en effet à éviter l'obstacle). L'utilisation de l'évitement d'obstacle pour faire arrêter le robot fournit un moyen d'interaction simple et intuitif ne nécessitant pas de dispositif spécifique à cette interaction. On utilise en ef-

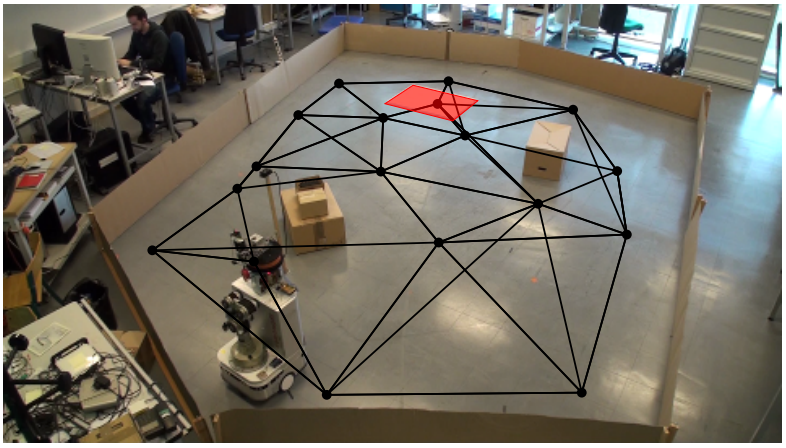


FIGURE 4.10 – Carte cognitive de l'environnement. Les points représentent les centres des cellules de lieu et les traits les transitions entre lieux. La zone de couleur est aussi un noeud de la carte.

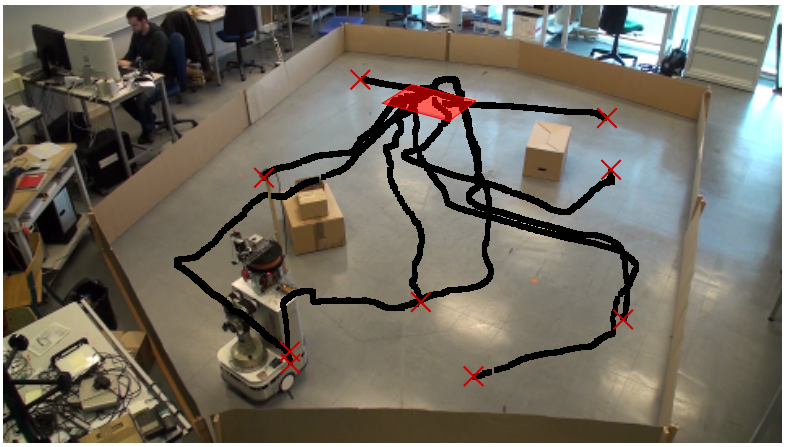


FIGURE 4.11 – Trajectoires obtenues durant les phases de navigation utilisant une carte cognitive pré-apprise pour rejoindre le lieu but (rectangle rouge) depuis divers points de départ (croix). Le but est atteint rapidement. Les chemins ne sont pas optimaux (lignes droites) car le système sélectionne le meilleur chemin en termes de nombre de transitions, pas de distance. La compétition ne sélectionnant qu'une transition, le robot a tendance à suivre les arêtes du graphe de la carte cognitive. Des trajectoires plus fines pourraient être obtenues en utilisant une compétition plus souple [Cuperlier, 2006] ou bien en augmentant la densité de cellules de lieu.

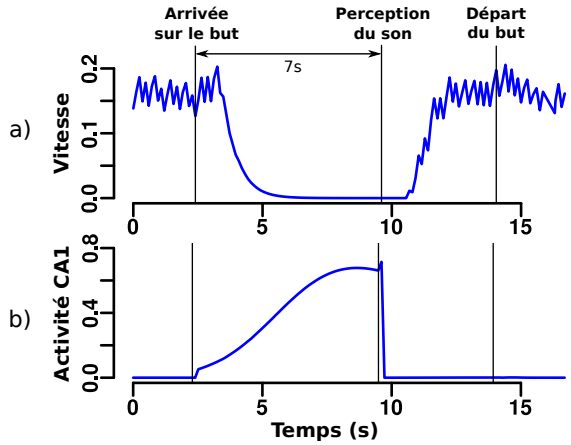


FIGURE 4.12 – a) Vitesse du robot **b)** Activité de prédiction du son au niveau du lieu but. La prédiction, déclenchant l'arrêt du mouvement à l'entrée du lieu but, atteint un pic juste avant la perception du son et s'arrête lorsque le son est émis. Le mouvement reprend alors et le robot quitte le lieu but.

fet les senseurs existants du robot pour interagir. Après 7 secondes d'immobilité sur la zone de couleur, un signal simulant la perception d'un son est envoyé au robot. Cette perception constitue pour le robot la satisfaction d'un but. Le système apprend alors le délai entre l'arrivée sur la zone de couleur et la perception du son. Sa proprioception lui fait alors associer l'action de ne pas bouger à cette transition. La perception du son inhibe la motivation correspondant à ce but, et le robot commence à explorer l'environnement jusqu'à ce que cette motivation remonte suffisamment pour qu'il utilise sa carte cognitive pour rejoindre le lieu but (c'est-à-dire la zone de couleur). Les trajectoires montrent la capacité du robot à rejoindre le lieu depuis n'importe quel endroit en utilisant la carte cognitive (fig. 4.11). Lorsqu'il détecte la couleur, le système prédit la perception du son et sélectionne l'action nécessaire pour atteindre cette prédiction : ne pas bouger (fig. 4.12). Le robot attend alors jusqu'à ce qu'il perçoive le son. Chaque nouvelle présentation du son renforce l'apprentissage de ce comportement.

4.3 Discussion et limitations

Le modèle présenté dans ce chapitre permet de rendre compte des activités observées dans l'hippocampe de rats réalisant la tâche de navigation continue. L'hypothèse de travail de ce modèle est que l'hippocampe apprend les relations temporelles entre des événements perceptifs consécutifs. La représentation suggérée des buts dans le cortex préfrontal nous a amené à prendre en compte un retour du PFC sur le cortex entorhinal menant à des apprentissages d'associations secondaires dans l'hippocampe. Ces associations secondaires prédisent la satisfaction d'un but. Elles seraient à l'origine des activités hors-champ enregistrées dans CA1. Nous postulons donc que le cortex préfrontal est nécessaire pour l'acquisition de ces prédictions et l'apprentissage de la tâche de navigation continue, ou de manière plus générale que le PFC est impliqué dans les tâches mettant en jeu des aspects d'inhibition motrice liée à des facteurs de délais temporels. Même une inactivation post-apprentissage du PFC n'entraîne pas d'effets visibles, son inactivation pre-apprentissage devrait fortement perturber l'apprentissage de la tâche et réduire les performances des rats. En revanche, d'autres stratégies de navigation pourraient par la suite

prendre le relais (voir chapitre 5). Notre modèle prédit également que les activités hors-champ devraient être plus fortes dans CA3 que dans CA1, puisque le contexte spatial provenant de EC devrait limiter les prédictions secondaires liées à la satisfaction du but dans CA1. Les résultats expérimentaux en simulation confirment que ce contexte mène à une réduction de la taille des champs de lieu entre CA3 et CA1.

Ce modèle suppose l'existence d'un retour du PFC sur EC. Il reste cependant à déterminer la raison du retour diffus de l'activité du PFC vers EC. Une explication plausible viendrait de la différence des codes utilisés dans le PFC et EC. Tandis que le cortex préfrontal possède un code représentant des buts et des contextes motivationnels, le cortex entorhinal code pour des événements perceptifs. La relation entre ces événement et le contexte auquel ils sont associés doit être apprise. Nous supposons en conséquence que cette diffusion de l'activité dans EC est un effet de bord de l'apprentissage hebbien de la topologie entre les population neuronale du PFC et EC. Cette diffusion permettrait de renforcer les synapses reliant un contexte motivationnel avec les événements pertinents pour la satisfaction du but associé. Ainsi le contexte motivationnel présent au niveau du PFC pourrait permettre la sélection de sous-ensembles dans EC afin de moduler les activités de prédiction de l'hippocampe en fonction de différents environnement, tâches, buts etc. On sait que les activités hippocampiques peuvent être modulées par un contexte motivationnel correspondant à un besoin de nourriture ou d'eau [Kennedy and Shapiro, 2009], les cellules de l'hippocampe répondant de manière différenciée selon le type de ressource recherchée. Les activités dans CA1 reflètent également le futur choix d'un tournant gauche/droite lors de différents épisodes de navigation dans un labyrinthe (codage prospectif), et marquent aussi une différence entre des épisodes d'exemple et de reproduction [Griffin et al., 2007]. Cette sélectivité dépendante du futur choix gauche/droite dans une tâche d'alternance se retrouve aussi au niveau entorhinal, et la particularité des cellules de EC d'être diffuses et bruitées pourrait tenir au fait qu'elles représentent des contextes [Lipton et al., 2007; Lipton and Eichenbaum, 2008]. Cette vue du cortex entorhinal fournissant un contexte temporel [Eichenbaum and Lipton, 2008] ou spatial [Redish and Touretzky, 1997] a été largement reprise. Dans notre modèle, nous supposons que l'apprentissage de ces contextes, au moins lorsqu'ils concernent des aspect motivationnels liés à des buts, se fait avec l'aide du cortex préfrontal. Ainsi, le contexte motivationnel sélectionnant des sous-ensembles de population dans EC et donc l'hippocampe, cela expliquerait pourquoi les activités hors-champ ne sont observables que pendant les phases de navigation orientées vers le but, et pas pendant les phases d'exploration aléatoire de l'environnement à la recherche de la nourriture. Cela pourrait également être un effet de mesure dû à la faible durée passée par le rat sur le lieu but pendant ces phases, qui ne laisse pas l'activité de prédiction secondaire se former suffisamment longtemps pour apparaître dans les enregistrements. Enfin, on pourrait imaginer que ce mécanisme puisse être utilisé dans de grands environnements, où l'utilisation d'un contexte visuel pourrait permettre de sélectionner des sous-parties locales de la représentation spatiale de l'environnement.

Le modèle présenté dans ce chapitre possède toutefois des limitations et laisse deux questions majeures en suspens. La première concerne la décision de s'arrêter. Le système actuel est capable d'associer une action effectuée avec sa conséquence en terme de transition entre événements (ex : le robot rentre dans le lieu A, va vers le nord et arrive dans le lieu B. La transition *AB* est alors associée avec l'action *mouvement vers le nord*). L'humain peut déclencher un arrêt du mouvement soit de manière explicite, en prenant le contrôle du robot, soit de manière implicite, en interagissant par le biais des senseurs du robot. Le robot découvre alors que la phase d'arrêt

qui lui a été imposée mène à la satisfaction d'un but, et l'intègre à son comportement. Dans le protocole expérimentale utilisé avec les rats, l'apprentissage de la tâche se fait très différemment. La position du lieu but et la phase d'attente sont apprises par *shaping*. Tout d'abord un lieu but plutôt large est défini, et la présence du rat sur ce lieu mène à un lâcher de nourriture sans délai. Le lieu but est ensuite progressivement réduit et le rat affine alors sa représentation spatiale du lieu but. Une fois la taille finale atteinte, le délai d'attente avant le lâcher de la nourriture est très progressivement introduit. Ce délai est augmenté petit à petit jusqu'à atteindre 2 secondes. Il est probable qu'en n'entendant pas la récompense habituelle arriver immédiatement, le rat apprenne progressivement qu'un temps d'attente est nécessaire avant de la recevoir. Le fait de quitter le lieu but de manière précoce en continuant de se déplacer empêche alors la réception de la nourriture, et doit motiver la décision de tenter de s'arrêter sur le lieu but. Pour apprendre de manière autonome ce genre de comportement, le robot doit être pourvu de deux capacités différentes :

1. La capacité à prédire l'arrivée d'un événement précisément dans le temps et à détecter son absence.
2. La capacité à utiliser ce signal de renforcement positif (perception de l'événement prédit) ou négatif (absence de l'événement prédit) pour modifier son comportement.

La deuxième question laissée en suspens est celle de la décision de reprendre le mouvement à la fin du délai d'attente si le son n'est pas perçu. Bien que le système présenté dans ce chapitre apprenne la relation temporelle entre l'entrée sur le lieu but et la perception du son, et soit capable de prédire précisément quand le son doit arriver, il n'utilise pour l'instant pas cette information. La prédiction de l'arrivée du son (indépendamment du timing) déclenche l'arrêt du mouvement, et le son déclenche la reprise en inhibant la motivation qui a mené à la sélection de cette transition au travers de la carte cognitive. De la même manière que pour la question de la décision d'arrêt du mouvement, il manque au modèle 2 composantes : la détection de l'absence d'un événement prédit (ici le son), et la capacité du robot à modifier son comportement en fonction de ces signaux (pour rependre son mouvement). Le système de détection de l'absence d'événements prédit, ou de détection de l'échec si on considère que l'action du robot qui devait mener à cet événement a échoué, sera présenté dans le chapitre 6. Nous nous intéresserons toutefois d'abord aux mécanismes qui permettent au robot d'adapter son comportement en fonction de signaux fournissant des feedbacks positifs ou négatifs sur ses actions dans le chapitre 5. Ce type de système s'inscrit dans le paradigme de l'apprentissage par renforcement. Une telle architecture pourrait non seulement apporter des pistes sur les mécanismes permettant au robot de prendre la décision de s'arrêter ou de repartir, mais pourrait aussi fournir la base d'un système de navigation de plus bas niveau utilisant les transitions. Cette stratégie de navigation pourrait permettre de comprendre comment la réalisation de la tâche peut être transférée, au fur et à mesure de l'apprentissage, de stratégies cognitives de haut niveau vers des mécanismes plus "automatiques" de navigation.

Publications personnelles

Hirel, J., Gaussier, P., Quoy, M., Banquet, J.-P., and Poucet, B. (2010c). Space and time-related firing in a model of hippocampo-cortical interactions. *BMC Neuroscience*, 11(Suppl 1):163

Hirel, J., Gaussier, P., Quoy, M., Banquet, J.-P., and Poucet, B. (2012). The hippocampo-cortical loop: Spatio-temporal learning and goal-oriented planning in navigation. En préparation pour soumission à Neural Computation

Les espèces qui survivent ne sont pas les espèces les plus fortes, ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements.

– Charles Darwin

CHAPITRE 5

Ganglions de la base et apprentissage par renforcement

Dans le chapitre 4, j'ai évoqué le passage progressif de stratégies cognitives de haut niveau à des mécanisme automatiques de plus bas niveau au fur et à mesure de la répétition d'une tâche. Ces mécanismes de bas niveau ne peuvent être appris qu'à travers de nombreuses répétitions de la tâche et des mêmes séquences d'actions menant à un but. Ils permettent alors au fur et à mesure de construire une sorte de réflexe de l'action à réaliser à chaque moment pour atteindre le but de la tâche. On parle alors de stratégie liée à des habitudes (*habits* en anglais). Les architectures d'apprentissage par renforcement rentrent généralement dans cette catégorie. Elles permettent d'associer un répertoire d'action avec des valeurs représentant une espérance des récompenses attendues pour chacune de ces actions. Il se développe donc, à travers les nombreux essais effectués, une estimation de la valeur de chaque action dans différentes circonstances. Dans le cadre de tâches de navigation complexes, où il est nécessaire d'utiliser le système de transitions pour effectuer un choix entre plusieurs chemins, ce type d'apprentissage peut proposer une alternative à la stratégie de carte cognitive. [Daw et al., 2006] suggèrent qu'un système sans-modèle (de l'environnement) basé sur un apprentissage par renforcement et correspondant à des habitudes existerait en parallèle d'un système avec-modèle basé sur des mécanismes de planification et une représentation de l'environnement. Ces deux systèmes, situés respectivement dans les ganglions de la base (BG) et dans le cortex préfrontal (PFC), coopéreraient pour fournir des propositions d'actions. Le choix de l'action à réaliser serait effectué en fonction de l'incertitude des options proposées par chacun des modèles, qui sont capables de s'auto-évaluer. C'est cette vue d'une coexistence de ces deux système, dans les ganglions de la base et le cortex préfrontal, que nous développerons et implémenterons ici pour notre propre modèle.

L'apprentissage par renforcement rentre dans la catégorie de l'apprentissage non supervisé. L'agent effectuant une tâche ne reçoit pas de retour direct à chaque instant sur les actions qu'il doit ou ne doit pas effectuer. Au lieu de cela, une récompense peut lui être attribuée à la fin de la tâche lorsque certaines conditions ont été remplies. Le système doit alors apprendre de manière autonome comment recevoir cette récompense. Une action pertinente qui mène à la réception d'une récompense a pu être suivie d'actions non pertinentes. Le problème, appelé *credit assignment problem*, est alors de discriminer les actions utiles pour recevoir la récompense des actions parasites. La plupart des algorithmes d'apprentissage par renforcement permettent statistiquement de faire cette distinction.

Je commencerai dans ce chapitre par présenter un état de l'art des modèles neuronaux bio-inspirés existants dans les domaines de la prédiction de récompenses, de l'apprentissage par renforcement et de la sélection de l'action. Je montrerai ensuite comment l'architecture de transitions utilisée dans l'équipe se prête particulièrement bien à une implémentation neuronale de l'algorithme de Q-learning. Un premier modèle algorithmique, suivi d'un modèle plus biologiquement plausible, seront présentés. Une expérience en simulation permettra de valider le fonctionnement du modèle dans le cadre de la navigation vers un but. Nous verrons finalement comment ce modèle est complémentaire avec celui de la carte cognitive, et comment ils peuvent être utilisés en parallèle dans un cadre de coopération. Des expériences avec lésions des structures cérébrales gérant ces différentes stratégies permettront de mettre en évidence leur rôle dans l'acquisition et la réalisation de tâches de navigation vers un but.

5.1 Etat de l'art des modèles neuronaux d'apprentissage par renforcement

Il existe de nombreux algorithmes d'apprentissage par renforcement. Les plus connus sont probablement le TD-learning [Sutton, 1988], pour *Temporal Difference learning*, et une de ses variantes, le Q-learning [Watkins and Dayan, 1992]. [Kaelbling et al., 1996] passent en revue un bon nombre de ces algorithmes, sans s'intéresser à une implémentation neuronale, et comparent leurs performances dans une tâche de navigation dans un monde "grille". De nombreux modèles neuronaux bio-inspirés utilisent l'apprentissage par renforcement pour la résolution de tâches diverses. Ces modèles impliquent généralement les ganglions de la base qui semblent être le siège du circuit de traitement des récompenses dans le cerveau du rat (voir section 1.3.2). Une description de l'anatomie et des connexions internes et externes des ganglions de la base a été faite dans la section 1.3.1. Pour une courte discussion des différents types de modèles impliquant les ganglions de la base, et de leurs caractéristiques communes et différences, se référer à [Beiser et al., 1997] et [DeLong and Wichmann, 2009] pour une discussion des nouvelles directions prises par ces modèles suite aux découvertes de cette dernière décennie.

5.1.1 Conditionnement et prédiction de récompense

Les premiers modèles visant à étudier le système de récompense chez l'animal se concentraient plutôt sur un aspect de conditionnement. L'étude se portait alors sur les mécanismes permettant au cerveau de prédire l'arrivée d'un stimulus, à la suite d'un premier stimulus. Ainsi, [Sutton and Barto, 1981] présentent un modèle de conditionnement classique basé sur un apprentissage avec une règle de Widrow-Hoff. Ce mécanisme de prédiction de l'arrivée d'un stimulus formera la base de leur futur modèle de TD-learning. De même, [Grossberg and Schmajuk, 1987; Grossberg et al., 1987] développent un modèle expliquant une variété de mécanismes de conditionnement et utilisant des signaux de renforcement. Ils utilisent le réseau ART (*Adaptive Resonance Theory*) pour apprendre une relation entre des signaux conditionnels et des représentations de besoins. L'aspect temporel fin des prédictions est modélisé par la suite de manière biologiquement plausible avec l'introduction du système de décomposition spectrale du temps [Grossberg and Schmajuk, 1989]. Un modèle de l'hippocampe avec les voies EC-DG-CA explique alors la capacité du système à traiter des récompenses distantes dans le temps [Grossberg and Merrill, 1992].

La découverte des activités de prédiction de récompense dans le circuit dopaminergique lié aux ganglions de la base [Schultz et al., 1993; Schultz, 1998] a orienté ces travaux de modélisation des conditionnements dans une nouvelle direction. En effet, ces activités présentent les caractéristiques d'une erreur de prédiction sur l'arrivée de récompense : une récompense inattendue provoque une excitation de ces neurones, tandis que l'absence d'une récompense attendue provoque leur inhibition. Le modèle de [Grossberg and Merrill, 1992] est alors adapté pour rendre compte de ces activités dans le circuit dopaminergique, et propose des hypothèses détaillées quant à l'origine de ces signaux [Brown et al., 1999]. D'autres modèles de prédiction de récompense dans les ganglions de la base sont développés en parallèle [Montague et al., 1996]. [Schultz et al., 1997] développent également leur propre modèle, en utilisant une modélisation simple du temps, qui est découpé en courtes périodes ayant chacune une représentation neuronale. Il finissent par intégrer également le modèle de décomposition spectrale, fournissant une alternative plus biologiquement plausible au système de découpage du temps utilisé auparavant [Contreras-Vidal and Schultz, 1999]. Tous ces modèles se contentent cependant de prédictions stimulus-récompense allant parfois jusqu'à modéliser des conditionnements secondaires mais ne traitent pas le *credit assignment problem* en propageant cette prédiction de récompense dans la chaîne des actions précédentes. C'est là qu'interviennent les modèles d'apprentissage par renforcement, et notamment celui de TD-learning, qui vont permettre de propager ce signal afin d'associer de longues séquences d'actions avec des prédictions de récompense.

5.1.2 TD-learning et modèles acteur-critique

Dans l'apprentissage par renforcement, l'environnement est généralement décrit comme un processus de décision markovien (MDP). L'agent passe d'un état à un autre à travers le choix d'une action. Le nombre d'actions est fini. Après avoir effectué une action, l'agent change d'état et reçoit éventuellement une récompense. L'algorithme de TD-learning essaie d'attribuer une valeur à chacun de ces états [Sutton, 1988] afin de maximiser la somme des récompenses attendues (5.1). Pour ce faire, la valeur de chaque récompense obtenue, en plus de la valeur de l'état courant, est utilisée pour mettre à jour la valeur de l'état précédent (5.2). Ainsi, au fil des actions effectuées, on peut rétropropager dans l'ensemble des états les valeurs de récompenses attendues dans le futur, permettant ainsi de construire pour chaque état un estimateur de la somme de ces récompenses. Les équations correspondantes sont les suivantes :

$$V_e(t) = \sum_{i=0}^{\infty} \gamma^i \cdot r(t+i) \tag{5.1}$$

$$\hat{V}_e(t+1) = \hat{V}_e(t) + \alpha(r(t+1) + \gamma \cdot \hat{V}_{e'}(t+1) - \hat{V}_e(t)) \tag{5.2}$$

où V_e est la somme des récompenses attendues dans le futur pour l'état e, γ un facteur de réduction permettant de moins prendre en compte les récompenses lointaines, $r(t)$ la valeur de la récompense obtenue à l'instant t, $\hat{V}$ un estimateur appris de la fonction de valeur, α la vitesse d'apprentissage et e' le nouvel état du système après exécution de l'action.

Ici, seul le dernier état est mis à jour lors de la réception d'une récompense, on appelle ce système le TD(0). Dans le TD(λ), les derniers états sont tous mis à jour avec un facteur de réduction λ diminuant l'importance des plus anciens états dans la réception de la récompense. Cela permet d'accélérer l'apprentissage et de résoudre plus rapidement le *credit assignment problem*. [Tesauro, 1992] fait une revue des capacités de l'algorithme de TD-learning. Dans une volonté

de vérifier son application possible à des problèmes complexes réels, il est adapté pour pouvoir apprendre à jouer au backgammon. Les résultats montrent que non seulement le programme arrive à atteindre un bon niveau intermédiaire de joueur de backgammon, mais qu'en plus il surpasse d'autres programmes conventionnels d'intelligence artificielle pour le backgammon.

Le Q-learning [Watkins and Dayan, 1992] est une variante de l'algorithme de TD-learning attribuant une valeur à des couples état-action plutôt qu'à l'état lui même (voir section 5.2.1 pour plus de détails sur son fonctionnement). [Prescott and Mayhew, 1992] utilisent un apprentissage par renforcement inspiré de cet algorithme pour donner un comportement d'évitement d'obstacles à un robot. Cet algorithme est utilisé pour renforcement les actions lui permettant d'éviter des collisions, en fonction de ses perceptions. [Fagg et al., 1994] proposent aussi un modèle neuronal implémentant le TD-learning pour l'apprentissage par un robot d'une une tâche d'exploration et d'évitement d'obstacles, en utilisant des capteurs sonar et un pare-chocs comme signaux d'entrée. L'implémentation neuronale la plus connue restera celle de [Barto, 1995] et leur modèle acteur-critique. Dans leur modèle, qui utilise le formalisme des réseaux de neurones mais ne fait pas de lien avec des structures cérébrales, l'apprentissage est séparé dans deux modules : l'acteur et le critique. Dans ce cadre, le rôle du critique est d'apprendre l'estimation $\hat{V}$ associée à chaque état du système d'après la règle (5.2). Le rôle de l'acteur est alors de sélectionner la meilleure action à effectuer en fonction de l'état courant. Le système apprend en fait à associer une action à chaque état en mettant à jour la valeur liée à une action après l'avoir effectuée. Cette mise à jour est faite en fonction de la valeur du nouvel état fournie par le critique.

Le critique met à jour ses estimations en calculant une erreur de prédiction sur la valeur de récompense attendue. La similarité entre ce fonctionnement et les activités enregistrées dans le système dopaminergique lié aux ganglions de la base laisse penser que les neurones prédicteurs de récompense observés dans l'aire tegmentale ventrale et la substance noire compacte pourraient servir à un tel calcul. Cette réflexion a donné le jour à de nombreux modèles bio-inspirés de l'apprentissage par renforcement ayant les ganglions de la base comme substrat neuronal. Dans un papier parallèle à celui présentant l'acteur-critique, [Houk et al., 1995] présentent une implémentation neuronale du modèle s'inspirant du circuit interne des ganglions de la base. Les signaux corticaux d'entrée du système arrivent dans le striatum. Depuis le striatum, les voies directes et indirectes (à travers le noyau sous-thalamique) permettent le calcul de l'erreur de prédiction dans la substance noire compacte. L'apprentissage est fait pour les précédents signaux actifs selon le principe du TD(λ). Au niveau neuronal, cela implique que les connexions synaptiques soient modifiées en fonction de l'activité post-synaptique pour des neurones présynaptiques qui ont été activés dans le passé. Ce mécanisme, appelé *trace d'éligibilité*, implique un apprentissage différé et est justifié par les mécanismes neuronaux impliquant des réactions chimiques successives introduisant des délais dans les modifications synaptiques. Le rôle de la protéine *CaM PK II* est mis en avant. Une simulation utilisant des signaux artificiels permet de vérifier les activités des différentes parties du système.

[Doya, 2000] propose également un modèle neuronal (non implémenté) de TD-learning dans les ganglions de la base. Il discute le rôle de cette structure dans l'apprentissage par renforcement (basé sur les récompenses), par opposition au cervelet qui serait impliqué dans l'apprentissage supervisé (basé sur un calcul d'erreur). [Joel et al., 2002] passent en revue les différents modèles acteur-critique de BG existants et discutent les limitations en terme de plausibilité neurobiologique (notamment en ce qui concerne la trace d'éligibilité et l'utilisation des voies indirectes et directes). Ils proposent un modèle du critique utilisant une approche évolutionnaire et adaptent

le modèle acteur-critique via l'utilisation de voies physiologiques thalamocorticales et des ganglions de la base pour le rendre plus plausible.

Ces modèles ont mené à de nombreuses implémentations en robotique, pour résoudre des tâches très diverses. La nécessité de catégoriser un environnement continu comme une série d'états peut être une difficulté. Dans un cadre de modélisation neuronale biologiquement plausible, les cellules de lieu fournissent le cadre idéal pour catégoriser une position spatiale en terme d'activité neuronale. Ainsi un état peut correspondre à une cellule de lieu gagnante ou une configuration de cellules de lieu. [Arleo and Gerstner, 2000] utilisent un modèle de cellules de lieu créées grâce à des informations visuelles et proprioceptive pour définir l'environnement comme un espace d'états. Dans leur expérience, le modèle neuronal est implémenté en simulation et utilise un algorithme de type Q-learning pour réaliser une tâche de navigation vers un but dans un environnement ouvert avec quelques obstacles. [Nakahara et al., 2001] utilisent le TD-learning pour apprendre des séquences d'actions visuo-motrices, via l'utilisation de boucles visuelles et motrices en parallèle. A l'aide d'un modèle acteur-critique, [Mannella and Baldassarre, 2007] reproduisent le comportement de poules recherchant de la nourriture dans un environnement ouvert en basant leur système sur des informations visuelles de perception de bords et hauteurs de murs. Enfin, un modèle de TD-learning utilisant des neurones à décharges est également utilisé par [Okatan, 2009] pour réaliser des tâches occulo-motrices. Le modèle réalise des associations entre des entrées sensorielles pertinentes pour la tâche et des prédiction de récompense dans l'hippocampe.

5.1.3 Sélection de l'action et stratégies multiples

La navigation spatiale peut impliquer une variété de stratégies différentes : un simple "Homing" où le robot navigue vers un amer visuel, des associations sensori-motrices entre une perception spatiale et une action, de la planification via l'utilisation d'une carte cognitive, l'optimisation des récompenses reçues à l'aide d'un apprentissage par renforcement, ou bien encore l'apprentissage d'une séquence d'actions motrices. Chez l'animal, toutes ces stratégies sont existantes et sont en coopération ou compétition permanente. La difficulté dans les modèles bio-inspirés à l'heure actuelle est de savoir comment ces stratégies, qui font appel à des circuits parallèles, cohabitent et comment une stratégie plutôt qu'une autre peut être amenée à être sélectionnée. [Arleo and Rondi-Reig, 2007] discutent de ces aspect multi-stratégies, en se focalisant sur les mécanismes intervenant dans les changements de stratégies qui peuvent aussi bien intervenir au cours de l'apprentissage que pendant un épisode de navigation.

[Guazzelli et al., 1998] ont conçu un modèle qui, bien que faiblement biologiquement plausible, utilise deux stratégies de navigation : une première stratégie égocentrique de préférence du robot pour certains angles de rotation et une deuxième stratégie de carte cognitive constituée de noeuds et arêtes représentant l'environnement. Les actions suggérées par ces diverses stratégies, codées sur des champs de neurones indiquant une direction à prendre, sont ensuite sommées. Aucun contrôleur gérant la sélection de l'action n'est présent, ce qui rend le modèle particulièrement sensible à la pondération correspondant à chacune des stratégies. Plus tard, [Foster et al., 2000] utilisent un système d'apprentissage par renforcement avec des activités de cellules de lieu simulées (par une fonction gaussienne de la position du robot simulé) dans deux expérience de navigation vers un but : une où le but est fixe et une où le but change régulièrement de place. En parallèle, un système apprend à associer des activités de cellules de lieu avec des coordonnées

dans l'espace. Il est aussi capable de mémoriser les coordonnées du lieu but. Un module, sans justification biologique donnée, permet alors de calculer un vecteur de direction en utilisant la différence entre les coordonnées du robot et du but. Le système d'apprentissage par renforcement seul permet de résoudre la tâche mais montre une adaptation très lente à des changements de lieu but, à l'opposé de ce qui est observé chez les rats. L'addition du second système permet, une fois les associations cellule de lieu/coordonnées bien apprises, de naviguer rapidement vers un nouveau lieu but. La sélection de l'action d'un système plutôt qu'un autre est faite par une simple addition d'activité sur les neurones d'actions avant compétition. Ce modèle rejoint donc une vue de deux systèmes parallèles, l'un permettant un apprentissage et une adaptation rapide, tandis que l'autre développe lentement des actions réflexes.

[Dollé et al., 2010] visent à modéliser le même paradigme de coopération entre stratégies. Dans leur modèle, une stratégie de taxon utilisant un modèle de Q-learning pour associer la perception d'amers visuels à des actions coexiste avec une stratégie de carte cognitive utilisant des cellules de lieu. Un renforcement permet d'apprendre la stratégie la plus efficace selon le contexte perceptif, en comparant les actions proposées par chaque stratégie avec l'action ayant permis de recevoir une récompense. Une expérience simulée avec un lieu but est conduite avec des agents possédant les deux stratégies ou uniquement celle de taxon. Les agents mono-stratégie montrent des performances meilleures lors du changement de position du but mais moins bonnes à long terme, après répétition de la tâche avec la nouvelle position du but. La représentation spatiale de la carte cognitive étant très dense, cette stratégie est utilisée pour la phase finale d'approche précise tandis que la stratégie de taxon est utilisée loin du but, pour donner une idée de la direction à prendre. Cependant les activités de cellules de lieu sont simulées comme une fonction gaussienne de la position du rat et leur centres sont prédéterminés, et aucun modèle biologiquement plausible de la carte cognitive n'est donné. De plus, on pourrait envisager une vue opposée où la carte cognitive posséderait une représentation spatiale grossière et interviendrait dans la navigation éloignée du but, tandis qu'à proximité les amers visuels proches serviraient de cadre de référence pour une navigation précise. On peut donc voir l'intérêt d'un arbitrage dynamique entre différentes stratégies de navigation.

[Doya et al., 2002] donne un formalisme mathématique pour la création d'un système d'apprentissage utilisant plusieurs modules de navigation, et apprenant à leur attribuer un signal de responsabilité. Ainsi la sélection de l'action entre les différentes propositions des modules serait apprise par renforcement. Dans [Khamassi et al., 2005], ce modèle (multiple acteurs et critiques) est implémenté, en comparaison avec des modèles plus simples d'acteur-critique (unique acteur-critique, unique acteur et multiples critiques). Une tâche de navigation réalisée en simulation permet de montrer que le modèle simple d'acteur-critique est incapable de résoudre certains aspects de la tâche. Cependant, la représentation en termes d'états utilisée dans cette expérience est très réduite (perception de la couleur et distance des murs dans un labyrinthe).

Si ces modèles traitent la problématique de la sélection d'une action parmi une variété de choix avec l'apprentissage par renforcement, d'autres modèles des ganglions de la base se focalisent sur les mécanismes de compétition qui permettraient d'effectuer cette sélection de l'action [Girard, 2003]. Ainsi [Gurney et al., 2001b,a] proposent un modèle de sélection de l'action basé sur les propriétés internes du circuit de BG. Ce modèle utilise des neurones intégrateurs à fuite et inclut des propriétés de persistance comportementale et de stabilité que ne possèdent pas d'autres mécanismes de compétition comme le WTA. Une simulation dans un monde grille permet de montrer des performances supérieures à celle obtenues avec un WTA [Girard et al.,

2003]. Dans cette tâche, le robot possède à chaque instant le choix entre un ensemble défini d'actions (consommer de la nourriture, explorer, éviter un obstacle etc.) et évolue dans un environnement où de la nourriture est éparpillée. Le modèle est plus tard adapté dans une tâche avec un vrai robot, qui possède une pince lui permettant de ramasser des cylindres représentant de la nourriture [Prescott et al., 2006]. Les données sont comparées avec des données neurophysiologiques et comportementales obtenus chez les animaux. Enfin, un autre modèle suggère que la sélection de l'action pourrait être faite au travers d'une réduction de la dimensionnalité des signaux entre l'entrée et la sortie de BG [Shah and Alexandre, 2011]. Ce mécanisme fonctionnerait de manière similaire à une analyse en composantes principales (ACP) et utiliserait de manière biologiquement plausible des champs de neurones dynamiques [Amari, 1977].

5.2 Modèle de Q-learning utilisant les neurones de transition

5.2.1 Implémentation neuronale

Nous avons vu dans la section 5.1.2 que l'algorithme de TD-learning se base sur la prédiction de l'espérance de la somme des futures récompenses. Cette valeur est apprise comme une fonction V_e des états du système. Ainsi, pour sélectionner une action parmi un éventail de possibilités, il faut connaître les états accessibles et sélectionner l'action permettant de rejoindre celui qui a la valeur $V(e)$ la plus haute. Une variante du TD-learning, le Q-learning [Watkins and Dayan, 1992], prédit cette même somme des futures récompenses mais comme une fonction $Q(e, a)$ d'un couple état/action (fig. 5.1). A partir d'un état actuel e et, en connaissant toutes les actions qu'il est possible de réaliser, on peut alors sélectionner l'action a correspondant à la valeur $Q(e, a)$ la plus haute. L'équation d'apprentissage des valeurs de Q, inspirée de (5.2), est la suivante :

$$Q_{e,a_1}(t+1) = (1-\alpha) \cdot Q_{e,a_1}(t) + \alpha \cdot (r(t+1) + \gamma \cdot \max_a Q_{e',a}(t+1)) \tag{5.3}$$

$$\text{ou } Q_{e,a_1}(t+1) = Q_{e,a_1}(t) + \alpha \cdot \delta(t+1) \tag{5.4}$$

$$\text{avec } \delta(t+1) = r(t+1) + \gamma \cdot \max_a Q_{e',a}(t+1) - Q_{e,a_1}(t) \tag{5.5}$$

où a_1 est l'action venant d'être effectuée pour passer de l'état e à e', $\delta(t+1)$ est le terme d'erreur d'estimation de Q, α la vitesse d'apprentissage, r la valeur de la récompense obtenue suite à l'action a_1 et γ un facteur de réduction permettant de moins prendre en compte les récompenses lointaines.

Certaines versions du modèle acteur-critique, qui est une implémentation neuronale de l'algorithme de TD-learning, utilisent les cellules de lieu comme représentation des états du robot [Arleo and Gerstner, 2000; Foster et al., 2000]. Ces neurones sont parfaitement adaptés pour jouer ce rôle de catégorisation de la position spatiale du robot. Dans notre modèle, les neurones de transition représentent un changement d'état (lorsqu'un événement correspondant à ce changement d'état est produit dans EC) et sont associés à l'action nécessaire pour effectuer ce changement. Il existe donc une analogie très forte entre les couples état/action utilisés dans le Q-learning et les transitions de notre modèle. C'est donc assez naturellement que j'ai conçu une implémentation neuronale du Q-learning utilisant les neurones de transitions.

