Université Pierre et Marie Curie, Paris 6

DEA de Sciences Cognitives

MEMOIRE

présenté en vue d'obtenir

Le DEA de Sciences Cognitives de l'Université de Paris VI

Sujet : *Un modèle d'apprentissage par renforcement dans une architecture de contrôle de la sélection de l'action chez le Rat Artificiel Psikharpax.*

Directeurs de stage *: Dr Agnès Guillot, AnimatLab, LIP6*
Pr Alain Berthoz, LPPA, Collège de France
Encadrants *: Benoît Girard, AnimatLab*
Dr Sidney Wiener, LPPA

REGUIGNE-KHAMASSI Mehdi

20 juin 2003

TABLE DES MATIERES

INTRODUCTION : OBJECTIF DU STAGE

Le travail de stage présenté dans ce mémoire participe au projet Psikharpax, un robot dont l'architecture de contrôle vise à synthétiser certains des mécanismes connus pour être impliqués dans la sélection de l'action et la navigation chez le rat. Ce projet réunit notamment le Laboratoire de Physiologie de la Perception et de l'Action et l'AnimatLab du Laboratoire d'Informatique de Paris 6 et présente deux objectifs principaux :

- D'un côté comprendre la nature des traitements cognitifs impliqués dans les processus de décision pour l'action chez le rat, et les mécanismes neuraux qui sous-tendent ces processus. La modélisation robotique peut aider à cette compréhension puisqu'elle permet d'observer les propriétés dynamiques de structures nerveuses dans leur ensemble alors que l'électrophysiologie ne permet à ce jour d'enregistrer qu'une centaine de neurones simultanément.
- De l'autre, concevoir des robots 'animats', c'est-à-dire inspirés des animaux, faisant preuve d'autonomie et d'adaptation face à des environnements changeants, mais pouvant également servir de modèles du comportement animal [Guillot & Meyer, 2002 ; Webb, 2001]. Alors que la robotique classique ne parvient généralement à implémenter que des robots monotâches, l'approche biomimétique pourrait, en s'inspirant des interactions entre les aires cérébrales et de leurs architectures, parvenir à concevoir des robots combinant plusieurs fonctions cognitives.

Le travail présenté ici consiste à doter l'architecture de contrôle de Psikharpax de certaines capacités d'apprentissage. Actuellement, cette architecture gère la sélection de ses actions en s'inspirant des connections de structures nerveuses présentes chez les vertébrés et supposées effectuer cette sélection : les ganglions de la base. Cet ensemble de noyaux sous corticaux, intimement relié au système dopaminergique, est considéré comme impliqué dans le contrôle des comportements locomoteurs [Graybiel, 1995 ; DeLong, 2000]. Le module de sélection de l'action de Psikharpax a ensuite été connecté à un module de navigation [Filliat, 2001 ; Filliat, Meyer, 2002] au cours d'une thèse soutenue en septembre 2003 par Benoît Girard. La modélisation de cette intégration s'inspire elle aussi des ganglions de la base. Ainsi simulé, le

robot peut explorer et mémoriser l'emplacement de plusieurs sources dont la nature – récompense ou punition – lui est donnée a priori, et alterner entre ces sources de façon à maximiser son temps de survie.

Or, l'hypothèse a été récemment émise que les ganglions de la base avaient un rôle important dans l'analyse contextuelle de l'environnement qui permettrait l'association entre la nature de certains stimuli et une éventuelle récompense donnée par l'environnement, et dans l'utilisation de cette information pour la formation de programmes moteurs [Houk, 1995]. Deux mécanismes d'apprentissage différents et siégeant dans deux parties différentes des ganglions de la base, le striatum dorsal et le striatum ventral, auraient été identifiés comme supportant cette formation :

- L'apprentissage stimulus-réponse (S-R), ou apprentissage des habitudes, s'apparentant au conditionnement instrumental et permettant la formation de procédures comportementales dont l'exécution serait supportée par le striatum dorsal ou moteur [Houk, 1995].
- L'apprentissage de la sélection de stratégies de navigation pour les comportements orientés vers un but, supposée siéger dans le noyau accumbens, recouvrant les parties associative, limbique et motrice du striatum ventral [Pennartz, Groenewegen & Lopes Da Silva, 1994 ; Ikemoto & Panksepp, 1999 ; Wiener, Tabuchi & Mulder, 2002].

Des découvertes récentes ont montré qu'un modèle de type Acteur-Critique (Actor-Critic), architecture utilisée en Intelligence Artificielle pour doter un robot de capacités d'apprentissage par renforcement et fondée sur un algorithme d'Apprentissage par Différence Temporelle (TD-Learning) [Sutton & Barto, 1998], permettait de bien décrire les mécanismes soulignant le premier apprentissage [Houk, Adams & Barto, 1995 ; Schultz, Dayan & Montague, 1997 ; Graybiel, 1998]. Cette hypothèse a été favorisée par un certain nombre de données neurophysiologiques montrant que la dopamine pourrait jouer le rôle d'un signal d'erreur de prédiction agissant sur les ganglions de la base, permettant ainsi de renforcer l'association entre un stimulus conditionnel – prédisant une récompense – et l'action menant à la récompense prédite par ce stimulus [Schultz, 1998, 2001 ; Schultz, Tremblay & Hollerman, 2000].

Partant de cette hypothèse, la partie Actor, supposée siéger dans la partie des ganglions de la base associée au striatum dorsal, a déjà été modélisée précédemment dans l'architecture de la sélection de l'action [Gurney, Prescott & Redgrave, 2001a,b ; Girard et al, 2002]. Mon travail de stage consiste à adjoindre une partie Critic à ce modèle Actor. Cette modélisation sera inspirée de modèles computationnels déjà existants, mais aura l'originalité d'intégrer une partie Actor détaillée, contrairement à celle des modèles développés jusqu'à présent (voir [Joel, Niv & Ruppin, 2002] pour une revue).

Ce mémoire décrit la démarche adoptée, les différents travaux effectués pendant le stage, ainsi que les résultats obtenus lors de la simulation du modèle sur le robot. Nous y verrons dans une première partie, un état de l'art sur la structure et les fonctions des ganglions de la base, et nous tenterons d'y montrer ce sur quoi se fondent ces hypothèses d'un apprentissage par renforcement supporté par la dopamine. Cette partie se terminera par l'analyse de données électrophysiologiques effectuée au LPPA sur des rats réalisant une tâche de recherche de récompenses dans un labyrinthe en croix, cette tâche impliquant les deux mécanismes d'apprentissage énoncés plus haut.

La deuxième partie de ce mémoire présentera ensuite un état de l'art de l'apprentissage par renforcement en Intelligence Artificielle ainsi qu'une brève revue des modèles Actor-Critic d'apprentissage par renforcement inspirés des ganglions de la base.

L'implémentation de l'un de ces modèles sera présentée dans une troisième partie, et simulée dans la même tâche que les rats du LPPA.

Enfin, nous ferons un bilan de ce travail et une discussion générale, afin de tirer des conclusions sur la validité du modèle, et d'ouvrir des perspectives sur le type de mécanismes qui peuvent être utilisés pour modéliser le deuxième apprentissage énoncé : l'apprentissage de la sélection de stratégies de navigation pour les comportements orientés vers un but. Nous nous demanderons notamment si un modèle dérivé du même type d'architecture Actor-Critic peut sous-tendre cet apprentissage.

I. DONNEES EXPERIMENTALES SUR L'APPRENTISSAGE PAR RENFORCEMENT DANS LES GANGLIONS DE LA BASE

A. Les ganglions de la base.

Les ganglions de la base (GB) sont un ensemble de noyaux subcorticaux interconnectés, éléments d'une boucle cortex-GB-thalamus-cortex et intimement lié au système dopaminergique mésencéphalique (figure 1). Un grand nombre de données expérimentales a été accumulé concernant la structure des ganglions de la base et la physiologie des neurones les composant [Houk, Davis & Beiser, 1995 ; Graybiel, 1998 ; DeLong, 2000]. Cet intérêt particulier est lié au fait que de nombreux désordres du mouvement (maladie de Parkinson, maladie de Huntington, syndronme de Tourette, etc.) chez l'humain sont dus à des maladies affectant les ganglions de la base.

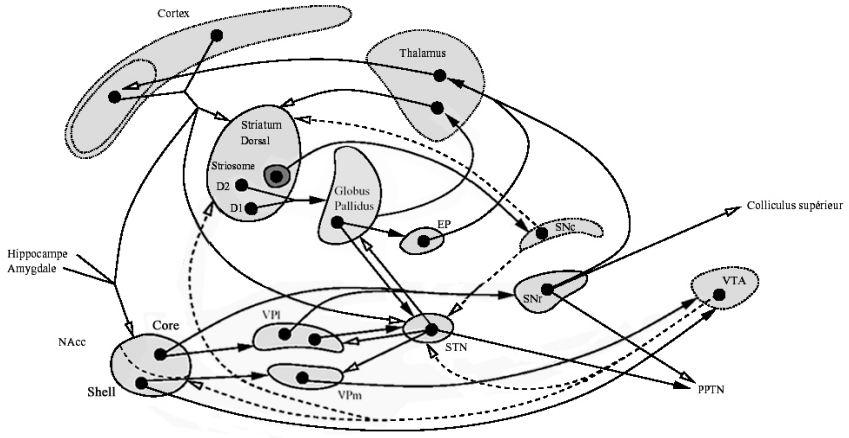

Figure 1 : Représentation simplifiée des principales connections des ganglions de la base, mettant en valeur la boucle cortex-GB-thalamus-cortex ainsi que les interactions avec le système dopaminergique. Les connections excitatrices (glutamatergiques) sont représentées par des flèches évidées, les connections inhibitrices (GABAergiques) par des flèches pleines, et les connections dopaminergiques par des flèches en pointillés. (Voir dans le texte pour la signification des différentes abréviations).

Pour la compréhension de la suite de ce document, il convient de décrire ici les principales caractéristiques qui vont servir dans la suite de l'exposé. Les noyaux des GB chez le rat sont les suivants (figure 1) :

- Les entrées incluent : le noyau subthalamique (STN), le striatum ventral (comprenant le noyau accumbens (NAcc) et la partie profonde du tubercule olfactif) et le striatum

dorsal, lui-même composé d'une grande région, les matrisomes, parsemée de compartiments bien délimités, appelés les striosomes. Cette distinction est importante pour le travail décrit ici puisque ces deux régions n'ont pas les mêmes efférences : les striosomes projettent sur les neurones dopaminergiques tandis que les matrisomes projettent sur les neurones pallidaux [Graybiel, 1998 ; Windels, 2001]. Nous verrons plus loin ce que cela implique.

- Les noyaux intermédiaires incluent : le globus pallidus (GP) et son extension ventrale, le pallidum ventral (VP), lui-même divisé en deux parties distinctes, l'une médiale et ventrale (abrégée VPm pour des raisons historiques), l'autre dorsale et latérale (VPl).
- Les sorties incluent : le noyau entopédonculaire (EP) et la substance noire réticulée (SNr).

Les GB sont liés au système dopaminergique mésencéphalique, composé principalement de la substance noire compacte (SNc) et de l'aire tegmentale ventrale (VTA). Les striosomes du striatum dorsal se projettent exclusivement sur la SNc, tandis que le noyau accumbens est divisé en une écore ventro-latérale appelée « shell », se projetant préférentiellement vers la VTA et la SNc, et un noyau dorso-médian, ou « core », qui projette principalement sur la SNc [Haber, Fudge & McFarland, 2000 ; Ikemoto, 2002]. SNc et VTA se projettent en retour sur les noyaux d'entrées des GB (figure 1).

D'autre part, il apparaît que les GB réalisent un ensemble d'intégrations effectuées dans des canaux fonctionnellement distincts, maintenus séparés depuis le cortex jusqu'aux noyaux de sortie des GB. Compte tenu de la connectivité des ganglions de la base, cette intégration pourrait s'apparenter à la sélection, parmi un grand nombre d'activations du cortex, de celles qui sont pertinentes en fonction du contexte [Joel & Weiner, 2000 ; Joel, Niv & Ruppin, 2002]. En effet, de nombreux travaux suggèrent que les GB sont en position, via la désinhibition du thalamus, de réguler l'activité du cortex, et que ceci leur permet d'avoir un rôle global de sélection : sélection des comportements, motivation, planification, apprentissage de la récompense associée à une action, mémoire de travail, classification (voir [Greenberg, 2001] pour une revue). Les GB sont alors considérés comme un interrupteur de type "qui perd gagne" : le canal d'entrée soumis à l'excitation corticale la plus forte est désinhibé, ce qui permet un renforcement de cette activation corticale maximale [Chevalier &

Deniau, 1990]. Alors que les GB étaient initialement considérés comme faisant partie du système moteur extrapyramidal, c'est-à-dire de la partie du système moteur en charge des aspects automatiques du mouvement, ces hypothèses suggèrent que le rôle des GB peut être étendu à des fonctions de plus haut niveau, voir cognitives.

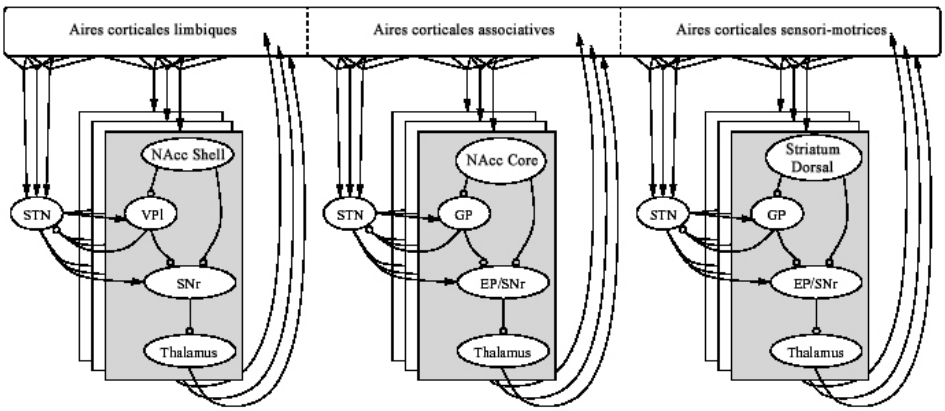

Figure 2 : Représentation schématique des trois grandes boucles cortex-GB-thalamus-cortex et des canaux qui les composent. Pour chacune des trois boucles, on a représenté trois canaux.

Enfin, il apparaît que la boucle cortex-GB-thalamus-cortex peut être subdivisée, chez le rat, en trois boucles principales aux fonctions différenciées – motrice, associative et limbique – (voir figure 2), au sein desquelles d'autres subdivision sont susceptibles d'exister [Alexander & Crutcher, 1990 ; Alexander, Crutcher & DeLong, 1990]. On considère ainsi que les GB sont en position de sélectionner :

- Les mouvements, qu'il s'agisse des mouvements des membres ou de saccades oculaires (boucle motrice).
- Les comportements en fonction des motivations (boucles motrices et limbiques).
- Les éléments à conserver en mémoire de travail (boucle associative).
- Les motivations en fonction des états internes (boucle limbique).

Notons que l'existence de ces boucles cortex-GB-thalamus-cortex pourrait mettre les GB en position non seulement de sélectionner des états corticaux à un instant donné, mais également d'intégrer et de contrôler leurs successions, et par là de générer des séquences d'actions [Berns & Sejnowski, 1996, 1998 ; Beiser & Houk, 1998 ;Nakahara, Doya & Hikosaka, 1999, 2001 ; Koechlin et al, 2002].

B. Dopamine et mécanismes d'apprentissage par renforcement dans les GB.

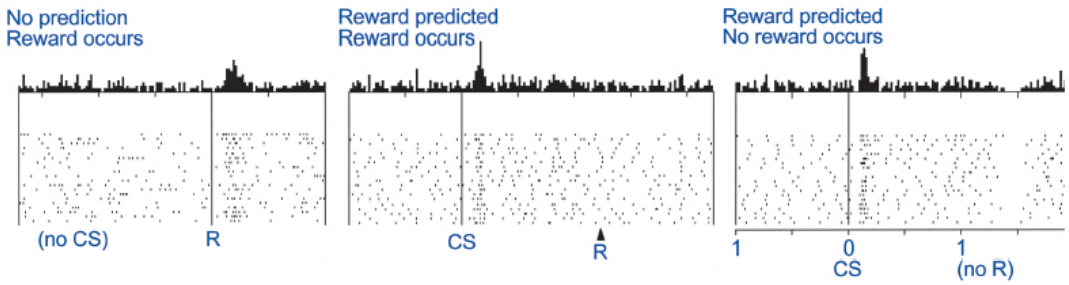

Figure 3 : Activités des neurones dopaminergiques mesurées pendant des expériences sur des singes apprenant des tâches comportementales [Schultz, 2001]. CS : Stimulus Conditionnel, R : Récompense.

L'idée que les ganglions de la base puissent être le siège d'un apprentissage par renforcement pour les associations stimulus-réponse (S-R), c'est-à-dire le choix d'une action en réponse à un contexte sensoriel, avec la dopamine comme signal de ce renforcement a été amplement développée grâce aux travaux de Wolfram Schultz et collègues [Schultz, Dayan & Montague, 1997]. Au cours de nombreuses enregistrements d'électrophysiologiques chez des singes apprenant des tâches comportementales, il a découvert que les neurones dopaminergiques répondaient de façon phasique à des récompenses inattendues (stimuli inconditionnels), et qu'au fur et à mesure de l'expérience les décharges de dopamine se décalaient dans le temps pour répondre à des stimuli prédisant ces récompenses (stimuli conditionnels), et ne plus répondre aux récompenses elles-mêmes [Schultz, 1998, 2001 ; Schultz, Tremblay & Hollerman, 2000]. De plus, ce système dopaminergique semble être très sensible au temps entre un stimulus et la récompense prévue, puisqu'on constate une dépression nette de l'activité des neurones dopaminergiques dans le cas où une récompense prédite ne se produit pas, comme le montre la figure 3. Enfin, des effets de potentiation et dépression à long terme (LTP et LTD respectivement) ont été observés sur les synapses cortico-striatales suite à des expositions à la dopamine, cette plasticité synaptique pouvant être la base d'un mécanisme d'apprentissage basé sur la dopamine. La dopamine est donc couramment considérée comme un signal d'erreur de prédiction jouant le rôle de médiateur du renforcement permettant la formation de procédures motrices, ou habitudes, dépendantes du contexte sensoriel envoyé du cortex vers le striatum dorsal [Houk, Adams & Barto, 1995 ; Schultz, Dayan & Montague, 1997].

