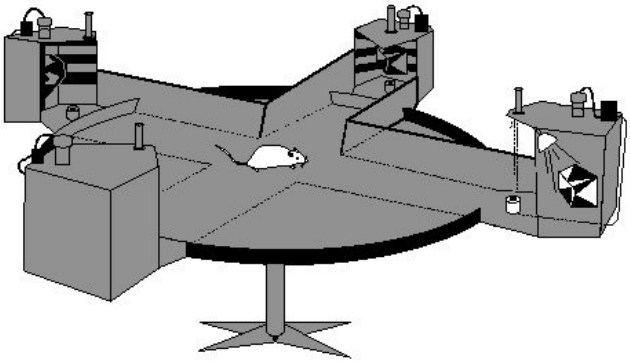
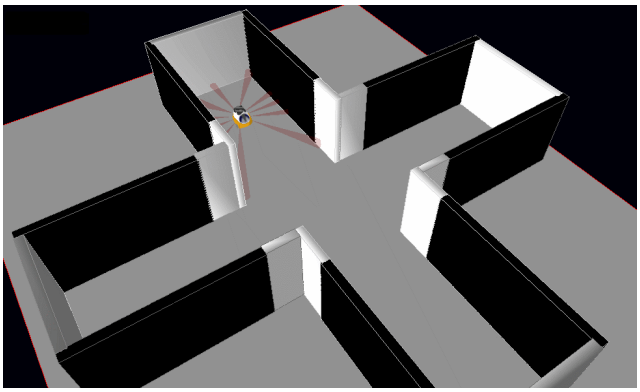
Complementary roles of the rat prefrontal cortex and striatum in reward-based learning and shifting navigation strategies:
Electrophysiological and computational studies, application to simulated autonomous robotics.

Mehdi KHAMASSI

PhD thesis – Université Pierre et Marie Curie (Paris Universitas) - 2007

Speciality
COGNITIVE SCIENCE





Presented on september 26th 2007, in presence of the jury:

Pr. Alain Berthoz	Examinator	LPPA, Collège de France
Pr. Philippe Bidaud	President of the jury	ISIR, Université Paris 6
Dr. Kenji Doya	Reviewer	IRP, Okinawa Institute of Sci. and Tech.
Dr. Agnès Guillot	Thesis director	LIP6/ISIR, Université Paris X
Pr. Cyriel Pennartz	Examinator	SILS, Universiteit van Amsterdam
Dr. Bruno Poucet	Reviewer	LNC, CNRS-Université de Provence
Dr. Sidney Wiener	Thesis director	LPPA, CNRS-Collège de France

TITLE

Complementary roles of the rat prefrontal cortex and striatum in reward-based learning and shifting navigation strategies.

ABSTRACT

Many mammals can behave according to different navigation behaviors, defined as « strategies » which, although not systematically requiring conscious processes, depend on the specific task they are required to solve. In certain cases, if a visual cue marks the goal location, the agent can rely on a simple stimulus-response (S-R) strategy. In contrast, other tasks require the animal to be endowed with a representation of space that allows it to locate itself and to locate goals in the environment. In order to efficiently navigate, the animal not only should be able to learn and exhibit these types of strategies, but it should also be able to select which strategy is the most appropriate to a given task conditions in order to shift from one strategy to the other to optimize outcomes.

The present work employs a multidisciplinary approach (*e.g.* behavior, neurophysiology, computational neuroscience and autonomous robotics) to study the roles of the rat prefrontal cortex and striatum in learning and shifting navigation strategies, and their possible application to robotics. It aims more particularly at investigating the respective roles of the medial prefrontal cortex (mPFC) and of different parts of the striatum (DLS :dorsolateral ; VS: ventral) in these processes, and the nature of their interactions.

The experimental work presented here consisted in :

(1) studying the role of the striatum in S-R learning by : (a) analyzing electrophysiological data recorded in the VS of rats performing a reward-seeking task in a plus-maze; (b) designing an Actor-Critic model of S-R learning where VS is the Critic which drives learning, whereas DLS is the Actor which memorizes S-R associations. This model is applied to robotics simulations, and compared with existing models in a virtual plus-maze;

(2) studying the role of mPFC in strategy shifting by means of electrophysiological recordings in the mPFC of rat performing a task requiring such kind of shifts.

The principal results of this work suggest that :

(1) In the S-R framework: (a) as in primates, the rat VS shows a reward anticipation activity coherent with the Actor-Critic theory; (b) these reward anticipations can be combined with self-organizing maps in an Actor-Critic model that gives a better performance than previous models in a virtual plus-maze, and that shows generalization abilities potentially applicable for the field of autonomous robotics;

(2) the rat mPFC seems to play a role when the animal's current strategy has poor reward yields, prompting learning of another strategy. Moreover, population activity in mPFC changes rapidly in correspondence with shifts in the animal's task-solving strategy, possibly underlying the contribution of this brain area to flexible selection of behavioral strategies.