On considère ici que la satisfaction d'un but est toujours liée à un événement dans EC (perception d'un son, présence d'une zone de couleur sous le robot, flash lumineux etc.). C'est cette

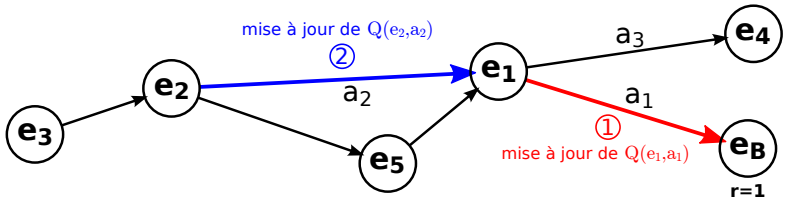


FIGURE 5.1 – Exemple d'apprentissage des Q-values dans l'algorithme de Q-learning. Supposons qu'un robot parte d'un état e_1 et arrive sur un état but e_B en effectuant une action a_1. A l'arrivée dans e_B, il reçoit une récompense $r = 1$ car il a atteint un but. La valeur de $Q(e_1, a_1)$ va alors être mise à jour grâce au terme $r(t+1)$ qui vaut 1. Le robot repart ensuite d'un autre point de départ, passe par une série d'états, et finit par passer d'un état e_2 à l'état e_1 précédemment rencontré en effectuant l'action a_2. Le terme de récompense r vaut 0 car l'état e_1 n'est pas un but. Cependant, $Q(e_1, a_1)$ est différent de 0 suite au premier apprentissage et donc $\max_a Q_{e_1,a}$ aussi. La valeur de $Q(e_2, a_2)$ va donc être mise à jour à son tour en intégrant un facteur de réduction γ. Au fur et à mesure des séquences d'états et d'actions réalisées pour rejoindre le but, les valeurs de Q vont être mises à jour pour refléter un espoir de récompense pour chaque couple état/action.

satisfaction du but qui déclenchera le signal de "récompense" $r(t)$ utilisé dans l'apprentissage par renforcement. Selon un mécanisme décrit par l'équation (5.3), cette valeur de récompense est retropropagée lorsqu'une action est faite, donc lorsqu'une transition est effectuée dans notre modèle. L'implémentation neuronale (fig. 5.2) du Q-learning apprend directement les valeurs de Q dans des poids synaptiques dont les neurones pré-synaptiques sont les neurones de transition. Ces valeurs sont ensuite utilisées pour biaiser le choix d'une transition prédite à l'aide d'une compétition de type WTA, comme cela est fait avec la carte cognitive. Les équations d'apprentissage des connexions synaptiques W^{TQ} correspondant aux valeurs de Q sont les suivantes :

$$\frac{dW_{ij}^{TQ}(t)}{dt} = \alpha \cdot T(t-dt) \cdot e_j(t) \cdot X^{\delta}(t) \cdot W^{\delta Q}(t) \tag{5.6}$$

$$\text{avec } e_j(t) = \begin{cases} \max(X_j^T, \lambda \cdot e_j(t-dt)) & \text{si } T(t) = 1 \\ e_j(t-dt) & \text{sinon} \end{cases} \tag{5.7}$$

α est la vitesse d'apprentissage, $T(t-dt)$ le signal indiquant qu'une transition vient d'être effectuée (l'apprentissage est retardé pour pouvoir utiliser les nouvelles prédictions de transitions liées à l'événement perçu), $e_j(t)$ la trace d'éligibilité d'une transition et λ le terme du TD(λ) permettant de calculer la trace d'éligibilité en mémorisant les transitions précédemment effectuées et en diminuant leur importance progressivement. X^{δ} est l'erreur d'estimation calculée de manière neuronale et le poids $W^{\delta Q}$ est fixé à 1.
L'équation de l'activité correspondant aux Q-values pour chaque transition est la suivante :

$$X_i^Q(t) = \sum_k X_k^T(t) \cdot W_{ik}^{TQ}(t) \tag{5.8}$$

Le modèle travaille avec des prédictions de transitions binarisées (donc non temporelles) et leur associe une valeur Q. Le terme d'erreur utilisé dans l'équation du Q-learning (5.3) est calculé en réalisant une compétition sur les valeurs de Q correspondant aux transitions prédites et en prenant la valeur du gagnant (ce qui revient à faire un $\max$). Puis la différence entre l'ancienne valeur de Q prédite pendant la réalisation de la transition et le gagnant de la compétition,

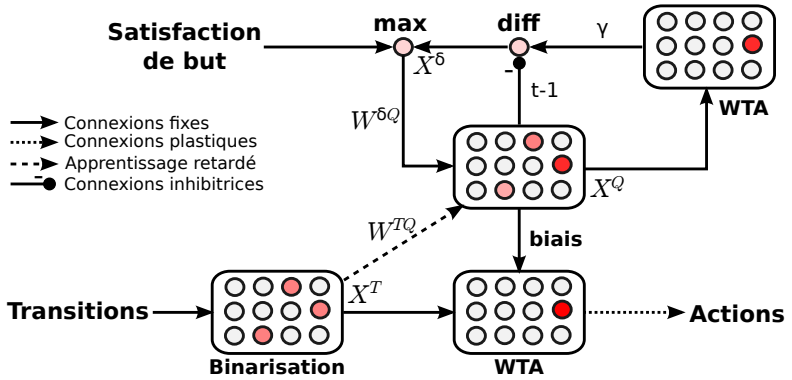


FIGURE 5.2 – Implémentation neuronale du Q-learning. L'erreur de prédiction est calculée par une boucle neuronale et réutilisée pour mettre à jour les poids synaptiques. Les activités résultantes servent à biaiser le choix d'une transition.

réduit d'un facteur γ, est faite. Un second $\max$ est ensuite réalisé entre le signal de récompense immédiat $r(t+1)$ et cette erreur de prédiction $\gamma \cdot \max_a Q_{e',a}(t+1)) - Q_{e,a_1}(t)$. Cette dernière utilisation d'un second $\max$ est une modification de l'équation originale du Q-learning qui utilise une somme. Le terme d'erreur est maintenant :

$$\delta(t+1) = \max(r(t+1), \gamma \cdot \max_a Q_{e',a}(t+1)) - Q_{e,a_1}(t)) \tag{5.9}$$

Cette modification fait suite aux observations faites lors des premiers tests expérimentaux en simulation. Lors de ces expériences, le robot navigue continuellement, rejoignant le but lorsque sa motivation est active, explorant l'environnement le reste du temps. Dans le cas d'une exploration complète de l'environnement, toutes les cellules de transition entre les différents lieux sont apprises. On apprend donc aussi bien les transitions menant au lieu but, que les transitions le quittant. Lors de la mise à jour de la valeur de prédiction de Q pour les transitions menant au but, le signal $r(t)$, donnant un renforcement pour avoir satisfait le but, va s'ajouter aux prédiction de récompenses futures. Or, le simple fait de quitter le lieu but pour y revenir immédiatement représente une récompense future. Ce rebouclage des activités de prédiction entraîne une dynamique de convergence de l'algorithme très lente. En effet chaque nouvelle satisfaction du but entraîne de manière récursive une mise à jour des transitions quittant le but et y revenant. Ce problème de temps de convergence n'est généralement pas abordé dans la littérature. Dans la plupart des expériences conduites, un essai s'arrête lorsque le robot atteint son but, et il est ensuite placé à un autre endroit d'où il doit de nouveau rejoindre le but. Cet aspect de navigation continue, où l'espace des états et actions est entièrement connecté et exploré, est alors caché par le protocole expérimental. En remplaçant la somme par une fonction $\max$, la convergence se fait beaucoup plus rapidement, puisqu'on élimine cet aspect de mise à jour récursive. Le comportement du robot n'est cependant plus le même. De la même manière qu'avec la carte cognitive, le robot va alors se diriger vers un unique but, considéré comme le plus intéressant (en termes de valeur et distance). L'équation originale, quant à elle, favorisait le choix d'un chemin maximisant les satisfactions de buts et pouvant faire un détour pour satisfaire un but minoritaire, sur le chemin d'un but prioritaire (fig. 5.3).

Un avantage de l'approche des transitions par rapport aux modèles traditionnels acteur-critique est qu'elle permet de séparer l'aspect de l'apprentissage par renforcement de l'aspect

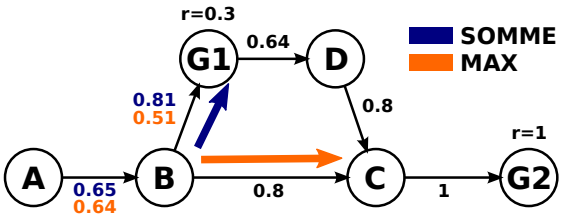


FIGURE 5.3 – Différence de comportement dans l'utilisation d'un opérateur somme ou max pour le calcul d'erreur de prédiction de Q. Dans le cas SOMME, le robot fera un petit détour pour satisfaire un but minoritaire G1 ($r = 0.3$) tandis que dans le cas MAX il se dirigera vers le but majoritaire G2 ($r = 1$) en utilisant le chemin le plus court. Paramètres : $\gamma = 0.8$

du codage des actions. En effet, une transition est sélectionnée et peut être associée à une action plus ou moins complexe, dont le codage est libre. De plus, les transitions non sélectionnées par le WTA sont aussi prédites, avec leurs valeurs de Q correspondantes et constituent donc des alternatives possibles pour la sélection d'une action. Dans l'implémentation neuronale d'un Q-learning réalisée par [Arleo and Gerstner, 2000] ou [Dollé et al., 2010], les valeurs de Q sont directement apprises sur les connexions entre cellules de lieu ou amers visuels et actions. Cependant, leur modèle suppose un ensemble prédéfini d'actions (Sud, Ouest, Nord, Est), chaque action étant représentée par un neurone. Ce type de modèle ne peut fonctionner que lorsque les représentations neuronales des actions sont parfaitement orthogonales, sans quoi des interférences apparaîtraient entre les différentes actions possibles depuis un même état. Dans notre modèle, les actions sont codées sous formes de directions préférées dans un champ de neurones dynamique [Amari, 1977]. Chaque neurone du champ dynamique (représentant les 360°) correspond donc à une direction, mais plusieurs d'entre eux peuvent être activés simultanément par une même transition avec des niveaux d'activité différents pour signaler des préférences de direction. Les associations entre les transitions et les directions préférées qui leur sont associées sont apprises par les poids synaptiques reliant les neurones de transition au champ dynamique, de manière complètement indépendante pour chaque transition. De plus, dans notre modèle, l'apprentissage des associations entre transitions et actions se fait de manière latente, indépendemment de l'apprentissage par renforcement et peut alors être affiné bien avant la découverte d'un but. L'apprentissage par renforcement permet ensuite d'effectuer un choix entre des actions représentant réellement la topologie de l'environnement et reposant sur une exploration, et pas un ensemble d'actions prédéterminées. Enfin, l'utilisation du Q-learning qui travaille directement avec les couples état-action permet de s'affranchir de la séparation du modèle en deux modules : un acteur travaillant avec les action, et un critique travaillant avec les états. Ainsi l'acteur et le critique sont réunis au sein du même module d'apprentissage des valeurs de Q. Cela évite une certaine recopie d'apprentissage dans l'acteur, pour les actions, des valeurs qui avaient été apprises par le critique, pour les états.

5.2.2 Modèle biologiquement plausible

Le substrat neuronal permettant le calcul de la différence entre les prédictions consécutives pour le signal d'erreur δ est sujet à de nombreuses discussions. [Houk et al., 1995] mettent en avant les connexions directes et indirectes entre le striatum et la substance noire compacte. La différence de temps de transmission de l'activité neuronale dans la voie excitatrice et la voie inhibitrice

permettrait ce calcul. Ce modèle est cependant très dépendant des caractéristiques temporelles liées à la dynamique des ces synapses, qui définissent le délai acceptable entre deux prédictions. De plus, l'hypothèse de différentiation des voies directes et indirectes semble en désaccord avec certaines données neurophysiologiques [Joel et al., 2002].

De plus, de nombreux modèles font appel à un mécanisme d'apprentissage différé utilisant une *trace d'éligibilité* [Barto, 1995; Foster et al., 2000; Arleo and Gerstner, 2000; Khamassi et al., 2005; Dollé et al., 2010]. Ce mécanisme suggérerait qu'une trace de l'activité neuronale passée est présente au niveau synaptique, et permettrait des modifications synaptiques pour des actions finies depuis longtemps. La plausibilité biologique de ce mécanisme reste à prouver. [Houk et al., 1995] proposent une explication impliquant les neurones épineux du striatum. Les propriétés d'une protéine (CaM PK II) et une cascade de signaux intracellulaires seraient responsables des délais dans les modifications synaptiques. Ce modèle est cependant très dépendant des dynamiques temporelles de ces réactions chimiques et ne permet pas de pouvoir expliquer comment le système peut travailler avec des délais variables entre la réalisation d'actions successives et la satisfaction de buts.

Le besoin de recourir à ces deux dynamiques temporelles vient de l'impossibilité pour le réseau de neurone de fournir de manière simultanée les valeurs de $\max_a Q_{e',a}(t+1)$ et $Q_{e,a_1}(t)$. A l'instant où une transition est réalisée, et où le système arrive dans un nouvel état, la prédiction associée à cette transition est produite par le réseau de neurones. L'instant d'après, le système commence à prédire les transitions possibles et la valeur max de ces prédictions peut être calculée. Cependant la dernière transition réalisée n'est plus active et un apprentissage différé doit être utilisé. Pour éviter le besoin de recourir à cet apprentissage différé, j'ai mis au point une version biologiquement plausible du modèle présenté dans la section précédente [Hirel et al., 2010b]. Une implémentation neuronale biologiquement plausible d'un $\max$ peut être faite en utilisant des inhibitions latérales [Yu et al., 2001]. Le modèle, appartenant à la famille des algorithmes de type TD(0), utilise un apprentissage en deux temps (fig. 5.4). Le système n'a pas vocation à modéliser de manière précise l'anatomie interne des ganglions de la base. Il se focalise plutôt sur les interactions entre structures cérébrales et montre comment un apprentissage par renforcement peut être réalisé en utilisant des règles d'apprentissage simples, sans contraintes temporelles.

Etape 1 : Une mémoire de travail dans le striatum stocke l'information de la dernière transition réalisée. Quand les prédictions de transitions possibles depuis le nouvel état sont disponibles, la valeur maximale entre ces prédictions et un éventuel signal de satisfaction de but est apprise et associée avec la transition mémorisée. Pour toute transition $e \rightarrow e'$, ce système apprend à prédire la valeur de $\max_a(\gamma \cdot Q_{e',a}(t+1))$. La règle d'apprentissage est celle du LMS :

$$\frac{dW_{ij}^{Q_p}(t)}{dt} = \alpha \cdot T(t-dt) \cdot (\gamma \cdot X^{max}(t) - X_i^{Q_p}(t)) \cdot X_j^{mem}(t) \tag{5.10}$$

α est la vitesse d'apprentissage, $T(t-dt)$ le signal de indiquant qu'une transition a été réalisée à l'instant $t-dt$ et que les prédictions de transitions pour le nouvel état sont maintenant disponibles, γ le facteur de réduction des prédictions de récompense et θ un seuil d'activité. X^{mem} est l'activité des neurones mémorisant la dernière transition effectuée et X^{mex} l'activité du neurone donnant la valeur maximale des Q-values pour les transitions prédites.
L'équation pour le calcul des activités neuronales X^{Qp} est celle-ci :

$$X_i^{Q_p}(t) = f(\sum_k W_{ik}^{Q_p}(t) \cdot X_k^{mem}(t)) \tag{5.11}$$

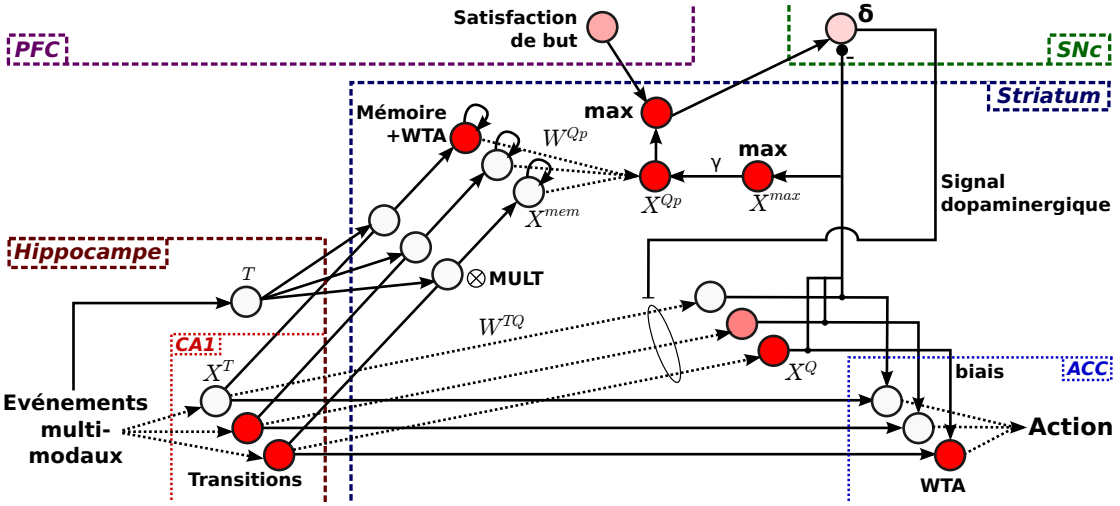


FIGURE 5.4 – Modèle neuronal de Q-learning biologiquement plausible. L'apprentissage des valeurs de Q associées à chaque transition est fait en deux temps.

Etape 2 : Les valeurs de Q sont apprises dans des poids synaptiques avec les transitions comme neurones pré-synaptiques. Quand l'agent entre dans un nouvel état, il commence à prédire toutes les transitions avec leurs valeurs de Q, utilisées pour sélectionner une transition à effectuer. Lorsqu'une transition est réalisée, le signal d'erreur δ est calculé à partir de la différence entre sa valeur directe Q et la prédiction Q_p apprise dans l'étape 1, et intégrée avec le signal instantané de satisfaction de but. Ce signal agit comme une modulation dopaminergique de l'apprentissage. L'équation d'apprentissage correspondante est la suivante :

$$\frac{dW_{ij}^{TQ}(t)}{dt} = \alpha \cdot T(t) \cdot \delta(t) \cdot X_j^T(t) \tag{5.12}$$

δ est la modulation dopaminergique correspondant au signal d'erreur de prédiction, $T(t)$ est le signal de transition réalisée et X^T est l'activité de prédiction des transitions. Cette équation permet donc de renforcer les connexions synaptiques provenant de neurones activés, en modulant ce renforcement en fonction de l'erreur de prédiction.
Le calcul de l'activité neuronale X^Q correspondant aux Q-values est :

$$X_i^Q(t) = f(\sum_k W_{ik}^{TQ}(t) \cdot X_k^T(t)) \tag{5.13}$$

Ce système permet de produire simultanément les activités Q_{e,a_1} apprise dans l'étape 2 et $\max_a(r(t+1), \gamma \cdot Q_{e',a}(t+1))$ apprise dans l'étape 1. Le système peut donc fonctionner avec de simples synapses excitatrices et inhibitrices sans dépendre de dynamiques temporelles fixées de manière ad-hoc. En revanche, cette implémentation d'un apprentissage de type TD(0) se fait en deux temps et réduit encore la vitesse de convergence de l'algorithme, qui est beaucoup plus lent qu'un TD(λ). Certains algorithmes de type Dyna [Sutton, 1991] utilisent un modèle de l'environnement pour accélérer l'apprentissage par renforcement du Q-learning en simulant des séquences d'états et d'actions. Une implémentation biologique de ce type d'algorithme pourrait faire intervenir la capacité de l'hippocampe et du cortex préfrontal à rejouer lors de phases de

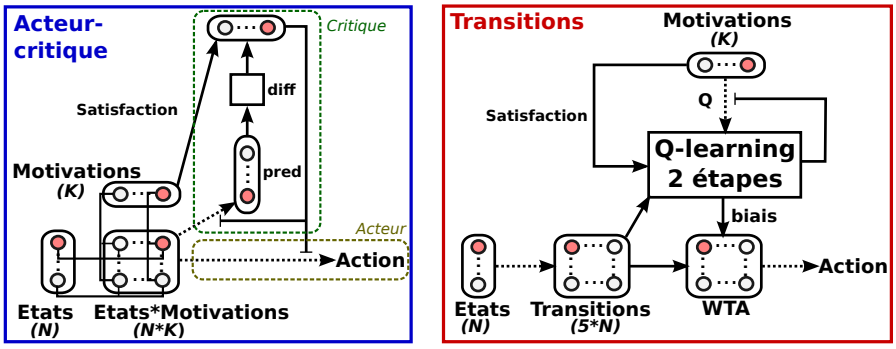


FIGURE 5.5 – Comparaison des implémentations neuronales des TD et Q-learning dans le cas de buts multiples. Quand le nombre de lieux et de buts augmentent, la solution utilisant les transitions devient de moins en moins coûteuse en termes de nombre de neurones face à un modèle acteur-critique.

repos de l'animal des séquences de lieux apprises [Karlsson and Frank, 2009; Peyrache et al., 2009]. Les phases de repos ou de sommeil pourrait alors servir à consolider et facilité l'apprentissage.

5.2.3 Navigation multi-buts

Dans un environnement simple avec un unique but, les systèmes basés sur des couples état-action ou transition-action marchent de manière similaire. En termes computationnels, l'architecture de transition nécessite environ $5N$ neurones en supplément des N neurones codant pour les états. Cependant, l'architecture de transitions montre sa force dans des environnements avec multiples buts et motivations. Supposons la présence de K types de buts dans l'environnement avec K motivations correspondantes. Dans le modèle acteur-critique, un état étant associé à une action, on ne peut effectuer de choix entre plusieurs actions selon le contexte motivationnel. Il est nécessaire d'introduire une couche intermédiaire de $K \cdot N$ neurones pour associer un couple état/motivation à une action [Arleo and Gerstner, 2000]. Dans notre modèle, pour fonctionner avec plusieurs motivations, quelques modifications sont apportées à l'architecture de la figure 5.4. En effet, les neurones de motivation vont déclencher les activités de prédiction Q_p et Q pour chacune des transitions. Les synapses W^Q et W^{Q_p} ont donc des neurones de motivation et non de transition comme origine. Les activités de transitions a^{mem} et a^T sont alors utilisées pour moduler localement l'apprentissage des poids synaptiques pour les transitions actives. Ce modèle ne marche cependant que sous certaines conditions : une seule motivation peut être active à la fois, sans quoi des interférences entre les prédictions liées aux diverses motivations auraient lieu lors de la propagation des valeurs de Q. De plus, cette propagation ne peut se faire que lorsque la motivation correspondante est active, ce qui limite les capacités d'apprentissage latent. Mais, au final, ce modèle requiert moins de neurones qu'une architecture acteur-critique pour des tâches ayant de multiples buts et sous-buts avec autant de motivations associées (fig. 5.5).

5.2.4 Expériences en simulation

Le modèle neuronal décrit dans la section 5.2.2 a été implémenté grâce au simulateur de neurones *Promethe* et utilisé dans une expérience en simulation [Hirel et al., 2010b]. L'environne-

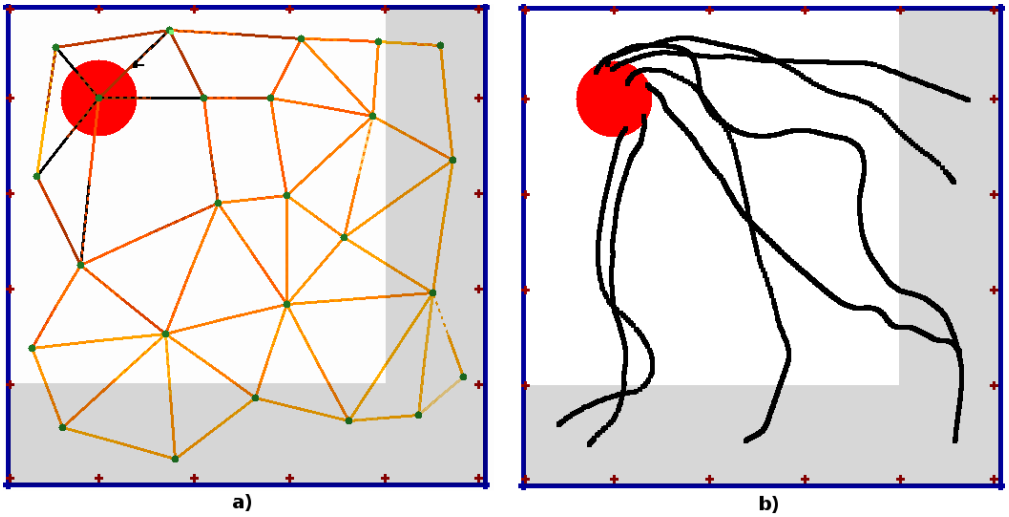
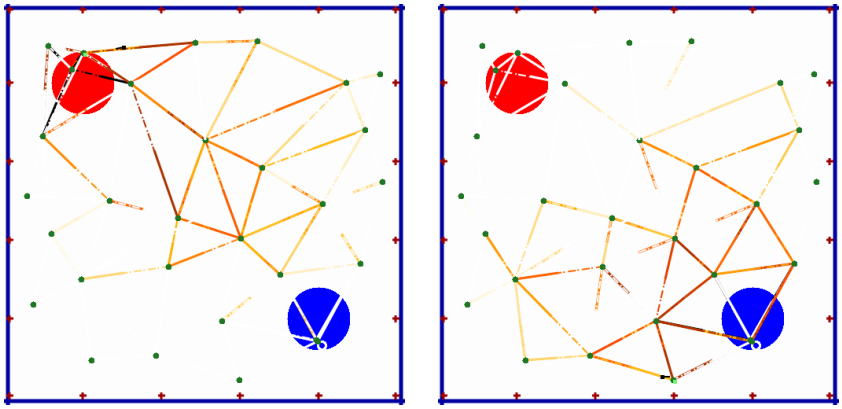


FIGURE 5.6 – a) Graphe des transitions apprises. Les couleurs de transitions plus sombres indiquent des valeurs de Q plus élevées. **b)** Trajectoires prises par le robot pendant la navigation vers le but. La partie grisée, contenant les points de départ, représente la zone où la motivation est réactivée après un essai réussi.

ment est ouvert, sans obstacles, et possède 20 amers visuels simulés le long des murs. Un but est placé dans le coin nord-ouest et signalé par une zone de couleur. Cette couleur peut être détectée lorsque le robot se situe dessus. Une première phase d'exploration aléatoire permet au robot de construire sa représentation de l'environnement : lieux, transitions et actions associées (voir chapitre 2). Pendant cette exploration, l'arrivée du robot sur la zone de couleur déclenche la satisfaction du but. L'architecture d'apprentissage par renforcement utilisant les transitions propage les valeurs de prédiction de cette satisfaction dans le graphe des transitions et le robot forme une représentation des chemins optimaux pour rejoindre le but (fig. 5.6).

Pendant la deuxième partie de l'expérience, les phases d'exploration et de navigation vers le but sont alternées. Le niveau de motivation gère le comportement du robot. Quand il est motivé, l'architecture de Q-learning est utilisée pour rejoindre le but. La satisfaction du but inhibe la motivation, et une phase d'exploration aléatoire commence. La motivation est réactivée lorsque le robot atteint une zone comprenant les extrémités est et sud de l'environnement, les plus éloignées du but. Les performances du robot dans la tâche dépendent de sa capacité à rejoindre rapidement le but depuis n'importe quel point dans cette zone. L'architecture de Q-learning permet, après apprentissage, de sélectionner les meilleures transitions pour rejoindre le but et donne des trajectoires efficaces (fig. 5.6). Les trajectoires ne sont pas des lignes droites car le robot suit la meilleure transition, et a tendance à suivre les arêtes du graphe. De plus le chemin le plus court est calculé en termes de nombre de transitions, ce qui ne représente pas toujours le chemin optimal dans l'espace euclidien, étant donné la géométrie variable des champs de lieu.

Enfin, un test a été réalisé pour valider l'utilisation de l'architecture décrite dans la section 5.2.3 dans un cadre multi-buts. Le cadre de simulation est similaire à l'expérience précédente mais l'environnement possède deux buts, associés à deux motivations différentes. Ici, plutôt que d'être réactivée dans une zone particulière, la motivation est réactivée après un délai fixe suivant la satisfaction du but qui lui est associé. Dans un premier temps, la première motivation est



(a) Motivation vers le but rouge active. **(b)** Motivation vers le but bleu active.

FIGURE 5.7 – Graphe des valeurs Q associées aux transitions dans deux contextes motivationnels différents après un apprentissage partiel.

activée et le robot construit sa représentation des valeurs de Q associées à cette motivation, en alternant les phases d'exploration et de navigation vers le premier but. Puis le même protocole est reproduit avec la deuxième motivation, permettant de construire un deuxième ensemble de valeurs de Q. On peut donc voir que le robot construit correctement des prédictions lui permettant de sélectionner des transitions dans deux contextes motivationnels différents (fig. 5.7). Cette alternance est nécessaire puisque les deux motivations ne peuvent être actives en même temps dans cette architecture.

5.3 Apprentissage par renforcement et carte cognitive : complémentarité et coopération

5.3.1 Caractéristiques des deux systèmes

J'ai présenté dans la section 5.2 un modèle de Q-learning utilisant les transitions. Ce modèle fonctionne de manière similaire à la carte cognitive en attribuant des valeurs aux transitions et en utilisant ces valeurs pour biaiser le choix d'une transition lors d'une compétition. Cette transition est alors sélectionnée et son action effectuée. Cependant, nous avons ici deux systèmes avec des particularités propres à chacun (tab. 5.1).

Le fait que ces systèmes puissent agir en parallèle et impliquent des boucles cérébrales différentes amène de la redondance et de la robustesse dans le système. Cela pourrait expliquer pourquoi certaines expériences de lésion n'affectent pas les performances de navigation de rats maîtrisant parfaitement une tâche. Le rôle du PFC dans l'acquisition rapide des capacités de résolution d'une tâche pourrait s'effacer au fur et à mesure que les ganglions de la base développent des habitudes. Cet aspect de complémentarité et de compétition est discuté par [Daw et al., 2006] et des expériences ont été menées sur le sujet [Foster et al., 2000; Dollé et al., 2010]. Je vais me concentrer ici sur les aspects de coopération entre ces stratégies et sur le transfert d'une stratégie à une autre au fur et à mesure de l'apprentissage, ou lors d'une défaillance d'une des deux stratégies. Je n'aborderai que brièvement la question de la compétition entre des

Carte cognitive	**Apprentissage par renforcement :**
Apprentissage en 1 coup. Propagation dans la carte dès la découverte d'un but (N itération de simulation pour la propagation d'un but à N noeuds de distance).	Apprentissage lent. Propagation des valeurs au fur et à mesure de la répétition de la tâche (N répétitions pour un but à N transitions de distance).
Apprentissage latent. Carte construite indépendamment de la découverte d'un but. Disponibilité des chemins vers le but immédiatement après sa découverte.	Apprentissage partiellement latent. Représentation des transitions et actions construite indépendamment de la découverte d'un but. Construction des chemins optimaux pour rejoindre le but au fur et à mesure de leur utilisation.
Gestion de plusieurs types de buts et motivations. Intègre les contraintes liées à plusieurs motivations actives.	Gestion de plusieurs types de buts et motivations. Sélectionne uniquement la motivation la plus active.
Coût computationnel élevé, calcul dynamique de la propagation de l'activité de motivation dans la carte.	Coût computationnel faible, association directe d'une transition avec sa valeur par les poids synaptiques.
Stratégie de haut niveau liée à des aspects cognitifs de gestion de buts.	Stratégie de bas niveau liée à des aspects d'habitudes et de réflexes.
Nécessité de connaître une solution précise au problème pour pouvoir rétro-propager la motivation jusqu'à l'état actuel.	"Intuition" des bonnes actions à faire (statistiquement) a priori efficaces sans besoin de connaître une solution précise.
Structures cérébrales impliquées : cortex préfrontal, cortex pariétal.	Structure cérébrale impliquée : ganglions de la base.

TABLE 5.1 – Comparaison des propriétés des stratégies de navigation par carte cognitive et apprentissage par renforcement.

stratégies multiples proposant des actions différentes, qui pourrait faire l'objet de recherches ultérieures.

5.3.2 Expériences de coopération et données de lésion

Comme nous l'avons vu précédemment, les architectures de navigation utilisant la carte cognitive et l'apprentissage par renforcement reposent sur le même socle d'apprentissage de transitions dans l'environnement. Le mécanisme de sélection d'une transition est aussi le même dans les deux cas : des valeurs associées à chaque transition sont utilisées pour biaiser le choix de l'une d'entre elles dans une compétition WTA. Afin de tester la compatibilité de ces deux approches et leur capacité à travailler en coopération, j'ai donc mis en place une série d'expériences de navigation vers un but. Dans ces expériences, le système de carte cognitive présenté dans le chapitre 2 et le système de Q-learning présenté dans ce chapitre fonctionnent en parallèle. La sélection d'une transition est réalisée par l'addition du biais provenant des deux mécanismes de navigation (fig. 5.8). Dans le cas où aucun des deux systèmes ne sait proposer une transition, car le but n'a pas été découvert ou bien la représentation de l'environnement est incomplète, une direction est choisie de manière aléatoire. De plus, un mécanisme de sélection de l'action de type

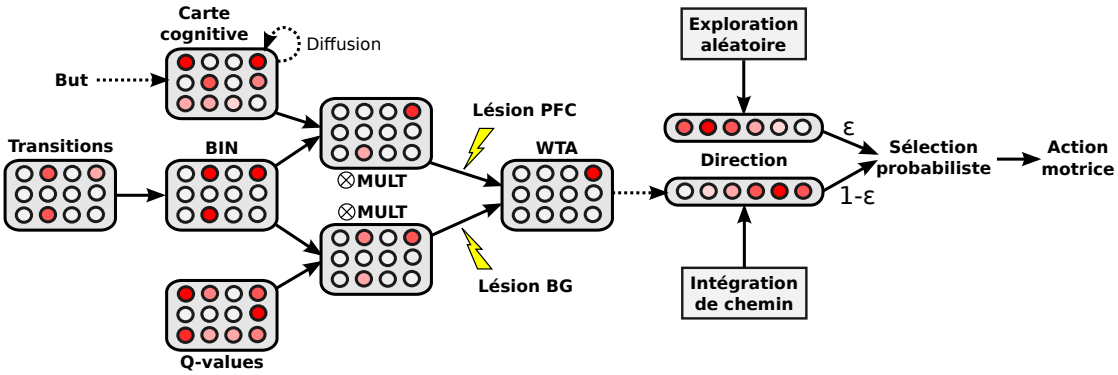


FIGURE 5.8 – Architecture neuronale de sélection de l'action pour la coopération entre les stratégies de navigation par carte cognitive et Q-learning. Les biais sur les prédictions de transitions proposés par chaque stratégie sont ajoutés et une compétition WTA a lieu. La direction proposée par les stratégies de navigation est choisie avec une probabilité $1 - \epsilon$, et une direction aléatoire avec une probabilité ϵ. Si les stratégies de navigation ne proposent aucune direction, le bruit dans la sélection de l'action entraîne un comportement d'exploration aléatoire. Des lésions du PFC ou de la partie de BG traitant les récompenses, correspondant respectivement aux stratégies de carte cognitive et Q-learning, peuvent être effectuées au niveau de la transmission des transitions biaisées.

ϵ-greedy est mis en place. Ce mécanisme correspond à un bruit dans la sélection de l'action. L'action associée à la transition proposée par le couple carte cognitive/Q-learning est choisie avec une probabilité $1 - \epsilon$ tandis qu'une direction aléatoire peut être prise avec une probabilité ϵ. Ce choix est effectué dans chaque nouvel état du robot. Ce mécanisme permet de ne pas tout le temps suivre les directions proposées par les systèmes de navigation, qui peuvent avoir une représentation incomplète de l'environnement et proposer un chemin sous-optimal. Avec cette faible probabilité d'exploration, le système peut alors découvrir tous les chemins menant au but, sans trop compromettre ses performances de navigation. Une valeur trop élevée pour ϵ entraînerait de mauvaises performances car le robot effectuerait la majorité de ses actions de manière aléatoire. A l'opposé, une valeur trop faible pousserait le robot à toujours utiliser la même solution pour atteindre le but alors qu'il pourrait découvrir des solutions beaucoup plus performantes en explorant son environnement. Une amélioration possible du système serait d'utiliser un ϵ qui diminue au fur et à mesure que l'apprentissage progresse afin de privilégier l'exploration au début de l'expérience et l'exploitation de l'apprentissage à la fin. La valeur de 0.1 utilisée ici offre un bon compromis entre exploration et exploitation.

Les expériences ont été faites dans un environnement ouvert avec un unique but placé dans le coin nord-ouest (voir fig. 5.6). A chaque essai, le robot démarre d'une position aléatoire dans une zone couvrant les parties est et sud de l'environnement. Le robot est motivé en permanence et cherche à rejoindre le but. Dès que le but est atteint, le robot est déplacé ("kidnappé") et replacé dans la zone de départ. Le kidnapping est ici utilisé pour permettre de comparer les vitesses d'apprentissage des différentes stratégies en mesurant le temps mis pour rejoindre le but. Le protocole expérimental mis en place dans la section 5.2.4 laissait une petite période d'exploration avant de reprendre la navigation vers le but, plutôt que de reprendre la navigation immédiatement. Dans le but de comparer l'évolution des temps mis pour rejoindre le but au travers de l'apprentissage, cette période d'exploration est éliminée du protocole expérimental pour ne pas biaiser les résultats du fait de l'apprentissage effectué pendant cette exploration.

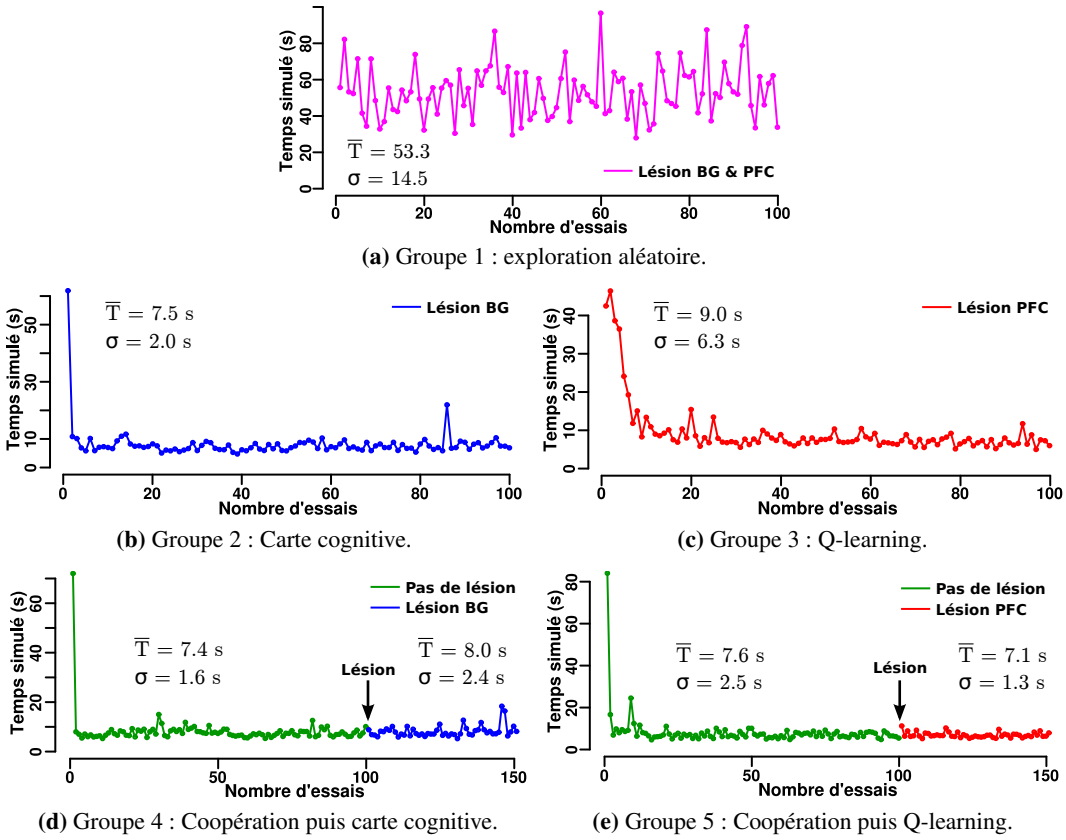
Groupe	Essais	Lésions
1	100	Lésions PFC et BG avant 1er essai. Exploration aléatoire uniquement.
2	100	Lésion BG avant 1er essai. Utilisation de la carte cognitive uniquement.
3	100	Lésion PFC avant 1er essai. Utilisation du Q-learning uniquement.
4	150	Lésion BG après le 100e essai. Utilisation des deux stratégies en coopération pour les 100 premiers essais, carte cognitive uniquement pour les 50 derniers essais.
5	150	Lésion PFC après le 100e essai. Utilisation des deux stratégies en coopération pour les 100 premiers essais, Q-learning uniquement pour les 50 derniers essais.

TABLE 5.2 – Groupes de robots pour l'expérience et lésions effectuées. Chaque groupe contient 10 robots.

La partie exploratoire de l'expérience se fait donc uniquement par la sélection probabiliste de directions aléatoires grâce au terme ϵ. Enfin, des lésions ont été effectuées pour inactiver la ou les stratégies de navigations. Dans ce cas, la stratégie lésée ne fournit plus de biais pour la sélection d'une transition.

Les expériences ont été réalisées avec 5 groupes différents (tab. 5.2). Chaque groupe est constitué de 10 robots qui effectuent l'expérience séparément. Chacun des robots simulés effectue une session de 100 essais (un essai étant terminé lorsque le but est atteint). Les robots démarrent la session sans connaissance a priori de l'environnement et doivent immédiatement réaliser la tâche. Lors du premier essai, les robots doivent d'abord trouver le lieu but en explorant. Un premier groupe a subit des lésions pré-apprentissage des deux structures liées aux stratégies de navigation (circuit dopaminergique des ganglions de la base et PFC). Ces 10 robots se basent donc uniquement sur l'exploration aléatoire pour rejoindre le but. Un second groupe a reçu des lésions pré-apprentissage du circuit dopaminergique des ganglions de la base (qui inactivent les mécanismes de prédiction de récompense nécessaires au Q-learning mais laissent les associations transition-action intactes et la sélection d'une transition par le PFC possible). Ces robots se basent donc uniquement sur la stratégie de carte cognitive pour naviguer. Le troisième groupe a subit des lésions frontales pré-apprentissage et se repose uniquement sur la stratégie de Q-learning pour naviguer. Les deux derniers groupes effectuent la session de 100 essais sans lésions, en utilisant les 2 stratégies. Puis une lésion post-apprentissage est réalisée, suivie d'une nouvelle série de 50 essais. Le quatrième groupe subit une lésion post-apprentissage des ganglions de la base et le cinquième du PFC. Dans les 50 derniers essais, la navigation repose donc sur une unique stratégie, dont l'apprentissage avait été effectué en coopération avec l'autre stratégie avant la lésion.

Dans tous les cas, excepté le premier groupe qui a subit des lésions multiples, le robot apprend à naviguer vers le but assez rapidement (fig. 5.9). De manière générale, de petites irrégularités dans la réalisation de la tâche sont causées par le mécanisme ϵ-greedy de sélection de l'action qui peut, de manière aléatoire, éloigner temporairement le robot du but. Cependant, cet effet est le même pour tous les groupes et ne biaise pas les résultats en faveur d'une stratégie ou d'une autre. La comparaison entre les groupes n'utilisant qu'une seule stratégie dès le départ de l'expérience montre que la stratégie de carte cognitive développe très rapidement une représentation de l'environnement permettant d'atteindre des performances optimales. Le Q-learning met plus de temps avant de stabiliser ses performances, car sa représentation des chemins se



(a) Groupe 1 : exploration aléatoire.

(b) Groupe 2 : Carte cognitive.

(c) Groupe 3 : Q-learning.

(d) Groupe 4 : Coopération puis carte cognitive.

(e) Groupe 5 : Coopération puis Q-learning.