Toutefois, nous avons vu dans le paragraphe précédent que les différentes parties du système dopaminergique ne reçoivent pas des afférences des mêmes structures. En particulier SNc reçoit principalement des afférences du striatum dorsal et du noyau accumbens « core » tandis que VTA en reçoit davantage du noyau accumbens « shell » (striatum ventral) [Haber, Fudge & MacFarland, 2000 ; Ikemoto, 2002]. On peut donc se demander si la dopamine peut coder le même type de signal dans le striatum dorsal et le striatum ventral, et si un même type d'apprentissage que le S-R, peut être envisagé dans les deux parties.

Ikemoto et Panksepp [Ikemoto & Panksepp, 1999] répondent plutôt par la négative puisque, tout en ne réfutant pas le rôle possible de la dopamine du SNc comme médiateur du renforcement, ils considèrent la dopamine de VTA comme signal de nouveauté incitant à la génération de comportements non stéréotypés de recherche de récompense. En effet, contrairement au striatum dorsal qui ne reçoit d'afférences que de la partie sensorielle du cortex, le noyau accumbens reçoit également des informations de l'hippocampe, de l'amygdale, des aires corticales préfrontales (associatives) et limbiques, ce qui le mettrait en position de prendre en compte davantage de paramètres pour le choix des comportements, et lui a valu l'hypothèse de jouer le rôle d'interface entre les systèmes limbiques et moteurs lui permettant de faire intervenir la motivation – par l'intermédiaire du « shell » – et les stratégies de navigation – par le « core » – dans le contrôle de comportements orientés vers un but [Pennartz, Groenewegen & Lopes Da Silva, 1994 ; Shibata et al, 2001 ; Wiener, Tabuchi & Mulder, 2002 ; Berthoz, 2003].

Il semblerait donc qu'il faille considérer deux types d'apprentissages différents dans les parties ventrale et dorsale des ganglions de la base :

- L'un ne tenant compte que du contexte sensoriel et permettant la formation de simples associations S-R (striatum dorsal).
- L'autre prenant en compte la motivation et les informations spatiales pour mettre en concurrence des stratégies de navigation spatiale vers la récompense (striatum ventral) [Wiener, Tabuchi & Mulder, 2002].

Une expérience spécifique d'électrophysiologie menée au LPPA permet de distinguer ces deux sortes d'apprentissage chez des rats recherchant la plus grande source de récompense

possible dans un labyrinthe en forme de croix (tâche du « Plus-Maze ») [Tabuchi, Mulder & Wiener, 2000, 2003]. Une partie de mon travail de stage a consisté en l'analyse de données mesurées dans le striatum ventral pendant cette tâche, résultats qui sont présentés dans le paragraphe suivant.

C. Analyse de données électrophysiologiques.

Dans une démarche de recherche des capacités d'adaptation du rat qui sont liées à la sélection de l'action et à la régulation des comportements orientés vers un but, il apparaissait important de comprendre quels mécanismes d'apprentissage sont mis en jeu dans la partie core du noyau accumbens. L'hypothèse à tester ici consiste à dire que l'accumbens core met en comparaison des informations égocentriques d'un côté, des informations allocentriques et la motivation de l'animal de l'autre, de façon à choisir la meilleure stratégie pour atteindre un but déterminé [Shibata et al, 2001 ; Wiener, Tabuchi & Mulder, 2002]. Ces stratégies appliquées à la tâche du Plus-Maze sont explicitées dans le tableau 1.

Stratégies employées par le rat pendant la tâche	Mécanismes supposés impliqués (parties du striatum impliquées)
Une stratégie égocentrique : se diriger vers un lieu visible.	Apprentissage S-R pour associer les caractéristiques visuelles de ce lieu avec l'éventuelle présence de récompense (striatum dorsal).
Une stratégie allocentrique : utiliser une représentation spatiale de l'environnement (ex : une carte cognitive [O'Keefe & Nadel, 1978]) et les indices visuels distaux (voir légende) pour localiser le lieu contenant la plus grande quantité de récompense.	Intégration des informations spatiales provenant de l'hippocampe, des informations visuelles provenant du cortex et des informations liées au caractère aversif ou attractif d'un stimulus provenant de l'amygdale, pour l'apprentissage de la meilleure stratégie de navigation vers ce lieu (core du noyau accumbens).

Table 1 : Stratégies alternées par le rat lors de la tâche du Plus-Maze [Tabuchi et al, 2000]. Concernant l'hypothèse que ce sont préférentiellement les indices visuels distaux qui sont utilisés pour mettre à jour les représentations spatiales du rat, se reporter à la littérature pour les cellules de lieu [O'Keefe & Dostrovsky, 1971 ; Cressant, Muller & Poucet, 1997] et pour les cellules de direction de la tête [Ranck, 1984 ; Zugaro, Berthoz & Wiener, 2001 ; Zugaro, 2002].

C.1 Description de la tâche expérimentale.

La tâche consiste à placer des rats partiellement privés d'eau dans un labyrinthe en croix contenant une source de récompense (un réservoir d'eau) à l'extrémité de chaque branche (figure 4). Chacun des quatre réservoirs contient une quantité différente de récompense (respectivement 1, 3, 5 et 7 gouttes) et le but de la manipulation est de faire effectuer aux rats un comportement dirigé par un but (trouver la plus source délivrant la plus forte récompense),

en combinant les deux types de mécanismes décrits dans le tableau 1, sachant qu'un réservoir ne délivre de la récompense que s'il est éclairé.

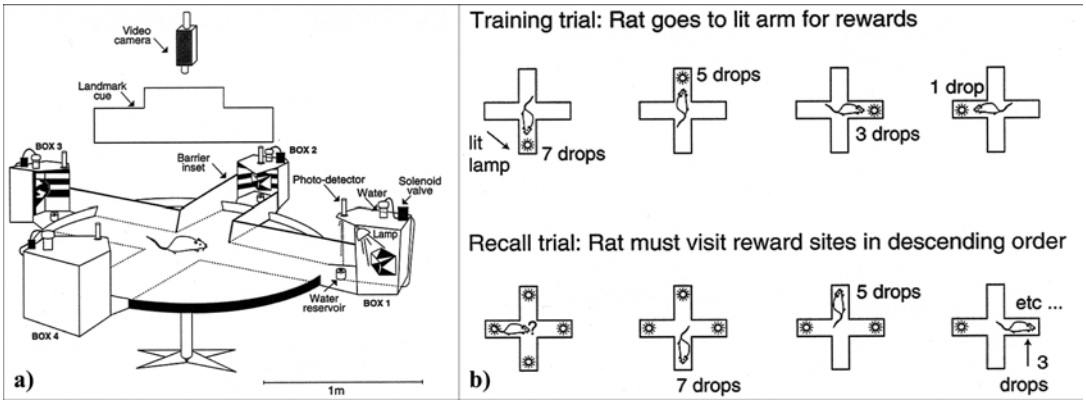

Figure 4 : Dispositif expérimental [Wiener, Tabuchi & Mulder, 2002]. a) Le labyrinthe vu de perspective. b) Les deux phases de l'expérience : La phase d'apprentissage où les réservoirs sont éclairés un par un dans l'ordre décroissant de leur quantité de récompense, et la phase de rappel où tous les réservoirs sont éclairés et où le rat doit retrouver l'ordre de visite précédent.

Dans une phase de pré entraînement, un seul des 4 réservoirs est éclairé à un moment donné, et les rats doivent apprendre à associer cet éclairage avec la présence d'eau dans ce réservoir et l'obscurité des autres avec l'absence de récompense. Cette phase relève d'un apprentissage S-R impliquant le striatum dorsal (tableau 1). Ensuite, il s'agit pour le rat d'apprendre à aller visiter les quatre réservoirs successivement dans un ordre bien défini : de la meilleure source de récompense à la moins bonne de façon à localiser l'emplacement de chacune des valeurs de récompenses (figure 4). Chaque jour, les rats sont exposés à une nouvelle distribution des différents volumes de récompense parmi les quatre réservoirs qu'ils visitent un par un pendant une série d'essais constituant la "phase d'entraînement". Puis, pendant une dernière phase appelée "phase de rappel", on teste la façon dont les rats se souviennent de cette distribution des quantités de récompense lorsqu'ils ont le choix entre n'importe lequel des quatre réservoirs. Les distributions sont ensuite changées et le rat doit recommencer une nouvelle série de phases d'apprentissage et de rappel pendant que les mêmes cellules sont de nouveau enregistrées. L'hypothèse étant que la combinaison des différents mécanismes impliqués est effectuée dans le noyau accumbens, les cellules qui ont été mesurées lors de cette tâche se situent toutes dans le striatum ventral (dans les parties accumbens core, noyau caudé ventro-médial et accumbens shell dorsolatéral).

C.2 Matériel et méthodes.

Une description détaillée du matériel utilisé dans cette expérience est présentée dans [Tabuchi, Mulder & Wiener, 2002]. Nous nous contenterons ici de résumer l'essentiel:

- Les 7 rats utilisés ont été mis sous régime de privation partielle d'eau.
- Ils ont été placés dans un labyrinthe en forme de croix d'1m70 de côté.
- Pour chaque performance de la tâche expérimentale, les activités de plusieurs neurones du striatum ventral ont été mesurées simultanément.
- De ces enregistrements, après amplification, filtrage, échantillonage et synchronisation avec les données comportementales, seuls les signaux dépassant un certain seuil donné ont été stockés pour une analyse ultérieure. Ainsi, le signal électrique continu est transformé en une suite d'événements discrets, un train de potentiels d'action appelé aussi « raster », support à partir duquel l'analyse présentée ici a été effectuée.

C.3 Résultats de l'analyse.

De façon à tester l'hypothèse selon laquelle l'accumbens intègre des informations spatiales, motivationnelles, et liées à la récompense pour contrôler les comportements orientés vers un but, et de façon à essayer de détecter des mécanismes pouvant relever d'un certain apprentissage, nous voulions tester statistiquement l'influence de 4 facteurs sur les décharges des cellules mesurées :

1) Les corrélations spatiales – différences dans les taux de décharges des cellules de l'Accumbens lorsque l'animal occupait des positions différentes dans les 4 branches du labyrinthe.
2) Les corrélations comportementales – comparaisons des taux de décharges pendant l'approche d'une récompense, pendant la consommation de cette récompense, ou pendant le départ vers un autre site.
3) Les corrélations avec les phénomènes d'apprentissage entre les différentes phases de l'expérience – comparaisons des taux de décharge entre les phases d'entraînement (nécessitant un stockage d'informations sur la distribution spatiale des sources de récompenses) et les phases de rappel.

4) Les variations de taux de décharges corrélées avec la prédiction ou l'attribution de récompense.

Parmi 170 cellules mesurées dans le Noyau Accumbens Core, le Noyau Caudé vendromédial et l'Accumbens Shell dorsolatéral, seules 117 cellules ont été catégorisées comme phasiques (neurones ayant des périodes de décharge distinctes et des périodes de silence). Pour la compréhension des analyses qui vont suivre, nous avons adopté trois types de représentations des données décrites sur la figure 6.

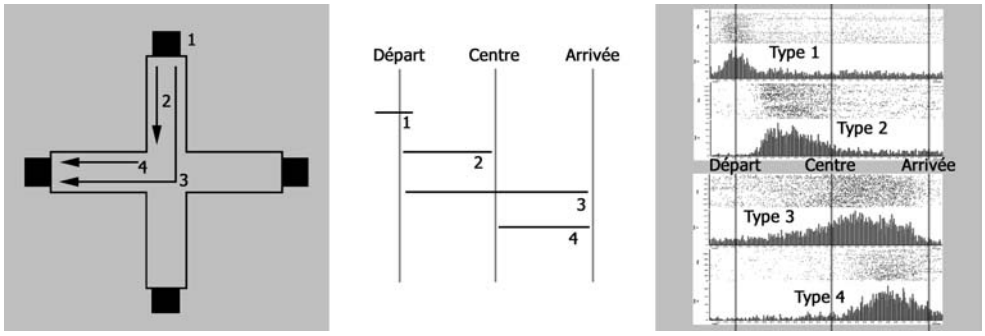

Figure 6 : Convention pour la représentation des données. Le schéma de gauche représente le labyrinthe vu de dessus comprenant une représentation spatiale des différents trajets possibles du rat. Dans le schéma du milieu, chaque ligne horizontale représente l'intervalle temporel de décharge d'une cellule par rapport aux trois points de repères temporels : le Départ d'un réservoir, l'Arrivée au Centre et l'Arrivée au réservoir suivant. Le schéma de droite propose une représentation en histogrammes des décharges des mêmes cellules par rapport aux trois repères temporels.

C.3.1 Corrélations comportementales.

Famille des cellules ayant des décharges ponctuelles	nb	Famille des cellules ayant des décharges prolongées	nb
Départ d'un réservoir	**16**	D'un réservoir jusqu'au centre	**05**
Arrivée au centre	**03**	Du centre jusqu'à un réservoir	**13**
Arrivée à un réservoir	**19**	D'un réservoir jusqu'au réservoir suivant	**16**
		Pendant la consommation de récompense	**19**

Table 2 : Catégorisation des cellules mesurées dans le Striatum Ventral. Ces cellules ont été regroupées en familles selon leur profil temporel de décharge. A côté de chaque catégorie est indiqué le nombre de cellules classées dans cette catégorie. 'nb' signifie nombre de cellules.

91 neurones phasiques (77% du total) ont révélé des changements significatifs de leur taux de décharge liés aux différentes phases comportementales de la tâche spatiale. Nous avons classés ces cellules en 7 grands groupes selon le moment de la tâche où elles ont déchargé. Le premier point intéressant à remarquer est que ces 7 groupes se répartissent sur l'ensemble de la tâche, et qu'ils se chevauchent, accompagnant ainsi chaque phase de la séquence comportementale effectuée par le rat pour aller d'une source de récompense à une autre en

passant par le centre. La figure 7 montre l'ensemble des cellules ayant révélé des corrélations comportementales significatives, disposées par rapport aux trois repères temporels.

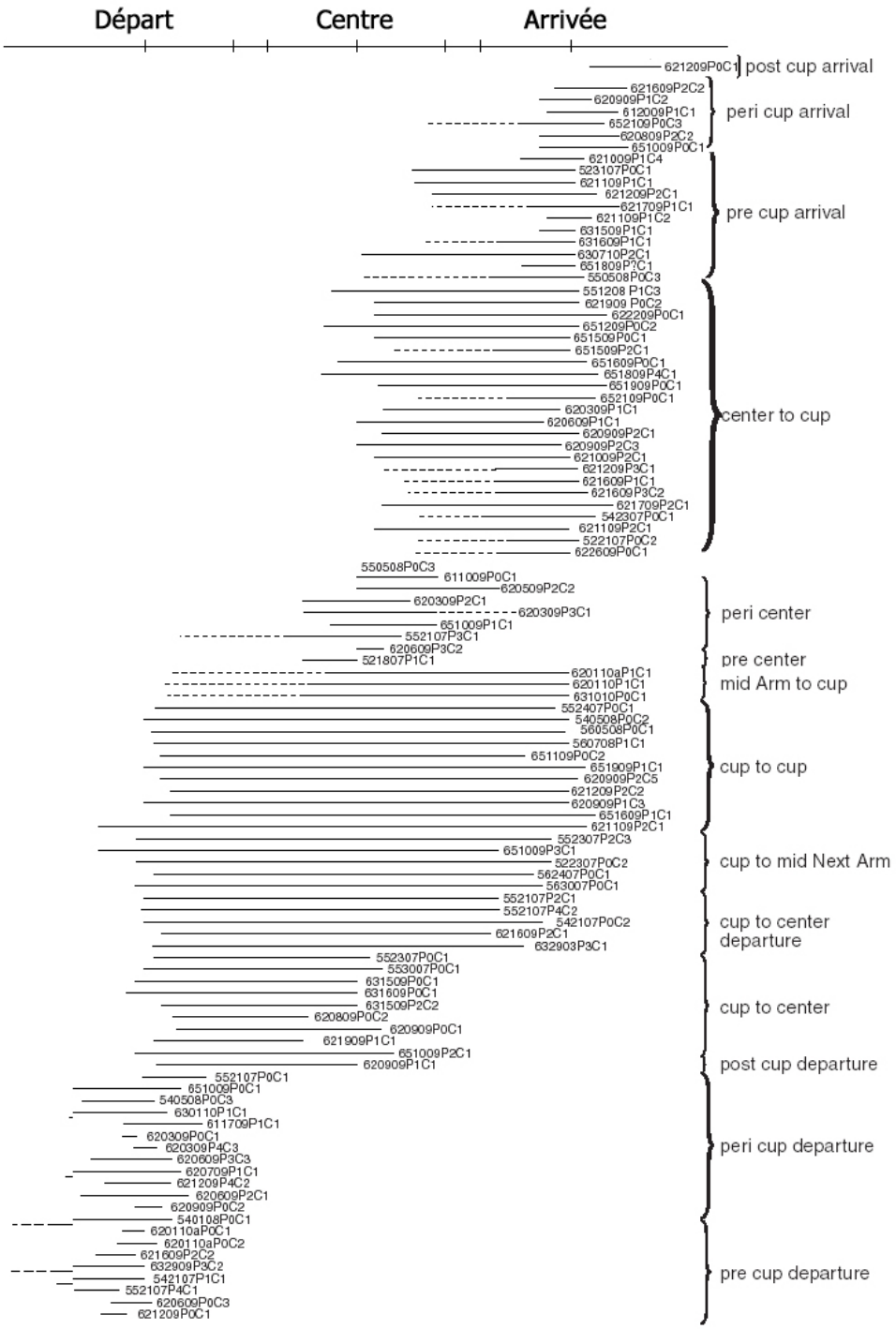

Figure 7 : Schéma des intervalles de décharges des différentes cellules du Striatum Ventral ayant révélé des corrélations comportementales significatives. Les trois points de repères temporels sont : le Départ d'une réservoir, l'Arrivée au centre et l'Arrivée au réservoir suivant. Chaque ligne horizontale est une cellule analysée.

D'autre part, les 7 groupes de cellules choisis pour la catégorisation se répartissent en 2 grandes familles : les cellules qui ont une décharge assez ponctuelle temporellement autour d'un évènement repère de la tâche (qui sert comme point de synchronisation pour les analyses), et les cellules qui ont une décharge qui s'étale dans le temps entre deux évènements repères. Le tableau 2 présente les 7 groupes de cellules classés par familles, avec le nombre de cellules classées dans chaque groupe.