In conclusion the results are discussed in the framework of previous behavioral, physiological and modeling studies. We propose a new architecture of the rat prefronto-striatal system, where sub-territories of the striatum learn concurrent navigation strategies, and where the medial prefrontal cortex helps decide at any given moment which strategy dominates for behavior.

Keywords: prefrontal cortex; striatum; navigation strategies; learning; shifting; TD-learning; reward; Actor-Critic model.

TITRE

Rôles complémentaires du cortex préfrontal et du striatum dans l'apprentissage et le changement de stratégies de navigation basées sur la récompense chez le rat.

RÉSUMÉ

Les mammifères ont la capacité de suivre différents comportements de navigation, définis comme des « stratégies » ne faisant pas forcément appel à des processus conscients, suivant la tâche spécifique qu'ils ont à résoudre. Dans certains cas où un indice visuel indique le but, ils peuvent suivre une simple stratégie stimulus-réponse (S-R). À l'opposé, d'autres tâches nécessitent que l'animal mette en oeuvre une stratégie plus complexe basée sur l'élaboration d'une certaine représentation de l'espace lui permettant de se localiser et de localiser le but dans l'environnement. De manière à se comporter de façon efficace, les animaux doivent non seulement être capables d'apprendre chacune de ces stratégies, mais ils doivent aussi pouvoir passer d'une stratégie à l'autre lorsque les exigences de l'environnement changent.

La thèse présentée ici adopte une approche pluridisciplinaire – comportement, neurophysiologie, neurosciences computationnelles et robotique autonome – de l'étude du rôle du striatum et du cortex préfrontal dans l'apprentissage et l'alternance de ces stratégies de navigation chez le rat, et leur application possible à la robotique. Elle vise notamment à préciser les rôles respectifs du cortex préfrontal médian (mPFC) et de différentes parties du striatum (DLS :dorsolateral ; VS : ventral) dans l'ensemble de ces processus, ainsi que la nature de leurs interactions.

Le travail expérimental effectué a consisté à :

(1) étudier le rôle du striatum dans l'apprentissage S-R en : (a) analysant des données électrophysiologiques enregistrées dans le VS chez le rat pendant une tâche de recherche de récompense dans un labyrinthe en croix ; (b) élaborant un modèle Actor-Critic de l'apprentissage S-R où le VS est le Critic qui guide l'apprentissage, tandis que le DLS est l'Actor qui mémorise les associations S-R. Ce modèle est étendu à la simulation robotique et ses performances sont comparées avec des modèles Actor-Critic existants dans un labyrinthe en croix virtuel ;

(2) Dans un deuxième temps, le rôle du striatum dans l'apprentissage de stratégies de type localisation étant supposé connu, nous nous sommes focalisés sur l'étude du rôle du mPFC dans l'alternance entre stratégies de navigation, en effectuant des enregistrements électrophysiologiques dans le mPFC du rat lors d'une tâche requiérant ce type d'alternance.

Les principaux résultats de ce travail suggèrent que :

(1) dans le cadre S-R : (a) comme chez le singe, le VS du rat élabore des anticipations de récompense cohérentes avec la théorie Actor-Critic ; (b) ces anticipations de récompense peuvent être combinées avec des cartes auto-organisatrices dans un modèle Actor-Critic obtenant de meilleures performances que des modèles existants dans un labyrinthe en croix virtuel, et disposant de capacités de généralisation intéressantes pour la robotique autonome ;

(2) le mPFC semble avoir un rôle important lorsque la performance de l'animal est basse et qu'il faut apprendre une nouvelle stratégie. D'autre part, l'activité de population dans le mPFC change rapidement, en correspondance avec les transitions de stratégies dans le comportement du rat, suggérant une contribution de cette partie du cerveau dans la sélection flexible des stratégies comportementales.

Nous concluons ce manuscrit par une discussion de nos résultats dans le cadre de travaux précédents en comportement, électrophysiologie et modélisation. Nous proposons une nouvelle architecture du système préfronto-striatal chez le rat dans laquelle des sous-parties du striatum apprennent différentes stratégies de navigation, et où le cortex préfrontal médian décide à chaque instant quelle stratégie devra régir le comportement du rat.

Mots clés : Cortex préfrontal ; striatum ; stratégies de navigation ; apprentissage ; alternance ; TD-learning ; récompense ; modèle Actor-Critic.

Acknowledgements

I wish to express my deep gratitude to the many people who, in one way or another, have contributed to the achievement of this thesis[1]. First of all, I would like to thank the members of the thesis committee, who accepted to allocate time for reading my manuscript, for making comments and corrections on my manuscript, and for coming to Paris to attend the oral defense: Prof. Philippe Bidaud, Dr. Kenji Doya, Prof. Cyriel M. Pennartz, and Dr. Bruno Poucet.