FIGURE 5.9 – Temps moyen nécessaire pour rejoindre le but en fonction du nombre d'essais déjà réalisés. Les temps sont moyennés sur la population de 10 robots pour chaque groupe. Le premier essai correspond au temps mis pour découvrir le but de façon aléatoire. La moyenne $\overline{T}$ et l'écart-type σ sont donnés (en excluant le premier essai). Paramètres : $\epsilon = 0.1, r = 1, \gamma = 0.8, \alpha = 0.5$

construit au fur et à mesure des essais. On peut cependant noter que dans le cadre de l'utilisation conjointe des deux stratégies, cette latence dans l'apprentissage du Q-learning est compensé par la carte cognitive lors des premiers essais. Ces expériences correspondent donc bien à certaines observations montrant le rôle du cortex préfrontal dans la facilitation de l'apprentissage d'une tâche, même si sa présence n'est pas indispensable. De plus, après cet apprentissage coopératif, la lésion d'une des structures correspondant à une stratégie de navigation ne diminue pas les performances du robot. Ces stratégies agissent donc en coopération et ce fonctionnement pourrait expliquer le peu d'effets de certaines lésions (notamment du PFC) sur les performances de navigation des rats après qu'ils aient atteint des niveaux asymptotiques de performances [Hok, 2007]. Enfin, on peut noter en regardant la moyenne des performances que les systèmes de navigation utilisés dans cette expérience fournissent des performances environ 7 fois supérieures à une simple exploration aléatoire de l'environnement. Cet écart devrait se montrer de plus en plus marqué dans des environnements plus grands et plus complexes.

5.4 Discussion

Dans ce chapitre, j'ai montré que l'architecture de transitions pouvait servir de base à une implémentation neuronale d'un algorithme d'apprentissage par renforcement : le Q-learning. L'apprentissage des transitions et des actions associées se fait de manière dissociée de la stratégie de navigation, et le Q-learning fonctionne alors en utilisant uniquement la représentation sous forme de transitions de l'environnement. Ce modèle s'éloigne des modèles traditionnels acteur-critique qui nécessitent à la fois un critique travaillant avec une représentation sous forme d'états et un acteur travaillant avec des actions. L'utilisation des transitions permet également à l'architecture d'avoir un socle commun avec d'autres stratégies de navigation comme la carte cognitive. On peut alors facilement mettre ces stratégies en parallèle et montrer qu'un apprentissage en coopération permet d'effacer certaines latences dans l'apprentissage liées à l'algorithme du Q-learning et de maintenir de bonnes performances grâces à la redondance du système lors de lésions de certaines structures. On rejoint ainsi des observations faites chez les animaux ne montrant pas d'effets notables d'une lésion du cortex préfrontal après apprentissage, mais le modèle prédit que sa lésion pré-apprentissage ralentirait les capacités du rat à acquérir la tâche, dans le cadre de tâches suffisamment complexes pour qu'elles ne puisse être apprises par d'autres stratégies très simples existant en parallèle. La fusion des informations fournies par les deux systèmes est ici réalisée de manière simpliste, en additionnant les biais correspondant aux transitions à effectuer. De futurs travaux pourraient intégrer un mécanisme de sélection de stratégie ou de méta-contrôle plus complexe [Damásio, 1994; LeDoux, 1998; Cañamero, 2005]. Celui-ci pourrait à la fois apprendre à sélectionner la meilleure stratégie en fonction du contexte dans lequel le robot se trouve, mais pourrait aussi détecter les dysfonctionnements d'une stratégie et l'inhiber pour en laisser une autre prendre la priorité [Hasson, 2011].

Dans les architectures présentées dans ce chapitre, la notion de la prédiction temporelle fine des transitions est absente. Le système travaille uniquement avec les informations de prédiction des transitions, sans se soucier du timing de ces transitions. La contrainte liée à cette absence d'informations temporelles est que le signal de satisfaction du but doit toujours être simultané avec un signal perceptif dans EC. Ainsi, lors de l'entrée du robot dans un nouvel état, on peut savoir si cet état est associé avec une satisfaction de but. L'algorithme de Q-learning permet de corriger ses valeurs de prédiction de récompense grâce à un calcul d'erreur. Ce calcul d'une différence entre des prédictions concernant les récompenses attendues et les récompenses reçues correspond bien aux activités neuronales enregistrées dans le système dopaminergique lié aux ganglions de la base [Schultz et al., 1993; Schultz, 1998]. Dès lors, si le robot venait à atteindre l'état du but sans recevoir de signal de satisfaction, il finirait par oublier ce but. Cependant, aucun mécanisme ne permet de corriger les prédictions lorsque l'état du but n'est plus atteignable (par exemple car l'action effectuée ne déclenche plus la perception de cet état). Il serait alors nécessaire d'utiliser un oubli passif ou actif, ou bien d'avoir un système de détection de l'échec. Il est possible pour le système de détecter qu'une action a mené le robot à un autre état que celui attendu (et que la satisfaction de but attendue n'a pas eu lieu) sans utiliser d'informations temporelles. En revanche, parfois une action n'entraîne pas de changement d'état, et la satisfaction du but peut être retardée par rapport à l'exécution de cette action (par exemple dans la tâche de navigation continue, l'action de ne pas bouger entraîne une satisfaction de but après 2 secondes ou, dans le cas d'essais d'extinction, n'entraîne aucun changement d'état). Il est donc nécessaire à ce moment que le calcul d'erreur de prédiction de récompenses comporte une composante tem-

porelle et soit capable de détecter le dépassement du délai normal d'attente. Ainsi l'utilisation de l'architecture de transition comme socle d'un algorithme d'apprentissage par renforcement amène un autre avantage : la capacité à pouvoir tirer parti des prédictions temporelles fines pouvant être données par le système de transitions. Nous verrons donc dans le chapitre 6 comment des prédictions temporelles peuvent être utilisées dans la modulation du comportement du robot. Nous avons vu dans ce chapitre que le modèle est cohérent avec la vue communément acceptée du système dopaminergique procédant à un calcul d'erreur sur des prédictions de récompenses. Je montrerai dans le chapitre suivant qu'il est aussi cohérent avec certaines hypothèses supposant le rôle des activités phasiques des neurones dopaminergiques pour l'apprentissage, dans les ganglions de la base, des contextes sensoriels et moteurs liés à la perception d'événements inattendus [Redgrave and Gurney, 2006].

Publications personnelles

Hirel, J., Gaussier, P., Quoy, M., and Banquet, J.-P. (2010b). Why and how hippocampal transition cells can be used in reinforcement learning. In *SAB '10: Proceedings of the 11th international conference on Simulation of Adaptive Behavior*. Springer-Verlag

Je n'ai pas échoué. J'ai simplement trouvé 10000 solutions qui ne fonctionnent pas.

– Thomas Edison

CHAPITRE 6

Utilisation de la détection de l'échec dans la modulation comportementale

Nous avons vu dans les chapitres précédents comment un robot peut utiliser une architecture neuronale biologiquement plausible pour apprendre des relations temporelles entre des événements perceptifs. Ces relations peuvent être prédites finement dans le temps et l'attente d'un événement particulier peut entraîner l'exécution de l'action nécessaire pour que cet événement se produise. Ces prédictions temporelles peuvent également être associées avec des signaux de satisfaction de buts. Par un mécanisme de rétropropagation emprunté au Q-learning, on peut alors construire des prédictions concernant la satisfaction d'un but dans le futur, en fonction des actions que le robot effectue. Les cadres d'utilisation de ce modèle montrés jusqu'à présent tirent parti des capacités de prédiction du modèle, mais pas de l'aspect d'estimation de l'écoulement du temps et de l'apprentissage de timings. En effet, dans les expériences réalisées jusqu'ici, la prédiction d'une transition particulière déclenche l'action correspondante, peu importe son timing. A la suite de l'exécution de cette action, un nouvel événement est perçu. Si tout s'est bien passé, l'événement perçu correspond à l'état d'arrivée de la transition. Dans le cas contraire, un événement non prédit ou bien ne correspondant pas à la transition sélectionnée est perçu. On peut alors dire que le système est en situation d'échec, puisque l'action qu'il a effectuée n'a pas mené à la conséquence attendue pour le robot. Un autre point de vue consisterait à dire que notre système a pour rôle l'apprentissage de régularités dans le comportement et les perceptions du robot. Ainsi un événement non prédit constituerait une irrégularité, ou encore une forme de *nouveauté*. Bien que notre modèle fournisse les outils nécessaires pour détecter ce genre de situation, cela n'a pas été réalisé jusqu'ici. L'arrivée d'un événement imprévu amène juste le système à planifier de nouveau ses actions pour rejoindre le but. Ainsi, dans l'état actuel, le modèle possède deux limitations majeures pour la résolution de tâches robotiques :

1. Si une action est effectuée dans l'attente d'un événement particulier (par exemple : attendre immobile pour entendre un son) et que cet événement n'arrive jamais, le robot peut être bloqué dans une situation d'impasse (attendre indéfiniment au même endroit).

2. Si une action effectuée ne mène pas à la conséquence attendue (par exemple : une direction prise par le robot ne mène pas au lieu visé, mais dans un autre), cet échec n'est pas détecté et n'a aucune influence sur le système.

Afin de répondre à la première limitation, il est nécessaire de connaître le timing de l'événement attendu afin de savoir quand le robot peut considérer que cet événement n'arrivera pas. C'est ici qu'intervient l'architecture de prédiction de transitions avec timing, qui remplit justement ce rôle. De même, cette architecture pourrait être utilisée pour détecter que la transition effectuée n'est pas celle qui avait été sélectionnée et offrir une solution à la deuxième limitation.

Je détaillerai donc dans ce chapitre comment les prédictions temporelles fournies par le système d'apprentissage de transitions peuvent être utilisées pour détecter des échecs dans la réalisation d'une tâche par des stratégies de sélection de transition. Je décrirai un réseaux de neurones permettant de signaler que le timing d'un événement attendu est dépassé puis je discuterai de la façon dont cette architecture s'intègre dans le modèle des interactions entre hippocampe, cortex préfrontal et ganglions de la base. Viendra ensuite la question de l'utilisation de ces signaux dans la modulation du comportement du robot. Nous verrons d'abord comment moduler immédiatement le comportement suite à un échec détecté aussi bien parce que le timing de l'événement prédit est passé que parce que l'événement perçu n'était pas celui prédit. Puis nous verrons comment moduler le comportement sur un plus long terme en rectifiant les apprentissages ayant mené à cet échec. Les capacités de généralisation de ce modèle seront démontrées à travers la réalisation de diverses tâches de navigation robotique, à la fois en simulation et sur robot réel. Je conclurai par une discussion sur les apports de ce système et les perspectives de travail.

6.1 Transitions temporelles et détection de l'échec

6.1.1 Modèle neuronal de détection de l'échec

Les prédictions temporelles de transitions fournissent une information sur le timing de la perception d'un prochain événement sous la forme d'une activité avec un pic anticipant ce timing. Ainsi la forme elle même de l'activité donne les informations sur la relation temporelle entre le dernier événement perceptif et celui qui est attendu. La hauteur du pic d'activité, dépendant de l'apprentissage répété de la transition, donne une information sur la *fiabilité* de la prédiction. Plus l'activité est forte et plus la prédiction est fiable car observée de nombreuses fois. La largeur du pic d'activité (ce qui correspondant à un pic d'activité plus ou moins loin dans le temps) donne une information sur la *précision* de la prédiction. Un pic étroit signifie une prédiction précise, qui résulte de l'apprentissage d'un timing peu variable à chaque répétition de la transition. Un pic plus large signifie une tendance à avoir un timing variable, ou bien un timing trop long pour que les structures cérébrales impliquées puissent le prédire avec précision (pour rappel, les apprentissages réalisés font appel à des mécanismes impliqués de manière plausible dans l'apprentissage de timings de l'ordre de centaines de millisecondes ou de secondes). Enfin, le *maximum d'activité* donne une information sur le timing lui même. Ce maximum précède la perception de l'événement attendu. Plus la prédiction est précise et le timing court, moins cette précession sera importante.

Nous allons donc nous intéresser ici aux moyens de détecter la non réalisation d'un état prédit. Il faut être suffisamment précis pour que le robot n'attende pas plus que nécessaire l'arrivée d'un événement qui ne viendra pas, mais aussi pouvoir prendre en compte la variabilité du timing de cet événement pour ne pas interrompre un comportement de manière précoce. Les prédictions temporelles ont été utilisées dans le cadre de l'apprentissage et la reproduction de séquences d'actions [Andry et al., 2001]. Dans ce cas, le pic d'activité anticipe le moment où la

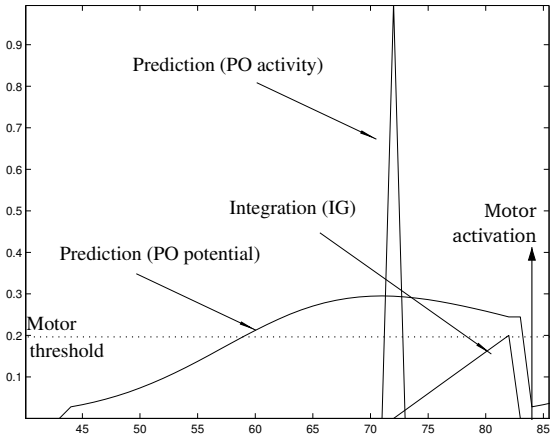


FIGURE 6.1 – Mécanisme de déclenchement d'une action basé sur une prédiction temporelle. La dérivée de la prédiction au niveau du maximum de potentiel devient négative et déclenche un pic d'activité excitant un neurone intégrateur. Une fois l'activité du neurone intégrateur dépassant un seuil fixé, l'action est déclenchée. Tiré de [Andry et al., 2001].

prochaine action doit être effectuée. Le déclenchement de cette action en fonction de l'activité de prédiction était fait de manière très simple : l'architecture détectait le maximum d'activité comme le point où la dérivée de la prédiction devenait négative, puis un intégrateur neuronal introduisait un petit délai entre ce signal et l'exécution de l'action pour compenser l'aspect d'anticipation de la prédiction (fig. 6.1).

Ce système permettait de détecter le moment où l'action devait être déclenchée, par rapport à ce qui avait été appris. Cependant, la vitesse d'intégration du neurone intégrateur, ainsi que le seuil, devaient être réglés manuellement. Ils permettaient de spécifier un délai fixe entre le pic de prédiction et le déclenchement de l'action, mais ne prenaient pas en compte l'anticipation variable de la prédiction, ou bien sa largeur. De plus, avec les améliorations apportées à l'architecture pendant cette thèse, un neurone de prédiction peut apprendre de manière continue à prédire le timing d'une transition. Ainsi, si une transition est réalisée de manière répétée avec deux timings différents, l'activité de prédiction possédera deux pics. Or, le système utilisé dans les séquences d'action se déclencherait invariablement après le premier pic, ne prenant pas en compte la possibilité que l'événement puisse arriver après le second pic. Il a donc fallu mettre au point un mécanisme de détection plus souple et moins ad-hoc.

Le réseau neuronal pour la détection de l'échec se base donc sur une comparaison entre l'activité maximale de la prédiction et l'activité actuelle (fig. 6.2). Une population de neurones X^{max} mémorise l'activité maximale depuis le début de la prédiction pour chaque transition. Lorsque la prédiction s'arrête, cette mémoire est remise à zéro. Ce calcul peut être formalisé sous la forme de l'équation suivante :

$$X^{max}(t) = f(H_\epsilon(X^T(t-dt)) \cdot \max(X^T(t), X^{max}(t-dt))) \tag{6.1}$$

X^T est l'activité de prédiction de transition. H est la fonction de Heavyside et ϵ est une valeur positive proche de 0 permettant de n'activer la mémoire que pour les transitions ayant des prédictions.

Cette mémoire de l'activité maximale est transmise vers d'autres neurones $X^{\overline{max}}$ qui ap-

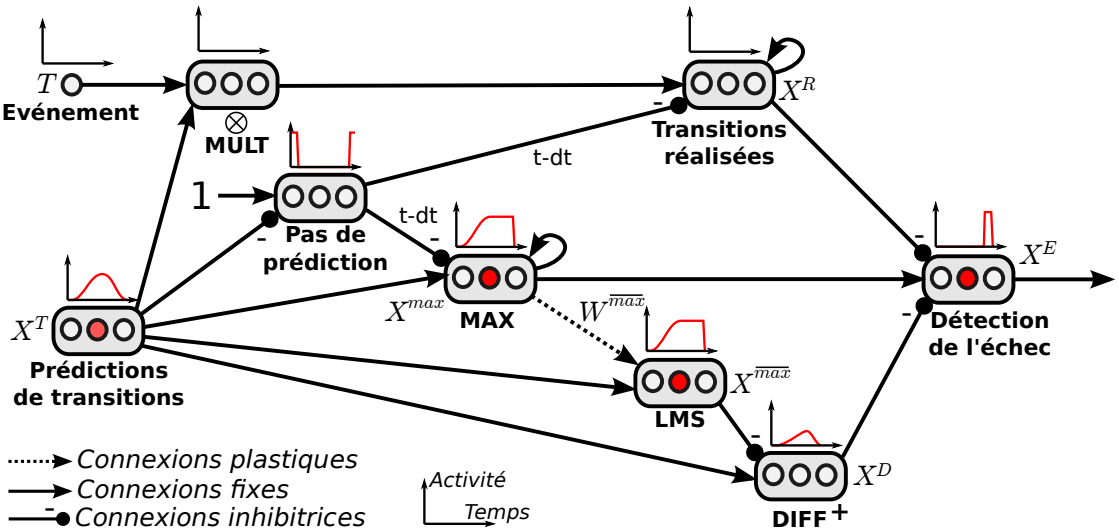


FIGURE 6.2 – Détails de l'implémentation neuronale du système de détection de l'échec.

prennent sa corrélation avec l'activité de prédiction de transition, par le biais d'une règle de LMS (3.9). En clair, pendant la phase montante de la prédiction, la mémoire de l'activité maximale est identique à l'activité de prédiction. Le LMS apprend donc la correspondance entre deux signaux identiques et les poids synaptiques $W^{\overline{max}}$ tendent vers 1. Ensuite, pendant la phase descendante, la mémoire a une activité supérieure à la prédiction. Le LMS tend alors à corriger cette erreur en diminuant les poids synaptiques. La conséquence de cet apprentissage est que les poids synaptiques entre la mémoire et le LMS vont tendre vers 1 quand l'événement se produit toujours au moment du pic de prédiction (puisque l'apprentissage ne s'effectuera que pendant la phase montante, avant que l'événement ne l'arrête en remettant à zéro les prédictions). Par contre, quand l'événement ne se produit pas systématiquement, ou que son timing est variable, l'apprentissage dans la phase descendante va faire converger les poids synaptiques vers une valeur de $W^{\overline{max}}$ inférieure à 1. Au final, l'activité $X^{\overline{max}}$ en sortie du LMS est l'activité X^{max} de la mémoire du maximum, multipliée par un facteur ≤ 1. Une différence X^D est alors calculée entre l'activité de prédiction des transitions et la sortie du LMS. Si cette différence est positive, c'est-à-dire que l'activité de prédiction est supérieure à $W^{\overline{max}}$ fois le maximum de prédiction, les neurones de différence sont activés. L'équation correspondante est la suivante :

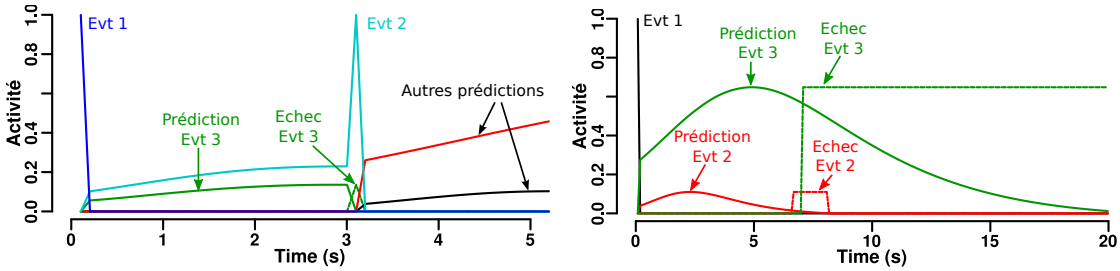
$$X^D(t) = f(X^T(t) - W^{\overline{max}}(t) \cdot X^{max}(t)) \tag{6.2}$$

Cette activité sert à inhiber le signal de détection de l'échec X^E. En d'autres termes, tant que l'activité de prédiction n'est pas retombée sous une fraction $W^{\overline{max}}$ de son activité maximale, le signal de détection de l'échec X^E n'est pas émis. De plus, une autre voie neuronale X^R garde en mémoire, jusqu'à ce que leurs prédictions s'arrêtent, les événements qui se sont produits. Cette mémoire peut être formalisée par l'équation suivante :

$$X^R(t) = H_\epsilon(X^T(t - dt)) \cdot f(T(t) \cdot X^T(t) + X^R(t - dt)) \tag{6.3}$$

Cette voie inhibe également l'émission du signal X^E de détection de l'échec. L'équation pour le signal d'échec est finalement :

$$X^E(t) = f(X^{max}(t) - H_\epsilon(X^D(t)) - H_\epsilon(X^R(t))) \tag{6.4}$$



(a) Cas où deux événements sont prédits et l'un est perçu. Lors de la réalisation de l'événement 2, un signal d'échec est émis pour l'événement 3 qui était aussi prédit. Cette exemple illustre la capacité du système à détecter que l'événement perçu n'est pas celui qui était prédit. Le signal d'échec a un niveau d'activité similaire à l'activité maximale de la prédiction correspondante.

(b) Cas où deux événements sont prédits mais rien n'est perçu. La transition prédisant l'événement 2 avait été apprise mais cet événement n'a été perçu qu'occasionnellement, d'où la faible activité de prédiction et la grande tolérance du système de détection d'échec pour cet événement. Le signal d'échec n'est déclenché qu'après que l'activité de prédiction soit retombée très bas. Pour l'événement 3, qui a été répété de manière régulière, le signal d'échec est émis peu après le maximum d'activité.

FIGURE 6.3 – Exemple de déclenchement du signal d'échec pour des activités temporelles de prédiction de transitions.

En conséquence, le signal de détection de l'échec s'active dans deux cas bien précis :

1. Si l'activité de prédiction a passé son pic et retombe sous une fraction $W^{\overline{max}}$ de son maximum, alors que l'événement prédit n'a toujours pas eu lieu.

2. Si un autre événement que celui prédit à remis à zéro les activités de prédiction. X^D devient alors nulle et l'utilisation de $X^T(t - dt)$ dans les équations permet de retarder l'inhibition des mémoires X^{max} et X^R.

Dans les deux cas, le signal de détection de l'échec est émis individuellement pour chaque transition dont l'activité de prédiction X^T est fournie au système (fig. 6.3). La force de ce système est que la valeur de $W^{\overline{max}}$ apprise représente une mesure de la variabilité du timing prédit. Plus l'événement prédit a tendance à ne pas être perçu, ou bien à être perçu avec un timing variable, plus ce poids synaptique va être faible et le système va être tolérant sur les retards acceptables par rapport aux prédictions avant de déclencher un signal d'échec. Ainsi un événement très régulier, se reproduisant systématiquement avec le même timing, sera signalé comme un échec si il n'a pas été perçu juste après le sommet du pic de prédiction. Pour un événement beaucoup plus irrégulier, il faudra attendre que la prédiction retombe vers une fraction basse de son activité maximale avant qu'un échec ne soit signalé.

6.1.2 Intégration dans le modèle des interactions hippocampe - cortex préfrontal - ganglions de la base

Le modèle neuronal développé dans la section précédente permet donc de détecter l'absence d'un événement sensoriel attendu. Cette capacité est primordiale pour détecter des irrégularités dans l'environnement du robot, par rapport à ce qui avait été appris. Lors de la réalisation d'une tâche, le comportement nécessaire pour atteindre le but est souvent répétitif et possède des régularités. La détection d'irrégularités dans la répétition d'un comportement bien acquis et menant

habituellement à la résolution de la tâche peut servir d'indicateur dans le cas où les conditions de satisfaction du but auraient changé. Ainsi le système développé peut donner des signaux indiquant des changements soit dans l'environnement, soit dans les conditions de résolution de la tâche et donc dans le comportement à adopter.

Le rôle du cortex préfrontal dans la capacité des mammifères à adapter leur comportement lors de changements dans les conditions de la tâche à été montré par de nombreuses études. Par exemple, chez des rats devant résoudre une tâche de navigation vers un but dans un labyrinthe en croix, deux stratégies peuvent être utilisées : une stratégie allocentrique faisant intervenir une représentation spatiale du labyrinthe avec laquelle le rat peut s'orienter vers un bras particulier du labyrinthe (ex : bras nord), ou une stratégie égocentrique où le rat apprend à tourner toujours dans la même direction une fois arrivé à l'intersection (ex : 90° à droite). Des lésions du cortex préfrontal, et plus précisément des régions infralimbiques et prélimbiques [Ragozzino et al., 1999], empêchent dans ce cas l'apprentissage d'un changement de stratégie pour la réalisation de la tâche (ex : passer d'une stratégie allocentrique à une stratégie égocentrique) mais pas d'un changement des conditions au sein d'une même stratégie (ex : passer du bras nord au bras sud dans le cas de la stratégie allocentrique). Ces données de lésion sont complétées par des enregistrements montrant des changements d'activités infralimbiques et prélimbiques lors de changements de stratégies mais pas lors de changements intra-stratégie [Rich and Shapiro, 2009]. Par ailleurs, dans une tâche où deux stimuli peuvent être associés à des valeurs de récompense différentes, le stimuli associé à la récompense la plus importante entraîne une vitesse de réaction plus élevée lorsque le rat doit appuyer sur un levier quelques instants plus tard pour recevoir cette récompense [Bohn et al., 2003]. Or, une lésion du PFC empêche le changement des associations entre stimuli et récompenses, même si les performances dans la tâche ne sont pas impactées. Dans ce cas, les stimuli continueront de fournir les mêmes vitesses de réaction même après que les valeurs de récompenses aient été changées. Les adaptations contrôlées par le PFC pourraient donc être liées à sa capacité à faire des associations entre stimuli, actions et récompenses.

Chez l'humain, le cortex préfrontal joue un rôle important dans la résolution de tâches cognitives demandant une adaptation à des règles changeantes. Ainsi, [Dehaene and Changeux, 1989] ont mis au point un modèle du cortex préfrontal pour la résolution de tâches de réponse différée. Dans ce modèle, un premier niveau d'associations sensori-motrices est accompagné d'un second niveau utilisant un apprentissage par renforcement pour adapter les associations en fonction des règles de l'expérience. Plus tard, [Dehaene and Changeux, 1991] ont développé un modèle neuronal du lobe frontal pour la résolution de la tâche "*Wisconsin Card Sorting*". Ce modèle inclut trois modules fondamentaux : une mémoire des règles déjà utilisées, une capacité à écarter certaines règles par raisonnement et un module de changement de stratégie se basant sur des signaux d'échec. Ces signaux d'échec, qui pourraient aussi correspondre à ceux émis par notre modèle, semblent être présents au niveau du cortex orbitofrontal humain, dans lequel des activités de traitement de l'absence de conséquences attendues ont été enregistrées [Schnider et al., 2007]. On peut recouper ces observations avec les activités des neurones dopaminergiques enregistrées dans l'aire tegmentale ventrale et la substance noire compacte qui montrent des capacités de prédiction de récompenses [Schultz et al., 1993; Schultz, 1998]. La capacité de ces neurones à détecter l'absence d'une récompense attendue en se basant sur la relation temporelle entre un stimulus et la récompense qui le suit, sous forme d'une dépression de l'activité neuronale, est fortement similaire au système de détection de l'échec. L'hypothèse la plus répandue est

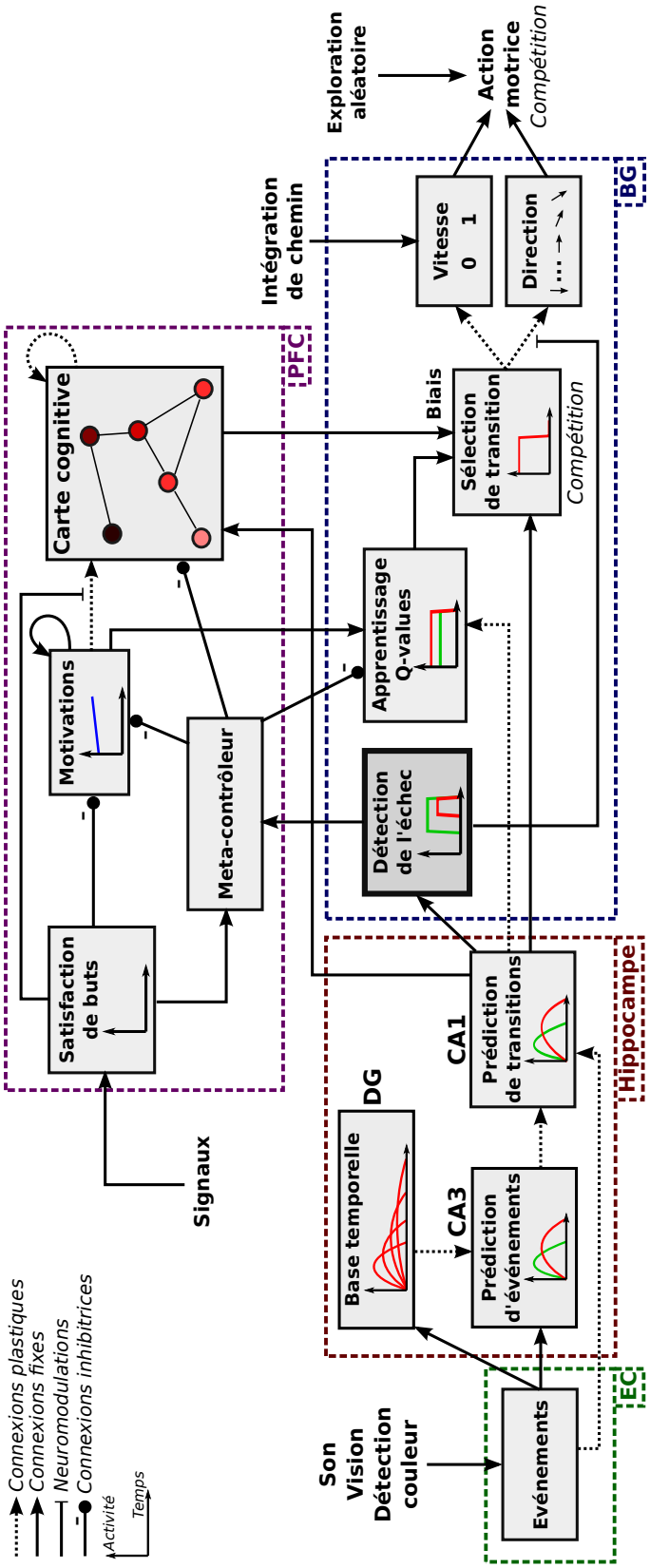


FIGURE 6.4 – Modèle intégré des interactions entre hippocampe, cortex préfrontal et ganglions de la base. Ce modèle inclut le sytème d'apprentissage de transitions, les associations transition-action, les stratégies de navigation par carte cognitive et Q-learning et le système de détection de l'échec fournissant des signaux à un méta-contrôleur situé dans le PFC. Ce méta-contrôleur peut moduler le comportement de manière immédiate, par exemple en inhibant les stratégies de navigation ou les motivations (voir section 6.2). Les signaux du système de détection de l'échec peuvent également servir à moduler l'apprentissage des associations transition-action pour modifier le comportement à plus long terme (voir section 6.3).

que ces neurones dopaminergiques encodent des signaux liés à la valeur des perceptions (ils sont excités par des perceptions associées à des récompenses mais inhibés par des stimuli aversifs). Cependant, [Matsumoto and Hikosaka, 2009] ont montré que seul un sous-ensemble des neurones dopaminergiques se comportait de cette façon. D'autres neurones sont excités aussi bien pour des stimuli positifs qu'aversifs, particulièrement lorsque ceux-ci sont imprédictibles. Cela montre bien, de manière cohérente avec notre modèle, une capacité de détection de nouveauté liée à la capacité de prévoir la perception de stimuli au niveau du système dopaminergique.

Notre hypothèse de modèle (fig. 6.4), résultant de ces observations, est que le traitement des prédictions temporelles et la détection de l'échec se font au niveau des ganglions de la base et du circuit dopaminergique. Les signaux résultants seraient alors transmis au PFC qui ferait office de méta-contrôleur en utilisant les signaux d'échec pour moduler le comportement, notamment dans le cadre de changements dans les conditions d'une tâche où il convient d'inhiber certains comportements pour éviter des effets de persévération. Ce méta-contrôleur pourrait agir de diverses manières pour modifier le comportement en inhibant diverses parties du modèle. Nous verrons certains exemples et leur applications robotiques dans la section 6.2. Les signaux dopaminergiques liés à la détection de l'échec pourraient également agir comme neuromodulateurs de certains apprentissages dans les ganglions de la base. Ainsi une modulation à plus long terme du comportement serait possible. Des exemples et applications robotiques de ces modulations sont présentés dans la section 6.3. La modélisation du méta-contrôleur et ses implications seront discutées dans la section 6.4.

6.2 Modulation comportementale immédiate

6.2.1 Tâche de navigation continue et essais d'extinction

La première application du signal de détection de l'échec est la modulation immédiate du comportement. Dans le cas où une action mène à un échec, il convient de modifier l'état du système qui a mené à la sélection de cette action. Sinon, le système pourrait être bloqué dans une impasse, en continuant à vouloir exécuter une action qui mène à des échecs répétés. Cette persévération de comportements qui ne sont plus pertinents dans un contexte donné est typique dans les expériences de lésion du cortex préfrontal. Si l'on revient à l'expérience de navigation continue menée dans la section 4.2.2, le robot avait appris à naviguer vers un lieu but à l'aide d'une carte cognitive. La reconnaissance de la zone de couleur avait été associée à une phase d'attente apprise par interaction avec l'humain. Ainsi le robot se maintenait immobile sur le but jusqu'à percevoir le son attendu. En revanche, la non présentation du son n'était pas détectée et le robot aurait continué à attendre le son dans un tel cas. Un système de frustration aurait pu permettre de changer d'action au bout d'un certain temps, mais un tel système devrait être paramétré manuellement en fonction de la tâche [Hasson and Gaussier, 2010]. C'est donc ici qu'intervient le réseau de neurones de détection de l'échec présenté dans ce chapitre. Je présente donc dans cette section un des fonctionnements possibles du méta-contrôleur utilisant les signaux d'échec (fig. 6.5). Grâce à ces signaux et à la modulation comportementale à travers l'inhibition des motivations du robot, il est possible de faire reprendre le mouvement du robot peu de temps après le délai de 7 secondes en cas de non présentation du son. Ainsi, le comportement du robot pourrait reproduire celui des rats aussi bien lors des essais normaux que des essais d'extinction.

L'idée derrière cette modulation comportementale est d'inhiber la motivation qui mène le

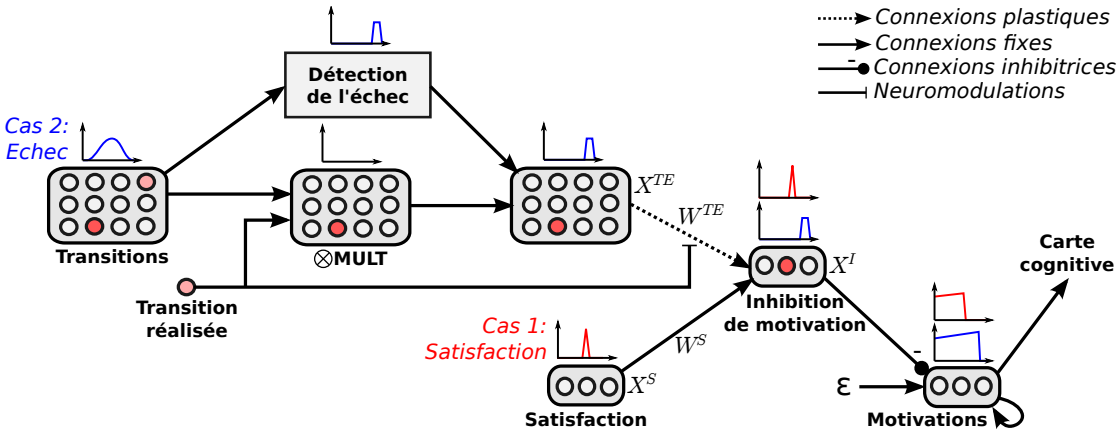


FIGURE 6.5 – Architecture neuronale d'inhibition de motivation en cas d'échec de la satisfaction d'un but. Le signal de satisfaction est associé à la transition qui lui correspond. La réalisation de cette transition, qui satisfait le but, inhibe la motivation. Dans les essais d'extinction, le signal d'échec pour cette transition inhibe aussi la motivation, peu après le pic de prédiction de la satisfaction. L'absence de motivation entraîne un changement de comportement du robot qui repart du but. Le système de détection de l'échec est celui présenté dans la figure 6.2.

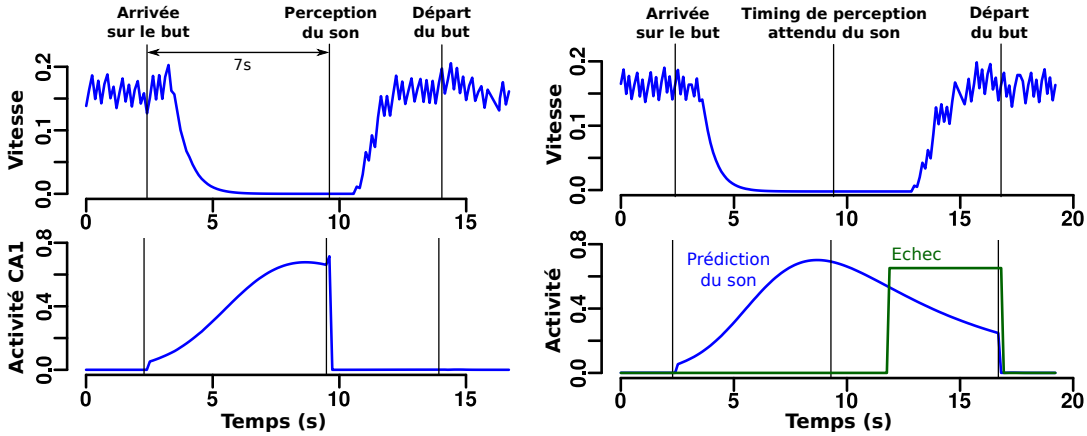
robot à tenter de rejoindre son but (le son). Pour ce faire, le méta-contrôleur apprend à associer la satisfaction du but (la perception du son) avec la transition *Lieu but→Son* qui l'a déclenchée. Les signaux de satisfaction peuvent correspondre à la consommation d'une ressource (nourriture, eau) satisfaisant un besoin particulier (faim, soif). Ils peuvent aussi être fournis directement par l'expérimentateur pour signaler au robot que l'état qu'il a atteint constitue un but de la tâche. Nous verrons dans le chapitre 7 que le signal peut aussi être fourni par l'interaction avec l'humain. Ici, le signal est perçu automatiquement par le robot lors de la perception du son. On considère qu'un conditionnement précédemment appris a permis l'association du son avec un but à atteindre. Etant donné que la transition est réalisée simultanément avec la satisfaction du but, un simple apprentissage associatif est réalisé. L'équation pour cet apprentissage est une règle de Widrow-Hoff à laquelle est ajoutée une modulation $T(t)$:

$$\frac{dW_{ij}^{TE}(t)}{dt} = \alpha \cdot T(t) \cdot (\sum_k W_{ik}^S(t) \cdot X_k^S(t) - X_i^I(t)) \cdot X_j^{TE}(t) \tag{6.5}$$

α est la vitesse d'apprentissage. $T(t)$ est le signal indiquant qu'une transition est réalisée et modulant l'apprentissage. X^{TE} est l'activité des neurones de transition ayant échoué, sauf lors de l'apprentissage ($T(t) = 1$) où le neurone de la transition réalisée est activé (afin de permettre l'association entre la transition et la satisfaction correspondante).
L'équation de calcul de l'activité neuronale est :

$$X_i^I(t) = f(\sum_k W_{ik}^{TE}(t) \cdot X_k^{TE}(t) - \theta) \tag{6.6}$$

En temps normal, la transition est réalisée, le son est perçu, et le déclenchement de la satisfaction inhibe la motivation correspondante, le but ayant été atteint. Cette motivation sera réactivée plus tard lorsque la phase d'exploration aléatoire sera terminée. Lors des essais d'extinction, le problème intervient lorsque le son est prédit, mais n'est jamais perçu, et l'action



(a) Vitesse du robot et activité de prédiction dans CA1 lors des essais normaux. Voir section 4.2.2.

(b) Vitesse du robot, activité de prédiction dans CA1 et signal de détection de l'échec lors des essais d'extinction. Quelques secondes après la prédiction du son, le signal d'échec inhibe la motivation et le robot repart avant de quitter le but.

FIGURE 6.6 – Résultats de l'expérience de navigation continue sur robot réel avec essais normaux et essais d'extinction.

associée à la prédiction du son (ne plus bouger) continue d'être exécutée. Sans aucun système de détection d'échec, le robot attend indéfiniment le son et ne satisfait jamais son but. Il est donc nécessaire d'avoir un mécanisme d'inhibition de la motivation qui prenne en compte l'échec de l'attente du robot. Le méta-contrôleur reçoit les signaux d'échec correspondant aux diverses transitions. L'association préliminaire des transitions avec la satisfaction du but permet de sélectionner uniquement le signal d'échec correspondant à la prédiction du son. Ce signal est alors utilisé pour inhiber la motivation. Or, c'est cette motivation qui, par l'intermédiaire de la carte cognitive, avait mené à la sélection de la transition *Lieu but→Son* et donc à l'exécution de l'action *Ne pas bouger*. Une fois la motivation inhibée, la carte cognitive ne propose plus de biais et la transition associée à l'arrêt du robot n'est plus sélectionnée. Le robot repart donc. Le système détecte que l'action qui devait le mener à la satisfaction d'un but n'aboutira pas et il inhibe donc la motivation qui le pousse à rechercher ce but. Par rapport à un essai où le son a été perçu, cette inhibition arrivera avec un certain retard par rapport au timing attendu de perception du son qui est représentatif d'une certaine tolérance à attendre l'événement prédit. Le robot se concentre ensuite sur d'autres objectifs s'il en a, ou bien explore son environnement.