C.3.2 Corrélations spatiales.

25 des 91 cellules ayant des corrélations comportementales (28%) ont une sélectivité spatiale pour un des 4 bras du labyrinthe. Parmi ces 25 cellules, 7 ont une absence de décharge significative pour certains bras, et 18 cellules présentent des modulations d'un bras à l'autre. La figure 8 présente un exemple de cellule ayant une sélectivité spatiale pour certains bras du labyrinthe comparée à une cellule n'ayant pas de corrélation spatiale.

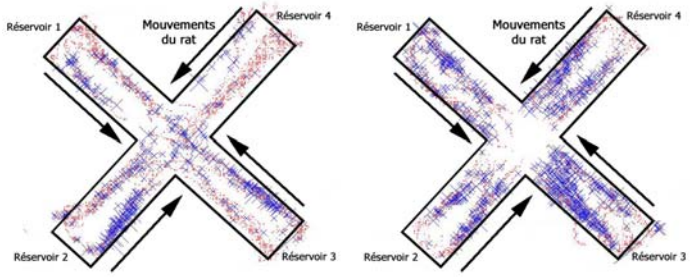

Figure 8 : Exemple de corrélation spatiale. Les petits points représentent les positions successives du rat dans le labyrinthe, et les croix les positions pour lesquelles les cellules ont eu une bouffée de décharges. A gauche, une cellule (ref. 522307P2C3) ayant une sélectivité spatiale pour certain des bras du labyrinthe. A droite, une cellule (ref. 553007P0C1) n'ayant pas de sélectivité spatiale. Les flèches montrent que les décharges de ces cellules n'ont été représentées que lorsque le rat se déplaçait d'un réservoir vers le centre.

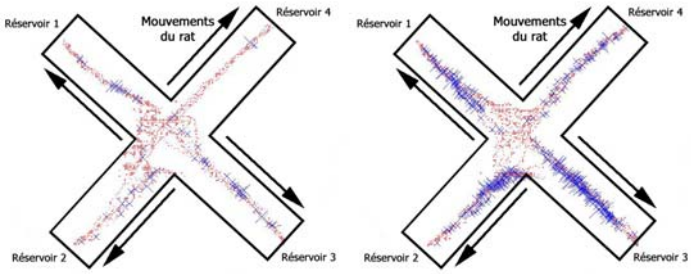

Figure 9 : Cellule comportementale déchargeant pour le déplacement de l'animal du centre vers les réservoirs et ayant des variations entraînement/rappel (ref. 620909P2C3). Cette cellule décharge plus pour les phases de rappel que pour les phases d'entraînement. A gauche, les décharges de la cellule pendant les phases d'entraînement, et à droite pendant les phases de rappel.

C.3.3 Variations entraînement/rappel.

Seulement 4 cellules sur 91 (4,4%) ont révélé des variations significatives de leur taux de décharge entre les phases d'entraînement et les phases de rappel. Les figures 9 et 10 présentent des exemples de ce type de cellules.

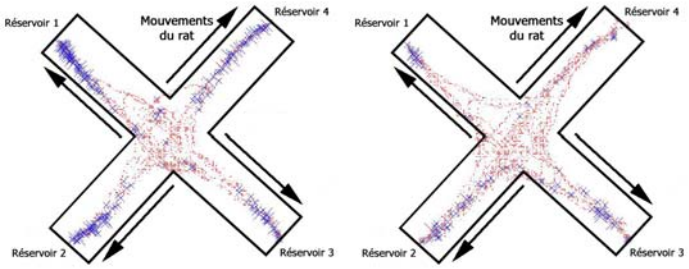

Figure 10 : Cellule comportementale déchargeant pour l'arrivée au réservoir et ayant des variations entraînement/rappel (ref. 621609P1C1). Cette cellule décharge plus pour les phases d'entraînement que pour les phases de rappel. A gauche, les décharges de la cellule pendant les phases d'entraînement, et à droite pendant les phases de rappel.

C.3.4 Corrélations à la récompense.

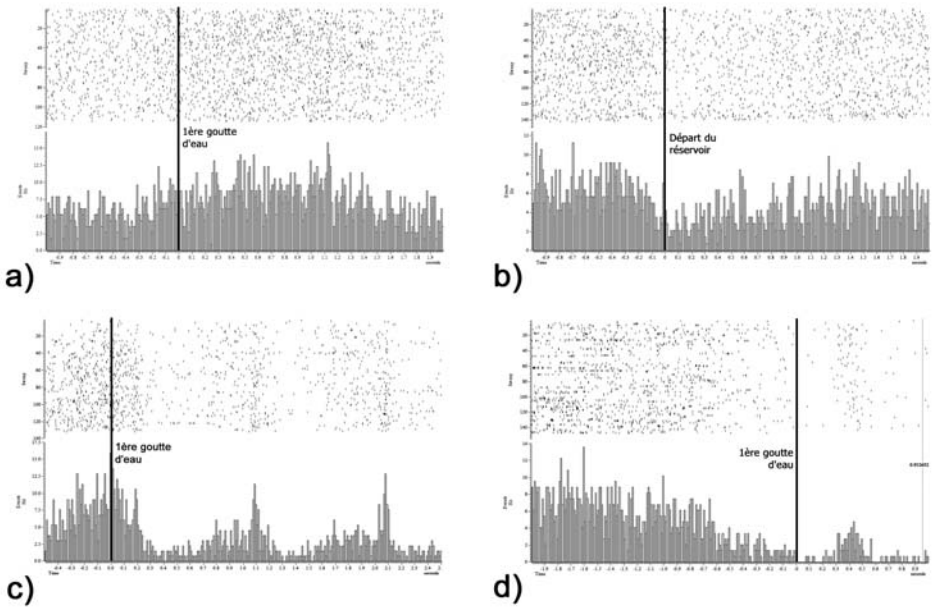

Figure 11 : Exemples de cellules ayant montré des décharges corrélées avec la consommation de récompense. a) Les décharges d'une cellule tonique (ref. 562307P2C1) ayant montré un taux de décharges moyen significativement plus important pendant la période de consommation de récompense. Les décharges sont représentées par rapport au repère temporel de l'arrivée au réservoir (réception de la 1ère goutte d'eau). b) Les décharges de la même cellule par rapport au départ du réservoir (fin de la consommation). c) Les décharges d'une cellule phasique (ref. 652409P0C1) ayant des bouffées marquées pour chaque goutte d'eau. Notons que la cellule commence à décharger en anticipation de la goutte d'eau puis atteint son maximum de décharge 100ms après celle-ci. d) Les décharges d'une cellule (ref. 552307P2C3) déchargeant systématiquement autour de 400ms après la réception de la première goutte d'eau (début de la récompense).

Parmi les 91 cellules classées comme comportementales, 27 peuvent être également catégorisées comme corrélées à l'attribution de récompense :

- 16 cellules phasiques déchargeant pendant la consommation de récompense.
- 3 cellules phasiques déchargeant à la réception de chaque goutte d'eau.
- 7 cellules phasiques déchargeant avant l'arrivée au réservoir, pouvant être interprétées comme anticipant la consommation de récompense.
- 1 cellule phasique déchargeant juste après le début de la récompense.

Parmi les cellules toniques, on peut également classer 11 cellules ayant une inhibition pendant l'attribution de récompense. Les figures 11 à 14 présentent des exemples de cellules dont les variations du taux de décharge étaient corrélées avec les périodes de consommation de la récompense.

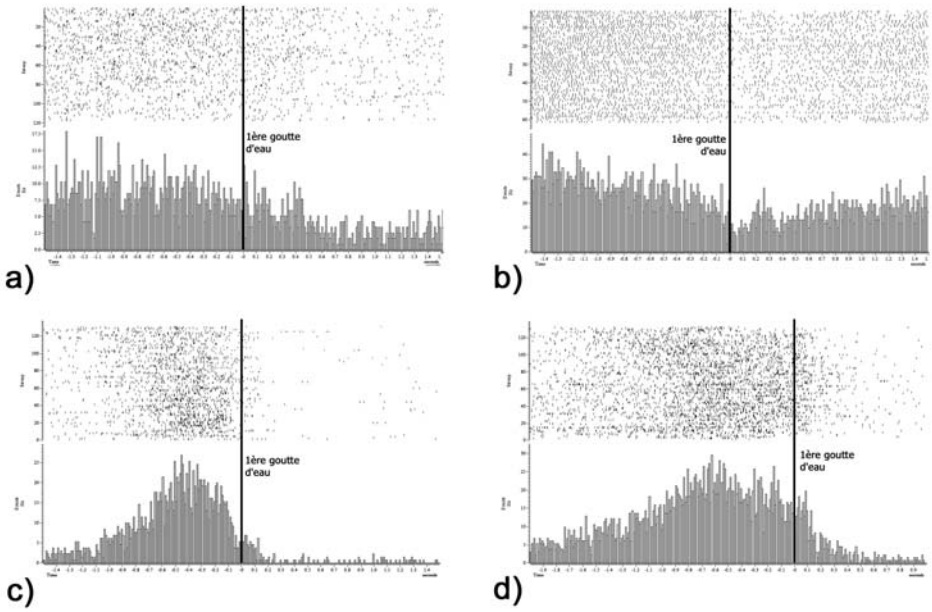

Figure 13 : Exemples de cellules ayant montré des décharges corrélées avec la consommation de récompense. a) et b) Cellules toniques (ref. 651509P4C1 et 630810P2C1) ayant montré une baisse significative de leur taux de décharge pendant la période de consommation de récompense. c) et d) cellules phasiques (ref. 620909P2C1 et 651209P0C2) ayant montré une augmentation significative de leur taux de décharge en anticipation de la première goutte d'eau reçue par le rat.

Notons toutefois que les méthodes utilisées ici ne permettent pas d'affirmer que les cellules de la figure 13.c) et 13.d) déchargent en anticipation de la récompense et pas pour la décélération de l'animal.

C.4 Discussion.

Ces résultats montrent qu'une grande partie (77%) des cellules analysées dans le striatum ventral avaient **des variations de leur taux de décharge corrélées significativement avec les différentes phases de la séquence comportementale** effectuées par le rat. Ceci va dans le sens de l'hypothèse que le noyau accumbens est impliqué dans les comportements orientés vers un but [Pennartz et al, 1994 ; Ikemoto & Panksepp, 1999]. Le chevauchement des activités des cellules tout au long de la séquence ressemble au codage par population envisagé classiquement pour le contrôle des comportements orientés vers des buts [Georgopoulos et al, 1986 ; Guigon & Burnod, 1995 ; Averbeck et al, 2002]. Il n'est cependant pas facile de distinguer un rôle de contrôle du comportement d'un rôle de suivi ou de représentation des éléments de ce comportement [Chang et al, 1996]. En d'autres termes, cela ne nous permet pas de conclure sur la signification du signal émis par le noyau accumbens vers les structures en aval. Correspond-il à une commande motrice ? A-t-il le même statut qu'un signal sensoriel qui reflète ce qui se passe ailleurs dans le système nerveux central ? Joue-t-il le rôle de superviseur de l'apprentissage S-R qui pourrait avoir lieu dans le striatum dorsal ?

L'analyse des corrélations spatiales de certaines de ces mêmes cellules nous donne déjà un élément de réponse. Nous avons vu en effet que 28% des neurones ayant une corrélation comportementale présentaient une sélectivité spatiale pour un ou plusieurs des 4 bras du labyrinthe. On peut donc écarter l'hypothèse que ces cellules codent seulement une commande motrice comme le fait d'avancer, mais considérer plutôt que **ces cellules sont sélectives à la position de l'animal dans le contexte de la tâche**. La corrélation première restant le comportement, on ne peut pas dire que ces cellules sont des cellules de lieux. Mais on peut toutefois y voir une modulation par des informations spatiales provenant de l'hippocampe et qui pourrait permettre au rat de déduire un chemin vers la meilleure récompense en fonction de ses représentations spatiales de l'environnement. Cela va donc dans le sens de l'hypothèse que nous voulions tester et qui consiste à dire que l'accumbens opère une mise en comparaison des stratégies allocentrées en fonction des repères visuels. Les méthodes employées ici ne nous permettent toutefois pas de trancher définitivement en faveur de cette hypothèse et d'autres études ont permis de montrer plus rigoureusement des réponses du noyau accumbens liées aux décharges de l'hippocampe et encrées sur le rythme theta de celui-ci [Albertin et al, 2000 ; Tabuchi, Mulder & Wiener, 2000].

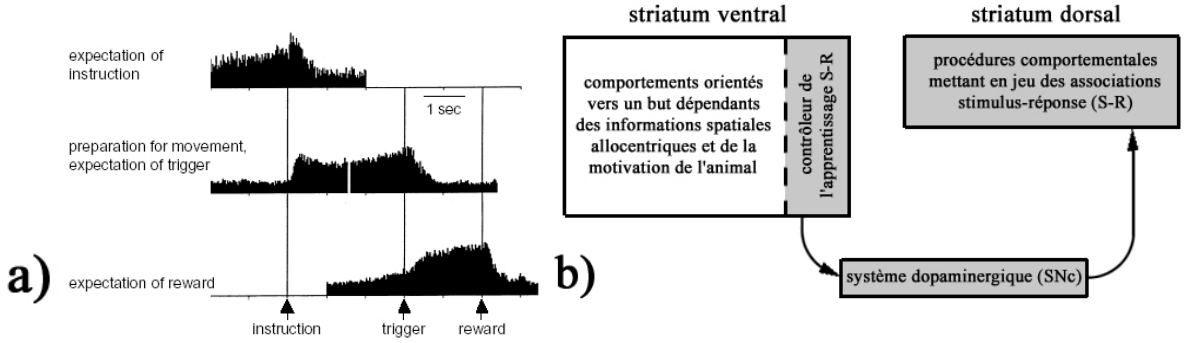

Figure 15 : a) Cellules mesurées dans le Striatum interprétées comme codant la prédiction d'évènements [Suri & Schultz, 2001]. La troisième ligne présente une cellule prédisant une récompense. b) Schéma simplifié représentant les zones pouvant être impliquées l'apprentissage S-R (zones en gris sur le schéma).

Pour terminer, parmi les cellules ayant montré des corrélations avec la récompense, les cellules qui avaient une inhibition pendant la consommation, ou avaient des bouffées d'activité pendant celle-ci peuvent être interprétées comme induisant **un renforcement**. De même, l'activité de neurones dans la période précédant l'obtention de récompense peut être interprétée comme codant **une certaine prédiction de cette récompense**, ce dernier type de cellules ayant également été observé dans le striatum au cours d'autres travaux [Martin & Ono, 2000 ; Suri & Schultz, 2001] comme le montre la figure 15.a qui montre une forte analogie avec la figure 14 des cellules que nous avons analysées. Or, il s'agit de caractéristiques tout à fait remarquable si l'on prend en compte le fait que les neurones de l'accumbens core projettent sur le système dopaminergique (SNc) qui projette en retour sur le striatum dorsal, et peut donc moduler les décharges des neurones dopaminergiques.

C.5 Conclusion.

Nous en arrivons à une conclusion essentielle pour le travail de modélisation à effectuer pendant ce stage. Non seulement l'accumbens core semble ne pas tenir compte uniquement du contexte sensoriel, puisqu'il intègre des informations spatiales pour le contrôle du comportement, ce qui sera une piste de recherche pour les mécanismes d'apprentissage pouvant avoir lieu dans le striatum ventral mais ne fera pas l'objet d'un modèle pendant le stage. Mais l'accumbens core montre également une activité neurale qui pourait contrôler les signaux dopaminergiques agissant sur le striatum dorsal et ainsi participer à l'apprentissage S-R comme décrit dans le paragraphe I.B de ce mémoire. La figure 15.b résume cette idée. Nous allons donc utiliser ces données pour la partie modélisation présentée dans le chapitre suivant.

II. L'APPRENTISSAGE PAR RENFORCEMENT BIOMIMETIQUE EN INTELLIGENCE ARTIFICIELLE : ETAT DE L'ART

A. L'apprentissage par Renforcement en Intelligence Artificielle.

A.1 Les processus de décision markoviens.

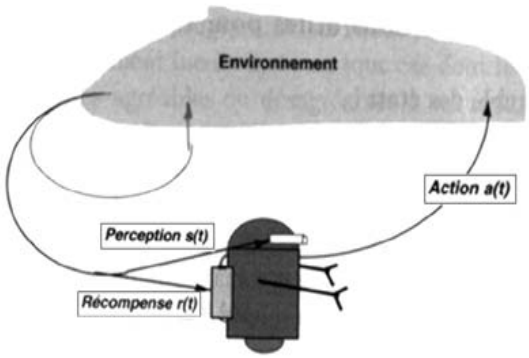

Figure 16 : Problématique de l'apprentissage par renforcement en Intelligence Artificielle. Un agent plongé dans un environnement effectue des actions qui lui rapportent ou non des récompenses. Il doit adapter sa politique de comportement pour maximiser la fréquence et la valeur de ces récompenses [Cornuéjols et Miclet, 2002].

En Intelligence Artificielle, le paradigme de l'apprentissage par renforcement est basé sur les processus de décision markoviens. Ces derniers considèrent un agent situé dans un environnement qu'il ne connaît pas, où le temps est discrétisé, et avec lequel il doit interagir : à chaque pas de temps t, l'agent est dans un état s^t, il effectue une action a^t pour se retrouver au pas de temps suivant dans un nouvel état s^{t+1}. Le comportement de l'agent s'assimile à sa politique, une fonction $\pi : S \times A \rightarrow \pi(A)$ qui indique pour tout état $s \in S$ la distribution de probabilités pour que l'agent choisisse les diverses actions $a \in A$ lorsqu'il se trouve dans cet état. Une fois son action choisie, l'agent dispose d'une fonction de transition T qui indique pour tout couple (état, action) la distribution de probabilité pour que l'agent se trouve au pas suivant dans chacun des états possibles quand il fait cette action dans cet état.

Dans certains couples (état, action), l'agent peut recevoir une récompense où une punition de l'environnement sous la forme d'un signal de renforcement. L'objectif global de l'agent est d'adopter un comportement qui lui permet de maximiser la fréquence et la valeur de ses récompenses. Il doit donc procéder à un certain apprentissage pour pouvoir établir un lien de

cause à effet entre ses actions et les récompenses de façon à effectuer les « meilleures » actions.