I am particularly grateful to my supervisors Dr. Agnès Guillot and Dr. Sidney Wiener, for effective and constant oversight, for strong and regular interactions, for shared joys and disappointments throughout the projects, for teaching me how to undertake experiments, how to design models, how to write articles, and how to deal with the aspects accompanying scientific investigations. Most of all, thank you for transmitting me your taste for science. Thanks to Professor Alain Berthoz for opening me the doors of his laboratory, for introducing me to the collaboration between the LPPA and the AnimatLab, and with whom interactions have systematically heighten my motivation. Thanks to Professor Jean-Arcady Meyer for accepting me in the AnimatLab team, for introducing me to the animat approach, for regularly feeding my ideas, and giving me the chance to be part of the fascinating ICEA project. Thanks to the young researchers that contributed to my supervision and introduced me to the techniques of computational modeling, electrophysiology, animal training, and neurophysiological data analysis: Dr. Angelo Arleo, Dr. Francesco Battaglia and Dr. Benoît Girard. I feel extremely lucky to have had the chance to cross your roads.

Thanks to all my collaborators, with whom I had repeated and fruitful scientific interactions. Discussions with you all have strongly contributed in refining my understanding of the brain, whether it is natural or artificial: Karim Benchenane, Eric Burguière, Vincent Douchamps, Luc Foubert, David Hopkins, Matthieu Lafon, Nizar Ouarti, Adrien Peyrache, Nathalie Rochefort, Patrick Tierney, Michael Zugaro and all from the CdF team; Thomas Degris, Laurent Dollé, Alexandra d'Erfurth, David Filliat, Loïc Lachèze, Louis-Emmanuel Martinet, Manuel Rolland, Olivier Sigaud, Paul Simard, Antony Truchet and all from the AnimatLab team; Antonius B. Mulder and Eichi Tabuchi from the accumbens team; Ujfalussy Balazs, Gianluca Baldassarre, Riad Benosman, Christophe Grand, Mark Humphries, Tamas Kiss, Francesco Mannella, Olivier Michel, Tony Morse, Patrick Pirim, Tony Prescott, Peter Redgrave, Massimiliano Schembri, Zoltan Somogyvari, Stefano Zappacosta, Tom Ziemke and the members of the consortium of the ICEA project; Ricardo Chavarriaga, Francis Colas, Sophie Denève, Jacques Droulez, Boris Gutkin, Nicolas Lebas and the members of the BACS consortium.

Thanks to members of the two laboratories I worked in during my thesis, who made this period enjoyable, confortable and easier: Adrien Angeli for his philosophy, Lionel Arnaud for his curiosity and enthousiasm, Gildas Bayard for his jokes, Valérie Blondel for her precious help, Christian Boucheny for his modesty and kindness, Christophe Bouder for his computer science skills, Julien Bourdaillet for simultaneous efforts in writing his thesis, Murielle Bourge for her kindness, Vincent Cuzin for his musical and cinematographic advices, Aline Delizy for her shoulders carrying the lab, Stéphane Doncieux and his wise advices, Suzette Doutremer for her kindness and magical skills in histology, Yves Dupraz and Michel Ehrette for their efficiency and modesty, Fabien Flacher for reinventing the world, Céline Fouquet for her imperial diplomacy, Michel Francheteau for his arabic lessons, Emilie Gaillard for her kindness, Patrick Gallinari for his mathematical advices, Pierre Gérard for giving me a first glance on what « writing a PhD thesis » means, Gabrielle Girardeau for blocking the access to the experimental room ;-), Julie Grèzes for her literary advices, Thierry

1 This work was first supported by a three year grant from the french Ministry of Research and Technology (allocation MRT), then for one year by the European Community Integrated Project ICEA.

Gourdin for the sound of his laugh, Stéphane Gourichon for his ability to build a constructive reasoning on any topic, Halim Hicheur for jokes about Tunisian people, Eric Horlait for his supervision of the LIP6, Kinga Igloi for her cultural advices, Isabelle Israël for her kindness, Jean-Daniel Kant for interesting discussions on the future of french universities, Rikako Kato for shared music, Gérard Krebs for his computer science skills, Thierry Lanfroy for his precious help, Yvette Lardemer for her kindness and help, Jean Laurens for his Bayesian equations, Pierre Leboucher for being the keymaster, Anne Le Séac'h for simultaneous efforts in writing her thesis, France Maloumian for her graphical skills, Ghislaine Mary for her precious help, Chantal Milleret for her humor, Camille Morvan for her french diplomacy, Matteo Mossio: So What ?, Jean-Baptiste Mouret for his electronics skills, Laurent Muratet for inventing the « Murge-la-Tête », Nicole Nardy for her help, Steve N'Guyen for his mustache, Panagiota Panagiotaki and her eloquent e-mails, Géraldine Petit and her simultaneous efforts on her PhD, Swann Pichon and his cultural advices, Annie Piton for her help, Nicole Quenech'Du for her help, Gabriel Robert for his outerspace jokes, Isabelle Romera for her help with missions, Laure Rondi-Reig for her advices, Marie-Annick Thomas for her help, Julien Velcin even if we only met when you left the lab, Mark Wexler for keeping the world stable despite strong head movements, Mohamed Zaoui for his help with European projects, Brigitte Zoundi for her help and kindness, and to the members of the team which took care of the animals: Bernard, Eddy, Gérard and Hassani.