L'expérience sur robot réel de la section 4.2.2 a été continuée afin de tester le fonctionnement des ajouts au modèle utilisé. On reprend alors le réseau de neurones précédent, avec l'ajout du système de détection de l'échec et du méta-contrôleur. Les cellules de lieux, transitions, actions et carte cognitive sont celles apprises dans l'expérience antérieure. Cette fois, les essais normaux, où le robot reproduit la tâche et perçoit le son après avoir attendu sur le but, sont suivis d'essais d'extinction où aucun son n'est émis. Le système de détection de l'échec s'active alors pour la transition de prédiction du son, peu après la fin du délai de 7 secondes. L'inhibition de la motivation permet au robot de reprendre son mouvement et de quitter le lieu but pour aller explorer l'environnement (fig. 6.6)

6.2.2 Apprentissage d'une ronde avec des points d'arrêt

Le protocole expérimental ayant déjà été mis en place et l'apprentissage effectué, la tâche de navigation continue était idéale pour tester les capacités du système de détection de l'échec. Cette expérience a permis de mettre en évidence le rôle de cette détection dans la résolution de la tâche, notamment pour les essais d'extinction qui ne peuvent être reproduits sans ce système. Cependant le modèle développé dans ce chapitre est générique et permet la résolution de tâches variées possédant des aspects de navigation avec des séquences de comportements sensori-moteurs et des contraintes temporelles. Une application réelle de ce genre de tâche est la mise en place de patrouilles de sécurité automatisées. Dans ce cadre, le robot devrait naviguer vers une série de points de passage. Le robot pourrait avoir à s'arrêter sur certains de ces points pour réaliser une action, comme regarder autour à la recherche d'intrus pendant un certain temps. A la fin de ce délai, il repartirait vers le prochain point.

En utilisant le même réseau de neurones que pour l'expérience de navigation continue, nous avons mis en place une ronde automatisée. Le robot apprend tout d'abord à suivre une certaine trajectoire. Pour cela, il est guidé par un humain grâce à une laisse attachée à un cou artificiel. Le robot ne possède aucune connaissance a priori de l'environnement. La trajectoire apprise forme une boucle, et le robot construit sa représentation de cette trajectoire en termes de lieux, appris grâce à un seuil sur le niveau d'activité des cellules de lieu, et transitions, qui sont liées entre elles dans la carte cognitive. Dès que la boucle est complétée, l'humain lâche la laisse et le robot commence à reproduire la trajectoire en se basant sur ses perceptions visuelles. Dans cette expérience, le bruit sur la sélection de l'action est supprimé ($\epsilon = 0$). En effet, on souhaite ici que le robot commence à reproduire la trajectoire apprise dès qu'on lâche la laisse. La suppression du bruit permet d'éviter que le robot ne quitte sa ronde pour aller explorer son environnement. Même en l'absence d'un biais de la carte cognitive, le système n'a appris qu'une séquence de transition et une seule transition est donc sélectionnée à chaque étape de la ronde. Le robot reproduit le seul chemin qu'il connaît, sans attribuer de valeur particulière aux transitions.

Ensuite, pendant la deuxième reproduction de la ronde, l'humain se place devant le robot pour le faire s'arrêter. L'opération est répétée en 3 endroits, avec des temps de pause différents. Le robot considère alors la cellule de lieu la plus proche comme un point d'arrêt. La précision spatiale d'un point d'arrêt dépend donc de la taille des champs de lieu après compétition. Quand l'humain se retire de devant le robot, celui-ci redémarre et reprend sa patrouille. Une différence fondamentale avec l'expérience de navigation continue est qu'aucun stimulus externe ne marque la fin de la période d'arrêt. Afin que le robot puisse associer cette phase d'arrêt avec une transition entre l'arrivée sur le lieu et un autre événement perceptif marquant la fin de cette phase, il nous faut ajouter une modalité au système d'apprentissage de transitions. Des informations proprioceptives venant de l'odométrie ont donc été rajoutées à l'architecture produisant des événements multi-modaux dans EC. Ici, un événement est perçu quand le robot reprend le mouvement après s'être arrêté. Ainsi, le système de transitions peut prédire la reprise du mouvement en fonction du temps écoulé depuis le dernier événement (par exemple l'entrée sur un lieu particulier). Lors de l'apprentissage des phases d'arrêt, étant donné que le robot était immobile pendant tout le délai d'attente, l'action de ne pas bouger est associée à la transition apprise. La durée du temps d'arrêt est apprise comme une prédiction temporelle. Enfin, chaque reprise de mouvement est accompagnée d'un signal de satisfaction spécifique à chacun des points d'arrêt. Des buts représentant le besoin de vérifier différents points de la patrouille sont donc appris. La carte cognitive sélectionnera donc la transition menant le robot à s'arrêter, et donc à satisfaire

son but, lors de l'entrée sur un point d'arrêt. Après le pic de prédiction, le système de détection d'échec, détectant que le robot n'a pas repris son mouvement, inhibera la motivation et le robot repartira satisfaire le prochain but. Le robot satisfait donc successivement ses buts en s'arrêtant à chaque point d'arrêt.

La figure 6.7 montre les lieux et transitions appris tandis que la figure 6.8 montre les trajectoires prises par le robot pendant les phases d'apprentissage et de reproduction de la ronde. La carte cognitive apprise est très incomplète, le robot ne connaissant que le chemin appris. En cas de kidnapping, le robot sera perdu si il est placé loin des lieux déjà appris. Il pourra cependant retrouver le chemin de la ronde en explorant ou bien en étant guidé à l'aide de la laisse, et intégrera ces informations à sa carte cognitive. Si on laisse le robot construire une carte exhaustive de l'environnement, il empruntera des chemins plus directs entre les points d'arrêt. Il n'y a pas de séquence définie dans le système pour la ronde. Au lieu de cela, le robot essayera de satisfaire de manière optimale ses buts, correspondant aux points d'arrêt. Ainsi, il optimisera sa trajectoire pour vérifier chaque point le plus souvent possible.

En plus de reproduire la trajectoire apprise, le robot s'arrête de manière autonome aux points d'arrêt appris. Le tableau 6.1 indique la durée d'arrêt du robot à chaque point. La durée d'arrêt indiquée pour l'apprentissage est le temps passé par l'humain devant le robot. Les valeurs de reproduction sont moyennées sur 5 reproductions. La première observation est que le robot arrive à reproduire individuellement chaque délai d'arrêt. La reproduction se fait avec des durées d'arrêt stables. Le robot a bien appris à s'arrêter lors de son entrée sur certains lieux et à reprendre le mouvement après un délai spécifique à chaque point d'arrêt. Cependant, on peut observer que les durées d'apprentissage et de reproduction peuvent être différentes. Il y a deux raisons à cela :

1. L'activité de prédiction a un pic qui anticipe le timing appris de reprise du mouvement. En revanche, le signal d'échec ne s'active que quand l'activité de prédiction retombe dans des valeurs plus basses. Cette tolérance à un retard de l'événement prédit déclenche la reprise du mouvement avec un retard sur le pic d'activité et peut retarder le départ du robot. A cela s'ajoute le temps mis par la commande motrice de reprise du mouvement pour être exécutée. Par ailleurs, nos prédictions suivent la loi de Weber [Grossberg and Schmajuk, 1989] : le pic d'activité précédant l'événement attendu est de moins en moins précis quand le délai s'allonge, comme chez les animaux et l'homme. Cela peut expliquer pourquoi les longs délais, qui sont prédit avec une grande anticipation par le système, ont tendance à avoir des temps de reproduction plus courts.

2. Notre système apprend à associer un lieu avec une période d'attente d'une durée spécifique. Le délai appris par le robot est la période entre l'entrée sur le lieu et la reprise du mouvement. Pendant l'apprentissage, tout le temps entre l'entrée sur le lieu et le moment où l'humain vient arrêter le robot est inclus dans le délai d'attente appris et s'ajoute au temps d'immobilité réelle du robot.

Cette seconde raison est liée à une simplification du modèle pour la tâche présentée ici. Elle cache en fait un problème fondamental. En effet, ici le système apprend simplement à lier des signaux d'entrée dans des lieux et des signaux de redémarrage de mouvement. Ainsi, pendant la reproduction, le robot décide de s'arrêter aux points d'arrêt appris afin de satisfaire son but : reprendre son mouvement. Paradoxalement, le robot a appris que l'action de s'arrêter devrait entraîner une reprise du mouvement après un certain délai. Cela n'étant évidemment pas le cas, le système signale l'échec de la reprise du mouvement et inhibe la motivation correspondante,

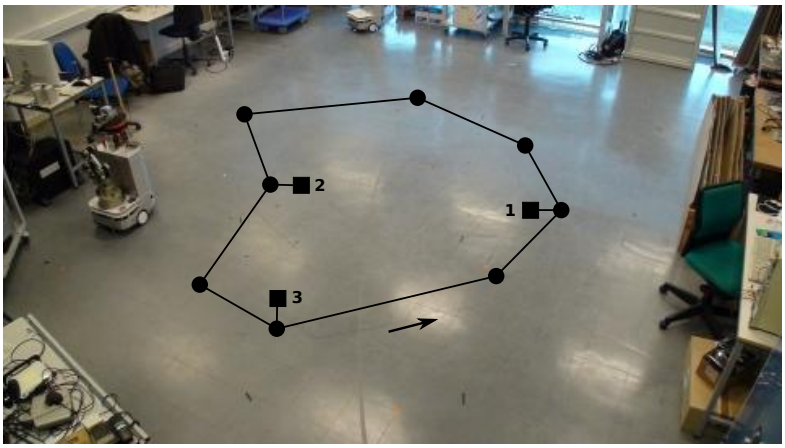


FIGURE 6.7 – Lieux et transitions appris pour la ronde. Les centres des champs de lieux sont indiqués par des disques noirs. Les carrés noirs représentent l'association entre un lieu et un événement proprioceptif. Ils indiquent les 3 points d'arrêt.

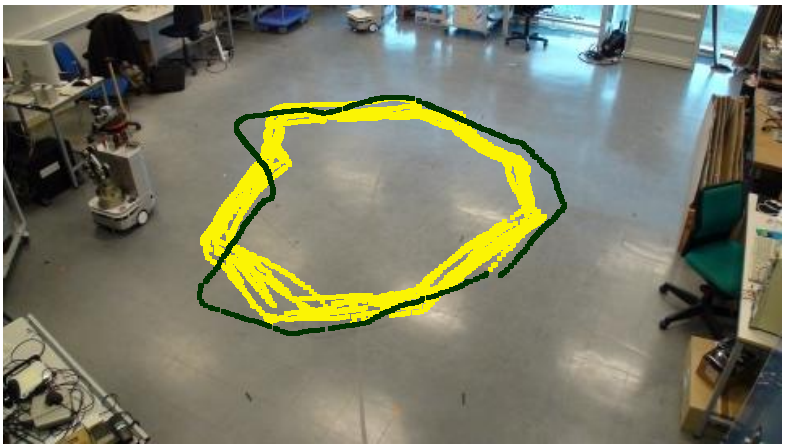


FIGURE 6.8 – Trajectoires du robot pendant la tâche. La trajectoire foncée correspond au chemin appris, quand le robot est guidé par l'humain. Les trajectoires claires sont la reproduction autonome. A cause de la taille des champs de lieu, la boucle reproduite peut être décalée par rapport à l'originale. C'est une conséquence des prédictions de transitions, qui déclenchent le prochain mouvement avant que le robot n'ait atteint le centre du lieu, où le robot avait été guidé. L'architecture offre un compromis entre la puissance de calcul nécessaire et la précision de la ronde en fonction du nombre de cellules de lieu apprises. Seuls les mouvement entre cellules de lieu sont appris. Un petit nombre de lieux entraînera donc une reproduction grossière de la ronde tandis qu'un grand nombre donnera une reproduction très précise mais très gourmande en calculs et mémoire, avec un apprentissage plus long.

TABLE 6.1 – Durée d'arrêt à chacun des points appris.

Point d'arrêt	**1**	**2**	**3**
Délai d'attente appris (s)	24.09	19.10	11.64
Délai reproduit moyen (s)	22.25	19.13	15.30
Ecart-type (s)	0.69	0.33	0.80

ce qui fait repartir le robot. Pour modéliser fidèlement la tâche et éviter de lier directement l'entrée sur le lieu à l'arrêt du mouvement, un événement marquant l'arrêt du robot devrait aussi être émis. Ainsi le robot apprendrait la chaîne d'événements *Lieu→Arrêt→Départ* et la durée d'arrêt apprise correspondrait réellement au temps d'immobilisation par l'humain. Le problème fondamental empêchant pour l'instant un tel fonctionnement est la modélisation des actions associées aux transitions. Dans le cadre de la navigation, l'action permettant de passer d'un lieu A à un lieu B est une direction prise par le robot. Cette direction est intégrée pendant toute la transition entre les lieux A et B. Lors de la reproduction, l'action est exécutée dès l'entrée sur le lieu A, le robot changeant d'orientation pour rejoindre le lieu B. Ici, dans le cadre du contrôle de l'immobilité du robot, on se place plutôt dans un paradigme de séquence d'actions motrices. Le robot exécute une action qui fait changer son état interne et déclenche un événement perceptif. Il apprend cette relation et tente plus tard de reproduire la séquence en exécutant de nouveau l'action. Or l'action (par exemple reprendre le mouvement) n'est exécutée qu'à la fin de la période d'attente. Deux questions se posent alors :

1. L'action associée à une transition (par exemple la transition *Arrêt→Départ*) doit-elle être représentative de l'action effectuée pendant la majorité de la période de transition (rester immobile) ou bien de l'action effectuée juste avant le moment de transition (reprise du mouvement).

2. Lors de la reproduction, l'action associée à une transition doit-elle être effectuée dès le début de la prédiction d'une transition (comme en navigation où l'on veut que le robot s'oriente vers le prochain but) ou bien au moment exact du délai appris pour la transition (comme dans le cas de la reproduction d'une séquence d'actions).

On a donc ici une incompatibilité entre le système de navigation qui fonctionne en intégrant une action moyenne pour une transition et dont la reproduction demande l'exécution de l'action dès le début de la prédiction de la transition, et le système de reproduction de séquences d'actions qui veut associer à une transition la toute dernière commande motrice et ne reproduire cette action qu'à l'instant du pic de prédiction représentant le timing de cette commande motrice. Afin de concilier ces deux approches, une solution serait d'aborder une représentation temporelle des actions, qui permettrait de représenter finement dans le temps l'action nécessaire pour effectuer une transition, plutôt que de devoir choisir entre deux modèles simplifiés de représentation des action qui sont incompatibles. On pourrait envisager l'utilisation d'un contrôle en force avec un homéostat pour fournir une solution à certains des problèmes rencontrés pour le codage temporel des actions.

Finalement, il convient d'ajouter que le système développé apprend en continu, pas uniquement pendant les premières rondes. Il est donc possible pour l'humain de modifier l'apprentissage du robot. De nouveaux points d'arrêt peuvent être ajoutés par l'humain en venant se placer devant le robot. Les délais déjà appris pour les points d'arrêt peuvent aussi être modifiés. On peut les raccourcir en utilisant la laisse pour faire repartir le robot de manière prématurée, ou bien les rallonger en bloquant le robot lorsqu'il tente de repartir. L'architecture va alors apprendre à adapter ses prédictions temporelles afin de prendre en compte ces changements.

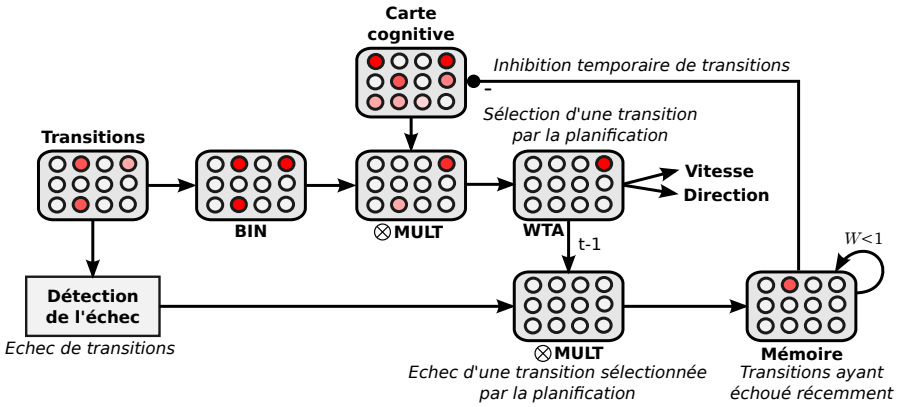


FIGURE 6.9 – Architecture d'inhibition de transitions dans la carte cognitive après échec de celles-ci. Une mémoire stocke le signal d'échec pour les transitions sélectionnées par la carte cognitive et entraîne leur inhibition de manière temporaire.

6.2.3 Navigation multi-buts avec inhibition de chemins

Dans les expériences précédentes, le système apprenait à associer des prédictions avec la satisfaction d'un but, et à inhiber la motivation associée à ce but lors d'un échec. On considérait donc que la stratégie menant à satisfaire ce but avait échoué, et que le but était, au moins temporairement, inaccessible. En inhibant la motivation, le robot pouvait alors s'occuper d'autres objectifs ou bien explorer son environnement. Dans le cas où il existe plusieurs moyens pour satisfaire un même but, ce fonctionnement suppose que l'on abandonne temporairement la volonté de satisfaire le but dès l'échec du premier moyen essayé. Or, le robot pourrait tenter de satisfaire son but en essayant une autre approche que celle qui a échoué. Plutôt que d'inhiber la motivation, on pourrait envisager d'inhiber la stratégie employée, ou encore tout simplement *le chemin qui a échoué*. C'est sur cette dernière méthode que je vais me pencher ici.

L'idée est donc, lors de l'échec d'une action effectuée dans le cadre d'une planification motivée, d'inhiber le maillon de la carte cognitive correspondant à cette action. Ce fonctionnement peut être ajouté de manière très simple au modèle utilisant la carte cognitive et le système de détection de l'échec (fig. 6.9). L'architecture récupère les signaux d'échec de toutes les transitions. Une population de neurones fait l'intégration entre ces signaux et le résultat de la compétition entre les transitions. Ces neurones s'activent lorsqu'ils sont co-activés (ce qui correspond à des neurones multiplicateurs des activités pré-synaptiques). Ainsi, l'activité neuronale de cette population représente le fait qu'une transition qui avait été sélectionnée, et donc dont l'action a été ou est en train d'être effectuée, a échoué. Une mémoire neuronale correspondant à cette transition est donc activée de manière temporaire. Cette mémoire a une influence inhibitrice sur la carte cognitive et va supprimer l'activité de la transition qui a échoué.

Cette inhibition permet donc de supprimer de manière temporaire des transitions dans la carte cognitive. L'activité de motivation de ces transitions est alors nulle. En conséquence, le calcul de la propagation de l'activité dans la carte s'en trouve modifié : l'inhibition temporaire d'une transition va modifier la propagation de l'activité de motivation dans tous les chemins qui menaient à cette transition. La carte cognitive va alors se réorganiser pour proposer des chemins alternatifs, n'utilisant pas la transition qui a échoué. Ainsi, si d'autres chemins vers le but existent, la carte cognitive biaisera le choix des transitions pour emprunter ces chemins.

Dans le cas contraire, si le seul chemin vers le but a échoué, d'autres stratégies de navigation pourront prendre le relais.

Afin de vérifier le fonctionnement de ce modèle, j'ai donc mis en place une version de la tâche de navigation continue indicée où deux lieux buts sont présents dans l'arène. Un but rouge se situe dans le coin nord-ouest tandis qu'un but bleu est dans le coin sud-est. L'expérience a été réalisée en simulation, dans un environnement ouvert. Dans un premier temps, le robot est autorisé à explorer de manière extensive son environnement pour former sa représentation en termes de lieux, transitions et carte cognitive. Il apprend aussi à lier les deux zones de couleur aux lieux voisins dans la carte cognitive. Ces deux zones, de couleurs différentes, sont représentées par des événements perceptifs distincts. Pendant cette première phase, le robot ne s'arrête pas sur les lieux buts et ne perçoit donc pas le son qui constitue son but. Il n'a donc pas fait l'association entre les zones de couleurs et un but quelconque.

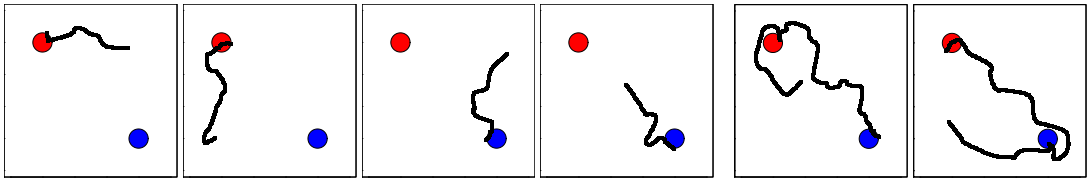
Ensuite, le robot apprend par supervision à rester immobile sur les deux buts. L'expérimentateur fait arrêter le robot sur les zones de couleur l'une après l'autre. Après 4s, un son est émis et le robot reçoit un signal de satisfaction de but. Il apprend donc à associer chaque zone de couleur avec la perception du son, et donc avec un but. L'action associée à cette prédiction est de ne pas bouger. Le son représentant le but est le même dans les deux zones. Il correspond donc au même but, et par extension à la même motivation. L'objectif du robot dans cette tâche est de percevoir le son. Le robot peut indifféremment se rendre sur l'une ou l'autre des zones et s'arrêter pour satisfaire ce but. Dans une première phase de navigation, les deux lieux buts permettent l'émission du son. La motivation monte progressivement jusqu'à atteindre un seuil activant la stratégie de planification, avant que la satisfaction du but ne l'inhibe et déclenche une phase d'exploration temporaire. La motivation est propagée dans la carte cognitive à partir de chaque zone (fig. 6.11a), et le robot va en général se diriger vers le but le plus proche lorsque sa motivation déclenche le comportement de planification (fig. 6.10a). De temps en temps, lorsque qu'une transition est sélectionnée et que la direction prise entraîne le robot dans un autre lieu que celui prédit, la transition va être temporairement inhibée. Cela peut éviter certaines situations où le robot tournerait en rond à cause de l'impossibilité de réaliser une transition car la topologie de l'environnement a changé (par exemple car un obstacle a été introduit).

Dans un second temps, le but rouge est inactivé. C'est-à-dire que le son n'est plus émis lorsque le robot reste sur celui-ci. Par contre, le but bleu est toujours actif. Deux cas peuvent alors se produire :

1. Le robot est plus proche du but bleu et se dirige vers celui-ci. Il s'arrête et perçoit le son, satisfaisant ainsi son but.

2. Le robot est plus proche du but rouge et va s'arrêter sur celui-ci. Peu après le délai de 4s, en l'absence de la perception du son, un signal d'échec est émis pour la transition *Lieu but rouge→Son*. Elle est alors inhibée dans la carte cognitive, qui se met à jour (fig. 6.11b). Le seul moyen de satisfaire le but est alors d'aller sur le but bleu. Le robot s'y dirige donc, s'y arrête 4s et satisfait son but (fig. 6.10b).

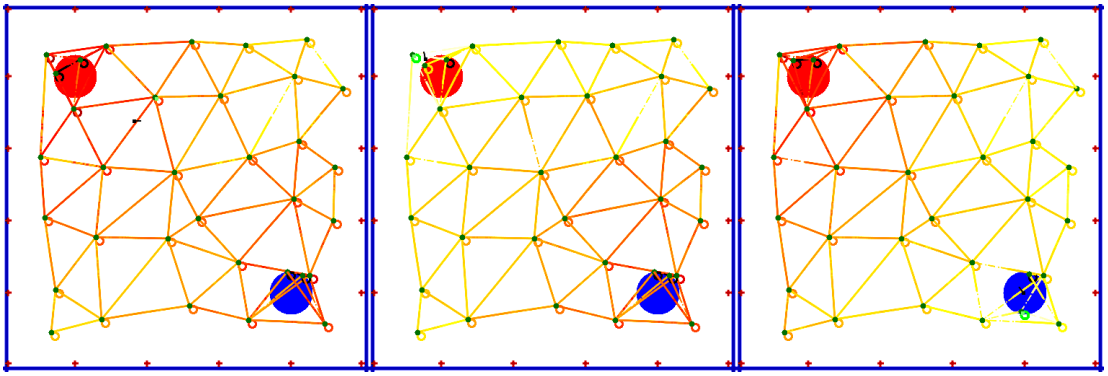
Puis les conditions sont inversées et c'est le but bleu qui est désactivé tandis que le rouge reste actif (fig. 6.11c). Dans ce cas, le système de détection de l'échec permet de moduler le comportement en inhibant les prédictions du son en fonction du lieu où il se trouve. Ainsi, lors d'un échec sur un des lieux, il peut détecter cette défaillance et changer de stratégie afin de satisfaire son but d'une autre façon, en allant sur l'autre zone de couleur. Si la motivation avait



(a) 4 essais avec des points de départ différents. Le robot se dirige immédiatement vers le but le plus proche et perçoit le son car le but est actif.

(b) 2 essais avec des points de départ différents. Le robot se dirige vers le but le plus proche, attend sans succès (but inactif) et repart vers le second but.

FIGURE 6.10 – Trajectoires de navigation motivée vers le but.



(a) Deux buts actifs. **(b)** But rouge désactivé. **(c)** But bleu désactivé.

FIGURE 6.11 – Carte cognitive utilisée par le robot dans le cas où les deux buts sont actifs et dans le cas où un seul but est actif, après la visite du but désactivé. L'échec de la prédiction de la transition associée à un but inactif l'inhibe temporairement dans la carte cognitive et les chemins sont recalculés. Dans le cas où les deux buts sont actifs, le gradient d'activité de motivation dans la carte se propage depuis les deux buts. Après l'inhibition temporaire d'une des transitions, on voit que le gradient ne se propage plus que du but toujours actif.

été inhibée, le système aurait cessé de vouloir satisfaire le but à chaque fois qu'il s'arrête sur un but désactivé. De plus, pendant les phases de navigation, l'échec d'une transition spatiale, si le chemin est obstrué ou si la direction associée à la transition n'est pas pertinente, entraîne une inhibition temporaire dans la carte cognitive. Ainsi le robot peut chercher un autre chemin pour rejoindre le but. La modulation du comportement se fait cependant ici uniquement à court terme, en inhibant une transition de manière temporaire. Le signal d'échec n'est pas directement utilisé pour modifier l'apprentissage afin de moduler le comportement à moyen et long terme. Ainsi le robot continuera de visiter un lieu but qui a été désactivé ou ne prendra pas en compte le fait que la direction associée à une transition le mène souvent ailleurs que sur le lieu attendu. Nous allons donc voir dans la section 6.3 comment ce genre de modulation à moyen terme peut être effectué.

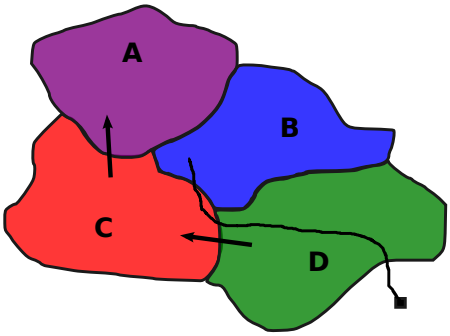


FIGURE 6.12 – Exemple où le robot suit une trajectoire donnée par la carte cognitive (DC, puis CA) mais les directions associées à chaque transition le mènent dans B au lieu de A. Dans le cas où un obstacle est présent entre B et A, les performances du robot peuvent être sensiblement réduites.

6.3 Modulation comportementale à moyen terme

6.3.1 Modulation des associations transition-action

Si la représentation spatiale du robot correspondait à un monde grille, où le robot a le choix entre 4 directions, la sélection d'une action mènerait toujours dans la case attendue. Cependant, les champs de lieux peuvent être irréguliers. De plus, le robot n'apprend qu'un mouvement moyen entre deux cellules de lieu. Il déclenche ce mouvement à partir de son entrée dans le premier lieu, qui peut se faire n'importe où sur le bord du champ de lieu, et la direction apprise ne le mène alors pas toujours vers le lieu visé. Dans certains cas, le robot reprend de manière répétée les mêmes chemins (par exemple pour rejoindre un but). Il suit alors une trajectoire dictée par la topologie apprise de l'environnement en termes de transitions et par la carte cognitive. Il passe alors souvent par la même succession de lieux. Il arrive que la direction apprise pour une transition, correspondant à une moyenne des directions possibles, ne soit pas adaptée et mène le robot de manière répétée vers un lieu inattendu (fig. 6.12).

Le système de détection de l'échec permet de détecter ce genre de situation. L'architecture décrite dans la section 6.2.3 permet d'utiliser le signal d'échec pour inhiber temporairement la transition défaillante dans la carte cognitive, afin que le robot cherche un autre chemin. Cependant aucune modification de l'action correspondant à la transition n'est associée à cet échec. Lors de l'échec d'une transition sélectionnée, il serait possible d'associer à la transition la direction prise comme une direction à éviter. Le système apprendrait alors que cette direction n'a pas permis de réaliser la transition sélectionnée. De manière plus générale, cette association "négative" peut même être faite dès qu'une transition prédite a échoué, qu'elle ait été sélectionnée ou non. Si la transition prédite est réalisée, la direction venant de l'intégration de chemin est associée positivement à la transition. Le robot apprend alors un mouvement qui permet de réaliser cette transition. En revanche, pour les transitions qui étaient prédites mais n'ont pas été réalisées, l'association est faite de manière négative. Le robot apprend alors les mouvements qui n'ont en général pas permis de réaliser la transition. Ainsi le robot apprend à construire à la fois une représentation des actions qui lui permettent de réaliser une transition, mais aussi des actions qui la font échouer.

On peut alors adapter notre modèle d'apprentissage des associations transition-action (voir fig. 6.15 dans la section 6.3.2). Lors de la réalisation d'une transition, l'apprentissage synaptique

est effectué entre le neurone de transition et les neurones codant pour la direction. Cependant les neurones de transition correspondant à des transitions qui ont échoué sont inhibés. Cette inhibition entraîne une dépression de l'activité neuronale, qui est modélisée par une activité négative. Cette activité négative entraîne ensuite le renforcement d'une connexion synaptique inhibitrice vers les neurones de direction associés. On peut noter qu'ici l'apprentissage d'associations négatives et positives est effectué sur un champ unique. Une implémentation plus biologiquement plausible pourrait être de séparer les voies excitatrices et inhibitrices afin de ne pas avoir à travailler avec des activités de neurones négatives. L'équation pour l'apprentissage est la suivante :

$$\frac{dW_{ij}^{TD}(t)}{dt} = \alpha \cdot T(t) \cdot (X_j^{TE}(t) \cdot \sum_k W_{ik}^{ID}(t) \cdot X_k^I(t) - |X_j^{TE}(t)| \cdot W_{ij}^{TD}(t)) \tag{6.7}$$

α est la vitesse d'apprentissage et $T(t)$ le signal indiquant qu'une transition est réalisée. X^{TE} est l'activité des neurones pour les transitions ayant échoué (activité négative permettant l'apprentissage d'une inhibition) ou pour les transitions réalisées (activité positive permettant l'apprentissage d'une excitation). X^I est l'activité des neurones du champ dynamique d'intégration de chemin donnant la direction prise par le robot. La première partie de l'équation correspond à un apprentissage hebbien lors de la co-activation des neurones de transition et de direction. Elle permet la diminution des poids synaptiques (pouvant aller dans des valeurs négatives) lorsque X^{TE} est négatif ou leur augmentation lorsque X^{TE} est positif. La seconde partie permet d'assurer la convergence des poids synaptiques aussi bien dans les valeurs positives que négatives. L'utilisation de la valeur absolue $|X^{TE}|$ dans l'équation permet d'éviter une divergence des poids dans les deux cas.
L'équation de calcul de l'activité correspondante est la suivante :

$$X_i^D(t) = \sum_k W_{ik}^{TD}(t) \cdot X_k^{TE}(t) \tag{6.8}$$

X^D est l'activité des neurones d'un champ représentant les directions associées aux transitions. L'activité peut être négative, pour signaler l'apprentissage d'une inhibition entre la transition et une mauvaise direction, ou positive pour signaler une direction préférée pour effectuer la transition.

Pour vérifier les propriétés de cet apprentissage, nous avons réalisé une expérience en simulation où le robot est laissé en exploration dans son environnement. Nous avons ensuite analysé les directions associées à quelques transitions, après une longue période d'apprentissage (fig. 6.13). Quelques différences sont notables entre l'apprentissage utilisant les signaux d'échec et un apprentissage LMS basé uniquement sur la réalisation des transitions, comme celui utilisé auparavant (fig. 6.14). Les bulles d'activité donnant la direction à prendre avec le LMS sont assez larges alors qu'en pratique elles mènent souvent à un échec de la transition. La dynamique de l'activité des champs de neurones avec le nouvel apprentissage est meilleure, montrant les directions à éviter avec une information sur les échecs liés à ces directions, plutôt qu'une absence d'information comme avec le LMS. L'exploration aléatoire permet ici de traverser les champs de lieux avec des mouvements très variés et d'avoir ainsi une représentation statistique non biaisée des directions à prendre pour effectuer une transition (ou échouer). Dans le cadre d'une navigation planifiée, cet apprentissage permettrait également de mettre à jour l'apprentissage en inhibant une direction prise de manière répétée et conduisant à un échec de la stratégie de navigation.

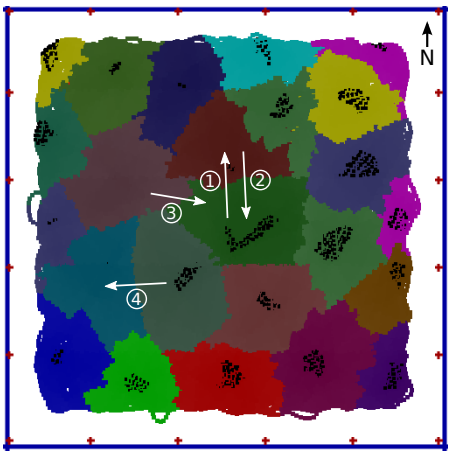


FIGURE 6.13 – Cellules de lieu apprises pendant l'exploration. Chaque couleur correspond à une cellule gagnante. Les points noirs correspondent aux endroits où la cellule répond de façon maximale. 4 transitions sont sélectionnées et une comparaison entre les champs de neurones appris pour la direction associée à ces transitions est montrée dans la figure 6.14.

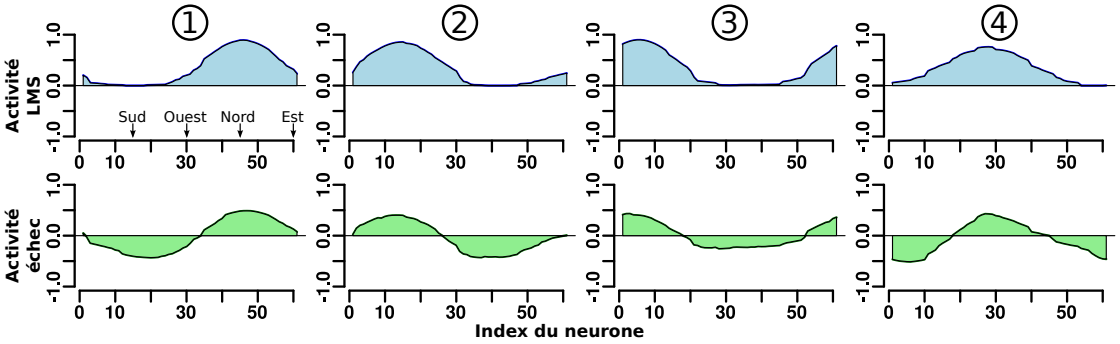


FIGURE 6.14 – Comparaison des directions apprises avec un LMS (en haut) et avec l'architecture intégrant les informations d'échec (en bas). 4 transitions sont montrées à titre d'exemple (voir fig. 6.13). La direction est codée sur un champ de 61 neurones indiquant une orientation préférée sur 360° (précision de 6° par neurone).

6.3.2 Apprentissage autonome de la tâche de navigation continue

Les signaux d'échec peuvent donc être utilisés pour associer une transition à des informations signalant des directions à éviter. Ces directions sont celles qui mènent à un échec de la transition. Cela peut inclure les directions qui mènent à la réalisation d'une autre transition prédite, ou encore les directions menant réalisation de la transition mais avec un temps beaucoup plus élevé qu'habituellement. Un mécanisme d'inhibition est alors appris entre la transition ratée et la direction prise. Cet apprentissage a été réalisé dans la section précédente pour les informations de direction liées à une transition, mais pas pour les informations de vitesse. Pourtant ces informations de vitesse constituent aussi une partie de l'action correspondant à une transition. Toutefois, le mécanisme fonctionnant pour la direction et utilisant le signal d'échec de toutes les transitions ne permettrait pas de faire des associations avec des informations de vitesse. En effet, la réalisation d'une transition en se déplaçant n'implique pas forcément que les autres transitions prédites mais non réalisées devraient être effectuées en restant immobile. Par contre, si une transition est sélectionnée et que son action en termes de vitesse et direction échoue, alors la commande en vitesse pourrait être responsable de l'échec au même titre que la direction. Une architecture où le signal d'échec des transitions *sélectionnées* est utilisé peut donc être envisagée. Dans ce cas, si l'action correspondant à la transition sélectionnée est effectuée et qu'elle échoue, alors on peut associer négativement les informations de vitesse, puisqu'elles n'ont pas permis de réaliser la transition.

L'apprentissage des directions utilise donc tous les signaux d'échec. Par contre, l'apprentissage des vitesses utilise uniquement les signaux d'échec des transitions sélectionnées (fig. 6.15). Dans une situation où le robot attend un événement prédit et effectue l'action associée, l'absence de perception de cet événement va associer négativement l'action effectuée avec la prédiction. Si l'action associée à une transition inclut un codage du mouvement du robot, alors le robot pourrait finir par décider de s'arrêter pour tenter de réaliser une transition connue, mais qui a échoué de manière répétée en se déplaçant. Ainsi ce mécanisme d'apprentissage pourrait constituer l'origine de la décision de s'arrêter dans la tâche de navigation continue. J'ai donc implémenté cette architecture neuronale dans le but d'apprendre à un robot à réaliser la tâche de navigation continue de manière complètement autonome. L'information de vitesse est alors codée par deux neurones : un représentant l'arrêt et l'autre le mouvement. L'association entre les transitions et les informations de vitesse est faite en utilisant les informations d'échec de la transition sélectionnée par la carte cognitive (association négative) en plus des informations sur les transitions réalisées (association positive). L'équation d'apprentissage pour le contrôle du mouvement est la même que pour la direction (6.7).

Une expérience est alors réalisée en simulation. L'environnement simulé possède un lieu but dans le coin nord-ouest. Pendant une première phase d'exploration, aucun son de satisfaction de but n'est émis. Le robot apprend les lieux, transitions et construit sa carte cognitive. Les actions associées aux transitions sont apprises suivant le mécanisme décrit dans la section 6.3.1. On considère que la stratégie par défaut du robot est de se déplacer dans son environnement (pour explorer, trouver de la nourriture etc.). Ainsi, en l'absence d'informations de vitesse provenant de la stratégie de navigation par carte cognitive, la compétition entre l'arrêt et le mouvement est légèrement biaisée en faveur du mouvement. Le robot apprend alors à associer avec toutes les transitions apprises pendant la phase d'exploration l'action de se déplacer (les transitions purement spatiales ne peuvent de toute manière pas être réalisées sans se déplacer). Après l'exploration de l'environnement, le système de simulation du son est activé et le robot apprend à

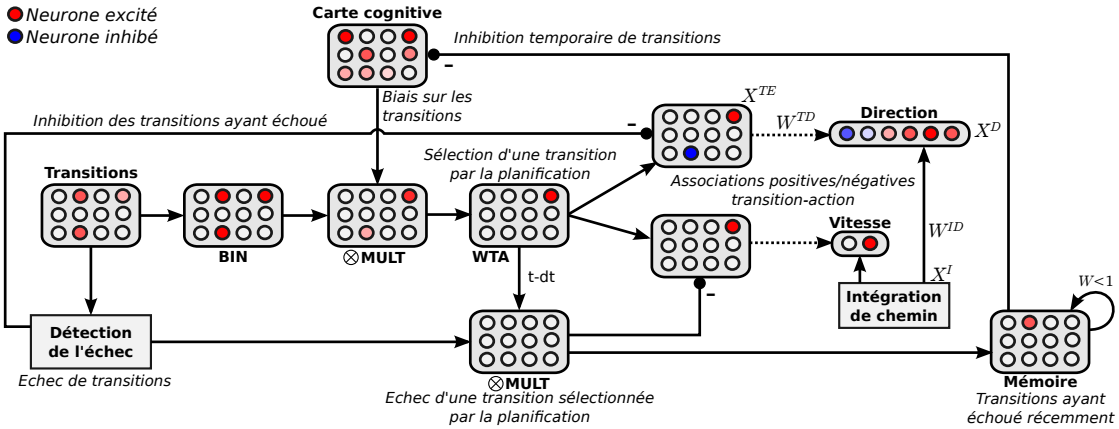
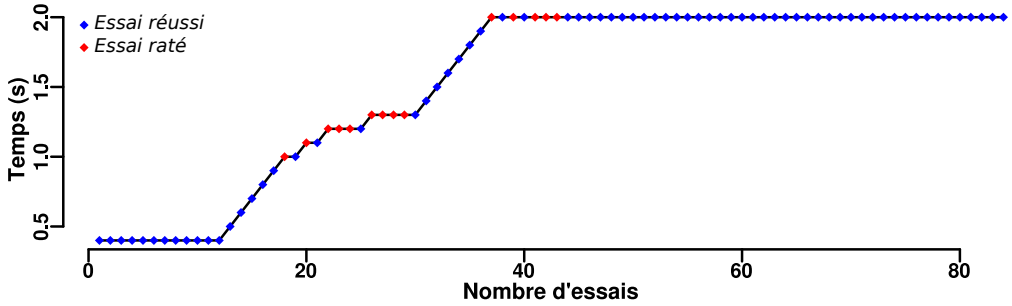


FIGURE 6.15 – Modèle de modulation comportementale utilisant la détection de l'échec. Une modulation immédiate se fait par l'inhibition de la transition ayant échoué dans la carte cognitive. Une autre modulation intervient au niveau de l'apprentissage des associations transition-action. Toutes les transitions ayant échoué se voient associées négativement avec la direction prise par le robot. Les transitions sélectionnées par la carte cognitive ayant échoué se voient associées négativement avec la vitesse du robot.