De façon à pouvoir évaluer la politique π de l'agent, on dispose d'une fonction valeur $V^{\pi}(s)$ qui associe à chaque état s une mesure de la récompense cumulée qu'un agent recevra s'il suit cette politique à partir de l'état s. Cette récompense cumulée s'assimile à la somme de tous les futurs signaux de renforcements et peut s'écrire :

$$R^{\pi}(t) = r^{t} + \gamma.r^{t+1} + \gamma^{2}.r^{t+2} + \ldots = \sum \gamma^{i-t}.r^{i} \quad \text{avec } 0<\gamma<1 \qquad \text{(I.A.1)}$$

De façon à ce que cette somme n'ait pas une valeur infinie, le terme γ, appelé « facteur de dévaluation », permet de prendre d'autant moins en compte les récompenses reçues que l'agent les recevra plus loin dans le temps. On peut alors définir la fonction valeur en s pour la politique π récursivement :

$$V^{\pi}(s) = \sum \pi(s,a).[R(s,a) + \gamma.\sum T(s,a,s').V^{\pi}(s')] \qquad \text{(I.A.2)}$$

L'équation I.A.1.2 est appelée équation de Bellman pour la politique π. Cette équation joue un rôle fondamental au cœur de toutes les méthodes d'optimisation qui permettent de définir des algorithmes d'apprentissage par renforcement.

A.2 La méthode de différence temporelle et les architectures Actor-Critic.

Il existe trois principales classes d'algorithmes permettant à un agent confronté à un processus de décision markovien de découvrir une politique optimale – une politique pour laquelle l'agent peut espérer recevoir la récompense cumulée maximale sur le long terme :

- Les algorithmes de programmation dynamique s'appliquent dans le cas où l'agent dispose d'un modèle de son environnement, c'est-à-dire lorsque les fonctions de transition T et de récompense R sont connues a priori.
- Les méthodes de Monte Carlo sont simples et ne présupposent pas la connaissance a priori d'un modèle, mais présentent le défaut de ne pas être incrémentales.
- Les méthodes de différence temporelle reposent sur une estimation incrémentale du modèle de l'environnement.

Nous ne développerons pas ici les deux premières classes qui sont expliquées dans les ouvrages [Sutton et Barto, 1998 ; Cornuéjols & Miclet, 2002], et nous nous restreindrons aux méthodes de différence temporelle qui sont les plus utilisées dans le cadre de l'apprentissage par renforcement car elles regroupent les points forts des deux autres : comme les algorithmes de programmation dynamique, elles sont incrémentales (la valeur estimée V(s) est mise à jour en fonction de la valeur estimée V(s)) ; comme les méthodes de Monte Carlo, elles n'ont pas besoin de modèle du monde car elles l'estiment sur la base de l'expérience de l'agent.

La méthode consiste à comparer deux estimations (ou prédictions) successives de la récompense cumulée que l'agent va recevoir.

$$P^{t-1} = r^t + \gamma.r^{t+1} + \gamma^2.r^{t+2} + \ldots \qquad \text{(I.A.3)}$$

$$P^t = r^{t+1} + \gamma.r^{t+2} + \gamma^2.r^{t+3} + \ldots \qquad \text{(I.A.4)}$$

Or, remarquons que l'équation I.A.3 peut s'écrire :

$$P^{t-1} = r^t + \gamma.(r^{t+1} + \gamma.r^{t+2} + \ldots) \qquad \text{(I.A.5)}$$

Ce qui, combiné à l'équation I.A.4, donne :

$$P^{t-1} = r^t + \gamma.P^t \qquad \text{(I.A.6)}$$

C'est la condition que doivent vérifier 2 prédictions consécutives correctes. Sutton appelle l'erreur dans la satisfaction de cette condition par 2 prédictions adjacentes la **Temporal Difference Error** (TD-Error) [Sutton, 1988] qui s'écrit donc : $\mathbf{r^t + \gamma.P^t - P^{t-1}}$.

L'apprentissage suivi ne consiste alors plus en l'attente du signal de renforcement à long terme mais en la modification de la fonction valeur $V^\pi(s)$ à chaque pas de temps en fonction de la TD-error entre deux prédictions consécutives par la procédure suivante :

$$V^\pi(s) \leftarrow V^\pi(s) + \beta.[r^t + \gamma.P^t - P^{t-1}] \qquad \text{(I.A.7)}$$

Où β joue le rôle de pas d'apprentissage. Barto décrit lui-même cette méthode par la phrase : « It is […] like the blind being led by the slightly less blind » [Barto, 1995]. Notons que cette méthode converge avec une probabilité de 1 [Dayan & Sejnowski, 1994].

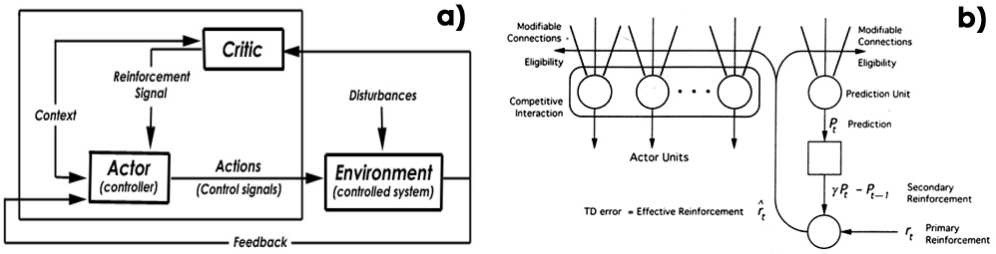

Figure 17 : Architecture d'un système de contrôle de type Actor-Critic [Barto, 1995]. a) Vision globale des interactions entre l'Actor, le Critic et l'environnement. A chaque nouveau pas de temps, le Critic envoie un signal de renforcement calculé à partir de l'erreur TD à l'Actor. b) Modèle plus détaillé de l'Actor (à gauche) et du Critic (à droite). On y voit chaque action représentée par un neurone dans la partie Actor, et un neurone jouant le rôle de Critic, calculant les prédictions de récompenses et agissant sur le système dopaminergique. Le signal dopaminergique est une résultante d'un renforcement primaire et d'un renforcement secondaire, dont la somme constitue la « TD error ».

L'idée est donc alors de concevoir une architecture de contrôle capable d'appliquer cette méthode. Pour cela, l'architecture la plus développée est le modèle Actor-Critic représenté à la figure 17.a [Barto, 1995]. D'un côté l'Actor est la zone mémoire qui mémorise la politique de l'agent et fait un choix d'action à effectuer sur l'environnement en fonction du contexte. De l'autre côté, le Critic est chargé d'évaluer à chaque instant la fonction valeur. Pour cela il fait une prédiction du résultat de l'action choisie par l'Actor, et calcule à l'instant suivant son erreur de prédiction en fonction de ce que renvoie l'environnement (la récompense positive, négative ou nulle). Si la récompense est meilleure que prévu (respectivement moins bonne que prévu), le Critic envoie un signal de renforcement positif (respectivement négatif) à l'Actor qui pourra donc à l'instant suivant choisir d'effectuer une action plus appropriée.

En Intelligence Artificielle, d'autres méthodes ont dérivé de l'apprentissage par différence temporelle. Il s'agit principalement des méthodes SARSA et Q-learning. Contrairement à la méthode des différences temporelles, l'algorithme SARSA travaille sur la qualité des couples (état, action) et non sur la valeur des états, ce qui oblige l'agent à prédire un pas de regard en avant quelle est l'action a^{t+1} qu'il réalisera lors du pas de temps suivant. L'algorithme du Q-learning, lui, n'a pas besoin de prédire quelle sera l'action effectuée au pas de temps suivant puisqu'il effectue des mises à jour de la politique de l'agent en fonction des actions optimales. Pour simplifier, il permet de connaître quelle sera l'action optimale au pas de temps suivant, ce qui lui évite d'avoir à faire un calcul pour prédire cette action. Cette propriété a value une grande réputation au Q-learning qui est sans doute l'algorithme d'apprentissage par renforcement le plus connu.

Mais dans la suite de ce mémoire, nous nous restreindrons à la méthode des différences temporelles qui, comme nous allons le voir dans le paragraphe suivant, ressemble fortement au comportement d'un neuromédiateur du système nerveux central : la dopamine.

B. L'analogie entre l'Erreur de Différence Temporelle et le signal de dopamine.

Comme nous l'avons vu au paragraphe I.B de ce mémoire, des données électrophysiologiques mesurées chez le singe montrent que la dopamine décharge pour des récompenses inattendues ou bien pour des stimuli conditionnels prédisant une récompense et pas pour la récompense ainsi prédite [Schultz, 1998, 2001]. Or, **cette activité postulée qui consiste à détecter des erreurs de prédiction ressemble fortement au comportement de l'erreur de Différence Temporelle et ainsi au signal de renforcement que nous avons vu dans les méthodes d'Intelligence Artificielle**. En effet, si l'on reconsidère l'équation de l'apprentissage par différence temporelle $\check{r}^t = r^t + \gamma.P^t - P^{t-1}$ où $\check{r}^t$ est le renforcement effectif à l'instant t (figure 17.b), on voit que :

- Dans le cas d'une récompense inattendue, on a $r^t>0$, $P^t=0$ et $P^{t-1}=0$ donc $\check{r}^t>0$.
- Dans le cas d'un stimulus prédisant un renforcement, on a $r^t=0$, $P^t>0$ et $P^{t-1}=0$ donc $\check{r}^t>0$.
- Dans le cas du renforcement prédit par ce stimulus, on a $r^t>0$, $P^t=0$ et $P^{t-1}>0$ donc $\check{r}^t=0$.
- Dans le cas de l'omission d'une récompense attendue, on a $r^t=0$, $P^t=0$ et $P^{t-1}>0$ donc $\check{r}^t<0$ qui se traduit par une interruption de l'activité des neurones dopaminergiques dans les modèles [Suri et Schultz, 1998], comme le présente la figure 3.

Ce constat a donné naissance à l'idée que **les ganglions de la base puissent abriter une structure de type Actor-Critic** [Houk et al, 1995 ; Suri et Schultz, 1998] **et être le siège d'un apprentissage par renforcement avec la dopamine comme signal de ce renforcement** [Schultz et al, 1997]. La connaissance de l'existence de boucles cortex-ganglions de la base-thalamus-cortex [Alexander et al, 1990] a elle aussi contribué à attribuer un rôle fonctionnel important aux ganglions de la base dans la sélection de l'action et dans l'apprentissage par renforcement. Le bouclage de l'information à travers les ganglions de la base ressemble en effet fortement aux itérations à chaque pas de temps utilisées dans la méthode de différence temporelle.

Partant de ces hypothèses, de nombreux modèles computationnels de type Actor-Critic (figure 17) se sont développés pour tenter de représenter le rôle fonctionnel des ganglions de la base dans la sélection de l'action et l'apprentissage par renforcement (voir [Joel, Niv & Ruppin, 2002] pour une revue).

C. Revue des modèles computationnels. Points de divergences et limites.

Nous allons voir ici qu'à partir d'un modèle de base dit « classique » et de ses limites [Houk, Adams & Barto, 1995] présenté à la figure 18, un grand nombre de modèles vont se succéder proposant chacun des solutions différentes pour tenter d'apporter une meilleure plausibilité biologique, un plus grand respect de l'anatomie des ganglions de la base, ou de meilleures caractéristiques fonctionnelles à l'architecture utilisée.

C.1 Le premier modèle Actor-Critic des GB : Houk, Adams & Barto, 1995.

Houk, Adams et Barto proposent un modèle où le striatum dorsal englobe la totalité de l'architecture Actor-Critic [Houk, Adams & Barto, 1995]. La partie matrix du putamen y joue le rôle de l'Actor, où des canaux maintenus ségrégés depuis le cortex représentent chacun une action qu'ils vont désinhiber dans le thalamus. A chaque pas de temps, le problème de la sélection de l'action est résolu par une opération de type « Winner-Takes-All » : les canaux s'inhibent les uns les autres pour qu'une seule action soit sélectionnée en sortie - celle qui a le plus fort taux de décharge (la plus forte salience). D'un autre côté, les neurones striosomaux calculent une prédiction de récompense à chaque pas de temps qui va servir d'excitation pour les neurones dopaminergiques par un chemin indirect via le noyau sub-thalamique. Selon ce modèle, la différence entre deux prédictions consécutives tient au fait qu'une prédiction à l'instant t va d'abord aller exciter les neurones dopaminergiques par le chemin indirect, puis va avec un peu de retard aller inhiber ces mêmes neurones par le chemin direct provenant du striatum (voir figure 18.a). Ainsi, l'inhibition aura lieu en même temps que l'excitation provoquée par l'excitation de la prédiction suivante à l'instant t+1. On obtient alors bien une différence entre les deux prédictions consécutives.

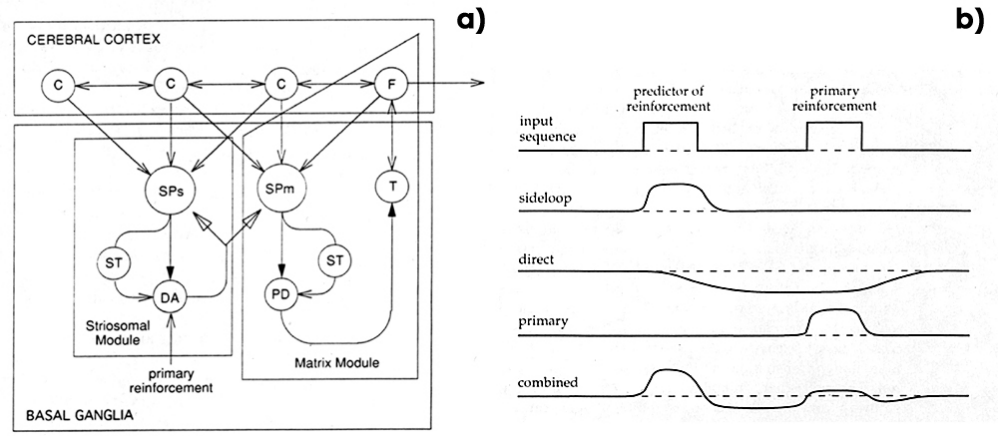

Figure 18 : Implémentation de l'Actor-Critic dans les Ganglions de la Base [Houk et al, 1995]. a) Architecture du modèle. C : Cortex, DA : Dopamine, PD : Pallidum, SPm : Neurone de la partie Matrisomale du Striatum, SPs : Neurone de la partie Striosomale du Striatum, ST : Noyau Sub-Thalamique, F : Cortex Préfrontal, T : Thalamus. b) Réponses des neurones dopaminergiques (libellée « combined ») aux différentes excitations et inhibitions reçues lors de récompenses et de stimuli prédisant ces récompenses. On voit que lors d'un stimulus « predictor of reinforcement », la dopamine décharge. Puis, une inhibition lente et prolongée due à l'apprentissage de l'association entre ce stimulus et la récompense empêche la dopamine de décharger au moment où l'animal obtient un « primary reinforcement » (la récompense effective).

Comme le montre la figure 18.b, cette opération n'a pas une grande précision temporelle car l'inhibition du striatum sur le système dopaminergique est lente et prolongée, ce qui ne permet pas d'expliquer la dépression de dopamine observée au moment précis où la récompense est attendue lorsque celle-ci est omise [Joel, Niv & Ruppin, 2002].

D'autre part, comme on peut le voir sur la figure 18.a, la partie Actor du modèle proposé par Houk et al est très simplifiée et très schématique. Ce qui ne permet pas à cette architecture de rendre compte des données anatomiques et physiologiques des ganglions de la base [Joel et al, 2002]. En effet, des données plus récentes ont mis en évidence les différences fonctionnelles entre les neurones du striatum dorsal ayant des récepteurs D1 à la dopamine, et ceux ayant des récepteurs D2 [Aizman et al, 2000]. On est donc en mesure de penser que cette partie du modèle doit avoir une architecture plus complexe qu'un ensemble de canaux fonctionnellement équivalents pour toutes les actions.

Enfin, le cortex transmet en entrée du modèle les stimuli bruts, sans composante temporelle, et nous verrons que l'un des enjeux les plus importants pour les modèles qui vont succéder sera de tenter de résoudre le problème de la représentation du temps entre un stimulus conditionnel et la récompense que ce stimulus prédit. En effet, de nombreuses études se sont attelées à dire que les ganglions de la base avait accès à un codage fin du temps, notamment

par l'intermédiaire de structures telles que le STN [Beurrier et al, 2002], ce qui n'est pas le cas dans le modèle de Houk, Adams et Barto, 1995.

C.2 Principaux modèles Actor-Critic tentant d'apporter une amélioration.

Les apports de la plupart des modèles Actor-Critic qui vont succéder au modèle de Houk et al diffèrent principalement sur les points suivants :

- Tout d'abord, **la façon dont le stimulus est représenté en entrée**. Pour obtenir une représentation précise du temps entre stimulus et récompense, un modèle va introduire une composante temporelle au stimulus, celle-ci permettant en quelque sorte à chaque instant de savoir combien de temps il s'est écoulé depuis la perception du stimulus [Montague et al, 1996]. Cette composante sera étendue à un mécanisme reproduisant une activité maintenue prolongée selon plusieurs constantes de temps différentes au niveau des striosomes, et représentant plusieurs stimuli utilisés dans une même tâche [Suri & Schultz, 1998]. Ce dernier modèle permet de reproduire de façon fiable la dépression de l'activité des neurones dopaminergiques en cas d'omission de la récompense et sera ensuite généralisé de façon à permettre au modèle de reproduire les réponses des neurones dopaminergiques à des nouveaux stimuli [Suri & Schultz, 1999, 2001]. Ce modèle a donc particulièrement retenu notre attention et nous l'utiliserons plus loin pour améliorer le modèle classique.

- **La façon dont la partie Critic calcule l'erreur de prédiction pour induire l'apprentissage**. En effet, le modèle de Houk et al suppose que ce calcul est effectué par les projections des striosomes du striatum dorsal vers le système doppaminergique, et que la différence entre 2 prédictions tient au fait que les projections indirectes excitatrices (via le STN) sont plus rapides que les projections directes inhibitrices. Ceci s'appuie sur quelques résultats expérimentaux [Kita & Kitai, 1991] mais est très contesté par les modèles suivants [Joel et al, 2002]. D'autres modèles vont proposer d'implémenter ce calcul toujours dans les striosomes, mais par la différence de dynamique temporelle entre les neurones projetant sur des neurones dopaminergiques ayant des récepteurs GABA-A, et ceux projetant sur les neurones dopaminergiques ayant des récepteurs GABA-B [Frank, Loughry & O'Reilly, 2001]. Mais la plupart des autres modèles vont supposer que les striosomes du striatum dorsal ne peuvent effectuer ce

calcul, et vont plutôt impliquer le cortex préfrontal ou le noyau accumbens [Suri & Schultz, 2001]. Nous avons vu dans la partie analyse de données que ceci concorde avec ce que l'on a pu observer des activités de certaines cellules dans l'accumbens core, et favorisera notre choix de nous inspirer du modèle de Suri et Schultz.