There is also an ensemble of scientists I had the opportunity to meet during my PhD, and who I would especially like to thank here: Arnaud Blanchard, Karim N'Diaye, Etienne Roesch and Thi Bich from the Arts&Sciences team; Mathieu Bertin, Makoto Ito, Katsuhiko and Kayoko Miyazaki, Tetsuro Morimura, Izumi Nagano, Eiji Uchibe, Junichiro Yoshimoto and everybody from Doya's unit; thanks to the Okinawan team, especially to Ahmed, Daniella Schiller, Ricardo Chavarriaga and Thomas Strösslin for trying to get a sound out of my japanese flute, Jeffrey Beck for trying to rebuild the world with a beer on the Okinawan beach, Sébastien Bouret and Michel Vidal-Naquet for an improvised a'capella concert on the Okinawan beach, Jadin Jackson, Stephen Cowen, Shi-Ichi Maeda, Jean-Claude Dreher, Emmanuel Procyk, Hirokazu Tanaka and Masami Tatsuno for great moments in Okinawa; Jean-Michel Deniau, Elodie Fino, Stéphane Germain, Yves Gioanni, Maritza Jabourian, Marie-Lou Kemel, Aude Milet, Jeanne Paz, Lucas Salomon, Marie Vandecasteele and all the members of the Institut de Biologie du Collège de France; Emmanuel Guigon, Etienne Koechlin, Thomas Jubault, Chrystèle Ody and everybody from EK's journal club; Etsuro Hori, Hisao Nishijo and Taketoshi Ono from Toyama University; Frederic Kaplan and Pierre-Yves Oudeyer from Sony CSL; Chloé Huetz, Marianne Leroy, Olivier Penelaud and everybody from the Concarneau team; Jean-Marc Edeline, Pascale Gisquet-Verrier from the NAMC in Orsay; Philippe Gaussier and Arnaud Revel from the ETIS-Neurcyber team; Jun Tani and the ISAB society; Jean-Louis Dessalles from the ENST. Thanks to all the people that contributed to the french network of students and young researchers in Cognitive Science, including those with whom I used to work within associations: Bahman Ahjang, Nicolas Baumard, Luca Bisognin, Aline Bompas, Stéphane Burton, Anne Caclin, Marie Chagnoux, Cyril Charron, Sylvain Charron, Maximilien Chaumon, Sylvain Chevallier, Barthélémy Durette, Julien Dutant, Naïma Ghaffari, Nicolas Gomond, Bastien Guerry, Luc Heintze, Vincent Jacob, Claire Landmann, Nicolas Larrousse, Anne-Laure Ligozat, Jean Lorenceau, Monique Maurice, Elsa Menghi, Nicole Morain, Jean-Pierre Nadal, Sandra Nogry, Cédric Paternotte, François-Xavier Penicaud, Camille Roth, François Vialatte. Thanks to Alain Desreumaux, Marie-Pierre Junier, Annette Koulakoff, Santiago Pita, François Tronche and to all participants to the french « Etats Généraux de la Recherche » in 2004.

Thanks to the members of Soyouz, who musically accompanied me during this period: Virgile Guihard, Laurent Gardivaud, David Kern, Vincent Gauthier, Bruno Porret.

Finally, I would like to keep the last lines of these acknowledgments for my brother, my parents, my families (the Khamassis, the Réguignes, the Robinets), my girlfriend, my roomates and all my

friends who supported me through the different stages of this period, who accepted the constraints of my wacky schedule, who did not get upset when facing my occasional mental indisponibility (due to over-concentration on my thesis subject...). Particularly to my closest family: Selim, Marianne, Martine, Hichem and Alexandra, and to my eternal friends: Alice, the Arnaults, the Attias, Anisse, Arnaud, Aurélie, Benj, Benoît, Chakib, Clarisse, the Dabbages, Delphine, Eden, Edith, Eric, Evren, Evelyne and the Embassy Team, Flora, Ghazi, Jennifer, Justine, Karim, Loïc, Lola, Luc, Manu, Marie-Jeanne, Martin, the Matthieus, Mitta and the Pailleron Team, the Morganes, Myrto, Naïma, Nathalie, Nicolas, Nizar, Olivier, Rodolphe, Seb, Solenne, Thomas, Virgile and Ziad.

Outline

INTRODUCTION : A PLURIDISCIPLINARY APPROACH IN THE FRAME OF COGNITIVE SCIENCES

This work is anchored in the field of Cognitive Science, a scientific domain defined by the meeting of an ensemble of disciplines bringing very different tools, methods of investigation, and languages. But they have in common the aim to better understand mechanisms of human, animal or artificial brain and thought, and more generally of any cognitive system, i.e. any information processing complex system able to acquire, to maintain and to transmit knowledges. These disciplines include Neuroscience, Psychology, Philosophy, Artificial Intelligence, Linguistics, Anthropology and others.