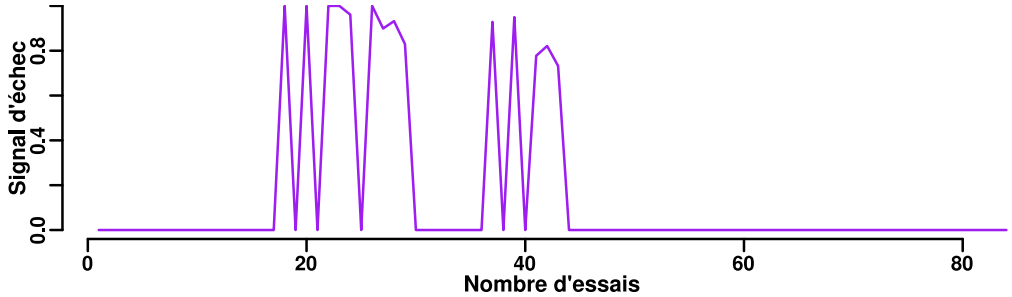
associer la zone de couleur avec la perception du son, qui satisfait son but. Dans un premier temps, le son est émis après 300ms passées sur le lieu but. Cela permet au robot de percevoir le son en restant suffisamment longtemps sur le lieu but même lorsqu'il se déplace. On laisse au robot 10 essais pour consolider ses prédictions sur l'association lieu/son. A ce niveau de l'expérience, le robot se déplace constamment, puisque sa stratégie de déplacement lui permet de satisfaire son but (exploration, satisfaction du but, exploration ... et ainsi de suite).

C'est ici qu'intervient le mécanisme de *shaping* employé par les neurobiologistes pour introduire le délai d'attente. Notre système de production du son est configuré pour augmenter progressivement la durée nécessaire de présence sur le lieu but avant le déclenchement du son. Après les 10 essais réussis à 300ms, cette durée est augmentée de 100ms pour chaque essai réussi, jusqu'à atteindre 2s (fig. 6.16a). On peut voir que jusqu'à 1s le robot parvient toujours à recevoir sa récompense en traversant le lieu but en se déplaçant. Au delà, il commence à échouer. Lorsque le robot est motivé, l'activité dans la carte cognitive permet la sélection de la transition de prédiction du son lorsqu'il arrive sur le lieu but. Mais l'action associée (se déplacer) ne lui permet parfois pas de rester pas assez longtemps dans le but pour percevoir le son. Ces essais sont alors considérés comme des échecs. Dans ce cas, la transition de prédiction du son est inhibée de manière temporaire grâce au système présenté dans la section 6.2.3 et le robot commence une phase temporaire d'exploration avant de retenter de satisfaire son but (car un seul but est présent dans l'environnement).

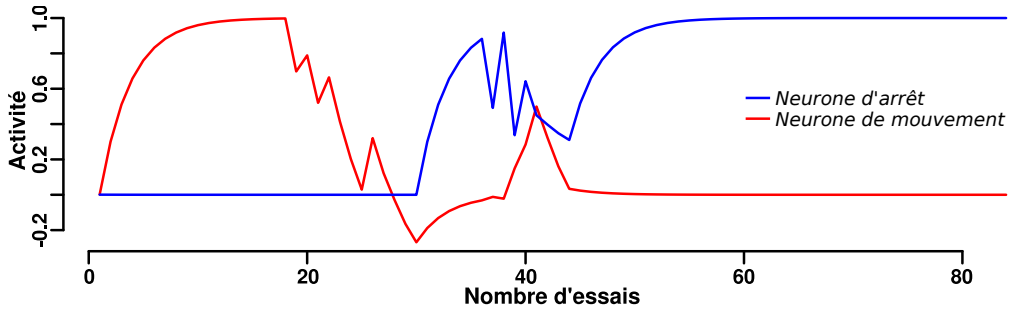
Ces essais ratés déclenchent le signal d'échec pour la transition de prédiction du son (fig. 6.16b). Les informations de vitesse et de direction provenant de l'intégration de chemin sont donc associées de manière négative à la prédiction. La direction est non pertinente dans ce contexte car elle n'influe pas sur le déclenchement du son. Le robot aura tout de même appris une orientation préférée pour attendre le son lorsqu'il est sur le lieu but. Par contre l'information de vitesse est pertinente. Ce signal d'échec va donc permettre l'apprentissage progressif d'une inhibition du neurone de déplacement par le neurone codant la transition de prédiction du son (fig. 6.16c). Au fur et à mesure des essais ratés (entre 1s et 1.4s), les poids synaptiques reliant



(a) Durée du délai nécessaire de présence sur le lieu but avant déclenchement du son. Après une première période de 10 essais à 300ms, le délai est augmenté de 100ms à chaque essai réussi, jusqu'à atteindre 2s.



(b) Signal d'échec pour la transition *Lieu but→Son*. Il correspond aux essais où le robot a quitté le lieu but sans avoir perçu le son, alors que la transition avait été sélectionnée.



(c) Activité des neurones commandant la vitesse du robot lors de la sélection de la transition *Lieu but→Son*. Une compétition WTA permet ensuite de sélectionner l'action à effectuer. La première phase sans délai pour la perception du son permet au robot de satisfaire son but en se déplaçant, d'où l'association entre la transition et le neurone de mouvement. Puis les signaux d'échecs successifs font retomber l'activité du neurone de mouvement, jusqu'à ce qu'il inhibe le comportement par défaut de mouvement. Les réussites suivantes permettent d'associer l'arrêt du mouvement à la transition, afin de satisfaire le but. Des échecs ont également lieu aux alentours de l'essai 40, dus à la tendance du robot à prédire de manière prématurée la perception du son car il n'a pas encore adapté son timing de prédiction au délai grandissant.

FIGURE 6.16 – Résultats expérimentaux de l'expérience de navigation continue apprise de manière complètement non supervisée. Chaque essai correspond à une entrée du robot sur le lieu but, avec la motivation active, et la transition de prédiction du son sélectionnée afin de tenter de satisfaire le but (donc non inhibée par le mécanisme de modulation temporaire). La perception du son marque un essai réussi.

la transition au déplacement vont diminuer. L'apprentissage précédent d'une excitation va être oublié puis la connexion va devenir inhibitrice. Dans un système biologique, le traitement des associations excitatrices et inhibitrices serait effectué dans des circuits parallèles. Ici, on modélise ces associations sur une connexion synaptique unique.

Une fois que l'aspect inhibiteur des connexions synaptiques est devenu assez fort, il va compenser le léger biais du système de sélection d'action pour la stratégie de mouvement. La compétition entre les actions d'arrêt et de mouvement va alors être en faveur de l'arrêt. Lors des prochains essais, l'action associée à la transition de prédiction du son est donc une forte inhibition du mouvement, qui déclenche l'arrêt du robot. Cet arrêt menant à la perception du son, l'action de ne pas bouger va alors être renforcée. C'est ainsi que le robot peut apprendre de manière autonome à s'arrêter : tout d'abord en inhibant la stratégie de mouvement car elle ne permet plus de mener à une satisfaction de but attendue, puis en renforçant l'arrêt du mouvement car celui-ci mène à cette satisfaction. Le shaping est indispensable pour cet apprentissage. En effet, si le délai de 2s était introduit brusquement ou si il était augmenté trop rapidement, le robot pourrait apprendre à s'arrêter mais le signal d'échec lié à ses prédictions temporelles le ferait repartir trop tôt en inhibant la transition de prédiction du son. C'est ce qui arrive dans cet expérience autour de l'essai 40, où le robot repart de manière prématurée car les prédictions temporelles n'ont pas eu le temps d'apprendre le nouveau timing de la transition. Dans ce cas, les timings précédents beaucoup plus courts qui ont été appris entraînent un déclenchement du signal d'échec qui fait repartir le robot prématurément. Il est donc primordial de faire évoluer le délai du son de manière suffisamment lente pour que les prédiction temporelles puissent s'adapter. Heureusement, après quelques échecs successifs, notre système de détection de l'échec devient plus tolérant sur les retards par rapport à la prédiction, permettant ainsi au robot de continuer à pouvoir réaliser la tâche, et lui laissant le temps d'adapter son apprentissage.

Au final, l'action d'arrêt du mouvement a été renforcée et associée à la prédiction du son. Le timing du délai de 2s est appris par une adaptation des prédictions temporelles et le robot réalise finalement parfaitement la tâche, sans essais ratés. Dans le cadre de cette expérience, le codage des actions est assez simple et adapté aux demandes de la tâche. Des associations directes sont faites entre les transitions et des neurones codant l'arrêt ou le mouvement du robot. En réalité, il est probable qu'un niveau intermédiaire existe. Des travaux futurs pourraient intégrer ce niveau intermédiaire afin de pouvoir associer une transition à des actions plus complexes. Un robot pourrait par exemple se déplacer à un endroit et prendre le temps de déplier un bras robotique avant de procéder à la suite de la tâche.

6.4 Discussion

Le système de détection de l'échec présenté dans ce chapitre permet de donner des capacités d'auto-évaluation au robot. Ainsi les prédicteurs que constituent les neurones de transition sont partie prenante du comportement mais servent aussi d'outils d'évaluation. Toute déviation des conséquences attendues de l'exécution d'une action est détectée comme un échec du comportement du robot. Nous avons montré que ce signal d'échec peut-être utilisé de multiples façons pour adapter le comportement en conséquence. On peut tout d'abord moduler immédiatement le comportement pour inhiber les facteurs (motivation, transition) ayant mené à la sélection de l'action qui a échoué. Ainsi, le robot ne reste pas bloqué dans l'exécution d'un comportement erroné. On peut aussi utiliser ce signal pour modifier l'apprentissage du robot, afin de supprimer

des associations apprises n'étant plus pertinentes. Dans ce cas, on adapte à plus long terme le comportement aux conditions changeantes d'une tâche. Le modèle inclut un mécanisme d'apprentissage par renforcement associant des valeurs de récompenses attendues aux transitions. Il permet donc d'expliquer le rôle des neurones dopaminergiques dans le calcul d'une erreur de prédiction. D'un autre côté, les hypothèses concernant le rôle des ganglions de la base dans la détection de nouveauté exprimées dans ce chapitre se rapprochent fortement de celles de [Redgrave and Gurney, 2006]. Dans cet article, les auteurs mettent en avant l'activation des neurones dopaminergiques dans le cas où des événements sont liés à des récompenses attendues ou inattendues, mais également lorsque des événements imprévus qui ne sont associés à aucune récompense surviennent. Il discutent la réception par les ganglions de la base de trois types d'informations contextuelles : une représentation sensorielle de l'événement inattendu, des informations liées à l'état sensoriel, métabolique et cognitif de l'animal et des informations sur les décisions comportementales et commandes motrices. Plutôt que de correspondre uniquement à une erreur de prédiction de récompenses, le rôle de l'activation phasique des neurones dopaminergiques serait alors de permettre l'apprentissage d'un contexte perceptif et comportemental ayant mené à un événement inattendu. Cet apprentissage serait facilité par l'alignement temporel de des différents signaux lors de la courte activation des neurones dopaminergiques. Cet apprentissage est supposé être utilisé aussi bien pour apprendre les comportements menant à des événements liés à des récompenses que pour éviter les comportement menant à des événements aversifs. Ces hypothèses se retrouvent dans notre modèle. L'apprentissage modulé par le signal dopaminergique permet d'associer des perceptions à des actions de manière positive ou négative. Les informations perceptives sont fournies par les neurones de transitions. L'activité des neurones dopaminergiques dépend de la capacité du système à détecter un événement inattendu, ou à détecter l'absence d'un événements attendu.

Plusieurs exemples d'inhibition immédiate du comportement utilisant cette détection ont été montré : inhibition de la motivation correspondant à la satisfaction de but qui a échoué, inhibition de la transition qui a échoué dans la carte cognitive. Dans le cadre d'une expérience où le robot disposerait de plusieurs stratégies de navigation pour se déplacer entre plusieurs buts de types variés (eau, nourriture etc.), les choix d'inhibition en cas d'échec seraient multiples pour le système : inhibition de la transition, motivation, but, stratégie de navigation etc. Dans les expériences présentées ici, le choix du type d'inhibition effectuée est fait de manière ad-hoc, en fonction de la tâche. Cependant, pour aller vers un système plus générique capable de résoudre une grande variété de tâches, ce choix devrait être fait par un système de méta-contrôle. C'est ce système qui prendrait en entrée les signaux d'échec et de satisfaction de but et qui déciderait de l'inhibition éventuelle d'une stratégie, but, motivation etc. Des travaux ont été réalisés dans cette optique par [Hasson and Gaussier, 2010], dans une tâche avec plusieurs buts de types différents et deux stratégies de navigation (proprioceptive et visuelle). Dans ces travaux, le système peut inhiber la stratégie de navigation, la motivation, ou le lieu but. Cependant, la détection de l'échec se base sur des informations de distance au but. Si la distance au but ne peut être réduite, une intégration temporelle finit par déclencher un signal de frustration. L'intégration temporelle est réglée par l'expérimentateur et représente une durée acceptable de persévération du robot avant de passer à un autre comportement. Avec le système développé dans cette thèse, les prédictions temporelles permettent de connaître exactement l'instant auquel le robot peut décider que le comportement est en échec, ainsi la "frustration" du robot peut être déclenchée en se basant sur ses expériences passées, et non sur une intégration temporelle fixée. De plus, la question du choix

de la cible de l'inhibition du méta-contrôleur reste en suspens. De futurs travaux pourraient se focaliser sur l'utilisation de mécanismes permettant d'évaluer individuellement chaque stratégie, but ou motivation afin de choisir intelligemment comment adapter le comportement.

Quand le signal d'échec est utilisé pour inhiber une transition dans la carte cognitive, cette dernière est remise à jour et s'adapte pour proposer de nouveaux chemins n'incluant pas la transition problématique. On peut concevoir ce mécanisme comme une "re-planification". Savoir quand re-planifier pour adapter le comportement du robot en fonction de changements dans l'environnement ou les conditions d'une tâche est une question qui touche de nombreux domaines de la robotique. En effet, la planification peut-être coûteuse en termes de temps et de puissance de calcul, et une re-planification constante est impensable dans de grands espaces d'états. En robotique classique, ces mécanismes de re-planification ont été implémentés sur des systèmes robotiques avec de nombreux degrés de liberté contrôlés par un ensemble de planificateurs *anytime* [Liu and Wan, 2010] ou bien sur des robots mobiles contrôlés par des cartes *growable costmaps* [Philippsen et al., 2008]. En revanche, dans ces systèmes des heuristiques définies par l'expérimentateur en fonction de la tâche sont nécessaires pour savoir quand re-planifier. D'un autre côté, notre système permet de fournir des informations sur le besoin d'une re-planification en fonction de ses expériences passées et de la détection de changements dans la réussite de son comportement. Ainsi, le besoin d'avoir un modèle explicite de la tâche est supprimé.

Enfin, notre système nous a permis d'accroître les capacités d'autonomie du robot. Il peut explorer son répertoire d'actions et adapter son comportement en fonction des succès ou échecs de celles-ci. L'absence de supervision (en dehors de la supervision indirecte de l'expérimentateur qui a conçu le mécanisme de shaping pour faciliter l'apprentissage) montre bien que le robot est capable d'acquérir les compétences de résolution de tâches différentes par lui-même. Jusqu'ici, le type de tâche (navigation vers un ou plusieurs but avec des périodes d'attentes) est resté relativement similaire et le système possédait uniquement des entrées pertinentes pour la tâche (détection de couleur, vision et détection d'amers, perception de son). Dans le but de rendre la modélisation du système vraiment indépendante de la tâche à effectuer et tirer parti des capacités de généralisation de l'architecture, le nombre d'entrées pourrait être multiplié mais celles-ci seraient sélectionnées en fonction de leur pertinence pour chaque tâche. Nous allons voir et discuter dans le chapitre 7 comment le modèle pourrait être ainsi généralisé.

Publications personnelles

Hirel, J., Gaussier, P., and Quoy, M. (2011a). Biologically inspired neural networks for spatio-temporal planning in robotic navigation tasks. Accepté pour publication à ROBIO 2011

Communications dans des ateliers de travail internationaux

Hirel, J., Quoy, M., and Gaussier, P. (2011b). Biologically plausible neural network for spatio-temporal robotic navigation. In *International Workshop on Bio-Inspired Robots*

Les machines un jour pourront résoudre tous les problèmes, mais jamais aucune d'entre elles ne pourra en poser un !

– Albert Einstein

CHAPITRE 7

Vers un modèle générique d'apprentissage de tâches robotiques

J'ai montré au travers des chapitres précédents que l'architecture développée pendant cette thèse est capable de résoudre une variété de tâches robotiques. Ces tâches ont jusqu'ici été inscrites dans le contexte de la navigation en robotique mobile que ce soit lorsque le robot doit naviguer vers un ou plusieurs buts de manière optimale, ou lorsqu'il doit suivre une trajectoire apprise. L'objectif dans ce chapitre sera de démontrer que cette architecture ne se limite pas à la résolution de tâches de navigation. En effet, le principe du modèle de transitions repose sur la capacité à prédire la relation temporelle entre des événements perceptifs. Ces événements peuvent résulter d'un changement d'état du robot basé sur des informations sensorielles ou proprioceptives. Ces prédictions sont ensuite utilisées par un système de planification qui permet au robot de sélectionner une action à effectuer. Le codage des actions peut être indépendant de la sélection des transitions. Ainsi même si, suivant le type de plate-forme robotique, le codage des états et actions du système peut être très variable, le principe du modèle de planification reste viable pour contrôler le comportement du robot afin d'apprendre à réaliser des tâches.

Au fur et à mesure que la complexité des plates-formes robotiques contrôlée par l'architecture augmente, le nombre de plus en plus élevé de signaux rend la catégorisation des états du système plus complexe. Les expériences réalisées jusqu'ici parvenaient à contourner cette complexité en n'utilisant que les signaux pertinents à la réalisation de la tâche, les autres signaux n'étant pas transmis au système de planification. Afin de pouvoir généraliser l'utilisation du modèle à des plates-formes robotiques complexes, il convient de doter le système d'une capacité à filtrer les informations pertinentes pour chaque tâche à réaliser. Cette capacité permettrait alors d'avoir une architecture vraiment générique, prenant de nombreuses entrées sensorielles et les sélectionnant en fonction de la tâche à réaliser.

Je montrerai donc dans ce chapitre la première étape d'un travail de recherche visant à rendre le modèle complètement générique et capable de résoudre une grande variété de tâches en utilisant diverses plates-formes robotiques. Dans un premier temps, je montrerai comment la planification utilisant l'apprentissage de transitions et la carte cognitive peut être utilisée pour contrôler un bras robotique manipulant des objets. Je discuterai ensuite des possibilités d'intégration d'un système de contrôle unifié d'un bras robotique sur un robot mobile en présentant des travaux préliminaires réalisés dans le cadre d'un suivi d'objet. Enfin je donnerai des pistes sur les mécanismes pouvant servir à sélectionner des signaux pertinents pour la construction des transitions,

notamment le *patterning* [Bellingham et al., 1985], avant de discuter des perspectives de travail et applications du modèle développé dans cette thèse.

7.1 Généralisation à des plates-formes robotiques variées

7.1.1 Planification avec un bras robotique

Les travaux présentés ici ont été réalisés en collaboration avec Antoine de Rengervé qui travaille sur des aspects de contrôle moteur d'un bras robotique et l'apprentissage de comportements par imitation visio-motrice immédiate ou différée [Rengervé et al., 2010]. L'objectif est de montrer que l'architecture pour l'apprentissage de transitions et de la carte cognitive peut être utilisée pour contrôler un bras robotique [Rengervé et al., 2011a]. La carte est alors utilisée pour planifier des trajectoires dans l'espace moteur du bras afin de trouver le chemin le plus court pour aller atteindre un but, tel qu'un objet à saisir.

La plate-forme robotique est ici nettement différente d'un robot mobile. Elle est constituée d'un bras Katana possédant 6 degrés de liberté et d'une caméra couleur montée sur un support fixe (voir annexe C). Cette caméra permet au robot de voir une partie de l'espace de travail du bras (l'espace de travail constitue la partie de l'espace que le bras peut occuper). Les commandes motrices liées au actions sont différentes d'un robot mobile (chaque degré de liberté du bras reçoit une commande en force). Ainsi, il a été nécessaire d'adapter une partie de l'architecture utilisée précédemment pour le contrôle du bras (fig. 7.1). Ces modifications concernent la catégorisation d'états en utilisant des entrées proprioceptives et le codage des actions, ainsi qu'un système de catégorisation visuelle pour les buts, mais la planification et l'apprentissage de transitions reste identique à celui utilisé en navigation. Le système de détection de l'échec a été mis de côté, afin de montrer la validité de l'approche avec une architecture de planification minimale dans un premier temps. Sans ce système, l'information précise du timing des transitions n'est donc pas utilisée ici.

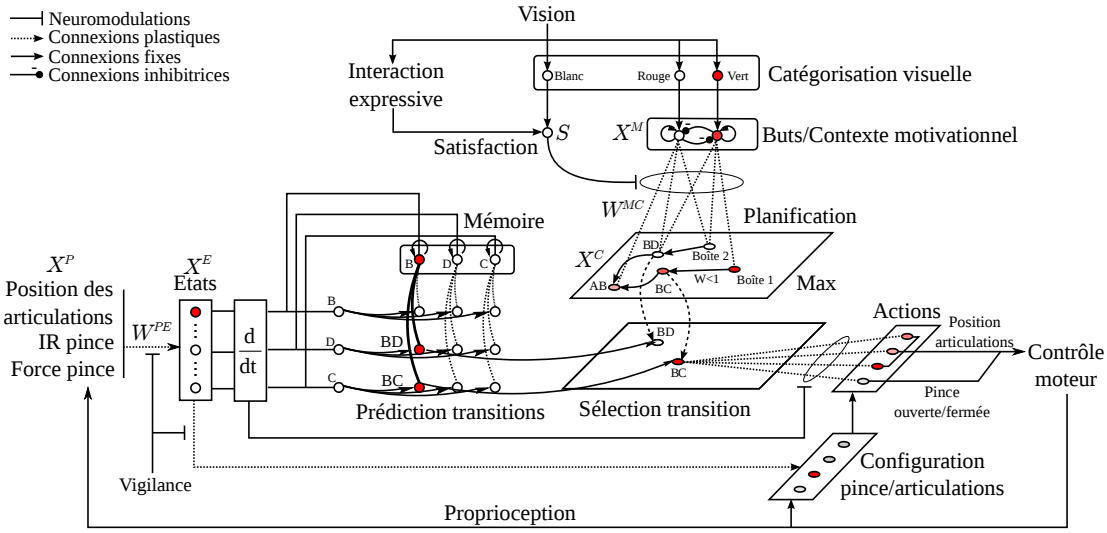


FIGURE 7.1 – Architecture globale de catégorisation d'états, apprentissage de transitions, catégorisation de buts, planification et sélection de l'action utilisée avec le bras robotique.

D'un point de vue neurobiologique, le modèle considère que certaines structures pourraient jouer un rôle à la fois dans la navigation et le contrôle moteur lié à la manipulation d'objets. La catégorisation des états ferait appel à des structures différentes dans le cas de la navigation (cortex entorhinal avec ses cellules de lieu) et du contrôle d'un bras (cortex moteur pour des états proprioceptifs). Cependant, l'hippocampe servirait dans les deux cas à détecter des changements d'états et les relations temporelles entre ces états. Les cortex préfrontaux et pariétaux seraient, quant à eux, impliqués dans la planification.

Les événements perceptifs à la base de l'apprentissage des transitions sont ici des changements d'états proprioceptifs. Les états sont catégorisés en fonction de signaux comportant les positions des articulations du bras, la détection d'un objet présent dans la pince grâce à un capteur infrarouge et des information de capteurs de force dans la pince permettant de savoir si un objet est saisi. Chacun de ces états est catégorisé grâce à un système de recrutement de neurones dépendant d'un seuil de vigilance. Cet algorithme est inspiré du système ART (Adaptive Resonance Theory) [Carpenter and Grossberg, 2002]. L'équation d'apprentissage est la suivante :

$$\frac{dW_{ij}^{PE}}{dt} = R_i^E(t) \cdot X_j^P(t) \tag{7.1}$$

R_i^E un signal de recrutement passant à 1 puis repassant à 0. X^P est l'activité des neurones procurant les informations proprioceptives.
L'équation de calcul de l'activité X^E des neurones d'états proprioceptifs est :

$$X_i^E(t) = 1 - \frac{1}{N} \sum_j |W_{ij}^{PE}(t) - X_j^P(t)| \tag{7.2}$$

N est la dimension de la couche d'entrée P. Le recrutement d'un état X_i^E dépend de la différence entre la position actuelle Θ_i de chaque articulation et la position θ_i de la configuration prototypique correspondant à l'état le mieux reconnu. Si la différence est au dessus d'un seuil de vigilance ν^E, le recrutement d'un nouveau neurone est déclenché. Le choix du seuil de vigilance, comme pour les cellules de lieu, détermine la granularité des états dans l'espace des configurations (proprioceptif ici). Lors de son recrutement, chaque état est associé à une configuration des articulations et de la pince. L'activité d'une neurone d'état est une mesure de distance par rapport à la configuration prototypique apprise pour cet état. Un événement perceptif, correspondant au nouvel état, est émis lorsque l'état le plus actif change. Le modèle de l'hippocampe apprend donc les transitions entre les états proprioceptifs du bras et la carte cognitive apprend à relier ces transitions.

Comme dans l'architecture utilisée en navigation, un signal de satisfaction peut être émis par le système. Le dernier événement perçu est alors associé à un contexte motivationnel. Plutôt que d'associer une motivation à un niveau de besoin augmentant avec le temps, le contexte motivationnel est ici régulé par le type d'objet manipulé. La vision focale du système robotique catégorise la scène en utilisant une détection de couleur. Deux types d'objets peuvent être reconnus : vert ou rouge. Quand un objet rouge est présenté au système visuel, le contexte motivationnel (ou but) rouge est activé jusqu'à ce qu'un autre objet soit présenté. Si le nouvel objet est vert, le contexte motivationnel rouge est inhibé et le vert activé. Ce mécanisme représente donc une mémoire de travail. Le système peut aussi reconnaître si une grande quantité de blanc est présente dans la scène visuelle. Quand l'humain présente à la caméra une feuille blanche sur

laquelle est dessiné un visage souriant, il signale que la dernière transition effectuée par le robot (et donc la dernière action) a permis de satisfaire le but actif (rouge ou vert). C'est un moyen d'utiliser une interaction visuelle pour fournir des signaux de satisfaction de but au système et d'utiliser ce renforcement pour associer les contextes motivationnels à des transitions dans la carte cognitive. Dans cette expérience, nous utilisons donc un système simple de catégorisation visuelle. Nous considérons que la catégorisation visuelle des objets en contextes et les signaux de satisfaction ont déjà été appris. L'apprentissage des associations entre buts et transitions dans la carte cognitive est régi par une équation de Hebb modifiée :

$$\frac{dW_{ij}^{MC}}{dt} = S(t) \cdot X_j^M(t) \cdot (\alpha \cdot (1 - W_{ij}^{MC}(t)) \cdot X_i^C(t) - \gamma) \tag{7.3}$$

α est une vitesse d'apprentissage, γ un facteur d'oubli actif et $S(t)$ le signal binaire indiquant qu'un but a été satisfait. Les X^M sont les activités des neurones de contexte motivationnel et X^C les activités des neurones de la carte cognitive. Le terme $(1 - W_{ij}^{MC}(t))$ permet la convergence des poids vers 1 quand les neurones de motivation et de la carte cognitive sont co-activés. Le terme d'oubli intervient lorsque la motivation est activée et que le but est satisfait, mais que le neurone X_i^C dans la carte cognitive n'est pas activé (une autre transition a mené à la satisfaction). Cette équation permet d'apprendre des associations de manière hebbienne et fait tendre les poids synaptiques vers 1. Elle permet également d'oublier des associations précédemment apprises mais ne se produisant plus.

L'activité d'un but se propage dans la carte cognitive qui propose alors les chemins optimaux pour rejoindre ce but, en termes de nombre de transitions dans l'espace articulaire. La sélection d'une transition par la carte cognitive entraîne l'exécution de l'action associée, qui est la configuration prototypique de l'état prédit par la transition. En effet, lors de la réalisation d'une transition, la configuration proprioceptive qui avait été apprise lors du recrutement du nouvel état est associée à la transition. Ainsi la sélection d'une transition fournira une configuration proprioceptive à atteindre au système de contrôle moteur (fig. 7.2).

L'espace proprioceptif du bras pourrait être exploré de manière autonome afin de construire une carte cognitive complète. A la différence du robot mobile qui évolue dans un espace 2D, le bras robotique évolue dans un espace proprioceptif 6D. L'exploration de cet espace peut donc être très longue et coûteuse. Elle peut mener à l'apprentissage d'un graphe complexe où le nombre de voisins de chaque état serait important. Un grand nombre de neurones de transition pourrait alors être nécessaire. Cependant, on peut utiliser le paradigme de l'apprentissage par démonstration [Billard et al., 2008] afin de profiter des capacités d'apprentissage rapide de la carte cognitive. L'interaction Homme-Machine permet alors de faciliter l'apprentissage de la tâche. Dans la section 6.2.2, un robot mobile apprenait en une seule démonstration comment suivre une trajectoire de ronde, en étant guidé par un humain avec une laisse. De même, avec un bras robotique, la manipulation passive du bras par l'humain peur servir à apprendre au système robotique une trajectoire dans son espace moteur. L'acquisition de cette trajectoire par le système de carte cognitive se fait alors dès la première démonstration. De manière identique, durant les premières années de l'enfance chez l'humain, la manipulation passive est un outil permettant aux parents de guider la perception et l'attention d'un enfant pour faciliter son apprentissage [Zukowgoldring and Arbib, 2007].

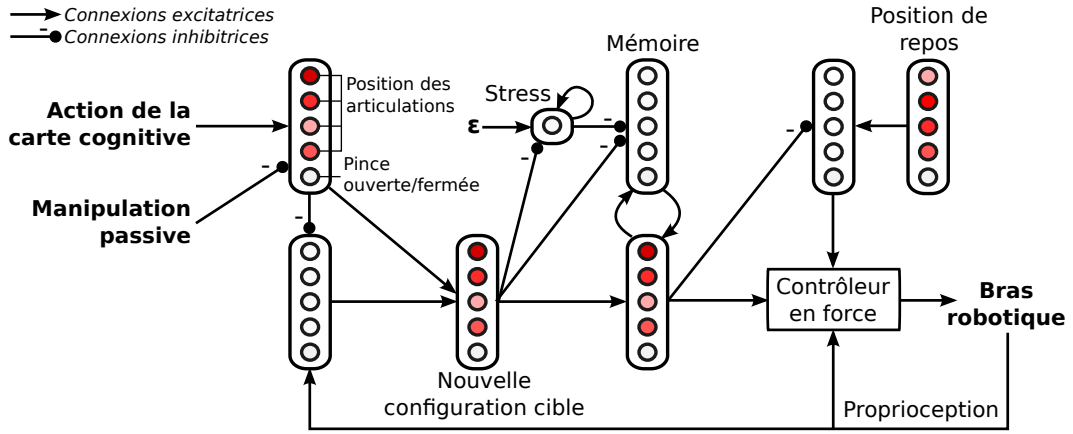


FIGURE 7.2 – Architecture détaillée du contrôle moteur. Le système de contrôle moteur reçoit des configurations cibles de la carte cognitive. Si le bras est manipulé par un humain (en saisissant le poignet du bras), cette action est inhibée. Sinon l'action est transmise à un contrôleur en force qui calcule les commandes motrices à envoyer au bras en comparant la configuration cible avec la configuration actuelle. Quand le bras n'est pas manipulé et que la carte cognitive ne permet pas de sélectionner une transition (et donc une action), le système maintient la dernière configuration cible stockée dans une mémoire. Tant que le système ne reçoit aucune commande, un niveau de stress augmente jusqu'à attendre un niveau d'activité suffisant pour inhiber la mémoire. Un processus réactif de bas-niveau va alors générer une commande en position pour renvoyer le bras vers une position de repos.

Apprentissage d'une tâche de tri d'objets

Le modèle a été implémenté pour réaliser une tâche de tri d'objets. Le système robotique est composé d'une caméra orientée vers un point de collecte et d'un bras robotique Katana à 6 degrés de liberté. Nous supposons que l'attention visuelle du robot est focalisée en permanence sur le point de collecte. Ce point est situé dans l'espace de travail du bras et une canette rouge ou verte peut y être déposée par l'humain. L'objectif est de saisir la canette et de la déposer dans une des deux boîtes situées à des positions fixes dans l'espace de travail du bras (fig. 7.3).

Dans cette tâche, le robot n'a pas de connaissance a priori de l'environnement. Il commence en position de repos et maintient cette position. L'humain saisit ensuite le poignet du robot et le bouge pour lui apprendre par démonstration comment atteindre le point de collecte avec sa pince. Le robot détecte l'effort sur son poignet et passe en mode passif tant que l'expérimentateur force sur le poignet. Pendant ce mouvement, des états proprioceptifs sont recrutés, les transitions entre ces états apprises et reliées dans la carte cognitive. Le premier objet est une canette rouge. Quand le robot détecte la présence d'un objet dans sa pince (ce qui mène à la catégorisation d'un nouvel état), un comportement réflexe de préhension se déclenche et le robot saisit l'objet. Les capteurs de force sur la pince détectent maintenant qu'un objet est saisi et un nouvel état est catégorisé. Le bras est ensuite de nouveau manipulé par l'expérimentateur et amené au dessus d'une des boîtes, apprenant ainsi la trajectoire montrée par l'humain entre le point de collecte et la boîte. Un capteur situé sur la pince commande son ouverture. Le capteur est utilisé pour simuler la manipulation passive de la pince par l'humain, qui forcerait son ouverture (des limitations mécaniques empêchent cette manipulation dans notre cas). Dès que l'objet a été lâché dans la boîte, un signal de satisfaction est transmis au robot en présentant une feuille blanche au système visuel. Ainsi, quand le robot lâche la canette rouge, le but correspon-

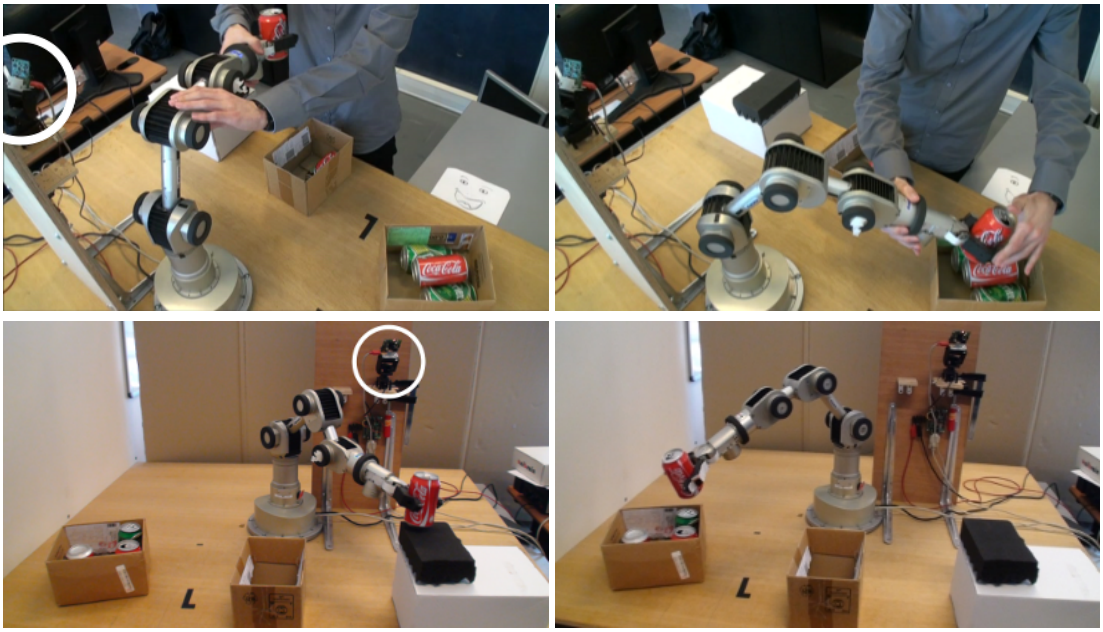


FIGURE 7.3 – Exemple d"une tâche de tri d'objets grâce à un guidage manuel du bras par l'humain. Suite à une démonstration par un humain pendant laquelle le robot est forcé à déposer des canettes rouges dans la boîte 1, le robot apprend une "carte cogntive" (séquence) qui lui permet de reproduire la tâche. La caméra est orientée vers le point de collecte pour reconnaître la couleur de la canette.

dant (rouge) est associé à la dernière transition réalisée (c'est-à-dire changer l'état de la pince de fermée à ouverte à ce point précis de l'espace articulaire). Le robot attend ensuite jusqu'à ce que l'humain lui montre comment retourner de la boîte à sa position de repos ou que son niveau de stress (qui s'accumule du fait qu'il ne reçoive aucun ordre du système de planification) l'y fasse retourner automatiquement. Dès que la carte cognitive a appris une boucle complète reliant la position de repos, le point de collecte et le point de dépôt, le robot peut planifier sa séquence de transitions pour atteindre le point de collecte où il avait précédemment saisi un objet. Le robot commence ainsi à reproduire la tâche sans intervention de l'humain. Etant donné qu'un des buts est toujours actif (en fonction de la couleur du dernier objet perçu), le robot va constamment essayer de satisfaire ce but, et donc déposer un objet dans la boîte correspondante.

Depuis la position de repos, la carte cognitive donne le chemin le plus court pour satisfaire le but (déposer l'objet à la bonne position). La première étape est de rejoindre le point de collecte. Si un objet est détecté, il est saisi et le robot peut continuer la phase suivante de planification. Sinon le robot attend qu'un objet lui soit donné. Ici, nous présentons de nouveau la canette rouge et laissons le robot reproduire la tâche. Le bras rejoint la position à laquelle il avait été guidé au dessus de la boîte 1 et ouvre sa pince, lâchant ainsi l'objet dans la boîte. Puis il retourne vers le point de collecte où une nouvelle canette a été placée. Le signal de satisfaction est présenté pendant 2 reproductions supplémentaires afin de renforcer l'association entre le but et la carte cognitive. Une fois que le premier but, déposer la canette rouge dans la boîte 1, a bien été appris, le même protocole est répété avec la canette verte et la boîte 2. L'humain montre d'abord comment lâcher la canette verte au dessus de la boîte puis laisse le robot reproduire la tâche, fournissant un signal de satisfaction pour les deux reproductions suivantes. On fournit ensuite

au robot une série de canettes rouges et vertes pour vérifier que le système a bien appris à trier les canettes dans les bonnes boîtes. La figure 7.4 résume les différentes étapes d'apprentissage de la tâche.

Si on observe la carte cognitive apprise pendant la tâche (fig. 7.5a), on peut remarquer qu'un plus grand nombre d'états a été appris près du point de collecte que pour les mouvements entre les boîtes. La représentation en 3D ne montre pas l'espace articulaire dans lequel ces états ont été appris et des positions proches dans l'espace cartésien peuvent être distantes dans l'espace moteur du bras. De plus, pendant la démonstration, les articulations ont été manipulées avec précision par l'humain pour obtenir la bonne approche du point de collecte permettant une prise de l'objet tandis que le mouvement pour rejoindre les boîtes est un mouvement simple ne faisant appel qu'à peu d'articulations. Le modèle de recrutement des états (7.1) explique donc les différences de densités d'états.

La reproduction des trajectoires apprises (fig. 7.5b), comme en navigation, peut être légèrement différente. Cela est dû à la reconnaissance d'un état avant l'arrivée dans la configuration prototypique qui lui est associée et donc à l'anticipation de la prochaine position à atteindre, un peu avant d'avoir fini le mouvement précédent. Ainsi la trajectoire reproduite tend à prendre de petits raccourcis par rapport à celle apprise. L'effet paraît plus important lorsque les états sont éloignés dans l'espace cartésien, l'anticipation étant alors plus grande.