- **La façon dont est représenté le signal de renforcement**. En effet, la plupart des modèles supposent que la dopamine joue le rôle de médiateur du renforcement. Or, des études ont révélé, lors de la présentation de stimuli, des réponses des neurones dopaminergiques à des latences d'environ 100ms environ, ce qui serait trop faible pour que le système nerveux central ait eu le temps d'identifier ces stimuli et de leur attribuer une certaine valeur de prédiction de récompense [Redgrave, Prescott & Gurney, 1999]. C'est pourquoi certains modèles, en s'appuyant également sur la difficulté d'implémenter dans le striatum le calcul de la différence entre deux prédictions, vont remettre en cause le rôle de la dopamine dans le renforcement [Gurney, Prescott & Redgrave, 2001a,b ; Pennartz, 1997 ; Pennartz, McNaughton & Mulder, 2000]. Certains proposeront une hypothèse alternative de renforcement par le glutamate impliquant des circuits corticaux et un calcul différent du signal de renforcement [Pennartz, McNaughton & Mulder, 2000].

- **La façon dont l'anatomie des ganglions de la base est respectée**. En effet, la plupart des modèles utilisent des parties Actor très simplifiées qui ne rendent pas compte de l'anatomie du striatum dorsal. Toutefois, des modèles vont tenter de préciser le rôle de certains autres noyaux que le striatum, comme le STN qui, par une afférence provenant du cortex pourrait favoriser un comportement déjà sélectionné ou l'interrompre [Berns & Sejnowski, 1996 ; 1998]. Gurney et collègues, n'acceptant pas le rôle de la dopamine dans le renforcement, vont se focaliser sur la partie Actor et tenteront de tenir compte de données plus complètes sur l'anatomie des ganglions de la base de façon à proposer un modèle GPR très détaillé [Gurney, Prescott & Redgrave, 2001a,b]. Ce dernier utilise la dopamine comme signal de transition entre les comportements et différencie le rôle de trois sous-parties : les matrisomes ayant des récepteurs D1 à la dopamine et opérant la sélection de l'action à proprement dite ; les matrisomes ayant des récepteurs D2 régulant l'activité d'ensemble, et le STN dotant le modèle de propriétés de persistance évitant les oscillations comportementales. Ce modèle étant à notre connaissance celui qui respecte

le plus finement ce que l'on sait de l'anatomie des GB, il a été choisi par Benoît Girard pour être implémenté dans Psikharpax. Mais ce modèle n'utilise pas de partie Critic, et il convient de déterminer si un modèle Critic tel que celui de Houk et collègues est compatible avec le GPR.

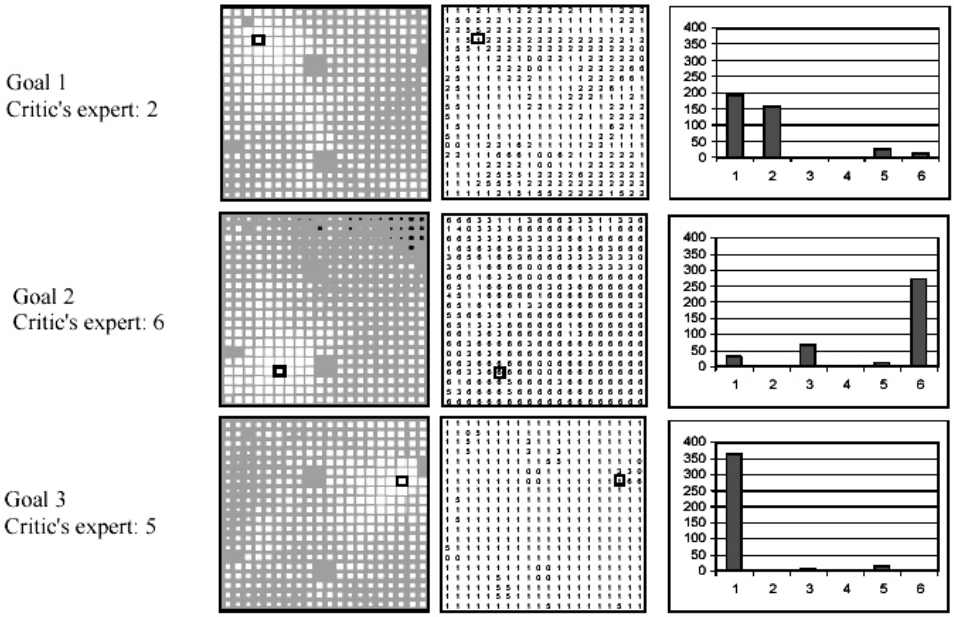

Figure 19 : Simulation des prédictions des Critic du modèle de Baldassarre [Baldassarre, 2002]. Pour chacun des 3 buts décrits par une zone de l'environnement, le graphique de droite affiche le niveau d'implication de chacun des 6 experts Critic dans la prédiction de récompense.

- **La façon dont ils permettent de construire une séquence comportementale**. En effet, si des modèles vont tenter de construire par apprentissage des séquences de réponses associées à un séquence de stimuli visuels [Nakahara, Doya & Hikosaka, 1999, 2001], l'utilisation d'un modèle Critic simple tel que le fait Houk et collègues montre vite ses limites de calcul. En particulier, un tel modèle, s'il n'utilise qu'une seule cellule de Critic, ne peut résoudre que des problèmes qualifiés mathématiquement de « linéairement séparables ». En pratique, le Critic ne peut alors pas adapter ses prédictions à des contextes sensoriels différents. Le modèle de Suri et Schultz tente de résoudre le problème en modélisant autant de cellules Critic que de stimuli utilisés dans la tâche de validation du modèle [Suri & Schultz, 2001]. Mais dans leur cas, c'est l'expérimentateur qui affecte arbitrairement chaque Critic à un stimulus donné avant le début de la tâche. Or, nous ne pouvons nous satisfaire d'une

telle solution si l'on veut pouvoir confronter le modèle à des stimuli nouveaux dans des environnements imprévisibles. Le modèle de Baldassarre adopte une solution différente [Baldassarre & Parisi, 2000 ; Baldassarre, 2002]. Il s'inspire de méthodes d'intelligence artificielle pour modéliser la partie Critic sous forme de « mixture d'experts » [Jacobs & Jordan,1991 ; Tani & Nolfi, 1999 ; Gourichon, 1999]. Ceci se présente sous la forme d'un ensemble de cellules apprenant à se spécialiser dans la description de différentes caractéristiques d'un environnement donné. Sur la figure 19, on voit pour trois zones « but » différentes de l'environnement, les différents experts qui sont impliqués dans la prédiction de récompense. On peut noter qu'il peut y avoir plusieurs Critic décrivant ensemble une zone. Le modèle de Baldassarre nous servira donc dans la partie modélisation pour tenter d'améliorer notre Critic.

Notons enfin que certains modèles vont être étendus à la mise en mémoire de travail dans le cortex préfrntal de séquences sensorielles pour la mise en œuvre des séquences de comportements [Berns & Sejnowski, 1996, 1998 ; Beiser & Houk, 1998 ; Frank, Loughry & O'Reilly, 2001]. L'analyse de ces différents modèles ouvre donc 5 principales questions pour notre travail présent et futur :

- Est-ce qu'une architecture de type Actor-Critic peut encore supporter un apprentissage par renforcement si l'on utilise une partie Actor très détaillée et plus fidèle à l'anatomie du striatum dorsal ?
- Est-ce que l'activité des neurones du striatum ventral peut permettre le calcul de la ***temporal difference error*** utilisée dans la partie Critic du modèle ?
- Est-ce que les informations transmises à la partie Critic par le cortex peuvent coder la représentation du stimulus avec composante temporelle utilisée par Montague et collègues puis par Suri et Schultz ?
- Est-ce qu'il est nécessaire de modéliser le Critic sous forme d'un réseau neuronal à plusieurs couches pour lui permettre de résoudre des problèmes non linéairement séparables ? Ou bien est-ce qu'une mixture d'experts contenant plusieurs cellules de Critic parallèles est suffisante ?
- Est-ce que l'on peut toujours envisager le rôle de la dopamine comme signal de renforcement ?

III. MODELISATION D'UNE PARTIE CRITIC POUR PSIKHARPAX

A. Implémentation du modèle de Houk, Adams & Barto, 1995.

La première question à laquelle nous devions répondre était de savoir si une architecture de type Actor-Critic peut encore permettre un apprentissage par renforcement des associations stimulus-réponse dans le cas où l'on utilise une partie Actor très détaillée, en l'occurrence le modèle GPR [Gurney, Prescott & Redgrave, 2001a,b]. Nous avons donc pour cela, intégrer le modèle classique de Critic [Houk, Adams & Barto, 1995] à l'Actor implémenté à partir du GPR par Benoît Girard dans Psikharpax.

Puis, pour vérifier que l'apprentissage S-R peut fonctionner ainsi, nous avons choisi comme tâche de validation la phase de Pre-training de la tâche du labyrinthe en croix effectuée par des rats au LPPA. En effet, nous avons vu dans la partie « analyse de données » de ce mémoire que cette phase impliquait un apprentissage procédural permettant l'association entre l'éclairage d'un réservoir et la présence de récompense dans ce réservoir, et permettant ainsi au rat de construire la séquence comportementale suivante :

1) Quand aucun réservoir n'est allumé, il faut retourner au centre pour déclencher l'éclairage.
2) Une fois que le robot arrive au centre et que le réservoir s'allume, il faut se diriger vers ce réservoir et s'y pencher pour déclencher l'émission d'eau et ainsi obtenir de la récompense.

Nous avons également vu dans cette analyse qu'il existait un certain nombre de cellules du noyau accumbens qui pouvaient être interprétées comme prédisant de la récompense. Or, c'est justement ce qu'est censée calculer une partie Critic comme nous l'avons vu au paragraphe II.B. L'architecture Actor-Critic semble donc être parfaitement appropriée pour cet apprentissage puisque comme nous l'avons vu dans la revue des modèles Actor-Critic, ces modèles permettent la construction d'une séquence comportementale grâce à des associations S-R. Cette partie Critic peut donc être implémentée dans le noyau accumbens, ce que nous avons fait pour cette intégration au modèle Actor GPR.

A.1 Description de la partie Actor implémentée par Benoît Girard à partir du modèle GPR.

Le modèle GPR (figure 20) base une partie de sa structure sur ce qui correspond au striatum dorsal. Il est constitué de plusieurs canaux symbolisant chacun un comportement possible du robot, et effectue de la sélection de l'action sur ces canaux par la combinaison de deux voies à travers les ganglions de la base :

- Une première voie, appelée BGI sur la figure 20 sélectionne le comportement qui a la plus forte salience.
- Une deuxième voie, appelée BGII régule le niveau général de l'activité dans la première voie, et effectue ainsi un contrôle sur la sélection.

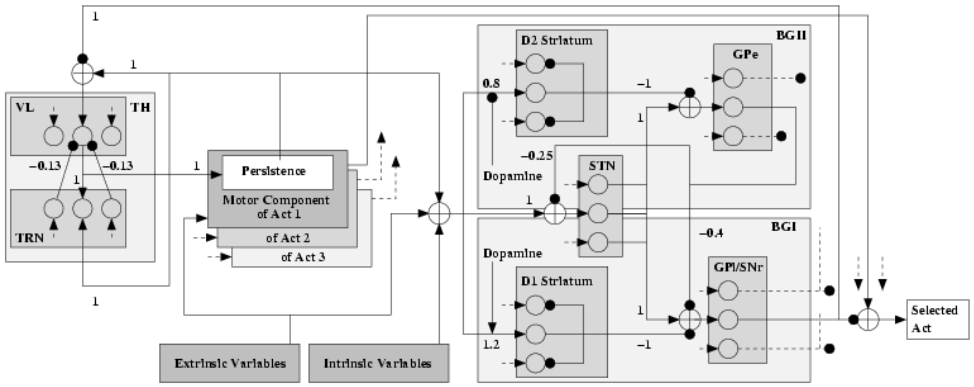

Figure 20 : Modèle GPR de la partie Actor du Striatum Dorsal tel qu'il a été présenté avant son implémentation dans Psikarpax [Gurney, Prescott & Redgrave, 2001a,b]. Chaque canal y représente un comportement. Une interaction entre trois modules permet d'opérer un "Winner-Takes-All" assurant la sélection d'une seule action à chaque pas de temps. Ces trois modules sont : la partie du Striatum contenant des récepteurs D1, celles contenant des récepteurs D2, et le Noyau Sub-Thalamique (STN).

Le modèle GPR a la particularité de doter le robot d'une certaine persistance dans son comportement par l'intermédiaire du bouclage de l'information à travers le thalamus, puis par la projection du cortex sur le STN. Avec ce modèle Actor détaillé, le robot a fait preuve d'une meilleure capacité de transition dans ses comportements et d'une meilleure conservation de son énergie qu'un modèle « Winner-Takes-All » [Girard et al, 2002].

A.2 Ajout de la partie Critic développée par Houk, Adams & Barto.

Nous avons donc utilisé le modèle Critic classique de Houk, Adams et Barto, 1995, et l'avons connecté à l'Actor de Psikharpax (figure 21). Notons que l'implémentation de la partie Critic

est ici supposée située dans le striatum ventral (accumbens core), conformément aux hypothèses formulées dans la partie analyse de données de ce mémoire (figure 15.b).

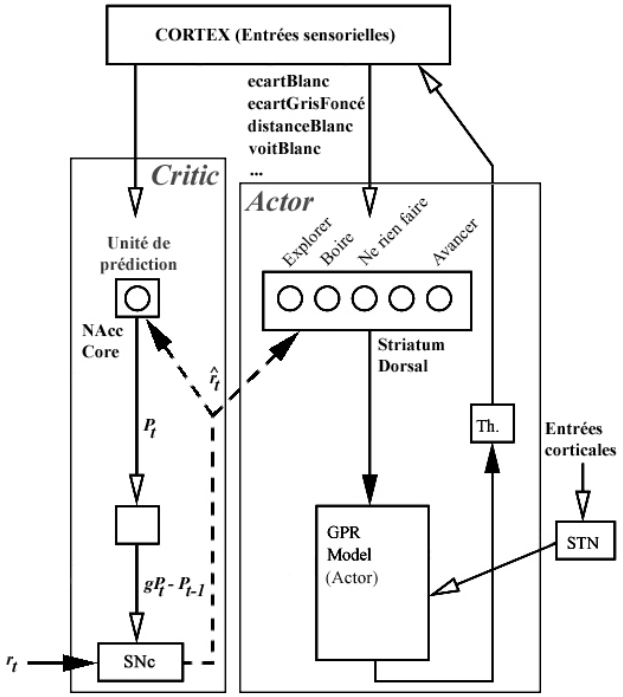

Figure 21 : Intégration du modèle Critic de Houk et al [Houk, Adams & Barto, 1995] avec le GPR [Gurney, Prescott & Redgrave, 2001a,b] de Psikharpax. g est ici le facteur de dévaluation γ de l'équation de Barto [Barto, 1995]. Nacc Core : Noyau de l'Accumbens ; Th : Thalamus ; STN : Noyau Sub-Thalamique ; VTA : Aire Ventrale Tegmentale ; SNc : Substance Noire compacta.

La partie Critic reçoit les mêmes données sensorielles d'entrée que la partie Actor. A partir de ces données délivrées à chaque instant t, le Critic calcule une prédiction P^t, et induit ensuite un signal dopaminergique à partir de l'erreur entre deux prédictions consécutives à partir de l'équation de l'erreur de différence temporelle de Barto [Barto, 1995] :

$$\breve{r}^t = r^t + \gamma.P^t - P^{t-1} \qquad \text{(III.A.1)}$$

A.3 Méthodes et adaptation à la tâche de validation.

A.3.1 Environnement de simulation.

L'environnement utilisé est un labyrinthe en croix contenant 4 réservoirs à ses extrémités comme dans la tâche du Plus-Maze. Les réservoirs considérés comme éteint sont simulés comme étant de couleur gris foncé. Le réservoir éclairé à un moment donné est de couleur blanche. Le simulateur de cet environnement a été dans un premier temps adapté pour que

celui-ci soit dynamique, c'est-à-dire pour qu'un réservoir puisse s'éteindre (sa couleur devient gris foncé) lorsque le robot y a bu toute la quantité d'eau disponible. De même, lorsque tous les réservoirs sont éteints et que le robot arrive au centre du labyrinthe (symbolisé par la couleur gris clair), le simulateur déclenche l'éclairage d'un nouveau réservoir. La figure 22 présente cet environnement.

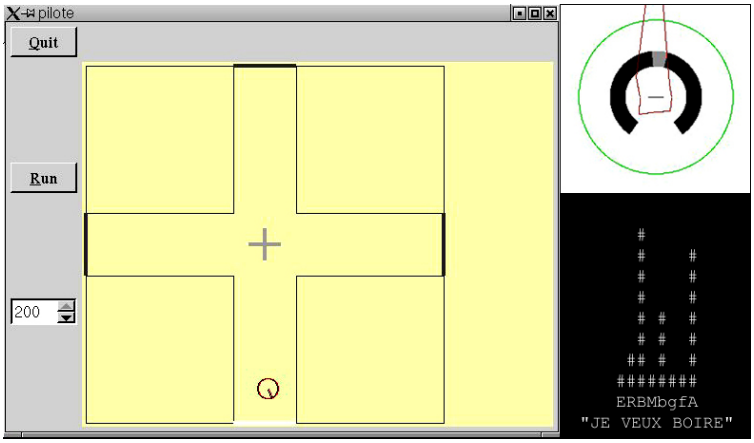

Figure 22 : Environnement de simulation du robot semblable à la tâche du Plus-Maze [Tabuchi et al, 2000]. La partie centrale représente l'environnement en 2D vu de dessus. Le robot est représenté par un cercle. Un rayon de ce cercle symbolise la direction de la tête du robot. En haut à droite est représenté la perception visuelle du robot (il voit du blanc vers le Sud, du gris clair vers le Nord et rien ailleurs). En bas à droite sont représentées les saliences de chacun des comportements. Le comportement choisi est mentionné dans la décision du robot. Abréviations des comportements : E,Explorer; R,ne Rien faire; B,Boire; M,Manger; b, se tourner vers un stimulus Blanc; g,se tourner vers un stimulus Gris clair ; f,se tourner vers un stimulus gris Foncé ; A,Avancer.