More practically, a cognitive science approach often takes the form of the interaction between some of the above-mentioned disciplines to study one particular cognitive function such as perception, learning, navigation, language, reasoning or even consciousness.

In the case of the PhD work presented here, the disciplines at stake include Neuroscience and Artificial Intelligence, and our investigations focused particularly on methods such as Behavior study, Neuropsychology, Neurophysiology, Computational Modeling and Autonomous Robotics to address the issue of reward-based navigation and related learning processes.

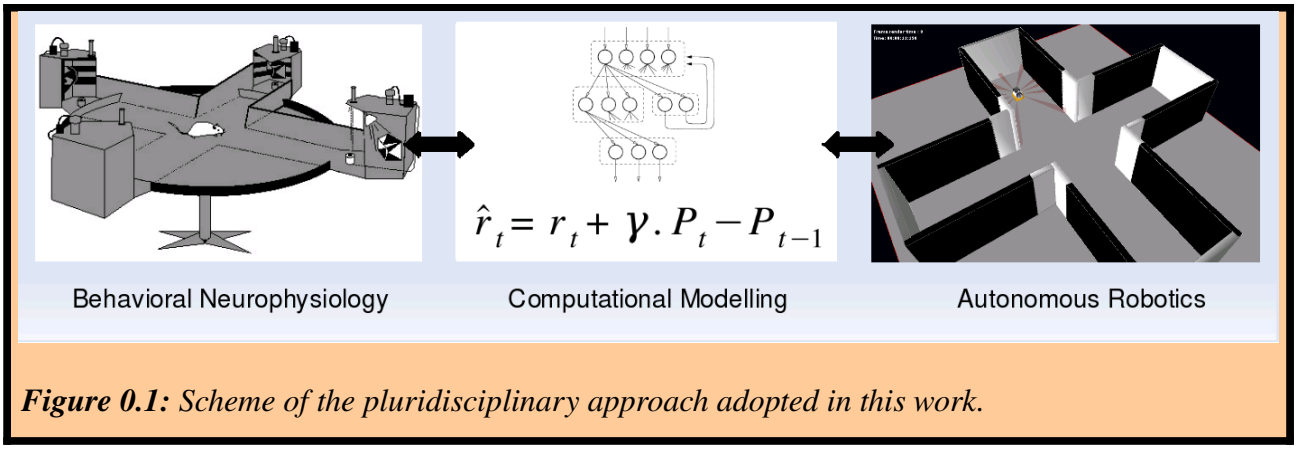


Figure 0.1: *Scheme of the pluridisciplinary approach adopted in this work.*

Why adopt a pluri-disciplinary approach ?

Studying brain functions such as navigation require complementary contributions from different fields (figure 0.1).

- *Behavior analyses* help understand the perimeter and limits of capacities of a given species: *e.g., rodents can learn to reach a goal cued by a landmark by means of stimulus-response associations (S-R learning);*
- *Neuropsychology*, including lesion studies or transient inactivation of a small part of the brain, investigate the neural substrate of the function by identifying which brain areas are necessary to subserve this function: *e.g., lesions of certain parts of the striatum – one of the subcortical nuclei called the basal ganglia –, impair S-R learning;*
- *Neurophysiology*, using electrodes, brain imaging or other techniques, permits to investigate how variables describing parts of the function are encoded and merged within a network of neural units: *e.g., in the previous S-R learning, dopaminergic neurons projecting to the striatum have an enhanced activation when an unexpected reward occurs, and a weaker*

response when a predicted reward is omitted;

- *Computational Modeling* aims at designing computational models to formalize and synthesize large quantities of empirical data related to the studied function, distilling them to a few simple notions. Furthermore, it can help establishing quantitative relationships between individual observations to generate predictions that can serve to validate current and future experiments (Nature Neuroscience Editorial, 2005): *e.g., a machine learning algorithm called temporal-difference (TD) learning, based on the comparison of two consecutive reward estimations for associating a sequence of actions leading to a given reward, seems to appropriately reproduce the error signal concerning rewards observed in dopaminergic neurons;*
- Finally, *Simulated Robotics* can provide further insights on models of a given function by studying their behavior while integrated with models of other brain functions, and while embedded within a simulated or physical body interacting with a realistic and natural environment. For example, integrating a model of reinforcement learning: *e.g. integrating the previous learning algorithm within a robotics platform, together with a model of vision providing inputs, can allow a robot to reproduce a S-R reward-seeking task in a simulated maze. However, the duration of the learning process and perceptual aliasing issues require more information from the above disciplines.*

Learning the methodologies and languages of several of these disciplines permits us to be at the interface of them, and to contribute in rendering the interaction fertile. Training pursued during this PhD training period aimed at learning to contribute to this interface.

What function is being studied here ?