Modification des conditions de la tâche

A ce point, le robot a correctement appris à réaliser la tâche. Notre système ne distingue cependant pas de phases d'apprentissage et de reproduction, il apprend tout le temps et ses associations peuvent être modifiées. L'humain peut donc, s'il le désire, modifier les conditions de la tâche. Pour ce faire, nous plaçons une canette rouge au point de collecte, que le robot saisit. Quand le robot commence son mouvement pour se diriger vers la boîte 1, l'humain saisi son poignet et le guide vers la boîte 2. Une fois que l'objet a été lâché dans la boîte, suite à l'utilisation du capteur sur la pince, un signal de satisfaction est présenté par l'humain. Grâce à l'équation d'apprentissage (7.3), l'action de lâcher l'objet au nouvel endroit est associée au but rouge et l'association entre ce but et l'ancien endroit est diminuée. La stabilité des associations but-transition apprises dépend du nombre de signaux de satisfaction présentés par l'humain. L'association initiale avait été renforcée 3 fois. Après la première démonstration des nouvelles conditions, les transitions menant à la nouvelle boîte sont toujours moins activées que celles menant à la boîte initiale. Après une nouvelle démonstration où l'humain corrige le mouvement du robot et fournit un signal de satisfaction, le nouveau comportement commence à remplacer l'ancien. Le robot peut maintenant reproduire la tâche avec les nouvelles conditions de manière autonome. Un dernier signal de satisfaction est fourni après cette reproduction pour stabiliser ce nouveau comportement (fig. 7.6).

Après cet apprentissage, le bras robotique place les deux types d'objet dans la boîte 2. Finalement, la procédure est répétée pour associer les canettes vertes avec la boîte 1. Les conditions de la tâche sont donc maintenant inversées par rapport aux conditions initiales. Ces nouveaux apprentissages utilisent les capacités de généralisation de la carte cognitive à leur avantage : seule la partie du mouvement du point de collecte vers la boîte a du être redémontrée par l'humain. Le robot utilise ses connaissances acquises lors de la première phase de la tâche pour retourner de la boîte vers le point de collecte une fois l'objet déposé.

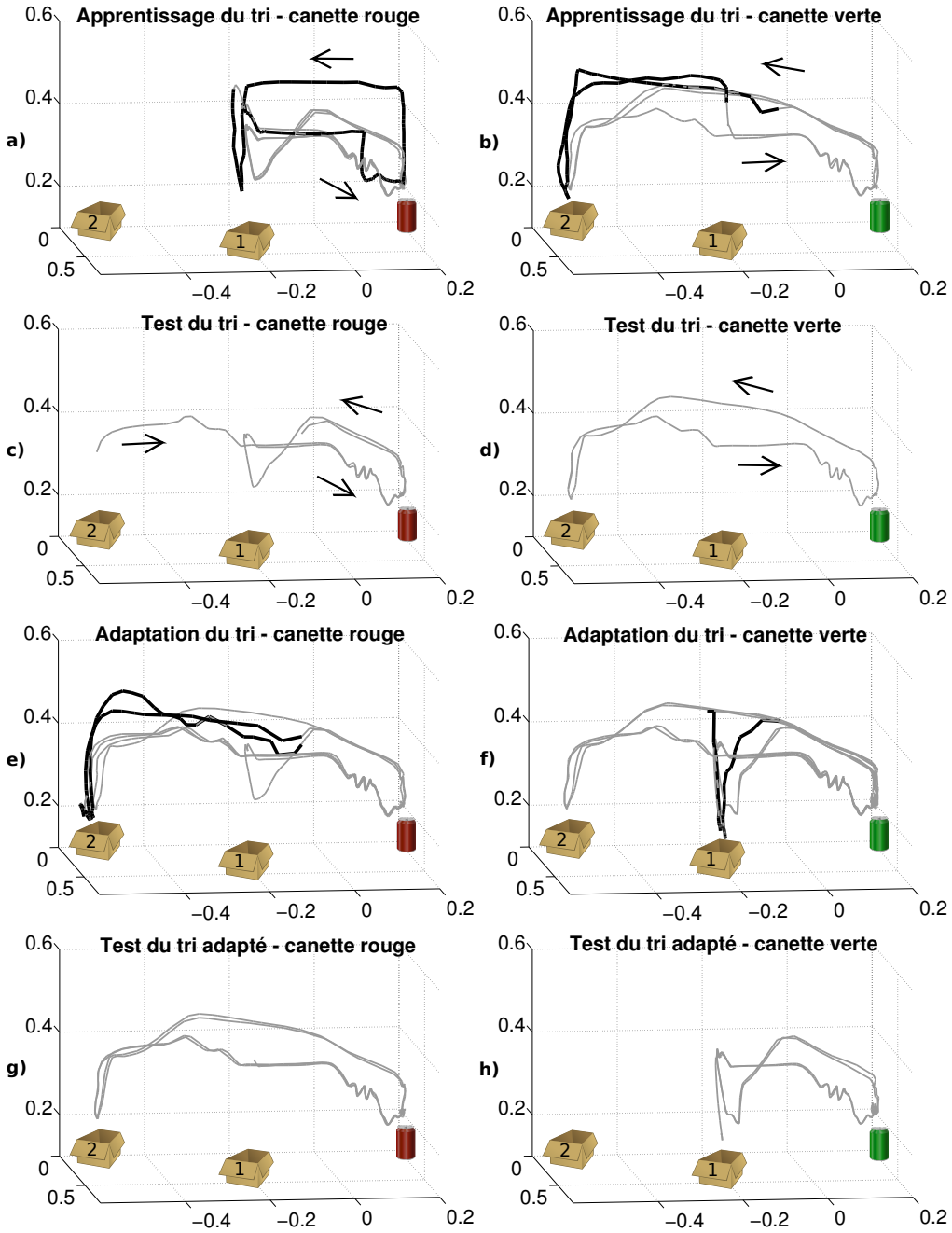
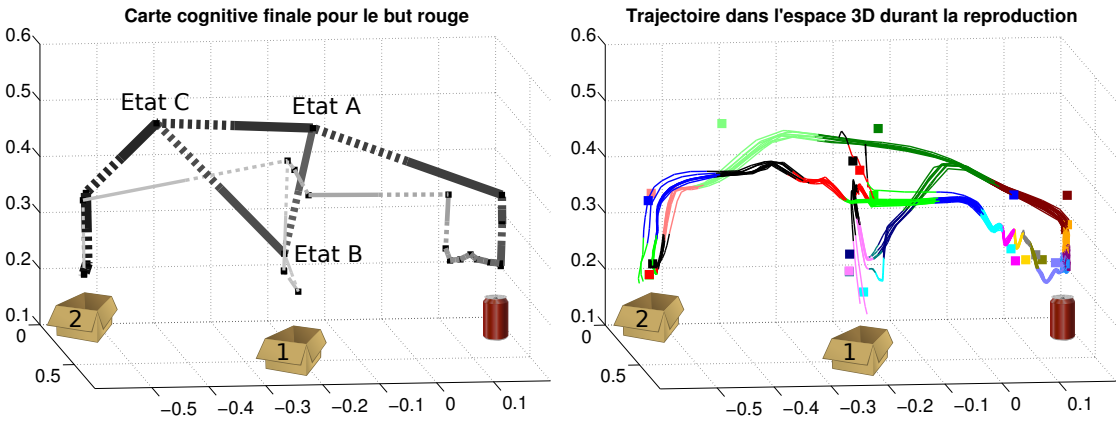


FIGURE 7.4 – Trajectoires de la pince dans l'espace cartésien 3D calculées à partir de la proprioception à l'aide d'un modèle inverse. L'expérience est découpée en plusieurs phases. **a) et b)** apprentissage du dépôt des canettes rouges dans la boîte 1 et des vertes dans la boîte 2. La ligne foncée et épaisse est la trajectoire de la pince pendant la manipulation passive par l'humain. **c) et d)** Validation de la reproduction de la tâche. **e) et f)** Apprentissage de nouvelles conditions : inversion des boîtes pour les canettes rouges et vertes. Le mouvement du bras est corrigé par l'humain qui le manipule pour lui montrer les nouveaux gestes attendus. 2 ou 3 répétitions et signaux de satisfactions sont nécessaires pour remplacer adapter le comportement aux nouvelles conditions. **g) et h)** Reproduction de la tâche modifiée.



(a) Propagation de l'activité de but dans la carte cognitive. L'épaisseur et la noirceur des lignes représente le niveau d'activation.

(b) Trajectoires du bras pendant la tâche. Les carrés sont les états appris et les couleurs représentent l'état le mieux reconnu en chaque point.

FIGURE 7.5 – Carte cognitive et trajectoires. Représentation de la position de la pince dans l'espace cartésien.

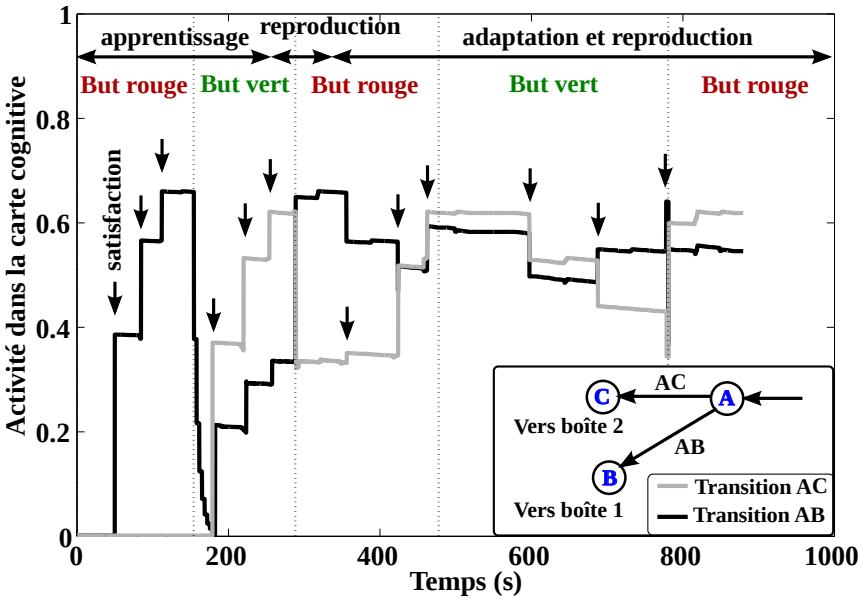


FIGURE 7.6 – Activité des transitions AB et AC dans la carte cognitive. Les états A, B et C représentent le point de bifurcation entre les chemins menant aux 2 boîtes (voir fig. 7.5a). Chaque signal de satisfaction renforce l'association entre le but actif et la dernière transition effectuée et diminue les autres associations. Ces modifications sont propagées pour les chemins menant à ces transitions.

Afin de tester la robustesse du système, nous l'avons perturbé lors de la dernière reproduction de la tâche. Dans un premier temps, nous avons pris l'objet de la pince du robot alors qu'il se dirigeait vers la boîte pour le lâcher. Cela peut avoir deux conséquences :

1. Si le bras est suffisamment proche d'une position à partir de laquelle il sait retourner au point de collecte, il partira rechercher un objet.

2. Sinon le bras est perdu et ne sait pas retourner au point de collecte. Il reste alors immobile jusqu'à ce que son niveau de stress s'accumule et le fasse retourner à sa position de repos, à partir de laquelle il sait retourner au point de collecte.

Pendant un autre essai, nous avons donné un objet au robot pendant qu'il se dirigeait vers le point de collecte. L'objet doit tout d'abord être présenté au système visuel pour être correctement catégorisé. Comme dans le cas précédent, le robot peut connaître un chemin partant de sa position actuelle vers la boîte appropriée ou bien être perdu. La position où l'objet a été donné va être reconnue comme un autre endroit où un objet peut être saisi, et donc comme un point de collecte. Après cela, le robot se dirigera donc vers le point de collecte le plus proche pour récupérer des objets. Des travaux futurs pourraient intégrer le système de détection de l'échec, couplé avec l'inhibition de transition, afin que le robot puisse détecter l'absence d'objet au point de collecte le plus proche et aller à un autre point pour récupérer un objet

Finalement, nous avons utilisé ici une catégorisation visuelle d'objets pour créer des buts, pouvant être considérés comme des contextes motivationnels. Cette catégorisation était supposée déjà apprise, différentes couleurs correspondent alors à des contextes bien définis. Une approche plus naturelle et plus plausible pourrait être employée dans le futur. Une tête expressive pourrait être ajoutée à la plate-forme robotique et reconnaître des expressions faciales pour faciliter l'interaction entre humain et robot [Hasson et al., 2010]. Avec un tel dispositif, l'humain pourrait directement donner des signaux de satisfaction, ou de punition, en affichant une expression positive (joie) ou négative (colère). Ce genre d'interaction a déjà été utilisée avec succès pour de la référenciation sociale d'objets [Boucenna et al., 2010]. L'expression faciale et la direction du regard de l'humain peut aussi permettre de guider l'attention du système visuel vers des objets intéressants [Boucenna et al., 2011] et aider à leur catégorisation. Le robot pourrait alors apprendre en interagissant avec l'humain quels objets sont intéressants et quels types d'objets doivent être catégorisés dans des contextes différents.

Il est possible que, dans le cadre de l'acquisition de comportements automatiques avec la répétition de la tâche, ces contextes permettent de sélectionner directement les états et transitions pertinents au niveau de l'hippocampe. Le retour du cortex préfrontal sur le cortex entorhinal avait été discuté dans le chapitre 4 pour établir des associations entre un contexte motivationnel et les états entorhinaux. Ici, les états proprioceptifs, possiblement codés dans le cortex moteur, pourraient être directement associés avec un contexte de but par un retour du préfrontal. La sélection des transitions se ferait alors en amont de l'hippocampe et permettrait d'expliquer les activités prospectives observées dans celui-ci.

Enfin, la carte cognitive a été utilisée ici pour apprendre rapidement une séquence de mouvements par l'interaction. Le système apprend en ligne en permanence. Il pourrait donc découvrir d'autres mouvements par l'exploration de son espace proprioceptif. La carte cognitive fournit toujours le chemin le plus court pour satisfaire les buts du robot. La découverte, par l'exploration, d'un raccourci pourrait donc amener le bras robotique à faire d'autres mouvements que ceux qui lui ont été enseignés pour réaliser la tâche. On pourrait cependant vouloir apprendre

au robot une séquence particulière. C'est cette séquence qu'il devrait effectuer pour réaliser la tâche, même si il connaît d'autres chemins plus courts. On pourrait alors imaginer que le robot possède plusieurs niveaux de cartes. Un niveau de carte générique, non spécifique d'une tâche, représenterait alors la connaissance du robot de la topologie de son environnement. Elle pourrait être apprise par exploration, interaction etc. Un second niveau de carte pourrait adopter une représentation spécifique à une ou plusieurs tâches qui ont été enseignées au robot. Un niveau de représentation spatiale moins précis serait utilisé et permettrait uniquement de mémoriser les aspects de séquence spécifiques à la tâche. On pourrait alors imaginer que cette seconde carte mémorise des séquences de sous-buts pour réaliser la tâche, tandis que la première carte fournirait une représentation spatiale précise permettant de passer d'un sous-but à un autre.

7.1.2 Intégration bras/robot mobile

Comme le montrent les travaux présentés dans la section précédente, les capacités de généralisation de l'architecture de planification lui permettent de contrôler différent types de plates-formes robotiques. Ainsi, elle a été utilisée pour contrôler un robot mobile (section 4.2.2) ou un bras robotique (section 7.1.1). L'étape suivante est alors de vérifier si l'on peut disposer d'une architecture unique pour contrôler une plate-forme composée d'un bras monté sur un robot mobile. De précédents travaux ont utilisé une telle plate-forme [D'halluin et al., 2010]. Dans cette expérience, un bras Katana était monté sur un robot mobile Robulab-10. Le robot apprenait à se déplacer vers un point de collecte en utilisant des cellules de lieu visuelles. Il saisissait un objet et se dirigeait vers un des deux points de dépôt dans l'expérience pour lâcher l'objet. Le choix du point de dépôt était fait en fonction de la taille de l'objet saisi. Un simple réseau de neurones apprenait par supervision à associer un état du système à un comportement de haut-niveau. Ces comportements correspondaient à des systèmes indépendants et alliaient navigation visuelle et manipulation. Des cellules de lieu étaient utilisées pour la navigation. Les mouvements pour la préhension d'objets étaient appris par démonstration en utilisant l'architecture de [Calinon et al., 2009]. Ainsi le réseau de neurones faisait la sélection entre plusieurs systèmes de contrôle indépendants ne travaillant pas du tout avec les mêmes représentations. Il serait donc intéressant de voir si une architecture unique pourrait contrôler à la fois le bras et la plate-forme mobile, en utilisant une représentation globale intégrant les divers comportements requis. Ce type d'architecture pourrait s'avérer indispensable afin d'utiliser des mécanismes de planification pouvant prévoir des séquences d'actions incluant aussi bien un contrôle du bras que des déplacements du robot dans l'espace.

L'intégration de l'architecture de planification sur la plate-forme bras+robot mobile reste une question en suspens. Cependant, des travaux préliminaires d'intégration ont été réalisés dans le cadre d'une tâche visant à atteindre et saisir un objet avec des comportements de bas niveau. Le cadre de ces expériences est alors de pouvoir repérer un objet dans l'environnement, se diriger vers lui et le saisir grâce au bras monté sur le robot. Ce genre de comportement forme la base d'un système de robot collecteur. Un tel comportement peut être obtenu par la somme de contrôleurs neuronaux simples gérant des aspects d'exploration, de suivi d'objets et de préhension.

Dans l'implémentation que nous avons réalisée, un objet particulier (une balle rose) est reconnu visuellement grâce à sa couleur. Le robot peut tourner sa tête grâce à une caméra montée sur un système pan-tilt. Le système pan-tilt essaie de maintenir l'objet au centre de l'image.

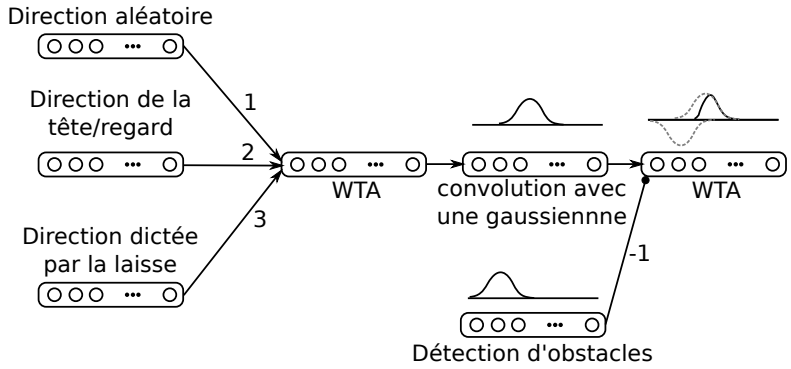


FIGURE 7.7 – Architecture de contrôle de l'orientation du robot. Les directions sont dans le référentiel égocentré. La direction aléatoire est générée périodiquement dans un référentiel absolu et convertie dans le référentiel égocentré à l'aide de la boussole. L'orientation de la caméra donne la direction de l'objet (qui est maintenu au centre de l'image). L'orientation dictée par l'humain tirant la laisse est donnée par une pression sur le cou artificiel.

Lorsque l'objet cible n'est pas détecté, la caméra prend des prises de vue à 360° pour tenter de repérer l'objet. Ces mêmes prises de vues pourraient également servir à la construction d'un code spatial sous forme de cellules de lieu. La direction du robot est, quant à elle, dictée par 3 sous-systèmes (fig. 7.7). Les sous-systèmes sont les suivants, par ordre de priorité décroissant :

1. Un contrôle direct par l'humain via une laisse, permettant de guider le robot dans une direction voulue.
2. L'alignement du corps du robot avec la caméra, lorsque celle-ci suit l'objet. Le robot essaie de minimiser la torsion de son cou et s'oriente ainsi vers l'objet.
3. Une direction aléatoire pour explorer son environnement.

Les poids synaptiques différents provenant des différents sous-systèmes permettent d'introduire un ordre de priorité pour une sélection avec une compétition WTA. Les commandes de l'humain sont prioritaires sur le suivi d'objet qui est prioritaire sur l'exploration aléatoire. Un système de détection d'obstacles vient inhiber les directions bloquées. La convolution préalable de la direction voulue avec une gaussienne permet de garder la direction non bloquée la plus proche possible de celle voulue.

D'un autre côté, la vitesse est régulée par la distance à l'objet perçue et par la détection d'obstacle. Le robot ralentit, voire s'arrête le temps de tourner, si des obstacles se trouvent directement sur son chemin. Sinon le robot garde une vitesse par défaut, qui est diminuée lorsque la taille perçue de l'objet augmente et donc lorsque l'objet se rapproche. La taille est détecté par un calcul du nombre de pixels de la couleur voulue. Le seuil pour l'arrêt du robot est fixé et correspond à une distance qui place l'objet dans l'espace de travail du bras. Cette valeur pourrait être apprise lors d'une exploration de l'espace de travail du bras avec la balle tenue dans la pince.

Les mouvements du bras sont dirigés par le contrôleur visio-moteur décrit dans [Rengervé et al., 2010] (fig. 7.8). Dans une première phase, le robot apprend des associations visio-motrices entre la position de sa pince dans son champ visuel et sa proprioception. La position de l'objet dans son champ visuel peut ensuite activer des attracteurs dans son espace moteur afin de générer un mouvement qui permet à la pince d'atteindre l'objet.

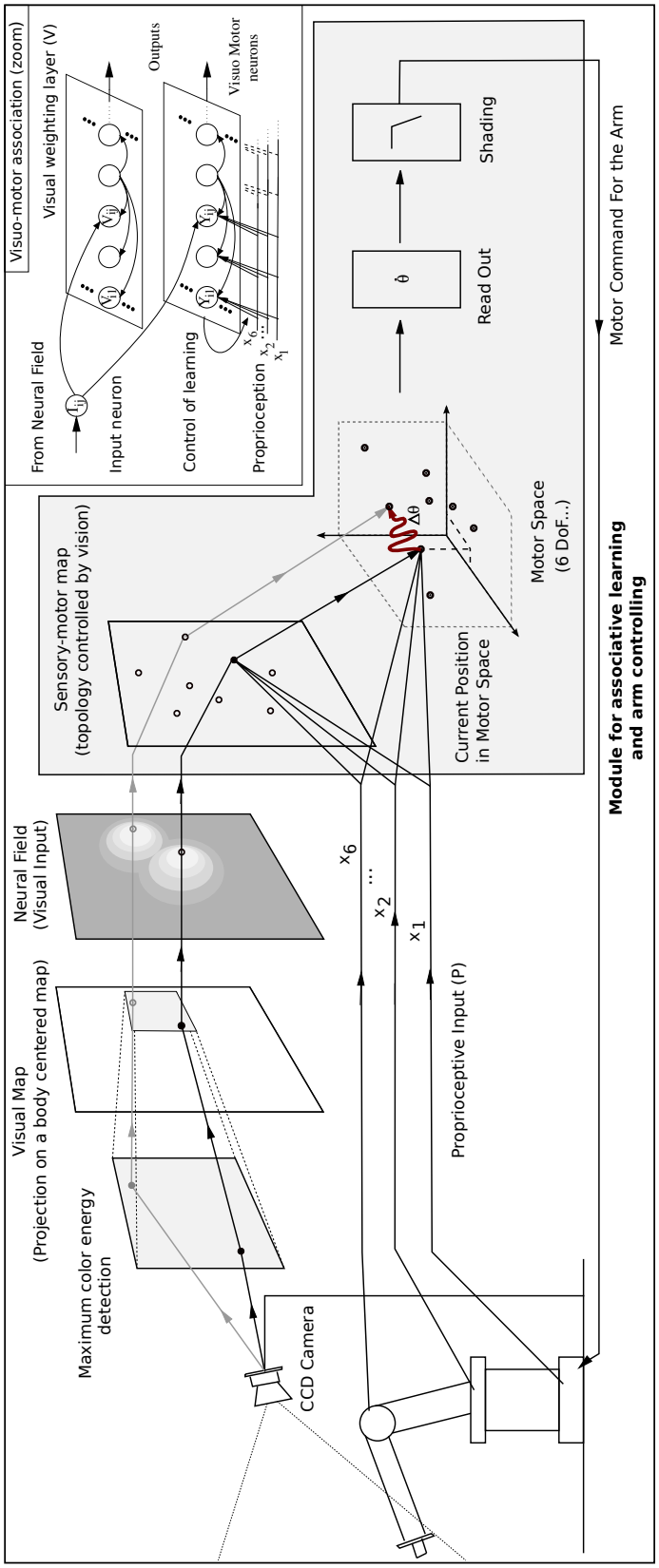


FIGURE 7.8 – Modèle d'apprentissage d'associations visio-motrices. Tiré de [Rengervé et al., 2010].

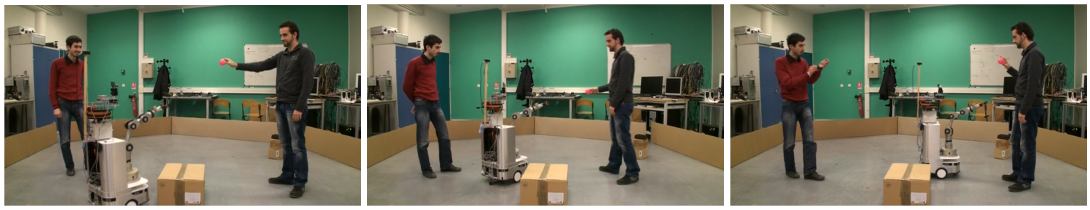


FIGURE 7.9 – Contournement d'obstacle par le robot tentant de rejoindre la balle tenue par l'humain.

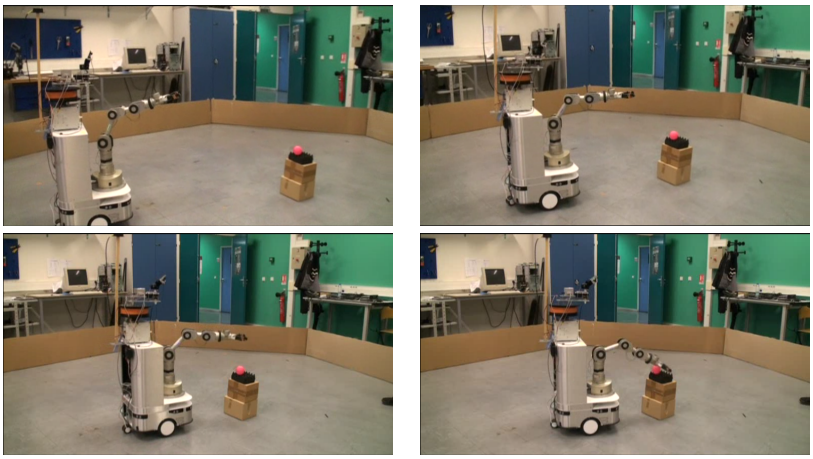


FIGURE 7.10 – Navigation du robot vers la balle posée sur une boîte. Le bras dirige constamment sa pince vers la balle.

Un test a été conduit pour vérifier le fonctionnement du système sur une plate-forme avec un bras Katana monté sur un robot Robulab-10. Le système de caméra pan-tilt est monté à l'avant du robot, au-dessus du bras et permet de détecter la présence d'une balle rose dans l'environnement. Des capteurs ultrasons fournissent les signaux nécessaires pour l'évitement d'obstacles. Dans un premier temps, le robot apprend les associations visio-motrices entre la position visuelle de sa pince et la proprioception du bras. La balle rose est tenue dans la pince, ce qui permet au système visuel de détecter la position de cette dernière.

Après apprentissage, le robot explore son environnement en se déplaçant aléatoirement et en regardant dans toutes les directions. Une fois que la balle est repérée, il se dirige vers celle-ci. Il contourne les obstacles présents sur son chemin (fig. 7.9). Une fois arrivé suffisamment près de l'objet, le robot s'arrête. La caméra donne la position de la balle dans son champ visuel et le robot envoie les commandes motrices correspondantes au bras, qui s'aligne alors avec la balle (fig. 7.10). Des objet placés sous la hauteur du système visuel du robot seront perçus comme plus bas lorsqu'ils sont proches que lorsqu'ils sont loin. Le bras du robot descend donc naturellement au fur et à mesure qu'il se rapproche de l'objet. Il est intéressant de noter que le système peut aussi bien être utilisé pour atteindre des objets statiques que pour guider le robot, étant donné que le robot suivra l'objet si un humain le tient dans sa main et se déplace.

Ce système réactif très simple permet au robot de se diriger vers un objet et de le saisir. Il ne permet cependant pas de planifier une trajectoire et échouerait dans un environnement plus complexe où le robot pourrait être coincé dans un cul-de-sac, ou bien dans lequel la prise de l'objet nécessiterait une approche particulière. Cependant, ce système constitue une étape vers

l'intégration de la planification pour le contrôle conjoint du bras et du robot mobile. La carte cognitive pourrait alors intégrer les états visuels et moteurs concernant le bras, tandis que les phases d'exploration visuelle fourniraient des informations permettant de créer des états spatiaux sous forme de cellules de lieu. Cette multiplication des états rend cependant le système de plus en plus complexe. Dans [D'halluin et al., 2010] les états globaux du système incluaient toutes les combinaisons possibles des différents états moteur/visuels/spatiaux etc. Avec une complexité grandissante des plates-formes robotiques et des tâches à effectuer, cette approche n'est pas viable. Elle n'est pas non plus compatible avec la volonté d'intégrer toutes les informations perceptives disponibles dans le système et de le laisser détecter celles qui sont pertinentes. Ainsi, il devient nécessaire de procéder à une catégorisation des états globaux (fusionnant des informations de types variés : visuelles, proprioceptives, etc.) qui se fasse de manière autonome et en fonction de leur nécessité pour la tâche à effectuer.

7.2 Généralisation des signaux traités

Actuellement, la catégorisation des transitions se fait entre le dernier événement perçu par le système et un nouvel événement. La mémoire temporelle estimant l'écoulement du temps ne prend en compte que l'événement le plus récent, et chaque nouvel événement efface la trace de l'événement passé. Ainsi, une relation temporelle entre deux événements fortement corrélés mais séparés par un autre événement peu pertinent ne pourrait être apprise. Jusqu'ici, les signaux fournis au système ont toujours eu un certain niveau de pertinence : informations visuelles, de détection de couleur et de son dans la navigation motivée ; informations de proprioception et d'état de la pince dans le contrôle du bras robotique etc. Cependant, si on veut tendre vers un système générique, la multiplication des signaux d'entrée du système ne garantira plus la pertinence de tous ces signaux pour une tâche donnée. L'apprentissage des transitions entre les événements réellement pertinents pour la réalisation d'une tâche pourrait alors être perturbé par des événements non pertinents. Il devient ainsi nécessaire d'utiliser un mécanisme de catégorisation des transitions capable de détecter des corrélations temporelles entre des événements généralement proches dans le temps.

Un exemple des limitations du système peut être donné en effectuant une expérience de navigation où deux zones de la même couleur sont présentes dans l'environnement (fig. 7.11). Dans cette expérience, les deux zones situées dans les coins nord-ouest (zone A) et sud-est (zone B) sont de la même couleur rouge. Cependant, seule la présence du robot dans la zone B peut déclencher un son au bout de 300ms. Une seule des deux zones représente donc un lieu but, puisque l'autre ne permet pas la perception du son et donc la satisfaction de but correspondante. Malgré tout, l'architecture de transitions apprend la relation entre les événements perceptifs consécutifs. A l'approche du but, le robot entre dans un lieu X, puis entre dans la zone de couleur, et enfin perçoit le son 300ms plus tard. Les deux transitions apprises sont donc *Lieu X→Couleur rouge→Son*. Le système apprend alors à prédire l'arrivée du son uniquement en fonction du temps écoulé depuis la perception de la zone de couleur. L'activité de prédiction sera donc aussi bien présente au niveau du lieu but réel (zone B), qu'au niveau de l'autre zone de couleur (zone A), ne produisant pourtant pas de son. En effet, la simple perception de la couleur rouge suffit à déclencher cette prédiction. De plus, dans la carte cognitive, le robot cherchera à atteindre la satisfaction de son but, donc le son. Pour cela, il sait qu'il doit d'abord atteindre une zone rouge. Il se dirigera alors vers la zone rouge la plus proche, malgré le fait que seule

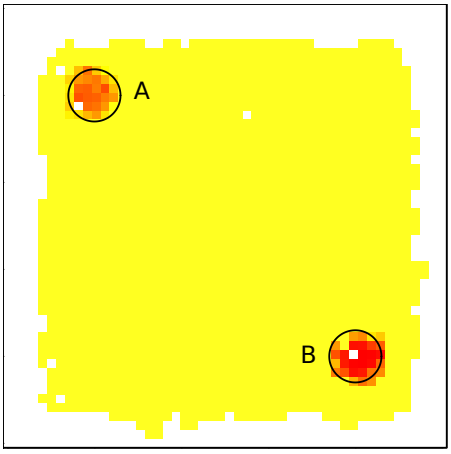


FIGURE 7.11 – Activité spatiale moyenne du neurone de transition prédisant la perception du son. La prédiction est basée sur la dernière perception qui est la détection d'une zone de couleur rouge, d'où l'indifférenciation des deux zones A et B. L'activité un peu plus forte dans la zone B est due à la perception du son elle-même.

la zone B corresponde à un but. Le problème est donc ici que la prédiction du son se base uniquement sur le dernier événement perceptif qu'est la détection de la zone de couleur. Une catégorisation plus complète de la perception prédictrice du son devrait inclure des informations visuelles également afin de différencier les lieux dans lesquels ces zones se trouvent. Il est alors nécessaire de pouvoir créer une forme composite "lieu B et couleur rouge".

La catégorisation automatique des événements perceptifs comme une combinaison de d'informations sensorielles multimodales se retrouve dans les concepts de *positive patterning* et *negative patterning*. Le positive patterning est l'apprentissage d'un conditionnement entre deux stimuli présentés de manière simultanée et une récompense, alors que la seule présentation de l'un ou l'autre des stimuli n'entraîne pas de récompense. Le negative patterning est le cas inverse, la présentation seule d'un des stimuli, mais pas leur présentation conjointe, est associée à une récompense. En termes informatiques, on pourrait dire que le positive patterning correspond à un ET logique, tandis que le negative patterning serait un XOR (OU exclusif). [Schmajuk and DiCarlo, 1992] présentent un modèle de conditionnement capable d'apprendre du positive et negative patterning. Le modèle, à l'instar d'un perceptron multi-couches, se base sur une couche de neurones d'entrée, une couche de sortie, et une couche cachée entre les deux. Un système de calcul d'erreur de prédiction est utilisé pour réaliser l'apprentissage dans la couche cachée.

Dans l'expérience réalisée dans la section précédente, un système de positive patterning est nécessaire. En effet, le son ne peut être perçu qu'en restant sur la zone rouge *ET* en étant dans la partie sud-est de l'environnement. Ainsi le système doit apprendre à caractériser le lieu but comme une combinaison d'informations de couleur et d'informations spatiales. Nous avons utilisé un LMS pour tenter de catégoriser l'état perceptif qui mène à la perception du son (fig. 7.12). Cet état prend en entrée les activités des cellules de lieu, ainsi que la détection de couleur et un signal non discriminant toujours actif. Le signal inconditionnel appris est la perception du son. Le signal non discriminant permet de montrer que le système laisse de côté les informations non pertinentes. De plus, lors de l'apprentissage, ce signal peut être appris comme inhibiteur et permettre au LMS d'apprendre à réaliser un positive patterning. En effet, l'inhibition tou-

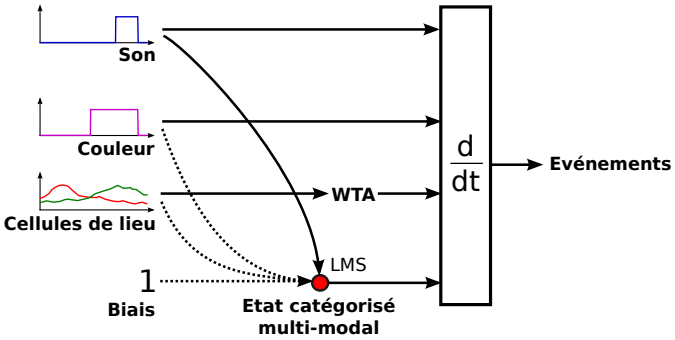


FIGURE 7.12 – Modèle de catégorisation d'un état multi-modal par rapport à la perception du son.

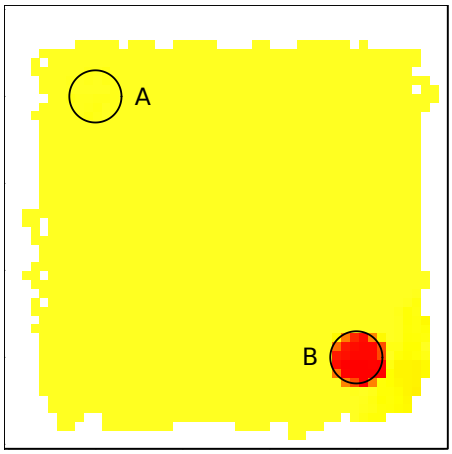


FIGURE 7.13 – Activité spatiale moyenne du neurone d'état multi-modal catégorisé par rapport à la perception du son. Le positive patterning est appris et le neurone n'est actif que dans le lieu B.

jours présente sur le LMS impose la coactivation de plusieurs signaux pour activer la sortie du neurone, tandis que l'activation individuelle de ces signaux n'est pas suffisante.

Au final, le neurone apprend à catégoriser un état qui est une coactivation de la détection de couleur et des activités des cellules de lieu qui sont actives à proximité du lieu but B. Son activité neuronale montre alors des corrélats spatiaux uniquement pour le lieu but actif (fig. 7.13). En effet, au terme d'un long apprentissage qui est le résultat de passages sur les deux lieux buts, le robot apprend à discriminer les échec de prédictions liés à l'utilisation de la détection de couleur lorsqu'il est dans la zone A. Il peut ainsi corriger son erreur et adapter ses prédictions.

Ce type de catégorisation pourrait avoir lieu dans le gyrus dentelé et caractériser les états perceptifs qui déclenchent une mémoire temporelle grâce aux cellules granulaires. [Almaguer-Melian et al., 2003] ont montré que l'apprentissage à long terme dans le gyrus dentelé pouvait être modulé par l'amygdale, lors d'épisodes ayant une forte saillance motivationnelle ou émotionnelle. Ces observations correspondraient donc très bien à la catégorisation automatique d'états perceptifs prédicteurs d'événements liés à la satisfaction d'un but tels que la perception du son dans cette expérience.

Le travaux présentés ici sont cependant très préliminaires et il reste beaucoup de questions à étudier avant leur intégration dans l'apprentissage de transitions. Tout d'abord, la catégorisa-

tion effectuée ici n'est pas automatique. L'architecture utilisée est très adaptée à la tâche. Un seul neurone est utilisé pour apprendre le patterning. On pourrait utiliser par la suite une population de neurones recrutés quand un patterning devient nécessaire. De plus, le système est conçu uniquement pour prévoir l'arrivée du son en fonction des autres signaux. L'importance de la modalité sonore par rapport aux autres est donc définie de manière ad-hoc. Un système plus générique permettrait de détecter quand un événement entorhinal unimodal n'est pas suffisant pour prédire le prochain événement. Le recrutement d'un neurone d'état correspondant à une conjonction de diverses modalités pourrait alors permettre de prédire plus précisément les événements futurs dans des cas de negative ou positive patterning.

Par ailleurs, il faudrait alors assouplir le mécanisme de remise à zéro de la mémoire temporelle dans le gyrus dentelé. En effet, lors de l'entrée sur la zone B, quelle batterie de cellules granulaires doit-on activer ? Celle correspondant uniquement à la perception de la couleur, ou celle correspondant à la perception de l'état multi-modal lieu+couleur ? Que se passe-t-il si certains événements peuvent être prédits en fonction du premier stimulus, et d'autres en fonction du deuxième ? Il serait alors intéressant de laisser une trace temporelle de tous les événements ayant eu lieu dans un passé proche, afin que leurs relations temporelles puissent être apprises statistiquement. On pourrait alors dissocier les événements pertinents des événements non pertinents dans la prédiction d'une future perception.

7.3 Discussion

Dans la première partie de ce chapitre j'ai démontré la capacité du système d'apprentissage de transitions et la planification par carte cognitive à contrôler plusieurs types de plates-formes robotiques. Un bras robotique a été utilisé pour apprendre une tâche de tri d'objet, en planifiant sa trajectoire entre un point de collecte et un point de dépôt dépendant du type d'objet. Ces travaux ont constitué une première étape vers la conception d'un réseau de neurone capable de contrôler simultanément différents sous-systèmes robotiques faisant partie d'une plate-forme robotique complexe. Un telle plate-forme pourrait être un robot mobile équipé d'un bras avec une pince ou un robot humanoïde complet. L'utilisation d'un réseau de neurones simple donnant au robot un comportement réactif de suivi et d'atteinte d'objets a montré la capacité d'une architecture neuronale à contrôler le robot dans son ensemble afin de lui donner un comportement cohérent.

Pour aller vers une architecture de contrôle unifiée, intégrant les travaux réalisés sur l'apprentissage de transitions et la planification, il faudrait revenir sur la caractérisation des états du système. La multitude de signaux disponibles lors du contrôle d'une plate-forme robotique complexe impose la catégorisation autonome d'états pertinents pour le système vis-à-vis de la tâche à réaliser. Une expérience préliminaire a été réalisée, montrant des résultats prometteurs sur une possible catégorisation par un mécanisme de positive et negative patterning. Cependant ces travaux ne représentent qu'une première étape pour que l'architecture de transitions puisse discriminer une suite de transitions pertinentes pour chaque tâche.