A.3.2 Entrées sensorielles du modèle et comportements du robot.

Le modèle Actor développé par Benoît Girard a été adapté de façon à ce que chaque comportement du modèle reçoive comme données d'entrée les variables résumées dans le tableau 3.

Les comportements possibles pour le robot sont : EXPLORER, NE RIEN FAIRE, BOIRE, MANGER, SE TOURNER VERS UN STIMULUS BLANC, SE TOURNER VERS UN STIMULUS GRIS FONCE, SE TOURNER VERS UN STIMULUS GRIS CLAIR, AVANCER. Certains comportements seront directement impliqués dans la tâche : BOIRE, SE TOURNER VERS UN STIMULUS BLANC, SE TOURNER VERS UN STIMULUS GRIS CLAIR et AVANCER. D'autres sont ajoutés pour vérifier que les comportements non impliqués dans la tâche ne sont pas associés dans les apprentissages : EXPLORER, NE RIEN FAIRE, SE TOURNER VERS UN STIMULUS GRIS FONCE ou MANGER. Dans le

modèle de l'Actor, la salience des comportements s'exprime donc en fonction des variables comme schématisé sur la figure 23 et comme formalisé par les équations III.A.1 à III.A.8.

Variable	Signification de la variable
Persist(i)	La persistance du comportement i (renvoyée uniquement au comportement i).
ecartBlanc, ecartGrisFoncé, ecartGrisClair	L'angle entre la direction de la tête du robot et le stimulus blanc (respectivement gris foncé, gris clair) lorsque celui-ci est perçu.
distanceBlanc, distanceGrisFoncé, distanceGrisClair	Distance estimée entre le robot et le stimulus blanc (resp. gris foncé, gris clair) si celui-ci est perçu.
proximBlanc, proximGrisFoncé, proximGrisClair	proximBlanc = 1 – distanceBlanc (de même pour les 2 autres variables)
voitBlanc, voitGrisFoncé, voitGrisClair	Vaut 0 si le robot ne perçoit pas de stimulus blanc (resp. gris foncé, gris clair), 1 sinon.
constante	Vaut toujours 1.

Table 3 : Variables utilisées en entrée du modèle. Elles sont utilisées aussi bien par la partie Actor que par la partie Critic. Ces variables sont constituées majoritairement de données sensorielles plus une variable 'persistance' pour chaque comportement et une constante toujours égale à 1.

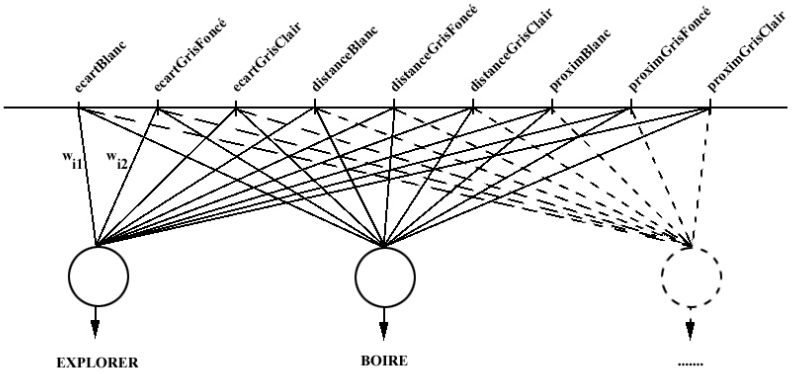

Figure 23 : Schéma des entrées par comportement dans la partie Actor du modèle. Pour chaque comportement i (8 comportements en tout), et pour chaque variable d'entrée j (14 variables en tout), un poids w(ij) détermine avec quelle force la variable j influe sur le comportement i.

$$\text{EXPLORER}(i=0)=w^{i0}.\text{persist}(i)+w^{i1}.\text{ecartBlanc}+\ldots+w^{i13}.\text{constante} \quad \text{(III.A.2)}$$

$$\text{NERIENFAIRE}(i=1)=w^{i0}.\text{persist}(i)+w^{i1}.\text{ecartBlanc}+\ldots+w^{i13}.\text{constante} \quad \text{(III.A.3)}$$

$$\text{BOIRE}(i=2)=w^{i0}.\text{persist}(i)+w^{i1}.\text{ecartBlanc}+\ldots+w^{i13}.\text{constante} \quad \text{(III.A.4)}$$

…

$$\text{AVANCER}(i=7)=w^{i0}.\text{persist}(i)+w^{i1}.\text{ecartBlanc}+\ldots+w^{i13}.\text{constante} \quad \text{(III.A.9)}$$

A.3.3 Adaptation du modèle.

Nous avons dans un premier temps utilisé une seule cellule de Critic, comme décrit dans l'article de Houk, Adams & Barto, 1995 (voir figure 21). Celui-ci n'exclue pas de pouvoir utiliser plusieurs cellules. Il n'explique toutefois pas comment ces cellules seraient coordonnées si le modèle en comprenait plusieurs.

Les stimuli (ou contextes sensoriels) sont envoyés en entrée du modèle sans composante temporelle et seulement à l'instant où ils sont perçus. La prédiction de récompense calculée à chaque pas de temps t à partir de ces données d'entrée s'écrit :

$$P^t = f\_sigmo(w'^1.ecartBlanc^t + \ldots + w'^{13}.constante^t) \qquad \text{(III.A.10)}$$

Où les w'^j sont les poids synaptiques de la cellule chargée de faire cette prédiction et f_sigmo une fonction sigmoïde introduisant un seuil de décharge du neurone de 0,5 et réduisant la sortie du neurone entre 0 et 1. A partir de cette prédiction est calculée l'erreur TD, c'est-à-dire le renforcement effectif sous forme de signal dopaminergique, selon l'équation III.A.1.

Nous avons choisi un facteur de dévaluation γ de 0,98. En n'étant pas trop petit devant 1, ce facteur permet au modèle de maintenir sa prédiction entre un stimulus et une récompense éloignés dans le temps (dans notre cas, une dizaine de secondes sont nécessaires au robot lorsque il n'a pas encore bien appris l'association (réservoir allumé)-(présence de récompense)). Au cours des essais que nous avons fait avec un facteur plus petit (γ = 0,9 voir 0,8 ou 0,75), le temps trop long entre le stimulus et la récompense laissait le signal de renforcement trop affaiblir la prédiction de récompense jusqu'à ce que le robot abandonne son approche du réservoir. Notons que dans leur tâche, Suri et Schultz ont utilisé un facteur de dévaluation γ égal à 0,99 pour des intervalles de 5 secondes entre le stimulus et la récompense, et égal à 0,85 pour des intervalles d'1 seconde [Suri & Schultz, 2001 ; Suri, 2002].

En ce qui concerne la mise à jour des poids synaptiques w^{ij} de l'Actor et du Critic, nous avons utilisé une vitesse d'apprentissage β (learning rate) de 0,2. Une telle vitesse a l'inconvénient de nécessiter de nombreux essais pour consolider l'apprentissage, mais elle a l'avantage de rendre l'apprentissage beaucoup plus stable qu'il ne le serait avec des vitesses plus importantes. L'équation de mise à jour des poids synaptiques s'écrit :

$$w^{ij} \leftarrow w^{ij} + E^{ijt}.\beta.\check{r}^{t} \qquad \text{(III.A.11)}$$

Dans cette équation, E^{ijt} est la valeur de l'entrée présynaptique de la synapse de poids w^{ij} à l'instant t (pour l'exemple de la variable binaire voitBlanc : $E^{ijt} = 1$ si le robot perçoit le stimulus blanc, $E^{ijt} = 0$ sinon).

Enfin, de façon à introduire un compromis exploration/exploitation, nous avons introduit un aléa dans la sélection des comportements faite par la partie Actor du modèle. Le compromis exploration/exploitation est en effet connu en Intelligence Artificielle comme condition nécessaire de ce type d'apprentissage, puisque si un agent ne fait qu'exploiter ce qu'il sait déjà, en l'occurrence reproduire les comportements tels qu'ils les a appris, il ne se trouvera jamais dans une situation nouvelle et ne pourra donc rien apprendre. La solution habituelle est d'introduire un hasard en sortie du modèle qui fait la sélection de l'action (l'Actor). Ici nous avons décidé d'adopter la procédure suivante : lorsque le robot fait la même chose depuis longtemps et que ça ne lui apporte pas de récompense, alors la salience d'un des comportements choisi au hasard est augmentée en entrée de l'Actor. Cette solution originale conserve la plausibilité biologique puisque ceci pourrait être obtenu par exemple en modifiant le taux de dopamine sur la population de neurones qui codent se comportement dans le modèle GPR. Ce modèle utilise en effet un taux de dopamine tonique maintenu constant pour permettre au robot de bien enchaîner les comportements : ni osciller perpétuellement entre deux comportements, ni persister sans cesse sur un même comportement [Gurney, Prescott & Redgrave, 2001a,b]. Or, si l'on prend en compte la modulation que l'accumbens shell peut exercer sur le striatum dorsal par l'intermédiaire de la dopamine phasique chez le rat [Haber et al, 2000] comme on peut le voir sur la figure 1 au début de ce mémoire, on peut envisager une modification comportementale en fonction de la motivation du robot. Cette solution sera évaluée plus amplement dans la discussion générale de ce mémoire.

A.4 Résultats.

Pour la représentation graphique des résultats, nous avons utilisé les conventions suivantes : nous appèlerons 'Itération' chaque traitement de l'information effectué par le modèle à chaque pas de temps (1 Itération ≈ 800 ms) ; nous appèlerons 'Essai' chaque période de temps démarrant au moment où le robot commence à chercher la récompense et se terminant au moment où il a fini de la consommer. Remarquons que plus le robot met de temps à atteindre

la récompense, plus il se produit d'itérations dans un même essai. Notons enfin que dans nos expériences les essais s'enchaînent en continu : la consommation de la récompense marque la fin d'un essai et le début du suivant.

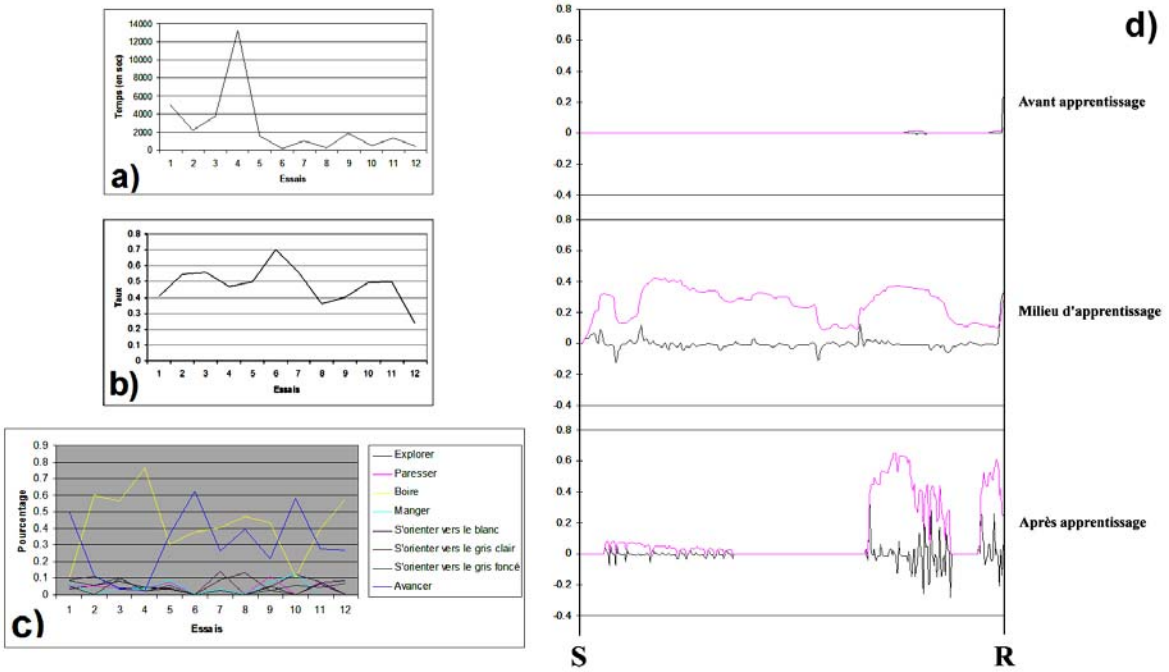

Figure 24 : Résultats de la simulation du modèle de Houk, Adams & Barto [Houk et al, 1995]. a) Courbe d'apprentissage du Critic. En abscisse, les essais jusqu'à chaque obtention de récompense, en ordonnées le temps en secondes. b) Représentation du renforcement (signal dopaminergique) ayant lieu pendant chaque consommation de récompense. En ordonnées le taux de renforcement. c) Pourcentage de sélections de chaque comportement pas essai de l'expérience. On voit qu'au début le comportement 'BOIRE' (en gris clair) est sélectionné environ 60% du temps. Puis à partir du 5ème essai, le comportement 'AVANCER' commence à être renforcé (c'est le début de la construction d'une séquence). Alors, ces deux comportements sont alternativement sélectionnés respectivement à 60% du temps et à 40%. d) Evolution de la prédiction de récompense faite par le Critic entre l'instant où le robot perçoit le stimulus (S), c'est-à-dire la lumière blanche (instant 0 sur l'axe des abscisses) et le moment où il consomme la récompense (R) au réservoir (point le plus à droite sur l'axe des abscisses). En gris clair, la prédiction de renforcement et en noir le signal de dopamine (le renforcement effectif). On peut constater sur ce schéma que la dopamine modélisée ici est un détecteur de fortes variations entre deux prédictions comme on pouvait s'y attendre au vu de l'équation III.A.1. En effet, on peut distinguer un pic de dopamine lorsque la prédiction augmente fortement, et une chute lorsqu'elle diminue fortement.

La figure 24.a montre sur une expérimentation, la courbe d'apprentissage du Critic, c'est-à-dire le temps que le robot a mis à chaque essai pour trouver la récompense et la consommer. On voit que le robot passe dès le 3ème essai par une phase où il le temps pour atteindre la récompense augmente considérablement. Ceci est du à un sur-apprentissage du comportement 'BOIRE'. En effet, comme on peut le voir sur la figure 24.c, les premières obtentions de récompense renforcent uniquement ce comportement dont la salience devient très forte : le robot sélectionne systématiquement le comportement 'BOIRE' puisque c'est ce qui lui

rapporte de la récompense. Mais comme ça ne lui donne pas de renforcement ailleurs qu'au réservoir allumé, ce comportement va être petit à petit affaibli dans les parties du labyrinthe loin de la lumière. Le comportement 'BOIRE' va alors devenir sélectif à un contexte sensoriel bien déterminé (proche du réservoir éclairé), ce qui va ainsi arrêter de bloquer le robot dans les autres zones de l'environnement. D'où la baisse du temps pour atteindre la récompense dans les essais suivants sur la figure24.a.

Toutefois, si l'on analyse l'apprentissage que fait la partie Critic du modèle pour prédire les renforcements, on peut voir ses limites rapidement atteintes. La figure 24.d montre l'évolution de la prédiction de récompense depuis l'instant où le robot perçoit le stimulus de lumière blanche jusqu'à la consommation de récompense. Cette courbe est représentée à trois phases de l'expérience : avant apprentissage, au milieu de l'apprentissage et après apprentissage. On peut voir qu'au début le modèle ne prédit pas de récompense, puis après 4 consommations de récompense (milieu d'apprentissage), il parvient à prédire un renforcement dès la perception du stimulus. Enfin, après 10 consommations de récompense (après apprentissage), le Critic a désappris à prédire de la récompense dès la perception du stimulus pour n'en prédire fortement que peu avant la consommation.

D'autre part, on peut voir sur la figure 24.b que le renforcement ayant lieu pendant la consommation de récompense reste fort pendant toute l'expérience. Or, l'apprentissage devrait faire disparaître tout renforcement pendant la récompense puisque après apprentissage, le Critic est censé ne plus faire d'erreur de prédiction pendant cette période. La persistance d'un renforcement fort pendant la récompense va contribuer à trop renforcer le comportement 'BOIRE' et va faire que ce comportement occupe une grande partie du temps du robot (figure 24.c). Le robot perd alors beaucoup de temps à sélectionner le comportement 'BOIRE' même lorsque celui-ci n'est pas nécessaire.

A.5 Discussion.

D'après les résultats présentés ci-dessus, on voit que le robot a bien appris qu'il fallait boire au réservoir éclairé et qu'il commençait même à apprendre une séquence à 2 éléments comportementaux comme suit :

1) A la perception d'un stimulus blanc (réservoir éclairé), sélectionner le comportement 'AVANCER'.
2) Puis lorsque le stimulus blanc est identifié comme suffisamment proche, sélectionner le comportement 'BOIRE'.

Mais nous avons vu que le Critic du modèle avait des capacités de prédiction de récompense limitées à une portion du labyrinthe et qu'il ne cessait d'être renforcé pendant la consommation de récompense.

Nous avons identifié 2 raisons à cette limitation du Critic. Tout d'abord, une seule cellule de Critic ne permet que de résoudre des problèmes linéairement séparables et il n'est donc mathématiquement pas possible pour cette seule cellule de décrire l'environnement sous forme de portions de droites. Le Critic se restreint donc rapidement à la description du voisinage du réservoir contenant de la récompense et il faudrait d'autres cellules pour décrire le reste du labyrinthe. D'autre part, le renforcement (le signal dopaminergique) reste élevé pendant la consommation de récompense (figure 24.b) alors que Houk, Adams & Barto prédisent qu'il va devenir fort au moment de la perception du stimulus et nul pendant la récompense. Dans la description du modèle par Barto, on peut lire qu'au moment du Renforcement Primaire (pendant la consommation de récompense) « Pt [la prédiction de renforcement] is still zero because it is predicting that zero primary reinforcement occurs after the trial » (Barto, 95). Or, ceci ne peut être le cas dans notre tâche pour deux raisons :

- Dans notre expérience, la consommation de récompense prend un certain temps, et si la prédiction de récompense est nulle pendant la consommation de la 1ère goutte d'eau, elle ne prédira pas qu'une 2ème goutte va arriver. Il y aura donc erreur de prédiction. Or, dans leurs simulations, Houk, Adams & Barto utilisent une récompense instantanée qui marque la fin d'un essai.
- Avec un Critic qui ne résout que des problèmes linéairement séparables, on ne peut prédire de la récompense juste avant de boire puis ne plus en prédire pendant la désaltération. En effet, ces deux situations sont décrites par un contexte sensoriel identique : la perception de la lumière blanche tout prêt du robot et la perception du centre du labyrinthe derrière lui.