The issue at stake here concerns navigation functions. Cognitive Neuroscience defines *navigation* as a capacity of determining and performing a path from a current position towards a desired location (Gallistel, 1990; Etienne and Jeffery, 2004). Navigation can be seen as a particular case of *goal-directed behavior*, that is a class of behaviors where decision of the action to perform is based on one's current motivational state and goal (one can be hungry and look for food, or one may be thirsty and look for water), one's knowledge about the consequences of candidate actions and whether or not this activity may bring one closer to attain the goal (Dickinson, 1980). However, as we will see later in the manuscript, there exist some navigational situations where a goal is not explicitly selected, and where navigation can be qualified as *reactive* or *habitual* (for example when one follows the same daily pathway to go to work). So many further efforts are needed to better characterize and understand rat behavior in the framework of restricted navigation paradigms. Several successive attempts have been made to classify different navigation behaviors *strategies* particularly in rodents and in biomimetic robots (Trullier et al., 1997; Redish, 1999; Franz and Mallot, 2000; Arleo and Rondi-Reig, 2007). These classifications will be discussed in this manuscript, and adapted to the work presented here.

Moreover, different brain pathways are called into action depending on the cues, signal processing and actions engaged to reach a resource – in other words, on how different navigation strategies are being performed. This is true in humans (Berthoz, 2003b; Berthoz et al., 2003; Hartley and Burgess, 2005) and in rodents (O'Keefe and Nadel, 1978; Redish, 1999). But the precise neural system that is engaged in each navigation strategy is not yet completely elaborated, and the way the brain learns, controls and coordinates these strategies is poorly understood. Notably, it is still an open question whether different brain structures are responsible for learning navigation strategies or for shifting from one to another, or whether the same structures can subserve these two functions (Devan and White, 1999). These are the kind of questions that we will address in the neurophysiological studies presented in this manuscript. More precisely, we will study the roles of two brain structures in the rat, the ventral striatum and the medial prefrontal cortex, which are assumed to be involved in these

learning and/or shifting processes.

Finally, an ensemble of bio-inspired models of navigation have been proposed to describe the involvement of particular brain areas in different strategies during navigation tasks (Burgess et al., 1994; Trullier and Meyer, 1997; Guazelli et al., 1998; Redish and Touretsky, 1998; Foster et al., 2000; Gaussier et al., 2002; Arleo et al., 2004; Banquet et al., 2005; Hasselmo, 2005; see Girard, 2003 or Chavarriaga, 2005 for reviews). These models propose contradictory solutions to describe the brain's involvement in navigation, and they can be improved both on the side of biological resemblance and computational efficiency. Results that will be presented in this thesis do not pretend to bring definitive solutions to the coordination of navigation strategies in these models. However, the approach employed participates in a collaborative manner to such models, and some Modelling work done during the PhD period contributes to the improvement of efficiency and biological plausibility in these types of rodent brain-inspired navigation systems.

Why study navigation in the rat ?

First, the rat is a good experimental model because it has many navigation abilities found in humans (Hartley and Burgess, 2005). They are able to learn different ways to reach a goal location in the environment as will be detailed and discussed below. These will include recognition of a place based on a configuration of objects, and building of a mental representation of the relative locations within the environment, that is a « cognitive map » (Tolman, 1948) which allows animal to plan detours and shortcuts. These diverse capacities give rise to discussion of navigation *strategies* in rats, bearing in mind that this does not systematically require conscious processes.

Furthermore, studying the rat brain and behavior in the framework of navigation can give clues towards the understanding of the same functions in humans. For instance, electrophysiological techniques enabled researchers to find the bases of a *cognitive map* in rodents by finding neurons called *place cells* that respond specifically when the animal occupies a particular location in space (O'Keefe and Dostrovsky, 1971; Muller et al., 1999). These results served as a basis for the later finding of such *place cells* in the human brain (Ekstrom et al., 2003).

Finally, the use of rats in laboratory experiments since 1856 has provided a huge database on their brain and behavior (Grobéty, 1990) which requires synthesis. Integrative neuroscience projects combining neurophysiology and robotics constitute a good tool to start this synthesis. One of these projects is the European Integrated Project ICEA (*Integrating Cognition Emotion and Autonomy*) (2006-2009), in the framework of which this PhD was pursued.

The ICEA project.

The ICEA project aims at designing an artificial rat, that is, a robot whose morphology, behavior and control architecture are as much as possible inspired by its natural counterpart. This project engages the *animat approach*, whose objective is to understand mechanisms of autonomy and adaptation in animals, and to import these mechanisms in bioinspired artefacts called *animats* (Meyer and Guillot, 1991; Wilson, 1991; Guillot and Meyer, 1994; Meyer, 1996; Ziemke, 2005, 2007), which in turn should be able to adapt to dynamic unpredictable environments. On the one hand, such a project provides an integrative approach to bring further insights into brain mechanisms, particularly by integrating models that have usually been tested separately. On the other hand, it aims at providing new brain-inspired algorithms to improve autonomy and adaptivity in autonomous robots, which is one of the potential fields of application of this kind of research.