L'aboutissement de ces travaux pourrait permettre la modélisation d'un mécanisme d'apprentissage encore manquant au modèle à l'heure actuelle. En effet, dans la tâche de navigation continue indicée, la couleur seule peut être prédictrice du son. Notre système de transitions permet de créer un bassin d'attraction des lieux voisins vers cette zone de couleur pour l'atteindre. Par contre, pour la tâche non indicée, il est nécessaire de se référer à des informations spatiales pour caractériser le lieu but. L'association du lieu but avec la cellule de lieu la plus proche est un

moyen simple de caractériser le lieu but mais c'est un moyen peu précis et très dépendant de la topologie des cellules de lieu. Un apprentissage plus efficace correspondrait au shaping utilisé dans les expériences de neurobiologie pour apprendre aux rats la taille du lieu but. Le shaping utilisé par les biologistes avec les rats consiste à réduire progressivement la taille du lieu but dans lequel l'animal déclenche un lâcher de nourriture. Un mécanisme de patterning pourrait reproduire cet apprentissage, en adaptant un état du système qui coderait pour ce lieu but en fonction des récompenses obtenues. Cet état intégrerait des informations sur les activités de cellules de lieux avoisinantes, des informations visuelles proprioceptives etc. Ainsi, grâce à ces signaux multiples, on pourrait construire une représentation spatiale de plus en plus précise. Il suffirait alors au système de transition d'apprendre ensuite la relation temporelle entre la perception de cet état multi-modal et la perception du son pour pouvoir apprendre le délai d'attente.

Finalement, ces travaux représentent une étape vers la conception d'une architecture générique de résolution de tâches. Le modèle n'aurait alors pas besoin d'être adapté à un type de tâche particulier. Le robot serait capable de construire lui même ses représentations de l'environnement et de ses buts pour réaliser divers types de tâches. De plus, l'architecture pourrait travailler en contrôlant plusieurs sous-systèmes robotiques de natures différentes. L'idée d'obtenir une architecture de contrôle générique en robotique pour la réalisation de tâches n'est pas nouvelle. Les travaux de [Alami et al., 1998; Paquier and Chatila, 2003] visent également à produire ce type d'architecture, indépendante de la tâche à réaliser et pouvant travailler avec des robots différents. Leur système construit de manière autonome ses représentation et des buts. Cela permet au robot d'acquérir les comportements nécessaire à la réalisation de tâches. Les travaux effectués dans cette thèse montrent qu'une telle architecture peut se concevoir dans le cadre d'une approche bio-inspirée. Dans ce cadre, la modélisation des structures cérébrales impliquées dans la résolution de tâches chez les animaux et chez l'homme peut permettre de fournir des pistes de modèle, permettant ensuite une application en robotique.

Publications personnelles

Rengervé, A., Hirel, J., Andry, P., Quoy, M., and Gaussier, P. (2011a). On-line learning and planning in a pick-and-place task demonstrated through body manipulation. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–6

Communications dans des ateliers de travail internationaux

Rengervé, A., Hirel, J., Quoy, M., Andry, P., and Gaussier, P. (2011b). A simple neural network controller merging different behaviors for collector robots. In *International Workshop on Bio-Inspired Robots*

Un expert, c'est quelqu'un qui a fait toutes ses erreurs dans un champ réduit d'applications.

– Niels Bohr

Conclusion

J'ai présenté dans cette thèse un modèle des interactions entre l'hippocampe, le cortex préfrontal et les ganglions de la base. Ce modèle a été mis au point dans le cadre de tâches de navigation mais il a aussi été testé sur un bras robotique dans le cadre d'une interaction Homme-Machine. Ces travaux se reposent sur un modèle d'apprentissage de transitions spatiales dans l'hippocampe, présenté dans le chapitre 2. Je montre dans le chapitre 3 que l'architecture correspondante permet la catégorisation d'états perceptifs multi-modaux en intégrant des informations sensorielles visuelles, proprioceptives, etc. Ces perceptions sont codées dans le cortex entorhinal. Dans le gyrus dentelé, des cellules granulaires répondant avec des dynamiques temporelles variées permettent de représenter l'écoulement du temps passé depuis un certain événement. C'est l'apprentissage de l'association entre cette mémoire du passé et de nouveaux événements perçus dans CA3 qui permet au système d'apprendre des relations temporelles précises entre ces perceptions multi-modales. Des neurones de transition dans l'hippocampe sont alors capables de prédire les futurs événements perceptifs possibles. Dans le cas spatial, une transition représente le passage d'un lieu à un autre.

Dans le chapitre 4, je discute l'utilisation des ces informations temporelles dans le cadre des interactions entre l'hippocampe et le cortex préfrontal. Un modèle de ces interactions, cohérent avec les observations et enregistrements faits chez le rat par les biologistes, est développé. Il permet à un robot de réaliser la tâche de navigation continue qui lui est enseignée par l'interaction avec un humain. Cette tâche nécessite que le robot puisse naviguer vers un lieu but non visible et attende pendant un délai fixe sur le lieu but sans bouger. Le modèle a pu reproduire certains enregistrements expérimentaux réalisés in-vivo chez le rat. Des activités hors-champ de cellules de lieu dans l'hippocampe semblent prédire la fin du délai d'attente sur le lieu but. Elles ont pu être reproduites grâce à l'apprentissage d'associations secondaires dans l'hippocampe. L'origine de ces associations serait un retour de l'activité du cortex préfrontal dans le cortex entorhinal. Bien que non modélisée dans cette thèse, nous émettons l'hypothèse que la fonction de ce retour est de pouvoir catégoriser des contextes liés à des buts ou d'autres informations à forte valeur motivationnelle traitées par le cortex préfrontal. Ainsi, cet apprentissage permettraient une sélection contextuelle des transitions pertinentes dès l'hippocampe. De futurs travaux pourraient se porter sur la modélisation de ce mécanisme, qui permettrait d'expliquer les activités prospectives observées dans l'hippocampe des rats. En effet, des cellules de lieu dans l'hippocampe déchargent de manière différenciée en fonction du prochain chemin pris dans un labyrinthe en T dans une tâche d'alternance [Wiener, non publié]. Une information de contexte au niveau du cortex entorhinal pourrait provenir d'une catégorisation des buts (gauche ou droite) au niveau préfrontal. Cela permettrait de sélectionner dès l'hippocampe la prochaine transition en fonction du but recherché. Ce genre de mécanisme serait un exemple de l'acquisition de comportements

automatiques, facilitée par le cortex préfrontal, lors de la répétition de tâches de navigation.

Dans le chapitre 5, j'ai montré que les informations de transitions peuvent servir de base à l'utilisation de différentes stratégies de navigation. Une première stratégie cognitive de haut niveau aurait pour siège le cortex préfrontal. Elle ferait appel à une planification des actions à effectuer pour rejoindre un but. Cette stratégie utilise une carte cognitive, stockée dans le cortex préfrontal et/ou pariétal. La carte relie les transitions pouvant être effectuées successivement. Son apprentissage peut être fait pendant une exploration de l'environnement, ou bien guidé par l'interaction avec un humain. Le robot peut alors utiliser sa carte pour planifier le chemin le plus court, en termes de nombre de transitions, pour rejoindre son but. Il peut ainsi sélectionner à chaque instant la transition la plus adéquate parmi les prédictions de transitions possibles qui sont fournies par l'hippocampe. Cette sélection d'une transition et l'association des transitions avec les actions qui ont permis de les effectuer seraient faites au niveau des ganglions de la base (notamment avec un rôle important du nucleus accumbens). La seconde stratégie de navigation correspondrait à un comportement automatique, résultat de nombreuses répétitions des mêmes actions. Elle est inspirée de l'apprentissage par renforcement et notamment du Q-learning. Cet algorithme se prête particulièrement bien à une implémentation neuronale utilisant une représentation sous forme de transitions. Cette stratégie permet au robot d'associer directement à chaque transition une valeur représentant un espoir de récompense. L'intérêt d'utiliser une représentation commune, les transitions, pour ces deux stratégies est qu'elles peuvent alors biaiser la sélection d'une action de manière parallèle en agissant sur la compétition entre les transitions proposées. On peut alors observer de manière expérimentale la coopération entre ces stratégies dans une expérience de navigation vers un but. La stratégie de carte cognitive est computationnellement coûteuse mais avec un apprentissage et une adaptation rapides. Elle permet de faciliter l'apprentissage par Q-learning, qui se révèle plus efficient (car moins coûteux) à long terme. Une des conclusions des études de lésions effectuées en simulation est que l'inactivation de l'une ou l'autre des stratégies après l'apprentissage d'une tâche est compensée par la redondance de ces systèmes. Cette lésion n'entraîne alors pas d'effets au niveau des performances pour des tâches de navigation motivée simples, comme observé chez les animaux.

Les simples prédictions, sans informations temporelles, des diverses transitions qui sont réalisables à chaque instant sont suffisantes pour établir une stratégie comportementale pour rejoindre un but. Cependant, l'apprentissage des timings précis de ces transitions devient nécessaire lorsque des changements interviennent dans l'environnement. Dans ce cas, le système peut détecter l'échec d'une transition en fonctionnant comme un détecteur de nouveauté. Ainsi la dynamique temporelle particulière de l'activité des neurones de transition a pu être utilisée pour prédire exactement le moment où un événement devrait être perçu. Dans le chapitre 6, j'ai décrit une architecture neuronale pouvant détecter l'absence d'un événement prédit. La détection se base sur le timing appris et sur la régularité de cet événement. Un signal d'échec, construit à partir de ses propres prédictions, servirait alors au robot à réguler et adapter son comportement. Cette adaptation se manifesterait de deux manières. Dans un premier temps, le robot pourrait inhiber les mécanismes ayant mené à la sélection de l'action ayant échouée. Dans un second temps, le robot pourrait adapter ses associations entre transitions et actions afin que des échecs répétés modifient son comportement à long terme. Le modèle a pu être utilisé afin de faire apprendre la tâche de navigation continue à un robot sans intervention de l'humain. Des mécanismes de dressage similaires à ceux utilisés avec les animaux ont alors du être utilisés. Ces mécanismes visent à modifier progressivement les conditions de réalisation de la tâche au fur et à mesure que

l'apprentissage progresse. Les capacités d'adaptation du robot sont donc primordiales dans ce contexte.

Notre modèle suppose que la détection de la nouveauté se ferait tout d'abord dans l'hippocampe, avec l'aide du septum. Elle permettrait la catégorisation automatique d'états correspondant à de nouveaux événements. Des informations sur les relations temporelles entre les différentes états codés seraient alors apprises sous la forme de transitions. Les signaux de prédiction liés aux transitions seraient ensuite traités au niveau des ganglions de la base. Les ganglions de la base pourraient donc se baser sur des prédictions temporelles pour détecter des événements inattendus correspondant à de la nouveauté, ou à un échec du comportement. Ces signaux d'échecs pourraient alors être utilisés directement en tant que modulation de l'apprentissage lié à la sélection de l'action. Ils seraient également transmis aux cortex préfrontal qui jouerait alors un rôle dans le méta-contrôle en venant inhiber des comportements appris précédemment. Cela permettrait d'éviter des effets de persévération du comportement lors d'échecs. En suivant ces hypothèses, on peut donc s'attendre à observer des activités liées à l'absence d'événements prédits (tels que la perception du son dans la tâche de navigation continue) aussi bien dans les ganglions de la base que dans le cortex préfrontal. Dans le cas de récompenses prédites, le système dopaminergique lié aux ganglions de la base a déjà montré de telles propriétés. Notre modèle nous permet d'émettre des prédictions sur la présence d'activités neuronales spécifiques aux essais d'extinction dans la tâche de navigation continue. Ces activités marqueraient la fin du délai d'attente et se propageraient des ganglions de la base vers le cortex préfrontal.

De plus, certaines prédictions au niveau comportemental peuvent être émises concernant des expériences de lésions. De précédentes expériences ont montré qu'une inactivation du cortex préfrontal après l'apprentissage de la tâche de navigation continue n'impactait pas les performances [Hok, 2007]. Ces observations sont cohérentes avec notre modèle, puisque d'autres stratégies de navigation pourraient prendre le relais. Cependant, le cortex préfrontal jouant un rôle de modulateur permettant d'inhiber certains comportements en cas d'échec, on s'attend à ce qu'il joue un rôle dans les essais d'extinction. Une lésion du cortex préfrontal après un apprentissage de la tâche où la récompense est systématiquement donnée pourrait entraîner des effets de persévération lors de l'introduction d'essais d'extinction. En effet, le rat pourrait alors attendre la récompense au delà du délai de 2 secondes, puisque la détection de l'échec ne déclencherait pas d'inhibition de son comportement d'attente si le PFC est inactivé. Ces effets de persévération devraient être particulièrement présents lors des premiers essais d'extinction à cause du caractère *nouveau* de l'absence du son. La présence du PFC serait alors nécessaire pour utiliser cette détection de nouveauté afin de modifier rapidement le comportement. Par la suite, l'animal pourrait progressivement s'habituer à des absences répétées de la récompense à la fin de la phase d'attente et donc ne plus détecter cette absence comme une nouveauté.

D'un autre côté, l'inactivation des projections synaptiques entre l'hippocampe et les ganglions de la base devrait empêcher l'acquisition par shaping du comportement d'attente par le rat pendant la phase d'apprentissage. Une lésion pré-apprentissage détruirait la capacité du rat à détecter un échec lors de l'introduction progressive d'un délai d'attente pour l'obtention de la récompense. Il ne pourrait alors plus explorer d'autres comportements tels que le fait d'attendre sans bouger. De plus, selon la zone couverte par les lésions, on pourrait totalement bloquer la transmission des informations de transition de l'hippocampe vers les ganglions de la base. Cela aurait pour effet d'empêcher l'association entre transitions et actions. Les capacités de navigation vers le lieu but pourraient toutefois être conservées par l'utilisation de stratégies sensori-

motrices de bas niveau (par exemple de type lieu-action) qui feraient appel à d'autres boucles cérébrales.

Pour finir, j'ai montré dans le chapitre 7 que le modèle présenté ici constituait la pierre angulaire d'un système de contrôle générique pour l'apprentissage de tâches en robotique. Le modèle a pu montrer sa capacité à contrôler aussi bien un robot mobile qu'un bras robotique dans une tâche de manipulation d'objets. La prochaine étape sera alors d'assurer un contrôle global des deux sous-systèmes. Dans le cadre de la manipulation d'objets dans une pièce, un objet doit parfois être pris sur une table et déposé dans un rangement à l'autre bout de la pièce. Le contrôle du bras robotique doit alors dépendre de celui du robot mobile et vice-versa. La planification d'une telle tâche inclut des étapes de navigation et de manipulation avec le bras qui sont regroupées dans une même séquence. On peut donc voir l'intérêt d'un système de planification unique, capable d'intégrer des informations multi-modales et de contrôler plusieurs types de plates-formes robotiques.

La question fondamentale restant en suspens est celle de la catégorisation automatique des états du système. La combinaison des états de plusieurs plates-formes crée un espace d'états de grande dimension. Il devient alors nécessaire pour l'architecture d'apprentissage de transitions de discriminer les transitions pertinentes. Cela peut tout d'abord passer par le recrutement de nouveaux états multi-modaux correspondant à une combinaison d'états liés à une seule modalité (proprioception, vision, etc.). Ainsi, l'architecture pourrait posséder des capacités de positive ou negative patterning. Dans un second temps, les neurones codant les transitions pourraient apprendre quels sont les événements du passé proche qui permettent statistiquement de bien prédire un nouvel événement, plutôt que de se baser uniquement sur le dernier événement perçu. On pourrait alors avoir un codage en transitions correspondant réellement aux besoins du robot pour résoudre une ou plusieurs tâches. On éviterait ainsi l'explosion combinatoire d'un codage en transitions tentant d'apprendre toutes les séquences d'événements réalisées qu'elles soient pertinentes ou non.

Enfin, cette question de la granularité du codage des états et des transitions se pose aussi pour la carte cognitive. Celle-ci doit-elle être une recopie du code hippocampique ou doit-elle, plus plausiblement, adopter un code spécifique, plus compact ? Lorsque l'on utilise un robot mobile en extérieur sur plusieurs kilomètres, des centaines, des milliers, voire des dizaines de milliers de lieux peuvent être appris. Il peut alors devenir nécessaire de rendre le code utilisé par la planification spatiale plus réduit. Une solution pourrait être l'adoption de cartes hiérarchiques avec des niveau de granularité différents [Martinet et al., 2011]. Dans ce cas, on pourrait imaginer un niveau de planification de faible granularité incluant des buts et sous-buts pour une tâche à réaliser. Un niveau plus fin de carte intégrerait les actions à réaliser pour passer d'un sous-but à un autre. Enfin la question de la multiplicité des cartes se pose aussi pour le contrôle commun de plates-formes robotiques diverses : existe-t-il une seule carte intégrant des états multi-modaux incluant des informations sur le bras robotique et le robot mobile, ou bien chaque sous-système possède-t-il une carte spécifique ? Dans ce dernier cas, il serait nécessaire de mettre en œuvre un système de planification capable de faire le lien entre de multiples cartes afin de planifier des séquences d'actions incluant des phases de navigation et de préhension d'objets.

Bibliographie

Acsády, L. and Káli, S. (2007). Models, structure, function : the transformation of cortical signals in the dentate gyrus. *Prog Brain Res*, 163 :577–599. 23

Alami, R., Chatila, R., Fleury, S., Ghallab, M., and Ingrand, F. (1998). An architecture for autonomy. *The International Journal of Robotics Research*, 17(4) :315–337. 163

Almaguer-Melian, W., Martínez-Martí, L., Frey, J. U., and Bergado, J. A. (2003). The amygdala is part of the behavioural reinforcement system modulating long-term potentiation in rat hippocampus. *Neuroscience*, 119(2) :319–322. 21, 79, 161

Alvernhe, A., Cauter, T. V., Save, E., and Poucet, B. (2008). Different ca1 and ca3 representations of novel routes in a shortcut situation. *J Neurosci*, 28(29) :7324–7333. 23, 53, 72

Amaral, D. G. and Witter, M. P. (1989). The three-dimensional organization of the hippocampal formation : a review of anatomical data. *Neuroscience*, 31(3) :571–591. 19

Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2) :77–87. 103, 106

Andry, P., Blanchard, A. J., and Gaussier, P. (2011). Using the rhythm of nonverbal human-robot interaction as a signal for learning. *IEEE T. Autonomous Mental Development*, 3(1) :30–42. 59, 74

Andry, P., Gaussier, P., Moga, S., Banquet, J.-P., and Nadel, J. (2001). Learning and communication in imitation : An autonomous robot perspective. *IEEE Transactions on Man, Systems and Cybernetics, Part A : Systems and humans*, 31(5) :431–442. 43, 55, 59, 120, 121

Andry, P., Gaussier, P., and Nadel, J. (2005). Autonomous learning and reproduction of complex sequences : a multimodal architecture for bootstraping imitation games. In *Proceedings of the Fifth International Workshop on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems*, volume 123, pages 97–100. 56

Arkin, R. (1998). *Behavior-based robotics*. Intelligent robots and autonomous agents. MIT Press. 15

Arleo, A. and Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation : a model of hippocampal place cell activity. *Biol Cybern*, 83(3) :287–299. 38, 101, 103, 106, 107, 109

Arleo, A. and Rondi-Reig, L. (2007). Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. *J Integr Neurosci*, 6(3) :327–366. 38, 66, 101

Baeg, E. H., Kim, Y. B., Huh, K., Mook-Jung, I., Kim, H. T., and Jung, M. W. (2003). Dynamics of population code for working memory in the prefrontal cortex. *Neuron*, 40(1) :177–188. 27

Banquet, J. P., Gaussier, P., Dreher, J. C., Joulain, C., Revel, A., and Gunther, W. (1997). Space-time, order and hierarchy in fronto-hippocamal system : A neural basis of personality. In *Cognitive Science Perspectives on Personality and Emotion*, pages 123–189. Elsevier Science BV. 38, 39, 40, 41, 43, 49, 50, 53

Banquet, J. P., Gaussier, P., Quoy, M., Revel, A., and Burnod, Y. (2005). A hierarchy of associations in hippocampo-cortical systems : cognitive maps and navigation strategies. *Neural Comput*, 17(6) :1339–1384. 40, 81

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In *Models of Information Processing in the Basal Ganglia*, pages 215–232. MIT Press. 100, 107

Bast, T., Wilson, I. A., Witter, M. P., and Morris, R. G. M. (2009). From rapid place learning to behavioral performance : a key role for the intermediate hippocampus. *PLoS Biol*, 7(4) :e1000089. 27, 52

Beiser, D. G., Hua, S. E., and Houk, J. C. (1997). Network models of the basal ganglia. *Curr Opin Neurobiol*, 7(2) :185–190. 98

Bellingham, W., Gillette-Bellingham, K., and Kehoe, E. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Learning & Behavior*, 13 :152–164. 10.3758/BF03199268. 146

Bertram, E. H. and Zhang, D. X. (1999). Thalamic excitation of hippocampal ca1 neurons : a comparison with the effects of ca3 stimulation. *Neuroscience*, 92(1) :15–26. 26, 81

Billard, A., Calinon, S., Dillmann, R., and Schaal, S. (2008). Robot programming by demonstration. In *Handbook of Robotics*, volume chapter 59, chapter 59. MIT Press. 148

Bohn, I., Giertler, C., and Hauber, W. (2003). Orbital prefrontal cortex and guidance of instrumental behaviour in rats under reversal conditions. *Behav Brain Res*, 143(1) :49–56. 124

Boucenna, S., Gaussier, P., and Andry, P. (2008). What should be taught first : the emotional expression or the face ? In *8th International conference on Epigenetic Robotics, EPIROB*. Lucs. 59, 74

Boucenna, S., Gaussier, P., and Hafemeister, L. (2011). Development of joint attention and social referencing. In *ICDL - Epirob 2011*, page In press. 154

Boucenna, S., Gaussier, P., Hafemeister, L., and Bard, K. (2010). Autonomous development of social referencing skills. In *From Animals to Animats 11*, volume 6226 of *Lecture Notes in Computer Science*, pages 628–638. Springer Berlin / Heidelberg. 154

Brooks, R. A. (1991). How to build complete creatures rather than isolated cognitive simulators. In *Architectures for Intelligence*, pages 225–239. Erlbaum. 15

Brown, J., Bullock, D., and Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *The Journal of Neuroscience*, 19(23) :10502–10511. 42, 99

Brun, V. H., Leutgeb, S., Wu, H.-Q., Schwarcz, R., Witter, M. P., Moser, E. I., and Moser, M.-B. (2008). Impaired spatial representation in ca1 after lesion of direct input from entorhinal cortex. *Neuron*, 57(2) :290–302. 22, 72

Buonomano, D. V. and Mauk, M. D. (1994). Neural network model of the cerebellum : temporal discrimination and the timing of motor responses. *Neural Comput.*, 6(1) :38–55. 43

Burgess, N., Donnett, J. G., Jeffery, K. J., and O'Keefe, J. (1997). Robotic and neuronal simulation of the hippocampus and rat navigation. *Philos Trans R Soc Lond B Biol Sci*, 352(1360) :1535–1543. 38, 39

Burgess, N., Recce, M., and O'Keefe, J. (1994). A model of hippocampal function. *Neural Netw.*, 7(6-7) :1065–1081. 37, 38, 39

Burton, B. G., Hok, V., Save, E., and Poucet, B. (2009). Lesion of the ventral and intermediate hippocampus abolishes anticipatory activity in the medial prefrontal cortex of the rat. *Behav Brain Res*, 199(2) :222–234. 34, 82, 83

Cajal, S. R. (1911). *Histologie du système nerveux de l'homme et des vertébrés*. Maloine. 20

Calinon, S., D'halluin, F., Caldwell, D. G., and Billard, A. (2009). Handling of multiple constraints and motion alternatives in a robot programming by demonstration framework. In *Proceedings of 2009 IEEE International Conference on Humanoid Robots*, pages 582–588. 155

Cañamero, L. (2005). Emotion understanding from the perspective of autonomous robots research. *Neural Networks*, 18(4) :445–455. 116

Carpenter, G. A. and Grossberg, S. (2002). Adaptive resonance theory. 147

Carter, R. M., Hofstotter, C., Tsuchiya, N., and Koch, C. (2003). Working memory and fear conditioning. *Proc Natl Acad Sci U S A*, 100(3) :1399–1404. 25

Chatila, R. and Laumond, J. (1985). Position referencing and consistent world modeling for mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE Computer Society Press. 13

Clark, R. E. and Squire, L. R. (1998). Classical conditioning and brain systems : the role of awareness. *Science*, 280(5360) :77–81. 25

Contreras-Vidal, J. L. and Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *J Comput Neurosci*, 6(3) :191–214. 42, 99

Corcoran, K. A. and Quirk, G. J. (2007). Recalling safety : cooperative functions of the ventromedial prefrontal cortex and the hippocampus in extinction. *CNS Spectr*, 12(3) :200–206. 25

Cun, Y. L. (1985). Une procédure d'apprentissage pour réseau à seuil asymétrique. In *COGNITIVA 85*. 14

Cuperlier, N. (2006). *Apprentissage et prédiction de séquences sensori-motrices : architecture neuro-mimétique pour la navigation et la planification d'un robot mobile.* PhD thesis, Université de Cergy-Pontoise. 53, 71, 92

Cuperlier, N., Quoy, M., and Gaussier, P. (2007). Neurobiologically inspired mobile robot navigation and planning. *Front Neurorobotics*, 1 :3. 40, 50

Cuperlier, N., Quoy, M., Giovannangeli, C., Gaussier, P., and Laroque, P. (2006). Transition cells for navigation and planning in an unknown environment. In *The Society For Adaptive Behavior SAB 2006*, pages 286–297, Rome. 51

Dalley, J. W., Cardinal, R. N., and Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents : neural and neurochemical substrates. *Neurosci Biobehav Rev*, 28(7) :771–784. 26, 27

Damásio, A. (1994). *Descartes' error : emotion, reason, and the human brain*. Quill. 116

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray. 14

Daw, N., Niv, Y., and Dayan, P. (2006). Actions, policies, values, and the basal ganglia. In *Recent Breakthroughs in Basal Ganglia Research*. Nova Science Publishers Inc. 97, 111

Dehaene, S. and Changeux, J.-P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *J. Cognitive Neuroscience*, 1 :244–261. 124

Dehaene, S. and Changeux, J. P. (1991). The wisconsin card sorting test : theoretical analysis and modeling in a neuronal network. *Cereb Cortex*, 1(1) :62–79. 124

Deller, T., Martinez, A., Nitsch, R., and Frotscher, M. (1996). A novel entorhinal projection to the rat dentate gyrus : direct innervation of proximal dendrites and cell bodies of granule cells and gabaergic neurons. *J Neurosci*, 16(10) :3322–3333. 21

DeLong, M. and Wichmann, T. (2009). Update on models of basal ganglia function and dysfunction. *Parkinsonism Relat Disord*, 15 Suppl 3 :S237–S240. 98

D'halluin, F., de Rengervé, A., Lagarde, M., Gaussier, P., Billard, A., and Andry, P. (2010). A state-action neural network supervising navigation and manipulation behaviors for complex task reproduction. In *Tenth International Conference on Epigenetic Robotics*. 155, 159

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., and Guillot, A. (2010). Path planning versus cue responding : a bio-inspired model of switching between navigation strategies. *Biol Cybern*, 103(4) :299–317. 39, 102, 106, 107, 111

Dorigo, M., Maniezzo, V., and Colorni, A. (1996). Ant system : optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern B Cybern*, 26(1) :29–41. 14

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr Opin Neurobiol*, 10(6) :732–739. 100

Doya, K., Samejima, K., ichi Katagiri, K., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Comput*, 14(6) :1347–1369. 102

Dragoi, G. and Buzsáki, G. (2006). Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron*, 50(1) :145–157. 23, 43, 53

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., and Tanila, H. (1999). The hippocampus, memory, and place cells : is it spatial memory or a memory space ? *Neuron*, 23(2) :209–226. 35, 56

Eichenbaum, H., Kuperstein, M., Fagan, A., and Nagode, J. (1987). Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task. *J Neurosci*, 7(3) :716–732. 24, 66

Eichenbaum, H. and Lipton, P. A. (2008). Towards a functional organization of the medial temporal lobe memory system : role of the parahippocampal and medial entorhinal cortical areas. *Hippocampus*, 18(12) :1314–1324. 25, 94

Fagg, A. H., Lotspeich, D., Hoff, J., and Bekey, G. A. (1994). Rapid reinforcement learning for reactive control policy design in autonomous robots. In *Proceedings of World Congress on Neural Networks*, pages 118–126. 100

Fenton, A. A., Lytton, W. W., Barry, J. M., Lenck-Santini, P.-P., Zinyuk, L. E., Kubík, S., Bures, J., Poucet, B., Muller, R. U., and Olypher, A. V. (2010). Attention-like modulation of hippocampus place cell discharge. *J Neurosci*, 30(13) :4613–4625. 24

Fleischer, J. G., Gally, J. A., Edelman, G. M., and Krichmar, J. L. (2007). Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device. *Proc Natl Acad Sci U S A*, 104(9) :3556–3561. 39

Foster, D. J., Morris, R. G., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10(1) :1–16. 39, 101, 103, 107, 111

Foster, T. C., Castro, C. A., and McNaughton, B. L. (1989). Spatial selectivity of rat hippocampal neurons : dependence on preparedness for movement. *Science*, 244(4912) :1580–1582. 24

Franz, M. O., Schölkopf, B., Georg, P., Mallot, H. A., and Bülthoff, H. H. (1997). Learning view graphs for robot navigation. *Autonomous Robots*, 5 :111–125. 40

Frezza-Buet, H. and Alexandre, F. (2002). From a biological to a computational model for the autonomous behavior of an animat. *Information Sciences*, 144(1-4) :1 – 43. 41, 43

Frezza-Buet, H., Rougier, N., and Alexandre, F. (2001). Integration of biologically inspired temporal mechanisms into a cortical framework for sequence processing. In *Sequence Learning*, volume 1828 of *Lecture Notes in Computer Science*, pages 321–348. Springer Berlin / Heidelberg. 41, 43

Gaussier, P., Banquet, J. P., Sargolini, F., Giovannangeli, C., Save, E., and Poucet, B. (2007). A model of grid cells involving extra hippocampal path integration, and the hippocampal loop. *J Integr Neurosci*, 6(3) :447–476. 47

Gaussier, P., Joulain, C., Banquet, J., Lepretre, S., and Revel, A. (2000). The visual homing problem : An example of robotics/biology cross fertilization. *Robotics and Autonomous Systems*, 30 :155–180. 38, 40, 44, 47, 48

Gaussier, P., Moga, S., Quoy, M., and Banquet, J. (1998). From perception-action loops to imitation processes : A bottom-up approach of learning by imitation. *Applied Artificial Intelligence*, 12 :701–727. 43, 53, 55

Gaussier, P., Revel, A., Banquet, J. P., and Babeau, V. (2002). From view cells and place cells to cognitive map learning : processing stages of the hippocampal system. *Biol Cybern*, 86(1) :15–28. 40, 50, 81

Gaussier, P. and Zrehen, S. (1995). Perac : A neural architecture to control artificial animals. *Robotics and Autonomous Systems*, 16(2-4) :291 – 320. 38, 39, 44

Giovannangeli, C. (2007). *Navigation autonome bio-inspirée en environnement intérieur et extérieur : Apprentissages sensori-moteurs et planification dans un cadre interactif.* PhD thesis, Université de Cergy-Pontoise. 44, 45, 46, 47

Giovannangeli, C. and Gaussier, P. (2007). Orientation system in robots : Merging allothetic and idiothetic estimations. In *13th International Conference on Advanced Robotics (ICAR07)*, pages 349–354. 45, 67, 193

Giovannangeli, C. and Gaussier, P. (2010). Interactive teaching for vision-based mobile robots : A sensory-motor approach. *IEEE Transactions on Systems, Man and Cybernetics, Part A : Systems and Humans*, 40(1) :13–28. 47

Girard, B. (2003). *Intégration de la navigation et de la sélection de l'action dans une architecture de contrôle inspirée des ganglions de la base*. Thèse de doctorat, spécialité informatique, LIP6/AnimatLab, Université Pierre et Marie Curie, Paris, France. 102

Girard, B., Cuzin, V., Guillot, A., Gurney, K. N., and Prescott, T. J. (2003). A basal ganglia inspired model of action selection evaluated in a robotic survival task. *J Integr Neurosci*, 2(2) :179–200. 102

Gorchetchnikov, A. and Grossberg, S. (2007). Space, time and learning in the hippocampus : how fine spatial and temporal scales are expanded into population codes for behavioral control. *Neural Netw*, 20(2) :182–193. 38

Gothard, K. M., Hoffman, K. L., Battaglia, F. P., and McNaughton, B. L. (2001). Dentate gyrus and ca1 ensemble activity during spatial reference frame shifts in the presence and absence of visual input. *J Neurosci*, 21(18) :7284–7292. 24, 66

Granon, S., Hardouin, J., Courtièr, A., and Poucet, B. (1998). Evidence for the involvement of the rat prefrontal cortex in sustained attention. *Q J Exp Psychol B*, 51(3) :219–233. 28

Granon, S. and Poucet, B. (1995). Medial prefrontal lesions in the rat and spatial navigation : evidence for impaired planning. *Behav Neurosci*, 109(3) :474–484. 28

Granon, S. and Poucet, B. (2000). Involvement of the rat prefrontal cortex in cognitive functions : A central role for the prelimbic area. *Psychobiology*, 28(2) :229–237. 28

Granon, S., Vidal, C., Thinus-Blanc, C., Changeux, J. P., and Poucet, B. (1994). Working memory, response selection, and effortful processing in rats with medial prefrontal lesions. *Behav Neurosci*, 108(5) :883–891. 27

Griffin, A. L., Eichenbaum, H., and Hasselmo, M. E. (2007). Spatial representations of hippocampal ca1 neurons are modulated by behavioral context in a hippocampus-dependent memory task. *J Neurosci*, 27(9) :2416–2423. 24, 94

Groenewegen, H. J. (2003). The basal ganglia and motor control. *Neural Plast*, 10(1-2) :107–120. 28, 29

Grossberg, S., Levine, D., and Schmajuk, N. (1987). Predictive regulation of associative learning in a neural network by reinforcement and attentive feedback. *Int J Neurol*, 21-22 :83–104. 42, 98

Grossberg, S. and Merrill, J. W. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Brain Res Cogn Brain Res*, 1(1) :3–38. 42, 98, 99

Grossberg, S. and Schmajuk, N. (1987). Neural dynamics of attentionally modulated pavlovian conditioning : conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, 15(3) :195–240. 42, 98

Grossberg, S. and Schmajuk, N. A. (1989). Neural dynamics of adaptive timing temporal discrimination during associative learning. *Neural Netw.*, 2(2) :79–102. 42, 43, 54, 55, 60, 98, 130

Guazzelli, A., Corbacho, F. J., Bota, M., and Arbib, M. A. (1998). Affordances, motivations, and the world graph theory. *Adapt. Behav.*, 6(3-4) :435–471. 101

Gurney, K., Prescott, T. J., and Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biol Cybern*, 84(6) :401–410. 102

Gurney, K., Prescott, T. J., and Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. ii. analysis and simulation of behaviour. *Biol Cybern*, 84(6) :411–423. 102

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052) :801–806. 23, 38

Hampson, R. E., Heyser, C. J., and Deadwyler, S. A. (1993). Hippocampal cell firing correlates of delayed-match-to-sample performance in the rat. *Behav Neurosci*, 107(5) :715–739. 24

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42 :335–346. 13

Hasselmo, M. E. (2005). A model of prefrontal cortical mechanisms for goal-directed behavior. *J Cogn Neurosci*, 17(7) :1115–1129. 41, 42

Hasselmo, M. E., Bodelón, C., and Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm : separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Comput*, 14(4) :793–817. 74

Hasselmo, M. E. and Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Netw*, 18(9) :1172–1190. 41, 43, 51, 73

Hasselmo, M. E. and Fehlau, B. P. (2001). Differences in time course of ach and gaba modulation of excitatory synaptic potentials in slices of rat hippocampus. *J Neurophysiol*, 86(4) :1792–1802. 24, 74

Hasselmo, M. E. and Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region ca1 : computational modeling and brain slice physiology. *J Neurosci*, 14(6) :3898–3914. 22, 25, 74

Hasson, C. (2011). *Modélisation des mécanismes émotionnels pour un robot autonome : perspective développementale et sociale*. PhD thesis, Université de Cergy-Pontoise. 116

Hasson, C., Boucenna, S., Gaussier, P., and Hafemeister, L. (2010). Using emotional interactions for visual navigation task learning. In *Proceedings of the International Conference on Kansei Engineering and Emotion Reasearch*, pages 1578–1587. 154

Hasson, C. and Gaussier, P. (2010). Frustration as a generical regulatory mechanism for motivated navigation. In *Proc. IEEE/RSJ Int Intelligent Robots and Systems (IROS) Conf*, pages 4704–4709. 48, 126, 143

Hirel, J., Gaussier, P., and Quoy, M. (2010a). Model of the hippocampal learning of spatio-temporal sequences. In *Artificial Neural Networks – ICANN 2010*, volume 6354 of *Lecture Notes in Computer Science*, pages 345–351. Springer Berlin / Heidelberg. 63, 67

Hirel, J., Gaussier, P., and Quoy, M. (2011a). Biologically inspired neural networks for spatio-temporal planning in robotic navigation tasks. Accepté pour publication à ROBIO 2011.

Hirel, J., Gaussier, P., Quoy, M., and Banquet, J.-P. (2010b). Why and how hippocampal transition cells can be used in reinforcement learning. In *SAB '10 : Proceedings of the 11th international conference on Simulation of Adaptive Behavior*. Springer-Verlag. 107, 109

Hirel, J., Gaussier, P., Quoy, M., Banquet, J.-P., and Poucet, B. (2010c). Space and time-related firing in a model of hippocampo-cortical interactions. *BMC Neuroscience*, 11(Suppl 1) :163. 82

Hirel, J., Gaussier, P., Quoy, M., Banquet, J.-P., and Poucet, B. (2012). The hippocampo-cortical loop : Spatio-temporal learning and goal-oriented planning in navigation. En préparation pour soumission à Neural Computation.

Hirel, J., Quoy, M., and Gaussier, P. (2011b). Biologically plausible neural network for spatio-temporal robotic navigation. In *International Workshop on Bio-Inspired Robots*.