La persistance d'un renforcement pendant la récompense fait que le Critic et le comportement 'BOIRE' ne cessent d'être renforcés. Ce qui fait que le comportement 'BOIRE' reste globalement fort par rapport aux autres comportements, qu'il est souvent sélectionné au dépriment des autres, ponctuant par exemple le déplacement du robot lorsque le comportement 'AVANCER' est sélectionné, et lui faisant ainsi perdre beaucoup de temps (figure 24.c).

Deux pistes vont êtres explorées pour résoudre ces limites. Tout d'abord, Suri et Schultz résolvent le problème de la prédiction pendant la récompense en introduisant une composante temporelle pour la description de tout événement (stimulus ou récompense) comme donnée d'entrée de leur modèle [Suri & Schultz, 2001]. Ainsi, au début de la consommation de la récompense, la prédiction de récompense est importante, puis la composante temporelle de cette récompense fait que la prédiction devient de plus en plus faible jusqu'à être nulle à la fin de la consommation, n'entraînant ainsi pas d'erreur de prédiction (voir figure 25).

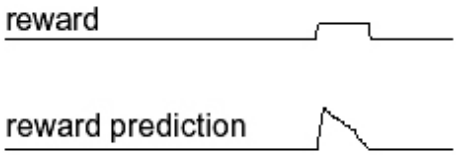

Figure 25 : Evolution de la prédiction de récompense pendant la consommation de récompense dans le modèle de Suri et Schultz [Suri & Schultz, 2001].

De plus, dans leur modèle; Suri et Schultz utilisent plusieurs cellules de Critic qui sont chacune assignée à la description d'un stimulus prédéterminé avant le début de l'expérience. Ainsi leur modèle peut résoudre des problèmes non linéairement séparables et ainsi décrire l'ensemble de l'environnement.

Nous ne pouvons directement adapter ces améliorations à notre modèle. En effet, dans le modèle de Suri et Schultz, c'est l'expérimentateur qui envoie à chaque pas de temps un signal au Critic pour dire que le stimulus A ou le stimulus B est identifié. Dans notre modèle, cela impliquerait un module à part entière qui filtre le contexte sensoriel et n'envoie des informations en aval que lorsqu'il détecte un des stimuli intéressants pour l'expérience. De plus, ce module devrait jouer le rôle de trieur pour envoyer à chaque instant les composantes temporelles du stimulus A uniquement au Critic A et les composantes du stimulus B au Critic

B. Dans notre cas, le Critic reçoit continuellement des données sur le contexte sensoriel et c'est lui seul qui, par apprentissage, parvient à ne répondre que dans un contexte donné. Il apprend ainsi lui-même à détecter ce qu'est un stimulus prédicteur de récompense. Nous ne pourrions pas non plus envoyer de composante temporelle à proprement parler au Critic car, à ce moment-là, toutes les cellules de Critic recevraient toutes les composantes temporelles d'un stimulus passé mêlées avec un stimulus présent et il deviendrait alors très difficile d'y discriminer les informations pertinentes pour l'apprentissage. Nous avons besoin de cellules Critic qui apprennent d'elles-mêmes à se répartir la description des différents stimuli de la tâche. Ce qui ne pourra être fait à partir du modèle de Suri et Schultz. Par contre, nous avons adapté la composante temporelle de Suri et Schultz. Ce qui est présenté dans le paragraphe suivant.

B. Amélioration du modèle à partir de Suri & Schultz, 2001.

Le modèle de Suri et Schultz a été décrit dans le paragraphe 2 de ce mémoire et nous nous restreindrons ici aux explications concernant la représentation temporelle du stimulus. En effet, pour notre deuxième modèle, nous avons utilisé une variante de la composante temporelle utilisée par Suri et Schultz. De la même façon que le signal d'entrée est de plus en plus faible en entrée du Critic dans leur modèle pendant la consommation de récompense, nous avons produit une inhibition de plus en plus forte sur la cellule de Critic pendant cette même période. Celle-ci fonctionne exactement comme la composante temporelle, mais elle permet d'être appliquée au cas où le Critic reçoit des entrées continues sur le contexte sensoriel : au début de la consommation de récompense, l'inhibition est nulle, puis elle devient de plus en plus forte à mesure que le robot a consommé de nombreuses gouttes d'eau. La prédiction de récompense devient ainsi de plus en plus faible pendant la consommation de récompense jusqu'à devenir nulle à la fin de la récompense. Mathématiquement, la seule différence par rapport au modèle précédent est la façon de calculer la prédiction de récompense à chaque pas de temps. L'équation III.A.9 devient :

$$P^t = f\_sigmo(w'^1.ecartBlanc^t + \ldots + w'^{13}.constante^t) - (\sum r^t) \qquad \text{(III.B.1)}$$

Nous n'avons pas changé les autres paramètres du modèle.

B.1 Résultats.

La figure 26 représente les résultats de la simulation du modèle de Houk, Adams et Barto amélioré à partir de Suri & Schultz, 2001. On peut voir que le temps mis par le robot pour atteindre la récompense baisse, signe que l'apprentissage est efficace (figure 26.a). On peut également remarquer que le renforcement ayant lieu pendant la consommation de récompense est de moins en moins fort au cours de l'expérience (figure 26.b), phénomène que l'on n'obtenait pas avec le modèle précédent (figure 24.b).

Le comportement 'BOIRE' n'est ainsi pas renforcé à outrance et le phénomène de sur apprentissage observé à la figure 24.a est atténué : comme on peut le voir sur la figure 26.a le robot met bien un peu plus de temps pour atteindre la récompense après 4 essais du fait qu'il sélectionne trop souvent le comportement 'BOIRE', mais le pic n'est pas aussi important qu'il ne l'est sur la figure 24.a.

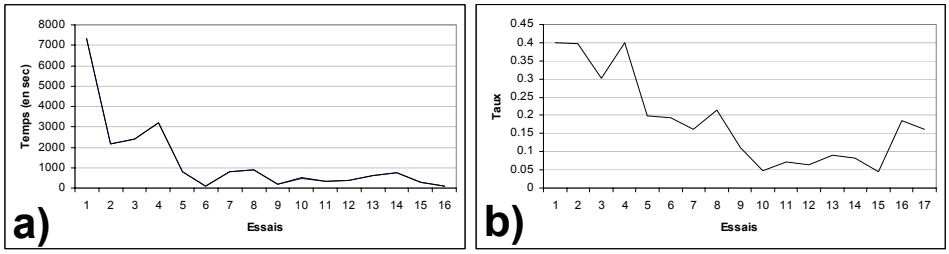

Figure 26 : Résultats de la simulation du modèle après ajout d'une inhibition de la prédiction pendant la récompense. a) Courbe d'apprentissage du Critic. b) Représentation du renforcement (signal dopaminergique) ayant lieu pendant chaque consommation de récompense.

Il faut néanmoins remarquer que le renforcement pendant la consommation recommence à augmenter à partir du 16ème essai (figure 26.b). Ce qui montre que de nouveau, le Critic a atteint ses limites de description de l'environnement.

B.2 Discussion.

L'amélioration du modèle par l'apport d'une inhibition de la prédiction pendant la récompense semblable à la composante temporelle de Suri et Schultz a permis d'améliorer les résultats. Avec cette amélioration, le robot commence à pouvoir construire une séquence de 3 éléments car le comportement 'SE TOURNER VERS LE STIMULUS BLANC' a commencé à être renforcé lorsque le robot passait au centre du labyrinthe. La séquence en construction est alors :

1) Au centre du labyrinthe et avec une perception de la lumière blanche sur le côté, sélectionner le comportement 'SE TOURNER VERS LE STIMULUS BLANC'.
2) Puis, toujours au centre, une fois le stimulus blanc centré en face du robot, sélectionner le comportement 'AVANCER'.
3) Enfin, une fois ce stimulus blanc identifié comme suffisamment proche, sélectionner le comportement 'BOIRE'.

L'amélioration du modèle a donc permis de construire une séquence comportementale plus longue. Mais étant donné que le modèle ne contient qu'une cellule Critic, cette dernière doit, pour commencer à renforcer le comportement 'SE TOURNER VERS LE STIMULUS BLANC' au centre du labyrinthe, délaisser la partie du labyrinthe près du réservoir contenant la récompense. A partir du 16ème essai, cette cellule a ainsi commencé à translater son domaine de prédiction vers le centre du labyrinthe et désapprend à bien prédire au niveau du réservoir éclairé. D'où la hausse du taux de renforcement pendant la consommation de récompense à partir du 16ème essai (figure 26.b).

Autres limites de la modification apportée du fait qu'elle ne reproduit pas exactement la composante temporelle de Suri et Schultz :

- Elle n'opère pas entre la perception du stimulus de lumière blanche et l'obtention de récompense, mais uniquement pendant la consommation de récompense. Ainsi, avec ce modèle, on ne peut toujours pas expliquer la forte dépression de dopamine au moment précis où la récompense est attendue lorsque celle-ci est omise. Toutefois, on observe quand même une baisse progressive du signal de renforcement dans ce cas.
- Elle nécessite une afférence inhibitrice vers la cellule de Critic, permettant d'atténuer son signal de prédiction de récompense. Or, le Cortex n'a que des afférences excitatrices vers le Striatum. Il faut donc examiner la plausibilité biologique de cette solution. L'idéal serait qu'elle provienne de l'amygdale ou d'une autre région limbique qui pourrait avoir pour fonction de délibérer une baisse de motivation pour la consommation de récompense ou une saturation de récompense indiquant que l'animal est rassasié. Mais cela risquerait de n'être pas suffisant pour cette tâche car dans l'expérience du Plus-Maze, les rats sont suffisamment privés d'eau pour ne pas être rassasiés après de

nombreuses consommations de récompense. Il faudrait donc examiner une autre solution biologique.

Pour tenter de répondre au problème de la limitation d'une seule cellule de Critic, nous avons vu que l'utilisation de plusieurs cellules de Critic dans le modèle de Suri et Schultz n'était pas adapté à notre tâche puisque dans leur cas, c'est l'expérimentateur qui affecte arbitrairement chaque stimulus à une cellule déterminée. Or, dans notre tâche, nous souhaiterions que de telles cellules apprennent seules à se répartir la description des différents stimuli de l'environnement. Un autre modèle répond à ce problème. Il s'agit du modèle de Baldassarre [Baldassarre, 2002] qui utilise un réseau de cellules de Critic sous la forme de "mixture d'experts". La prochaine étape va donc consister en l'implémentation d'une mixture d'experts de Critic à la place de la seule cellule que nous avons utilisé jusqu'à maintenant [Baldassarre, 2002].

C. Amélioration du modèle à partir de Baldassarre, 2002.

Le modèle de Baldassare a ceci de particulier que non seulement il utilise un ensemble d'experts Critic, mais il implémente également un réseau d'experts chargés de moduler l'activité des Critic. Ces modulateurs apprennent à accentuer les différences entre les prédictions calculées par chacun des Critic de façon à les aider à mieux se spécialiser. Pour cela, la contribution de chaque Critic aux différentes erreurs de prédiction est calculée, et la formule de mise à jour des poids des Critic est une modification de la règle de Widrow-Hoff :

$$w^{kij} \leftarrow w^{kij} + E^{ijt}.\beta.\check{r}^{kt}.h^{k} \qquad \text{(III.3.1)}$$

Où h^k est la contribution du Critic k à l'erreur de prédiction. Notons enfin que cette mise à jour ne dépend pas de l'erreur globale du modèle mais uniquement de l'erreur faite par le Critic k dans ses prédictions.

Le modèle est en cours de simulation avec différentes quantités d'experts Critic utilisés. Nous présenterons ici des résultats obtenus pour un modèle à 2 experts.

C.1 Résultats.

La figure 27 présente les résultats obtenus sur 29 essais dans le labyrinthe en croix.

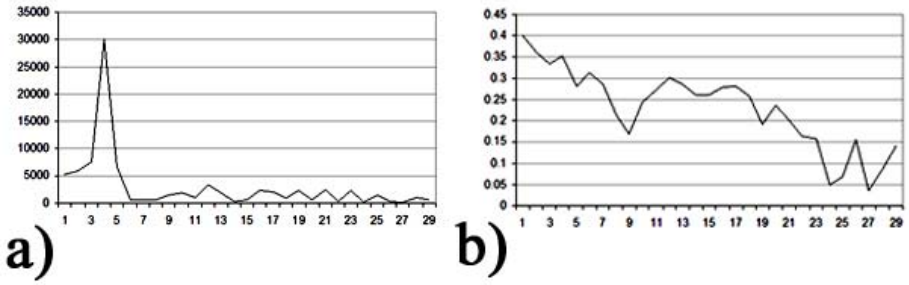

Figure 27 : Résultats de la simulation du modèle de Baldassarre avec 2 experts Critic. a) Courbe d'apprentissage du Critic. b) Représentation du renforcement (signal dopaminergique) ayant lieu pendant chaque consommation de récompense.

Comme avec les autres modèles, on constate un pic au bout de 4 essais dans le temps mis par le robot pour atteindre la récompense (figure 27.a). A partir du 6ème essai, ce temps diminue considérablement et demeure faible jusqu'à la fin de l'expérience. Sur la figure 27.b, on peut voir que le renforcement reste globalement fort pendant la consommation de récompense, même si la courbe affiche une tendance à la décroissance.

C.2 Discussion.

Le modèle ainsi simulé à permis au robot d'apprendre la même séquence à 3 éléments que le modèle de Suri & Schultz (paragraphe III.B). On peut voir d'après la courbe d'apprentissage du Critic que cet apprentissage est relativement stable à partir de 6ème essai (figure 27.a). Toutefois, on peut voir que le modèle met beaucoup de temps à diminuer ses erreurs de prédictions pendant la consommation de récompense (figure 27.b).

Or, nous avons constaté que les poids des experts ne différaient que faiblement, et uniquement sur les variables 'proximBlanc' et 'distanceBlanc'. L'un des deux experts avait un léger avantage sur l'autre dans sa contribution à la prédiction globale du Critic prêt du réservoir blanc, l'autre dominait ailleurs dans le labyrinthe. Les experts utilisés par Baldassarre se spécialisent difficilement et ceci est du au fait qu'ils ne sont mathématiquement pas indépendants. En conséquence, les deux experts apprennent simultanément et cela ralenti l'apprentissage. C'est pourquoi nous envisageons par la suite d'améliorer le modèle à partir de méthodes de mixture d'experts utilisées en Intelligence Artificielle, celles-ci ayant la caractéristique d'utiliser des experts indépendants qui sont soumis à apprentissage pendant des périodes distinctes [Tani et al, 2002 ; Gourichon, 2002].

CONCLUSION : BILAN GENERAL ET PERSPECTIVES

Le travail de modélisation et de simulation robotique présenté dans ce mémoire s'est focalisé sur l'apprentissage stimulus-réponse supposé siéger dans le striatum dorsal chez le rat [Graybiel, 1998]. Nous avons pour cela utilisé une architecture Actor-Critc mettant en œuvre un mécanisme d'apprentissage par renforcement [Sutton & Barto, 1998], et les résultats de simulations donnent à penser que ce mécanisme permet bien la construction de séquences comportementales, même lorsqu'il utilise un module Critic couplé à une partie Actor très détaillée telle que le modèle GPR [Gurney, Presscott & Redgrave, 2001a,b]. Toutefois, les résultats obtenus jusqu'à présent ne permettaient pas à l'apprentissage S-R de s'étendre à l'ensemble de l'environnement utilisé dans la tâche. Un premier travail à réaliser au cours d'une thèse est donc envisagé pour proposer un nouveau modèle Critic à partir d'améliorations inspirées de méthodes d'Intelligence Artificielle sur les mixtures d'experts. Ce travail fera l'objet d'une présentation de poster à une conférence nationale d'intelligence artificelle (Plateforme AFIA 2003) le 3 juillet prochain [Khamassi et al, 2003].

Toutefois, l'état de l'art sur l'apprentissage par renforcement biomimétique nous a permis de constater qu'il existait des modèles qui remettaient en question les architectures Actor-Critic pour l'apprentissage S-R, ainsi que le rôle de la dopamine comme substrat neural de cet apprentissage [Pennartz, McNaughton & Mulder, 2000]. Ces modèles proposent une hypothèse alternative d'un renforcement par le glutamate, donc mettant en oeuvre des circuits et mécanismes différents, basés sur une règle hébbienne à deux termes. Nous envisageons donc, comme deuxième travail à effectuer, l'implémentation d'un de ces modèles de façon à :

- D'une part, tester sa compatibilité avec l'architecture de sélection de l'action implémentée dans Psikharpax, et ainsi confronter cette hypothèse avec ce que l'on suppose de l'anatomie des ganglions de la base [Gurney, Prescott & Redgrave, 2001].
- D'autre part, comparer ses capacités d'apprentissage par renforcement avec ceux du modèle actuellement implémenté et utilisant la dopamine comme médiateur du renforcement, ceci à partir des performances obtenues en simulation.

Ce travail de modélisation serait à réaliser au LPPA et à l'AnimatLab en collaboration avec le Dr. Cyriel Pennartz du Netherlands Brain Research Institute dans le cadre d'une ACI « Neurosciences Intégratives » du Ministère de la Recherche qui vient d'être acceptée pour une durée de 1 an.