Previous work on the topic topic started in 2002 as a national project called « Psikharpax » (Filliat et al., 2004; Meyer et al., 2005), supported by the LIP6 and the CNRS/Robea interdisciplinary program, and involving a collaboration between the AnimatLab team at the Laboratoire d'Informatique de Paris 6 and the Laboratoire de Physiologie de la Perception et de l'Action at the Collège de France. A PhD thesis prepared by Benoît Girard within the framework of this project

proposed a first architecture of brain-inspired action selection integrating several navigation strategies, yet without reinforcement learning capabilities (Girard, 2003).

This project extended to the international level by involving eight European research teams and two private companies. It took the name of ICEA and received the financial support of the European Commission running through 2009. Within this new project, my PhD work particularly aims at recording and analysing new neurophysiological data about brain learning mechanisms involved in navigation behavior (experimental designs, animal training and data analysis at the LPPA), and at improving the existing architecture of action selection and navigation based on these results (at the AnimatLab/LIP6/ISIR).

What are the possible applications ?

On the one hand, such integrative neuroscience researches can contribute to our comprehension of human brain mechanisms in navigation: *How do we solve navigation tasks ? What makes us feel disoriented ? How do we learn to adapt to novel environments ?*

On the other hand, such researches can contribute to the field of autonomous robots and agents, by bringing complementary contributions to classical Artificial Intelligence approaches (Brooks, 1991, 1998; Guillot and Meyer, 2003). Until today, the nature has produced the best autonomous agents in terms of adaptation, flexibility, precision, robustness to noise or to damage to part of the system, energy saving and generalization to novel situations (Guillot and Meyer, 2001; Webb and Consi, 2001; Doya, 2001). So it is worthwhile taking inspiration from the natural brain to design autonomous artefacts. In the future, autonomous robots could be useful to perform tasks dangerous for humans, to explore space or the submarine world. They can also serve as interactive toys or for helping people in everyday tasks (Bidaud, 2000; Arleo, 2005; Meyer and Guillot, In press).

Roadmap of this manuscript

This thesis dissertation presents our contributions to the understanding of the rat striatum and medial prefrontal cortex (mPFC) in navigation strategies learning and shifting. For this purpose, experiments were designed, where:

> * rats had to *learn different reward-seeking tasks* and to *encode various sensorimotor associations* to achieve them – *i.e.* to *perform different strategies* for navigating towards goals: go towards a light, turn left, reach a particular position in space...
>
> * rats had to *detect changes* in the task rule imposed without any explicit signal. This requires to recall which previously learned strategy is the best for the new situation, or, if none is appropriate, to proceed with a new *learning process*.

More precisely, investigations in these experiments consisted in:
(1) studying the role of the striatum in Stimulus-Response (S-R) learning in a plus-maze by:

(a) **analyzing electrophysiological data** recorded in the Ventral Striatum (VS) of rats performing a reward-seeking task;

(b) **designing a bioinspired computational model of S-R learning** where VS drives learning, whereas the DorsoLateral Striatum (DLS) memorizes S-R associations. This model is applied to robotics simulations, and compared with existing models in a virtual plus-maze;

(2) studying **the role of mPFC in strategy shifting** by means of **electrophysiological recordings** in the mPFC of rats performing a Y-maze task requiring such kind of shifts.

The manuscript is organized in four chapters:

(i) the *state of the art* introducing navigation strategies and their selection in rodents: behavioral evidence, the neural substrates for their support, and the corresponding bioinspired computational

models;

(ii) a presentation of our work for studying the *role of the striatum in learning navigation strategies*, using electrophysiological, computational modeling and simulated robotics techniques;

(iii) a presentation of our work for studying the *role of the medial prefrontal cortex in navigation strategies shifting*, using electrophysiological and behavior modeling techniques;

(iv) a *discussion* synthesizing these results into a framework integrating the scientific background, trying to sketch an integrated architecture involving both the striatum and the mPFC in the coordination of navigation strategies.

Each chapter begins with a short introduction that outlines the content of the chapter, and provides a self-contained description of the theoretical and experimental concepts related to its main topic. Some of them include full papers already published or submitted.

CHAPTER 1 : BEHAVIORAL, NEURAL AND COMPUTATIONAL MODELS OF NAVIGATION STRATEGIES IN RODENTS

In this chapter, we review the main scientific background concerning the possible involvements of the medial prefrontal cortex (mPFC) and the striatum in reward-based learning and shifting navigation strategies. In the first section, we will present the behavioral evidence for the existence of different navigation strategies in the rat and the latter's capacity of shifting between them. Then, we will present the neuropsychological and neurophysiological literature concerning the involvement of the mPFC and striatum in these strategies. Finally, we will present recent contributions from computational modeling for the understanding of the role of the prefronto-striatal system in learning and shifting strategies.

To do so, we first have to provide a few points of emphasis:

1) Within the framework of navigation, here we are more interested by action selection mechanisms, and the learning mechanisms used to adapt action selection, rather than by the mechanisms of elaboration of spatial information employed in navigation – mainly because the mPFC and striatum may play a critical role in the former, while the hippocampal system is more implicated in the latter as we will see in the neurophysiological section.