Hok, V. (2007). *Bases neurales des comportements orientés vers un but : Etude des corrélats de l'activité unitaire préfrontale et hippocampique dans une tâche de navigation.* PhD thesis, Aix-Marseille Université. 19, 26, 34, 82, 115, 167

Hok, V., Lenck-Santini, P.-P., Roux, S., Save, E., Muller, R. U., and Poucet, B. (2007a). Goal-related activity in hippocampal place cells. *J Neurosci*, 27(3) :472–482. 32

Hok, V., Lenck-Santini, P.-P., Save, E., Gaussier, P., Banquet, J.-P., and Poucet, B. (2007b). A test of the time estimation hypothesis of place cell goal-related activity. *J Integr Neurosci*, 6(3) :367–378. 32, 33, 78

Hok, V., Save, E., Lenck-Santini, P. P., and Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proc Natl Acad Sci U S A*, 102(12) :4602–4607. 31, 80

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press. 14

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia*, pages 215–232. MIT Press. 100, 106, 107

Insausti, R. and Amaral, D. G. (2004). Hippocampal formation. In *The Human Nervous System*. Elsevier. 19

Jay, T. M. and Witter, M. P. (1991). Distribution of hippocampal ca1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of phaseolus vulgaris-leucoagglutinin. *J Comp Neurol*, 313(4) :574–586. 21

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia : new anatomical and computational perspectives. *Neural Netw*, 15(4-6) :535–547. 100, 107

Jung, M. W. and McNaughton, B. L. (1993). Spatial selectivity of unit activity in the hippocampal granular layer. *Hippocampus*, 3(2) :165–182. 23

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning : A survey. *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH*, 4 :237–285. 98

Karlsson, M. P. and Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nat Neurosci*, 12(7) :913–918. 109

Kennedy, P. J. and Shapiro, M. L. (2009). Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *Proc Natl Acad Sci U S A*, 106(26) :10805–10810. 24, 53, 94

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., and Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia : From natural to artificial rats. *Adaptive Behavior*, 13(2) :131–148. 102, 107

Koene, A., Baldassarre, G., Mannella, F., and Prescott, T. (2009). Distal place recognition based navigation control inspired by hippocampus - amygdala interaction. In *9th International Conference on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems*. 40

Koene, R. A., Gorchetchnikov, A., Cannon, R. C., and Hasselmo, M. E. (2003). Modeling goal-directed spatial navigation in the rat based on physiological data from the hippocampal formation. *Neural Netw*, 16(5-6) :577–584. 41, 74

Koene, R. A. and Hasselmo, M. E. (2005). An integrate-and-fire model of prefrontal cortex neuronal activity during performance of goal-directed decision making. *Cereb Cortex*, 15(12) :1964–1981. 42

Kohonen, T. (1989). *Self-organization and associative memory : 3rd edition*. Springer-Verlag New York, Inc., New York, NY, USA. 40

Káli, S. and Dayan, P. (2000). The involvement of recurrent connections in area ca3 in establishing the properties of place fields : a model. *J Neurosci*, 20(19) :7463–7477. 39

Lagarde, M., Andry, P., and Gaussier, P. (2007). The role of internal oscillators for the one-shot learning of complex temporal sequences. In *Proceedings of the 17th international conference on Artificial neural networks*, ICANN'07, pages 934–943, Berlin, Heidelberg. Springer-Verlag. 56, 59, 69

Lagarde, M., Andry, P., and Gaussier, P. (2008a). Distributed real time neural networks in interactive complex systems. In *CSTST '08 : Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 95–100, New York, NY, USA. ACM. 67, 189

Lagarde, M., Andry, P., Gaussier, P., and Giovannangeli, C. (2008b). Learning new behaviors : Toward a control architecture merging spatial and temporal modalities. In *Workshop on Interactive Robot Learning - International Conference on Robotics : Science and Systems (RSS 2008)*. 56, 66

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 473–480, New York, NY, USA. ACM. 14

LeDoux, J. (1998). *The emotional brain : the mysterious underpinnings of emotional life*. A Touchstone book. Simon & Schuster. 116

Lee, I., Griffin, A. L., Zilli, E. A., Eichenbaum, H., and Hasselmo, M. E. (2006). Gradual translocation of spatial correlates of neuronal firing in the hippocampus toward prospective reward locations. *Neuron*, 51(5) :639–650. 22

Lee, I., Rao, G., and Knierim, J. J. (2004). A double dissociation between hippocampal subfields : differential time course of ca3 and ca1 place cells for processing changed environments. *Neuron*, 42(5) :803–815. 23, 72

Lenck-Santini, P.-P., Muller, R. U., Save, E., and Poucet, B. (2002). Relationships between place cell firing fields and navigational decisions by rats. *J Neurosci*, 22(20) :9035–9047. 22, 24

Lenck-Santini, P.-P., Rivard, B., Muller, R. U., and Poucet, B. (2005). Study of ca1 place cell activity and exploratory behavior following spatial and nonspatial changes in the environment. *Hippocampus*, 15(3) :356–369. 24, 45

Leonard, J. J. and Durrant-Whyte, H. F. (1991). Simultaneous map building and localization for an autonomous mobile robot. In *lEEE/RSJ International Workshop on Intelligent Robots and Systems IROS'91*. 13

Levy, W. B. (1996). A sequence predicting ca3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, 6(6) :579–590. 43

Lipton, P. A. and Eichenbaum, H. (2008). Complementary roles of hippocampus and medial entorhinal cortex in episodic memory. *Neural Plast*, 2008 :258467. 24, 53, 94

Lipton, P. A., White, J. A., and Eichenbaum, H. (2007). Disambiguation of overlapping experiences by neurons in the medial entorhinal cortex. *J Neurosci*, 27(21) :5787–5795. 24, 94

Liu, H. and Wan, W. (2010). Adaptive replanning in hard changing environments. In *Proc. IEEE/RSJ Int Intelligent Robots and Systems (IROS) Conf*, pages 5912–5918. 144

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol*, 67(1) :145–163. 29

Mallot, H. A., Bülthoff, H. H., Georg, P., Schölkopf, B., and Yasuhara, K. (1995). View-based cognitive map learning by an autonomous robot. In *Proc. ICANN'95, Int. Conf. on Artificial Neural Networks*, volume II, pages 381–386. EC2. 40

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., and Robbins, T. (2002). Decision-making processes following damage to the prefrontal cortex. *Brain*, 125(Pt 3) :624–639. 28

Mannella, F. and Baldassarre, G. (2007). A neural-network reinforcement-learning model of domestic chicks that learn to localize the centre of closed arenas. *Philos Trans R Soc Lond B Biol Sci*, 362(1479) :383–401. 101

Manns, J. R., Howard, M. W., and Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, 56(3) :530–540. 24, 25, 53

Martinet, L.-E., Sheynikhovich, D., Benchenane, K., and Arleo, A. (2011). Spatial learning and action planning in a prefrontal cortical network model. *PLoS Comput Biol*, 7(5) :e1002045. 42, 51, 168

Matsumoto, M. and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248) :837–841. 126

Mauk, M. D. and Buonomano, D. V. (2004). The neural basis of temporal processing. *Annu Rev Neurosci*, 27 :307–340. 25

Maurice, N., Deniau, J. M., Glowinski, J., and Thierry, A. M. (1999). Relationships between the prefrontal cortex and the basal ganglia in the rat : physiology of the cortico-nigral circuits. *J Neurosci*, 19(11) :4674–4681. 28

Meeter, M., Murre, J. M. J., and Talamini, L. M. (2004). Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, 14(6) :722–741. 74

Mehta, M. R., Barnes, C. A., and McNaughton, B. L. (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc Natl Acad Sci U S A*, 94(16) :8918–8921. 22, 53

Mehta, M. R., Quirk, M. C., and Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, 25(3) :707–715. 22, 53

Middleton, F. A. and Strick, P. L. (2000). Basal ganglia output and cognition : evidence from anatomical, behavioral, and clinical studies. *Brain Cogn*, 42(2) :183–200. 28

Milad, M. R., Rauch, S. L., Pitman, R. K., and Quirk, G. J. (2006). Fear extinction in rats : implications for human brain imaging and anxiety disorders. *Biol Psychol*, 73(1) :61–71. 25

Milford, M. J. and Wyeth, G. F. (2008). Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24(5) :1038–1053. 39

Moga, S. (2000). *Apprendre par imitation : une nouvelle voie d'apprentissage pour les robots autonomes*. PhD thesis, Université de Cergy-Pontoise. 62, 187

Moga, S., Gaussier, P., and Banquet, J.-P. (2003). Sequence learning using the neural coding. In *IWANN'03 : Proceedings of the Artificial and natural neural networks 7th international conference on Computational methods in neural modeling*, pages 198–205, Berlin, Heidelberg. Springer-Verlag. 53, 55

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci*, 16(5) :1936–1947. 99

Morales, G. J., Ramcharan, E. J., Sundararaman, N., Morgera, S. D., and Vertes, R. P. (2007). Analysis of the actions of nucleus reuniens and the entorhinal cortex on eeg and evoked population behavior of the hippocampus. *Conf Proc IEEE Eng Med Biol Soc*, 2007 :2480–2484. 21

Moutarlier, P. and Chatila, R. (1990). An experimental system for incremental environment modelling by an autonomous mobile robot. In *Experimental Robotics I*, volume 139 of *Lecture Notes in Control and Information Sciences*, pages 327–346. Springer Berlin / Heidelberg. 13

Mulder, A. B., Tabuchi, E., and Wiener, S. I. (2004). Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur J Neurosci*, 19(7) :1923–1932. 23

Muller, R., Stead, M., and Pach, J. (1996). The hippocampus as a cognitive graph. *J Gen Physiol.*, 107(6) :663—-694. 40, 51

Muller, R. U. and Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J Neurosci*, 7(7) :1951–1968. 24

Nagumo, J. (1967). A learning method for system identification. *IEEE Trans. Autom. Control*, 12(3) :282–287. 63

Nakahara, H., Doya, K., and Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J Cogn Neurosci*, 13(5) :626–647. 101

Narayanan, N. S. and Laubach, M. (2009). Delay activity in rodent frontal cortex during a simple reaction time task. *J Neurophysiol*, 101(6) :2859–2871. 27

Oderfeld-Nowak, B., Potempska, A., and Roskoski, R. (1980). Acetylcholine levels increase in rat hippocampus following acute septal lesions : evidence for interactions between cholinergic and noncholinergic neurons. *Neuroscience*, 5(10) :1699–1703. 22

Okatan, M. (2009). Correlates of reward-predictive value in learning-related hippocampal neural activity. *Hippocampus*, 19(5) :487–506. 101

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Res*, 34(1) :171–175. 22

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press. 22, 24, 37, 39, 66

Paquier, W. and Chatila, R. (2003). Learning new representations and goals for autonomous robots. In *ICRA*, pages 803–808. IEEE. 163

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat Neurosci*, 12(7) :919–926. 27, 109

Philippsen, R., Kolski, S., Maček, K., and Jensen, B. (2008). Mobile robot planning in dynamic environments and on growable costmaps. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 144

Poucet, B. (1997). Searching for spatial unit firing in the prelimbic area of the rat medial prefrontal cortex. *Behav Brain Res*, 84(1-2) :151–159. 31

Prescott, T. J., González, F. M. M., Gurney, K., Humphries, M. D., and Redgrave, P. (2006). A robot model of the basal ganglia : behavior and intrinsic processing. *Neural Netw*, 19(1) :31–61. 103

Prescott, T. J. and Mayhew, J. E. W. (1992). Obstacle avoidance through reinforcement learning. *Advances in Neural Information Processing Systems*, 4 :523–530. 100

Prisco, G. V. D. and Vertes, R. P. (2006). Excitatory actions of the ventral midline thalamus (rhomboid/reuniens) on the medial prefrontal cortex in the rat. *Synapse*, 60(1) :45–55. 26

Quirk, G. J., Muller, R. U., and Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *J Neurosci*, 10(6) :2008–2017. 24

Quirk, G. J., Muller, R. U., Kubie, J. L., and Ranck, J. B. (1992). The positional firing properties of medial entorhinal neurons : description and comparison with hippocampal place cells. *J Neurosci*, 12(5) :1945–1963. 23

Quirk, G. J. and Vidal-Gonzalez, I. (2006). Keeping the memories flowing. *Nat Neurosci*, 9(10) :1199–1200. 21

Quoy, M., Moga, S., Gaussier, P., and Revel, A. (2000). Parallelization of neural networks using pvm. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, volume 1908, pages 289–296. Springer Berlin / Heidelberg. 67, 189

Ragozzino, M. E., Detrick, S., and Kesner, R. P. (1999). Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. *J Neurosci*, 19(11) :4585–4594. 27, 124

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal : a role in discovering novel actions ? *Nat Rev Neurosci*, 7(12) :967–975. 117, 143

Redish, A. D. and Touretzky, D. S. (1997). Cognitive maps beyond the hippocampus. *Hippocampus*, 7(1) :15–35. 38, 94

Redish, A. D. and Touretzky, D. S. (1998). The role of the hippocampus in solving the morris water maze. *Neural Comput*, 10(1) :73–111. 40, 43, 51

Rengervé, A., Boucenna, S., Andry, P., and Gaussier, P. (2010). Emergent imitative behavior on a robotic arm based on visuo-motor associative memories. In *Proc. IEEE/RSJ Int Intelligent Robots and Systems (IROS) Conf*, pages 1754–1759. 43, 55, 146, 156, 157

Rengervé, A., Hirel, J., Andry, P., Quoy, M., and Gaussier, P. (2011a). On-line learning and planning in a pick-and-place task demonstrated through body manipulation. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–6. 146

Rengervé, A., Hirel, J., Quoy, M., Andry, P., and Gaussier, P. (2011b). A simple neural network controller merging different behaviors for collector robots. In *International Workshop on Bio-Inspired Robots*.

Rich, E. L. and Shapiro, M. (2009). Rat prefrontal cortical neurons selectively code strategy switches. *J Neurosci*, 29(22) :7208–7219. 27, 124

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cereb Cortex*, 10(3) :284–294. 29

Rolls, E. T., Stringer, S. M., and Elliot, T. (2006). Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Network*, 17(4) :447–465. 38

Rondi-Reig, L., Petit, G. H., Tobin, C., Tonegawa, S., Mariani, J., and Berthoz, A. (2006). Impaired sequential egocentric and allocentric memories in forebrain-specific-nmda receptor knock-out mice during a new task dissociating strategies of navigation. *J Neurosci*, 26(15) :4071–4081. 25, 70

Rosenblatt, F. (1962). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Washington, Spartan Books. 14

Rossier, J., Kaminsky, Y., Schenk, F., and Bures, J. (2000). The place preference task : a new tool for studying the relation between behavior and place cell activity in rats. *Behav Neurosci*, 114(2) :273–284. 30

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Learning internal representations by error propagation*, pages 318–362. MIT Press, Cambridge, MA, USA. 14

Samsonovich, A. and McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *J Neurosci*, 17(15) :5900–5920. 39

Save, E., Nerad, L., and Poucet, B. (2000). Contribution of multiple sensory information to place field stability in hippocampal place cells. *Hippocampus*, 10(1) :64–76. 24

Schmajuk, N. A. and DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychol Rev*, 99(2) :268–305. 75, 160

Schmajuk, N. A. and Thieme, A. D. (1992). Purposive behavior and cognitive mapping : a neural network model. *Biol Cybern*, 67(2) :165–174. 40

Schnider, A., Mohr, C., Morand, S., and Michel, C. M. (2007). Early cortical response to behaviorally relevant absence of anticipated outcomes : a human event-related potential study. *Neuroimage*, 35(3) :1348–1355. 29, 124

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80(1) :1–27. 29, 99, 116, 124

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J Neurosci*, 13(3) :900–913. 29, 99, 116, 124

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306) :1593–1599. 99

Schultz, W., Tremblay, L., and Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cereb Cortex*, 10(3) :272–284. 29

Shah, N. and Alexandre, F. (2011). Reinforcement learning and dimensionality reduction : a model in computational neuroscience. In *International Joint Conference on Neural Networks IJCNN 2011*. 103

Skaggs, W. E., McNaughton, B. L., Wilson, M. A., and Barnes, C. A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6(2) :149–172. 23

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci*, 3(9) :919–926. 15

Sotres-Bayon, F., Cain, C. K., and LeDoux, J. E. (2006). Brain mechanisms of fear extinction : historical perspectives on the contribution of prefrontal cortex. *Biol Psychiatry*, 60(4) :329–336. 25

Steels, L. (2008). The symbol grounding problem has been solved. so what's next ? In de Vega, M., editor, *Symbols and Embodiment : Debates on Meaning and Cognition*, chapter 12. Oxford University Press, Oxford. 13

Stewart, M. and Fox, S. E. (1990). Do septal neurons pace the hippocampal theta rhythm ? *Trends Neurosci*, 13(5) :163–168. 22

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3 :9–44. 98, 99

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bull.*, 2 :160–163. 108

Sutton, R. S. and Barto, A. G. (1981). Toward a modern theory of adaptive networks : expectation and prediction. *Psychol Rev*, 88(2) :135–170. 98

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *J Neurosci*, 10(2) :420–435. 23

Tesauro, G. (1992). Practical issues in temporal difference learning. *Machine Learning*, 8(3-4) :257–277. 99

Thorpe, C., Floresco, S., Carr, J., and Wilkie, D. (2002). Alterations in time-place learning induced by lesions to the rat medial prefrontal cortex. *Behav Processes*, 59(2) :87. 27

Thorpe, C. M., Hallett, D., and Wilkie, D. M. (2007). The role of spatial and temporal information in learning interval time-place tasks. *Behav Processes*, 75(1) :55–65. 27

Thorpe, C. M. and Wilkie, D. M. (2002). Unequal interval time-place learning. *Behav Processes*, 58(3) :157–166. 27

Thorpe, C. M. and Wilkie, D. M. (2006). Rats' performance on an interval time-place task : increasing sequence complexity. *Learn Behav*, 34(3) :248–254. 27

Tijsseling, A. G. and Berthouze, L. (2003). A neural network architecture for the categorization of temporal information. 43

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol Rev*, 55(4) :189–208. 39

Varela, F., Thompson, E., and Rosch, E. (1991). *The embodied mind : cognitive science and human experience*. MIT Press. 13

Vertes, R. P., Hoover, W. B., Szigeti-Buck, K., and Leranth, C. (2007). Nucleus reuniens of the midline thalamus : link between the medial prefrontal cortex and the hippocampus. *Brain Res Bull*, 71(6) :601–609. 26, 81

Voicu, H. and Schmajuk, N. (2000). Exploration, navigation and cognitive mapping. *Adaptive Behavior*, 8(3/4) :207–224. 40, 41, 43

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3) :279–292. 98, 100, 103

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291(5504) :599–600. 17

Whishaw, I. Q. and Maaswinkel, H. (1998). Rats with fimbria-fornix lesions are impaired in path integration : a role for the hippocampus in "sense of direction". *J Neurosci*, 18(8) :3050–3058. 25

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *IRE WESCON Convention Record*, pages 123–134, Cambridge, MA, USA. MIT Press. 62

Wiener, S. I., Paul, C. A., and Eichenbaum, H. (1989). Spatial and behavioral correlates of hippocampal neuronal activity. *J Neurosci*, 9(8) :2737–2763. 24

Wilson, S. W. (1991). The animat path to ai. In *Proceedings of the first international conference on simulation of adaptative behavior : From animals to animats*, pages 15–20. 14

Witter, M. P. (1993). Organization of the entorhinal-hippocampal system : a review of current anatomical data. *Hippocampus*, 3 Spec No :33–44. 19

Yamashita, Y. and Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model : a humanoid robot experiment. *PLoS Comput Biol*, 4(11) :e1000220. 43

Yaniv, D., Vouimba, R. M., Diamond, D. M., and Richter-Levin, G. (2003). Simultaneous induction of long-term potentiation in the hippocampus and the amygdala by entorhinal cortex activation : mechanistic and temporal profiles. *Neuroscience*, 120(4) :1125–1135. 20

Yu, A. J., Giese, M. A., and Poggio, T. A. (2001). Biologically plausible neural circuits for realization of maximum operations. 107

Zukowgoldring, P. and Arbib, M. (2007). Affordances, effectivities, and assisted imitation : Caregivers and the directing of attention. *Neurocomputing*, 70(13-15) :2181–2193. 148

ANNEXE A

Apprentissage temporel de signaux continus

Le but des travaux présentés dans cette annexe est de vérifier la capacité d'un LMS à apprendre à prédire l'évolution de signaux continus en fonction du temps passé depuis le dernier événement perceptif détecté par le système (voir chapitre 3 pour une description de l'architecture utilisée). Une analyse de l'effet de différents paramètres du système a été conduite. L'expérience reprend celle décrite dans la section 3.2.1 avec quelques différences. Un robot simulé suit une trajectoire définie avec une vitesse constante (fig. A.1). Il doit uniquement apprendre à prédire l'évolution des différentes activités de cellules de lieu en fonction du temps passé dans le lieu actuel. L'estimation temporelle du délai écoulé depuis l'entrée dans le lieu est donnée par les batteries de cellules granulaires dans le gyrus dentelé, dont les activités sont définies par l'équation (3.5). L'équation d'apprentissage (3.9) au niveau de CA3 est la règle de Widrow-Hoff originale (LMS).

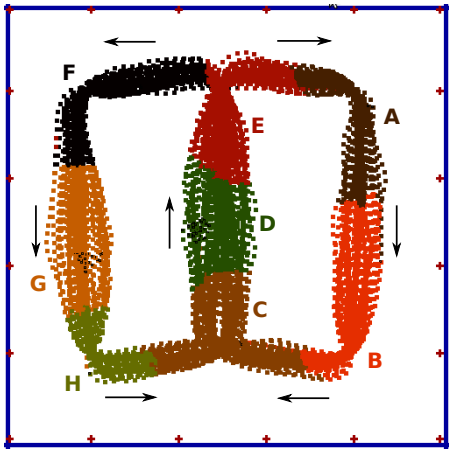


FIGURE A.1 – Trajectoire suivie par le robot simulé. Chaque couleur correspond à une cellule de lieu gagnante. Le robot apprend à prédire l'évolution de tous les signaux des cellules de lieu en fonction du temps passé dans le lieu actuel.

Les batteries de cellules granulaires ont des activités régulières, bien adaptées à la prédiction de signaux continus. L'objectif de l'expérience est de vérifier la nécessité de rendre les instants d'apprentissage du système aléatoires. En effet, de précédents travaux avaient montré que l'apprentissage devrait être réalisé de manière aléatoire pour éviter des effets de récence [Moga,

Essai	Paramètres			MSE (prédiction)
1	$C = 0.06$	$f = 25$	$\alpha = 0.5$	0.01519466
2	$C = 0.06$	$f = 15$	$\alpha = 0.5$	0.01727539
3	$C = 0.06$	$f = 10$	$\alpha = 0.5$	0.01331468
4	$C = 0.06$	$f = 5$	$\alpha = 0.5$	0.00947678
5	$C = 0.06$	$f = 1$	$\alpha = 0.5$	0.01059164
6	$C = 0.06$	$f = 1$	$\alpha = 0.9$	0.01543765
7	$C = 0.06$	$f = 1$	$\alpha = 0.9999$	0.02465519
8	$C = 0.06$	$f = 1$	$\alpha = 0.2$	0.00730408
9	$C = 0.06$	$f = 1$	$\alpha = 0.05$	0.01258609
10	$C = 0.04$	$f = 1$	$\alpha = 0.5$	0.01320566
11	$C = 0.08$	$f = 1$	$\alpha = 0.5$	0.01524083

TABLE A.1 – *Mean Squared Error* (MSE) de la prédiction de l'évolution des activités de cellules de lieu en fonction des paramètres C, f et α. Autres paramètres fixés : $T_0 = 0.1, n_C = 15, \Delta_t = 10, M = 0$

2000], qui favoriseraient l'apprentissage des valeurs les plus récentes des signaux de cellules de lieu et entraîneraient des erreurs de prédiction pour les valeurs plus anciennes. Plusieurs expériences ont donc été réalisées (tab. A.1). Les paramètres du système qui ont été variés sont la période moyenne d'apprentissage f donnant une probabilité $\frac{1}{f}$ d'apprendre à chaque itération, le paramètre C réglant la variance des gaussiennes utilisées pour simuler l'activité des cellules granulaires (voir équation 3.5) et la vitesse d'apprentissage α du LMS.

Les résultats sont calculés après que le robot a parcouru la trajectoire plusieurs fois et appris à prédire les signaux des cellules de lieu. L'apprentissage est alors désactivé et on compare les prédictions avec les signaux réels pendant plusieurs tours. L'erreur moyenne des prédictions est donnée. On peut voir qu'un apprentissage très fréquent ($f = 1$ ou $f = 5$) produit de meilleurs résultats qu'avec des fréquences d'apprentissage moins élevées (ce qui peut être dû à un plus grand nombre d'apprentissages), cela va à encontre d'un effet de récence qui dégraderait les performances. Le paramètre C a été choisi pour éviter un trop grand recouvrement des activités des cellules granulaires dans DG, ce qui peut expliquer pourquoi l'effet de récence ne se montre pas et les bonnes performances de l'apprentissage non aléatoire. Ces effets de récence commencent à dégrader les performances dans le cas où la variance des gaussiennes est augmentée (C plus grand) ou dans le cas où la vitesse d'apprentissage est trop élevée (α grand, ce qui rend le LMS instable). On peut donc conclure qu'un apprentissage aléatoire n'est pas nécessaire dans le cas où les activités des cellules granulaires dans DG sont bien temporellement séparées. On peut alors réduire grandement la vitesse d'apprentissage du système. Toutefois, ce type de modélisation demande un grande nombre de neurones si on veut couvrir des périodes de temps longues. L'utilisation de cellules granulaires ayant des périodes d'activité courtes et bien temporellement séparées n'est nécessaire que lorsqu'on veut prédire précisément des signaux continus. Dans la plupart des expériences robotiques conduites dans cette thèse, nous prédisons des événements ponctuels. Nous utilisons donc un modèle avec des gaussiennes de variances croissantes et d'amplitudes décroissantes qui permet d'avoir de bonnes précision de prédiction sur les timings courts. Pour les timings longs, les prédictions sont moins précises mais ne requièrent qu'un petit nombre de cellules granulaires dans le modèle.

ANNEXE **B**

Outils de simulation de réseaux de neurones

Les travaux de l'équipe utilisent en grande majorité deux outils de conception et simulation de réseaux de neurones : *Coeos* et *Promethe* [Lagarde et al., 2008a; Quoy et al., 2000]. Ces outils sont développés dans l'équipe depuis une vingtaine d'années et évoluent au fur et à mesure des besoins de modélisation et de contrôle robotique. J'ai eu l'occasion, au cours de ma thèse, d'utiliser ces outils pour développer mes architectures neuronales, mais aussi de participer au développement de ces outils écrits en C avec une approche orientée objet. Ainsi, j'ai pu améliorer les fonctionnalités existantes et en ajouter de nouvelles pour répondre à des problèmes particuliers pour lesquels aucune solution n'avait été implémentée, pour optimiser des temps de calcul ou encore pour faciliter la conception de réseaux de neurones. Je présenterai donc ici des exemples d'architectures neuronales conçues avec Coeos et s'exécutant dans Promethe. Je détaillerai également quelques contributions majeures au développement de ces outils.

B.1 Interface de conception : Coeos

Coeos est une interface graphique de conception de réseaux de neurones distribués. Elle permet de constituer des fichiers détaillant la structure du réseau de neurones, qui peuvent ensuite être lus par Promethe pour leur simulation informatique. Une architecture neuronale globale peut être découpée en plusieurs parties, qui s'exécutent sur des machines différentes et communiquent entre elles par un protocole réseau (fig. B.1). Ce fonctionnement permet de répartir la charge de calcul sur plusieurs machines. Il permet aussi de rendre la modélisation des réseaux de neurones plus modulaire, en ayant par exemple une partie s'occupant des traitements visuels tandis qu'une autre s'occupe du contrôle moteur du robot.

Chaque sous-partie de l'architecture globale représente un réseau de neurones à part entière, avec des entrées et sorties qui sont reçues et envoyées par des communications réseau. La conception graphique des réseaux de neurones adopte une approche modulaire par groupes de neurones. La visualisation est celle d'un graphe orienté qui définit des groupes de neurones comme noeuds du graphes et les connexions synaptiques entre ces groupes comme les arêtes (fig. B.2). Chaque groupe *neuronal* représente un ensemble de neurones avec les mêmes propriétés (mêmes équations d'activation et d'apprentissage). Les liens entre les groupes représentent un type de connectivité (plastique, fixe, un-vers-tous, un-vers-un etc.). Des groupes *algorithmiques* existent également pour traiter certaines parties algorithmiques de l'architecture (comme la communication avec le matériel embarqué, le contrôle moteur du robot etc.).

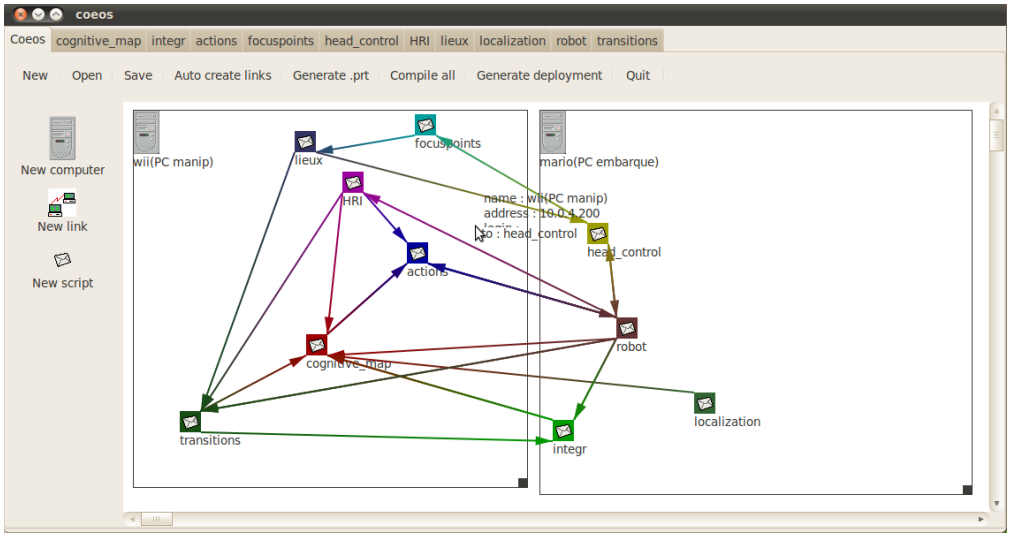


FIGURE B.1 – Capture d'écran de l'interface de Coeos montrant la répartition des sous-parties de réseaux de neurones sur 2 machines : un PC embarqué sur le robot et un poste de travail fixe. Les communications entre les 2 machines se font par Wifi. Les flèches représentent les échanges d'informations entre les différentes sous-parties.

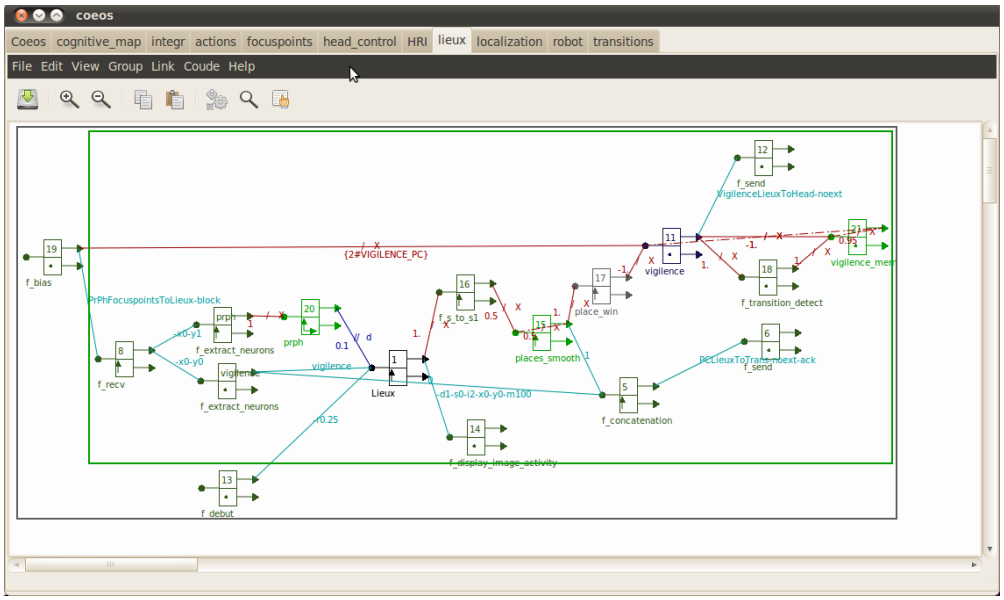


FIGURE B.2 – Capture d'écran de l'interface de Coeos montrant l'architecture neuronale d'une des sous-parties correspondant à l'apprentissage des cellules de lieu.

Les travaux de développement que j'ai pu effectuer sur Coeos incluent la refonte du code d'affichage des groupes de neurones, afin de le rendre plus performant et plus lisible, et la possibilité de sélectionner plusieurs groupes simultanément et de procéder à des éditions ou des déplacements de ces ensembles de groupes.

B.2 Simulateur temps-réel distribué : Promethe

Les réseaux de neurones conçus avec l'aide de Coeos peuvent ensuite être utilisés avec Promethe. Ce simulateur gère l'exécution des fonctions associées aux différentes groupes de neurones (activation et apprentissage pour les groupes neuronaux, fonction en C implémentée par l'utilisateur pour les groupes algorithmiques). Un séquenceur s'occupe de l'ordonnancement de l'exécution des différentes groupes et de l'optimisation de cette exécution en fonction de la vitesse de simulation. En pratique, l'exécution des groupes se fait en suivant le principe d'un réseau de Petri dans lequel des jetons sont propagés pour marquer l'exécution d'un groupe. Ainsi, un groupe ne peut être exécuté que si tous les groupes lui transmettant des informations ont été exécutés. La mise à jour de l'ensemble du réseau de neurones se fait alors par vagues.

On peut également concevoir différentes boucles s'exécutant avec des constantes de temps différentes par l'utilisation d'échelle de temps, qui fonctionnent comme des boucles *for* imbriquées. L'aspect d'exécution temps-réel peut être géré par des groupes émettant des jetons d'exécution avec une période définie, et s'assurant que les contraintes temps-réel sont bien respectées pour la mise à jour du réseau de neurones. Enfin, certains groupes algorithmiques peuvent gérer des aspects de communication réseau en recevant et transmettant des activités neuronales à d'autres parties de l'architecture. Une interface de contrôle permet d'afficher des informations de debug (fig. B.3) telles que l'activité des neurones de divers groupes, des informations sur la position du robot etc.

La conception d'architectures neuronales pour Promethe pendant ma thèse a impliqué le développement de multiples fonctions algorithmiques et de plusieurs groupes neuronaux pour l'implémentation des équations d'apprentissage mises au point pendant mes recherches. J'ai également travaillé sur le développement du simulateur lui-même : librairies de communication avec le matériel (commandes d'un système pan-tilt Biclops, librairie de captures d'images avec une caméra en utilisant Video4Linux2, récupération des informations d'un joystick etc.), intégration du simulateur de robot et d'environnement virtuels dans la librairie matériel du simulateur pour rendre le passage sur robot réel aussi transparent que possible, gestion de la compilation du simulateur etc.

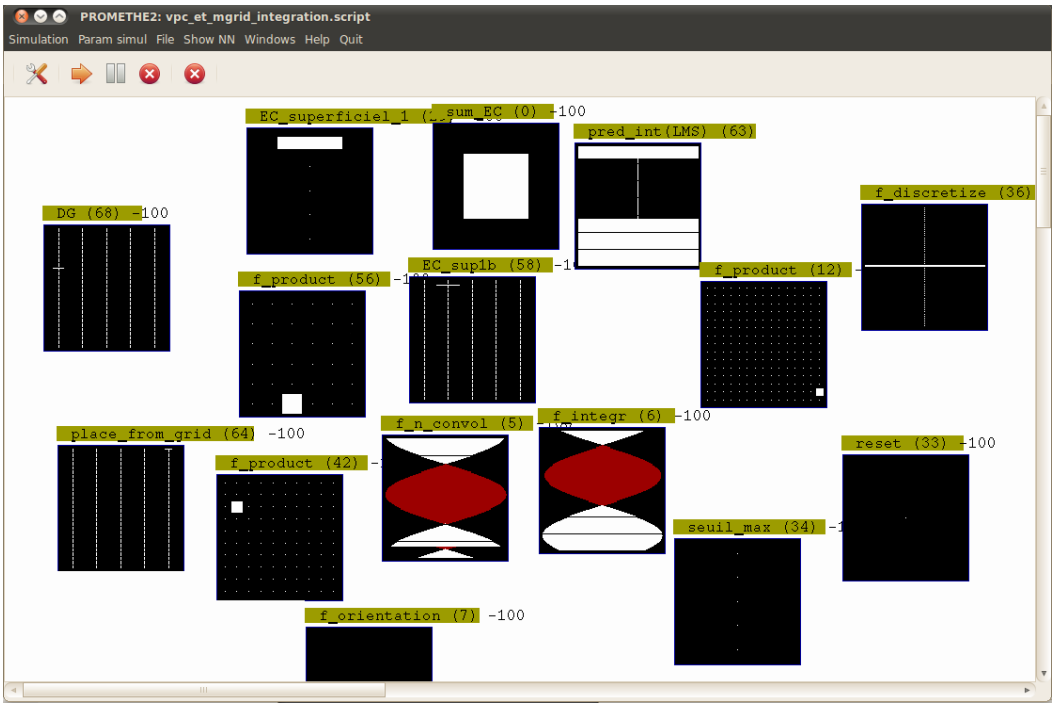


FIGURE B.3 – Interface de contrôle de Promethe. L'activité des neurones de plusieurs groupes est visible. Les neurones sont représentés par des rectangles blancs (ou rouges pour des activités négatives) dont la taille représente le niveau d'activité. Les groupes de neurones peuvent avoir différentes topologies (neurone unique, vecteur horizontal ou vertical, matrice).

ANNEXE C

Plates-formes robotiques

C.1 Robot mobile d'intérieur

Le robulab 10 de chez Robosoft est le robot mobile pour la navigation en intérieur utilisé dans l'équipe (fig. C.1). Il possède une base mobile équipée de 2 roues motrices et 2 roues libres. 9 capteurs ultrason sont répartis autour de sa base. Un coffre situé sur la base peut accueillir le matériel nécessaire pour le contrôle et la navigation visuelle du robot. Une caméra montée sur un système pan-tilt permet au robot de suivre un objet du regard ou bien de faire des panoramas à 360°de son environnement. Une boussole électronique permet au robot de connaître son orientation. Cette boussole matérielle pourrait être remplacée par l'utilisation d'une boussole neuronale intégrant des aspects de vision et de proprioception [Giovannangeli and Gaussier, 2007]. Un système de localisation utilisant des repères au plafond permet de connaître la position exacte du robot. Cette information n'est pas utilisée dans les architectures neuronales développées mais sert uniquement à la représentation des données acquises pendant les expériences. Un système de cou artificiel, utilisant un joystick intégré et des capteurs de pression, permet de guider le robot en tirant sur une laisse qui y est attachée. Un capteur permet de détecter la couleur du sol directement sous le robot, ce qui permet de simuler la présence de ressources par des zones de couleur. Enfin, un PC embarqué permet de faire tourner une partie de l'architecture directement sur le robot et de communiquer avec le matériel et le robot lui-même. Un routeur wifi permet d'échanger des informations avec d'autres parties de l'architecture pouvant s'exécuter sur un certain nombre de machines distantes. La base mobile possède un espace qui peut permettre d'y monter un bras robotique. Les dernières versions du robulab possèdent un bras Katana directement intégré à la base (fig. C.2).

C.2 Robot mobile d'extérieur

Le roburoc 4 de chez Robosoft est un robot d'extérieur doté de 4 roues motrices (fig. C.3). Il peut se déplacer jusqu'à 20km/h et embarquer une charge utile de 100kg. L'équipement monté sur le roburoc est relativement identique à celui du robulab. Le système pan-tilt avec caméra est monté sur une plate-forme de stabilisation afin de le maintenir à l'horizontale sur terrain accidenté. Il dispose également d'un laser pour une détection d'obstacles plus précise.

Lors de ma thèse, j'ai été amené à adapté des architectures de navigation sur roburoc 4. J'ai ainsi participé à des expériences préliminaires d'apprentissage de cellules de lieu devant mener

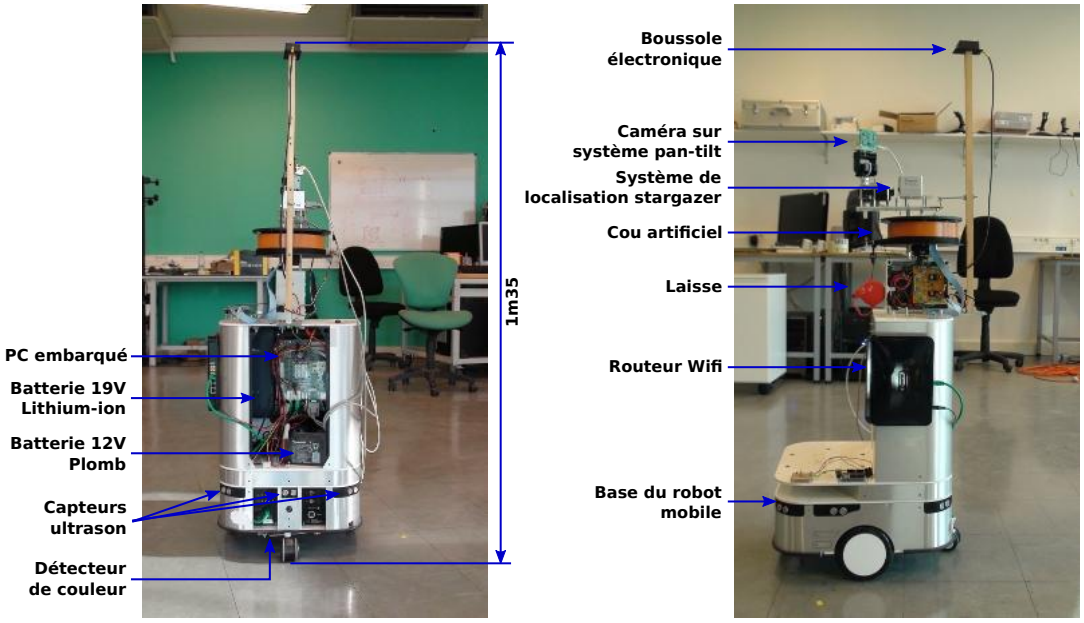


FIGURE C.1 – Robot robulab 10 de chez Robosoft, équipé pour des expériences de navigation visuelle en intérieur.

FIGURE C.2 – Robot robulab de dernière génération avec un bras Katana intégré à la base.

FIGURE C.3 – Robot roburoc 4 de chez Robosoft, équipé pour des expériences de navigation visuelle en extérieur.

par la suite à des expériences de navigation en extérieur dans de grands environnements.

C.3 Bras robotique

Outre son utilisation possible lorsqu'il est monté sur un robulab, j'ai également travaillé avec un bras robotique fixé sur un plan de travail (fig. C.4). Le bras robotique est un Katana de chez Neuronics. Il possède 6 degrés de liberté (rotation de la base, 3 articulations sur le bras, rotation du poignet et fermeture/ouverture de la pince). Une caméra montée sur un système pan-tilt est située au dessus du plan de travail afin de pouvoir faire de la reconnaissance et du suivi d'objets ainsi que pour apprendre des associations visuo-motrices.

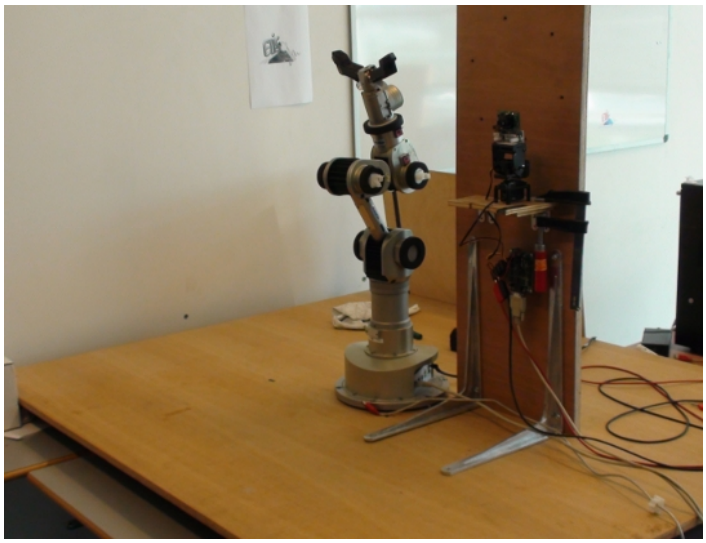


FIGURE C.4 – Table de manipulation d'objets avec un bras Katana de chez Neuronics et un système de caméra sur pan-tilt.