D'un autre côté, de façon à permettre au robot Psikharpax de réaliser la tâche du Plus-Maze dans son ensemble, et de comparer son comportement avec celui des rats réels au LPPA, il convient de doter le robot d'un mécanisme lui permettant d'apprendre à moduler son comportement en fonction des informations spatiales dont il dispose, de façon à pouvoir localiser, parmi plusieurs réservoirs éclairés du labyrinthe, celui qui contient la plus forte quantité de récompense, distinction qui ne peut être réalisée par un apprentissage S-R. Nous avons vu qu'une hypothèse consistait à dire que ce mécanisme pouvait s'apparenter à un apprentissage de la sélection des stratégies de navigation et qu'il pouvait mettre en œuvre le striatum ventral. L'analyse de données du noyau accumbens effectuée au LPPA a permis de donner quelques pistes de départ. Notamment, nous avons vu que cette intégration des informations spatiales pouvait être envisagée autrement que par le simple choix parmi des directions cardinales de déplacement. Or il apparaît que les modèles de navigation utilisant une carte cognitive élaborée par l'hippocampe, et intégrant ces informations spatiales pour le comportement dans le noyau accumbens, utilisent pour la plupart des "cellules de but" ou des « cellules d'action » symbolisées par des directions cardinales – Nord, Sud, Est, Ouest [Burgess, Recce & O'Keefe, 1994 ; Trullier, 1997 ; Trullier & Meyer, 1997 ; Arleo, 2000 ; Arleo & Gerstner, 2000]. Le troisième travail qui est envisagé pour une thèse consiste donc à élaborer un modèle du noyau accumbens intégrant la navigation par apprentissage sous forme de sélection de chemins vers des buts. Ce travail impliquerait une collaboration avec le Dr. Angelo Arleo, du LPPA-Collège de France, et mettrait en œuvre des comparaisons par simulation avec les modèles existants.

Le robot, ainsi doté de deux mécanismes d'apprentissage parallèles et complémentaires, l'un pour les comportements procéduraux et l'autre pour les stratégies de navigation, pourrait faire l'objet de simulations sur la globalité de la tâche du Plus-Maze. Notons qu'il serait alors possible, à partir des solutions que nous avons utilisées pour permettre au robot d'effectuer de nouveaux comportements et ainsi induire l'apprentissage (voir paragraphe III.A.3.3), d'envisager un modèle du rôle de modulateur et de favorisateur de certains comportements ou

stratégies que pourrait avoir la dopamine issue du shell du noyau accumbens. On disposerait ainsi d'une architecture globale utilisant plusieurs signaux dopaminergiques de natures différentes en fonction de l'endroit d'où ils sont émis [Di Chiara, 2002 ; Haber, Fudge & MacFarland, 2000 ; Ikemoto, 2002].

Après la validation du modèle sur la tâche du Plus-Maze, il conviendra enfin de le tester dans un environnement non expérimental mais « naturel » et donc imprévisible, au cours d'une tâche de survie afin de le comparer avec les modèles existants en robotique classique, ces derniers ayant la particularité de n'utiliser généralement qu'un seul mécanisme d'apprentissage à la fois.

BIBLIOGRAPHIE

Aizman O, Brismar H, Uhlen P, Zettergren E, Levey A.I, Forsberg H, Greengard P. & Aperia A. (2000) Anatomical and Physiological Evidence for D1 and D2 Dopamine Receptor Colocalized in Neostriatal Neurons. *Nature Neuroscience*, 3:226-230.

Albertin S.V, Mulder A.B, Tabuchi E.T, Zugaro M.B. & Wiener S.I. (2000) Lesions of the Medial Shell of the Nucleus Accumbens Impair Rats in Finding Larger Rewards, but Spare Reward-seeking Behavior. *Behavioral Brain Research*, 117:173-183.

Alexander G.E. & Crutcher M.D. (1990) Functional Architecture of Basal Ganglia Circuits: Neural Substrates of Parallel Processing. *Trends in Neurosciences*, 13:266-271.

Alexander G.E, Crutcher M.D. & DeLong M.R. (1990) Basal Ganglia-Thalamocortical Circuits: Parallel Substrates for Motor, Oculomotor, "Prefrontal" and "Limbic" Functions. *Progress in Brain Research*, 85:119-146.

Arleo A. (2000) *Spatial Learning and Navigation in Neuro-Mimetic Systems. Modeling the Rat Hippocampus*. PhD Thesis, Swiss Federal Institute of Technology, EPFL, Switzerland.

Arleo A. & Gerstner W. (2000) Spatial Cognition and Neuro-Mimetic Navigation: a Model of Hippocampal Place Cell Activity. *Biological Cybernetics, Special Issue on Navigation in Biological and Artificial Systems*, (83):287-299.

Averbeck B.B, Chafee M.V, Crowe D.A. & Georgopoulos A.P. (2002) Parallel Processing of Serial Mouvements in Prefrontal Cortex. *PNAS*, 99:13172-13177.

Baldassarre G. (2002) A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. *Journal of Cognitive Systems Research*, 3:5-13.

Baldassarre G. & Parisi D. (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In Meyer J.-A, Berthoz A, Floreano D, Roitblat H. & Wilson S.W. (Eds.), *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behaviour - SAB2000 - Supplement Volume*, pp. 131-139. Honolulu: International Society for Adaptive Behaviour.

Barto A.G. (1995) Adaptive Critics and the Basal Ganglia. In Houk J.C, Davis J.L. & Beiser D.G. (Eds.), *Models of Information Processing in the Basal Ganglia*, pp. 215-232, The MIT Press, Cambridge Massachusetts.

Beiser D.G. & Houk J.C. (1998) Model of Cortico-Basal Ganglionic Processing: Encoding the Serial Order of Sensory Events. *Journal of Neurophysiology*, 79:3168-3188.

Berns G.S. & Sejnowski T.J. (1996) *The Neurobiology of Decision Making*, chapter How The Basal Ganglia Make Decision, pp. 101-113, Springer-Verlag, Berlin.

Berns G.S. & Sejnowski T.J. (1998) A Computational Model of how the Basal Ganglia Produce Sequences. *Journal of Cognitive Neuroscience*, 10(1):108-121.

Berthoz A. (2002). *La Décision*. Eds Odile Jacob, Paris, France.

Beurrier C, Garcia L, Bioulac B. & Hammond C. (2002) Subthalamic Nucleus: a Clock Inside the Basal Ganglia? *Thalamus & Related Systems*, 2:1-8.

Burgess N, Recce M. & O'Keefe J. (1994) A Model of Hippocampal Function. *Neural Networks*, 7(6/7):1065-1081.

Chang J.-Y, Paris J.M, Sawyer S.F, Kirillov A.B. & Woodward D.J. (1996) Neuronal Spike Activity in Rat Nucleus Accumbens during Cocaine Self-Administration under Different Fixed-Ratio Schedules. *Neuroscience*, 74(2):483-497.

Chevalier G. & Deniau M. (1990) Disinhibition as a Basic Process of Striatal Functions. *Trends in Neurosciences*, 13:277-280.

Cornuéjols A. & Miclet L. (2002) *Apprentissage Artificiel: Concepts et Algorithmes*, chapter Apprentissage de réflexes par renforcement, pp. 483-510, Eds. Eyrolles, France.

Cressant A, Muller R.U. & Poucet B. (1997) Failure of Centrally Placed Objects to Control the Firing Fields of Hippocampal Place Cells. *Journal of Neuroscience*, 17(7):2531-3542.

Dayan P. & Sejnowski T.J. (1994) TD converge with probability 1. *Machine Learning*, 14:295-301.

DeLong M.R. (2000) The Basal Ganglia. In Kandel E.R, Schwartz J.H. & Jessel T.M. (Eds), *Principles of Neural Science 4th ed.* (pp. 853-867, Ch. 43) New Jersey: Prentice Hall.

Di Chiara G. (2002) Nucleus Accumbens Shell and Core Dopamine: Differential Role in Behavior and Addiction. *Behavioral Brain Research*, 137(1-2):75-114.

Filliat D. (2001) *Cartographie et Estimation Globale de la Position pour un Robot Mobile Autonome*. Thèse de Doctorat, LIP6-AnimatLab, Université Paris VI, France.

Filliat D. & Meyer J.-A. (2002) Global Localization and Topological Map Learning for Robot Navigation. In Hallam B, Floreano D, Hallam J, Hayes G. & Meuer J.-A. (Eds) *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, pp. 131-140, The MIT Press, Cambridge MA.

Frank J.F, Loughry B. & O'Reilly R.C. (2001) Interactions between Frontal Cortex and Basal Ganglia in Working Memory: a Computational Model. *Cognitive, Affective and Behavioral Neuroscience*, 1:137-160.

Georgopoulos A.P, Schwartz A.B. & Kettner R.E. (1986) Neuronal Population Coding of Movement Direction. *Science*, 233:1416-1419.

Girard B, Cuzin V, Guillot A, Gurney K. & Prescott T. (2002) Comparing a Bio-inspired Robot Action Selection Mechanism with Winner-Takes-All. In Hallam B, Floreano D, Hallam J, Hayes G. & Meuer J.-A. (Eds) *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, pp. 75-84. The MIT Press, Cambridge MA.

Gourichon S. (1999) Catégorisation Distribuée et Hiérarchisée de son Environnement par un Robot Mobile. In Drogoul A. & Meyer J.-A. (Eds.), *Intelligence Artificielle Située*, pp.87-108, Hermès Science Publications.

Graybiel A.M. (1995) Building Action Repertoires: Memory and Learning Functions of the Basal Ganglia. *Current Opinion in Neurobiology*, 5:733-741.

Graybiel A.M. (1998) The Basal Ganglia and Chunking of Action Repertoires. *Neurobiology of Learning and Memory*, 70:119-136.

Greenberg N. (2001) *The Neuroethology of Paul MacLean: Frontiers and Convergences*, chapter The Past and Future of the Basal Ganglia, Eds. Praeger.

Guigon E. & Burnod Y. (1995) Modeling the Acquisition of Goal-Directed Behaviors by Populations of Neurons. *International Journal of Psychophysiology*, 19:103-113.

Guillot A. & Meyer J.-A. (2001) The Animat Contribution to Cognitive Systems Research. *Journal of Cognitive Systems Research*, 2(2):157-165.

Gurney K, Prescott T. & Redgrave P. (2001a) A Computational Model of Action Selection in the Basal Ganglia. i. A New Functional Anatomy. *Biological Cybernetics*, 84:-410.

Gurney K, Prescott T. & Redgrave P. (2001b) A Computational Model of Action Selection in The Basal Ganglia. ii. Analysis and Simulation of Behaviour. *Biological Cybernetics*, 84:411-423.

Haber S.N, Fudge J.L. & McFarland N.R. (2000) Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum. *The Journal of Neuroscience*, 20(6):2369-2382.

Houk J.C, Adams J.L. & Barto A.G. (1995) A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In Houk, J.C., Davis J.L. & Beiser D.G. (Eds.), *Models of Information Processing in the Basal Ganglia*, pp. 215-232, The MIT Press, Cambridge MA.

Houk, J.C., Davis J.L. & Beiser D.G. (1995) *Models of Information Processing in the Basal Ganglia*. The MIT Press, Cambridge MA.

Ikemoto S. (2002) Ventral Striatal Anatomy of Locomotor Activity Induced by Cocaine, d-Amphetamine, Dopamine and D1/D2 Agonists. *Neuroscience*, 113(4):939-955.

Ikemoto S. & Panksepp J. (1999) The Role of Nucleus Accumbens Dopamine in Motivated Behavior: a Unifying Interpretation with Special Reference to Reward-Seeking. *Brain Research Reviews*, 31:6-41.

Jacobs R.A. & Jordan M.L. (1991) Adaptive Mixture of Local Experts. *Neural Computation*, 3(1):79-87.

Joel D, Niv Y. & Ruppin E. (2002) Actor-Critic Models of the Basal Ganglia: New Anatomical and Computational Prespectives. *Neural Networks*, 15:535-547.

Joel D. & Weiner I. (2000) The Connections of the Dopaminergic System with the Striatum in Rats and Primates: an Analysis with Respect to the Functional and Compartmental Organization of the Striatum. *Neuroscience*, 96(3):452-474.

Khamassi M, Girard B, Berthoz A. & Guillot A. (Sous presse) Mécanismes neuromimétiques d'apprentissage par renforcement dans l'architecture de contrôle du rat artificiel Psikharpax. *Plateforme AFIA 2003*, Laval, France.

Kita H. & Kitai S.T. (1991) Intracellular Study of Rat Globus Pallidus Neurons: Membrane Properties and Responses to Neostriatal, Subthalamic and Nigral Stimulation. *Brain Research*, 564:296-305.

Koechlin E, Danek A, Burnod Y. & Grafman J. (2002) Medial Prefrontal and Subcortical Mechanisms Underlying the Acquisition of Motor and Cognitive Sequences in Humans. *Neuron*, 35:371-381.

Martin P.D. & Ono T. (2000) Effects of Reward Anticipation, Reward Presentation, and Spatial Parameters on the Firing of Single Neurons Recorded in the Subiculum and Nucleus Accumbens of Freely Moving Rats. *Behavioral Brain Research*, 116(1):23-38.

Montague P.R, Dayan P. & Sejnowski T.J. (1996) A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. *The Journal of Neuroscience*, 16(5):1936-1947.

Nakahara H, Doya K. & Hikosaka O. (1999). Benefit of Multiple Representations for Motor Sequence Control in the Basal Ganglia Loops. *BSIS Technical Report*, No.98-5.

Nakahara H, Doya K. & Hikosaka O. (2001) Parallel Cortico-Basal Ganglia Mechanisms for Acquistion and Execution of Visuomotor Sequences: a Computational Approach. *Journal of Cognitive Neuroscience*, 13(5):626-47.

O'Keefe J. & Dostrovsky J. (1971) The Hippocampus as a Spatial Map: Preliminary Evidence from Unit Activity in the Freely-Moving Rat. *Brain Research*, 34:171-175.

O'Keefe J. & Nadel L. (1978) *The Hippocampus as a Spatial Map*. Clarendon Press, Oxford MA.

Pennartz C.M.A. (1997) Reinforcement Learning by Hebbian Synapses with Adaptive Threshold. *Neuroscience*, 81:303-319.

Pennartz C.M.A, Groenewegen H.J. & Lopes Da Silva F.H. (1994) The Nucleus Accumbens as a Complex of Functionally Distinct Neuronal Ensembles: an Integration of Behavioral, Electrophysiological and Anatomical Data. *Progress in Neurobiology*, 42:719-761.

Pennartz C.M.A, McNaughton B.L. & Mulder A.B. (2000) The Glutamate Hypothesis of Reinforcement Learning. *Progress in Brain Research*, 126:231-253.

Ranck J.B.J. (1984) Head-Direction cells in the Deep Cell Layers of Dorsal Presubiculum in Freely Moving Rats. *Society of Neuroscience Abstracts*, 10:599.

Redgrave P, Prescott T. & Gurney P. (1999) Is the Short Latency Dopamine Burst too Short to Signal Reward Error? *Trends in Neurosciences*, 22:146-151.

Tabuchi E.T, Mulder A.B. & Wiener S.I. (2000) Position and Behavioral Modulation of Synchronization of Hippocampal and Accumbens Neuronal Discharges in Freely Moving Rats. *Hippocampus*, 10(6):717-28.

Tabuchi E.T, Mulder A.B. & Wiener S.I. (2003) Reward Value Invariant Place Responses And Reward Site Associated Activity in Hippocampal neurons of Behaving Rats. *Hippocampus*, 13(1):117-32.

Tani J. & Nolfi S. (1999) Extracting Reglarities in Space and Time Through a Cascade of Prediction Networks: The Case of a Mobile Robot Navigating in a Structured Environment. *Connection Science*, 11(2):125-148.

Tremblay L, Hollerman J.R. & Schultz W. (1998) Modifications of Reward Expectation-Related Neuronal Activity during Learning in Primate Striatum. *Journal of Neurophysiology*, 80:964-977.

Trullier O. (1998) Elaboration et Traitement des Représentations Spatiales Servant à la Navigation chez le Rat. Thèse de Dostorat, AnimatLab/LPPA, Université Paris VI, France.

Trullier O. & Meyer J.-A. (1997) Place Sequence Learning for Navigation. In Gerstner W, Germond A, Hasler M. & Nicoud J.D. (Eds.) *Artificial Neural Networks - ICANN'97, 7th International Conference*, pp.757-762, Lausanne, Switzerland.

Schultz W. (1998) Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80:1-27.

Schultz W. (2001) Reward Signaling by Dopamine Neurons. *Neuroscientist*, 7(4):293-302.

Schultz W, Dayan P. & Montague P.R. (1997) A Neural Substrate of Prediction and Reward. *Science*, 275:1593-1599.

Shibata R, Mulder A.B, Trullier O. & Wiener S. (2001) Position Sensitivity in Phasically Discharging Nucleus Accumbens Neurons of Rats Alternating Between Tasks Requiring Complementary Types of Spatial Cues. *Neuroscience*, 108(3):391-411.

Suri R.E. (2002) TD Models of Reward Predictive Responses in Dopamine Neurons. *Neural Networks*, 15:523-533.

Suri R.E. & Schultz W. (1998) Learning of Sequential Movements by Neural Network Model with Dopamine-like Reinforcement Signal. *Experimental Brain Research*, 121:350-354.

Suri R.E. & Schultz W. (1999) A Neural Network Learns Spatial Delayed Response Task with Dopamine-like Reinforcement Signal. *Neuroscience*, 91(3):871-890.

Suri R.E. & Schultz W. (2001) Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation*, 13:841-862.

Sutton R.S. (1988) Learning to Predict by the Method of Temporal Differences. *Machine Learning*, 3:9-44.

Sutton R.S. & Barto A.G. (1998) *Reinforcement Learning: an Introduction*. The MIT Press, Cambridge, MA.

Webb B. (2001) Can Robots Make Good Models of Biological Behaviour? *Behavioral and Brain Sciences*, 24(6):1033-1050.

Wiener S.I, Tabuchi E. & Mulder A.B. (2002) Neural representations of goal directed action in ventral caudate and nucleus accumbens. *Society of Neuroscience Abstracts*, 32:477.10.

Windels F. (2001) Etude par Microdialyse des Variations de Glutamate et de GABA dans la Substance Noire et le Globus Pallidus Induites par la Stimulation Electrique du Noyau Subthalamique chez le Rat. Thèse de Doctorat, Université Joseph Fourier, Grenoble 1, France.

Zugaro M.B. (2002) *Influences des Signaux Multisensoriels et Moteurs dans l'Elaboration des Réponses des Cellules de Direction de la Tête chez le Rat*. Thèse de Doctorat, LPPA-Collège de France, Université de Paris VI, France.

Zugaro M.B, Berthoz A. & Wiener S.I. (2001) Background, but not Foreground, Spatial Cues are Taken as References for Head Direction Responses by Rat Anterodorsal Thalamus Neurons. *Journal of Neuroscience*, 21(14):RC154.