2) As we will try to stress in the first section, while existing classifications of navigation strategies in the rat rely upon distinctions of the different types of information that are used in each strategy (simple sensory cues, spatial maps of the environment, etc...), they have some discrepancies concerning the types of action selection mechanisms at stake, and this bears upon the behavioral flexibility which these mechanisms manifest. We will see that certain strategies which have been categorized separately could indeed rely on similar action selection mechanisms, while certain strategies regrouped in a single category appear to be distinguishable by different action selection mechanisms.

3) Moreover, whereas part of the neurobiological data on the mPFC and striatum that we will review comes from the navigation community, another part comes from the instrumental conditioning community, which has its own classification of *behavioral strategies*. Indeed, there are similarities between both kinds of strategies. They distinguish so-called « goal-directed behaviors » which are flexible and rely on the use of a representation of the possible consequences of actions – e.g. Action-Outcome (A-O) associations – and « habits » which are slow to acquire and are assumed not to rely on A-O associations (Dickinson, 1980; Dickinson and Balleine, 1994).

4) Finally, some computational work modelling the roles of the mPFC and striatum in action selection and reward-based learning is grounded on the *Reinforcement Learning* framework (Sutton and Barto, 1998), and proposes a dichotomy of learning algorithms which has been recently shown to parallel the *goal-directed behaviors* / *habits* dichotomy made in the instrumental conditioning community (Daw et al., 2005, 2006; Samejima and Doya, 2007). Indeed, they distinguish *model-based* reinforcement learning, which relies on a model of the *transition function* providing the information concerning the consequences of actions; and *model-free* (or *direct*) reinforcement learning where this *transition function* is neither learned nor used (Sutton, 1990; Sutton et al., 1992; see Kaelbling et al., 1996; Atkeson and Santamaria, 1997 for reviews).

As a consequence, in order to integrate the different scientific backgrounds addressed in this thesis, we will start by reviewing existing classifications of navigation strategies, trying to reconcile them with the *model-based* / *model-free* dichotomy. A few precautions before starting: This attempt will be simplified for the understanding of this thesis, and would require more work before possibly

bringing some contribution to the navigation community. Moreover, the word « model » will be used as a terminology, and does not mean that rodents necessarily have a « model » in their brain. Finally, the strategies that we will consider as « model-free » just assume that their action selection mechanism is model-free, while not addressing the way they elaborate spatial representations.

1. Behavioral evidence for navigation strategies in the rat

In the following sections, we will first list the main categories employed in usual classifications of navigation strategies in rodents (*section 1.1*). Descriptions of each strategy constituting these categories will be accompanied with explanations of possible ambiguities on terminology and classification concerning action selection mechanisms. Then, we will try to bring some elements of clarification from the field of *instrumental conditioning*, and propose a synthetic classification that will help explain the motivation for the navigation strategies in the current experimental designs (*section 1.2*). The section will finish by a presentation of the different modes of alternation (or *shifts*) between strategies a rat can perform, accompanied with behavioral evidence for such shifts (*section 1.3*).

1.1 Classifications of navigation strategies

Evidence for different navigation strategies in the rat comes from behavioral studies showing that they are able to rely on different information to localize themselves in the environment, and to use this information in different manners to reach a certain location in space (Krech, 1932; Restle, 1957; O'Keefe and Nadel, 1978).
Different classifications of navigation strategies have been proposed (O'Keefe and Nadel, 1978; Gallistel, 1990; Trullier et al., 1997; Redish, 1999; Franz and Mallot, 2000; Arleo and Rondi-Reig, In press). These classifications usually point out a series of criteria, some of them overlapping, to differentiate navigation strategies:

- **the type of information** required (sensory, proprioceptive, internal, ...). A distinction is usually made between *idiothetic* cues (internal information such as vestibular, proprioceptive, kinesthesic cues or efferent copies of motor commands) versus *allothetic* cues (external information provided by the environment such as visual, auditory, olfactive cues). In addition, some authors refer to the *dimension* of the stimulus that triggers a certain strategy, discriminating different sensorial modalities of stimuli or configuration of stimuli such as places in the environment – i.e. precise localizations encoded by the animal independently from its body orientation (Birrell and Brown, 2000; Colacicco et al., 2002; Raggozino et al., 2003);
- **the reference frame**: *egocentric,* centered on the subject; versus *allocentric*, centered on point(s) in the environment (single points, places, cue configurations, or place plus other contextual cues).
- **the type of memory** at stake (*procedural* memory, that is, memory of *how* to do; versus *declarative* memory, that is, memory of *what* to do), which is tightly related to:

  *** the kind of action selection** that is involved, which has an impact on learning mechanisms. One of the main distinctions is between *reactive* choices of a behavioral response versus *planned* responses. The precise difference will be explained later.

  *** the time necessary to acquire** each strategy. Some require a gradual or incremental learning process while others support a rapid one-trial learning process, the former being assumed to be less flexible than the latter (Sherry and Schacter, 1987).

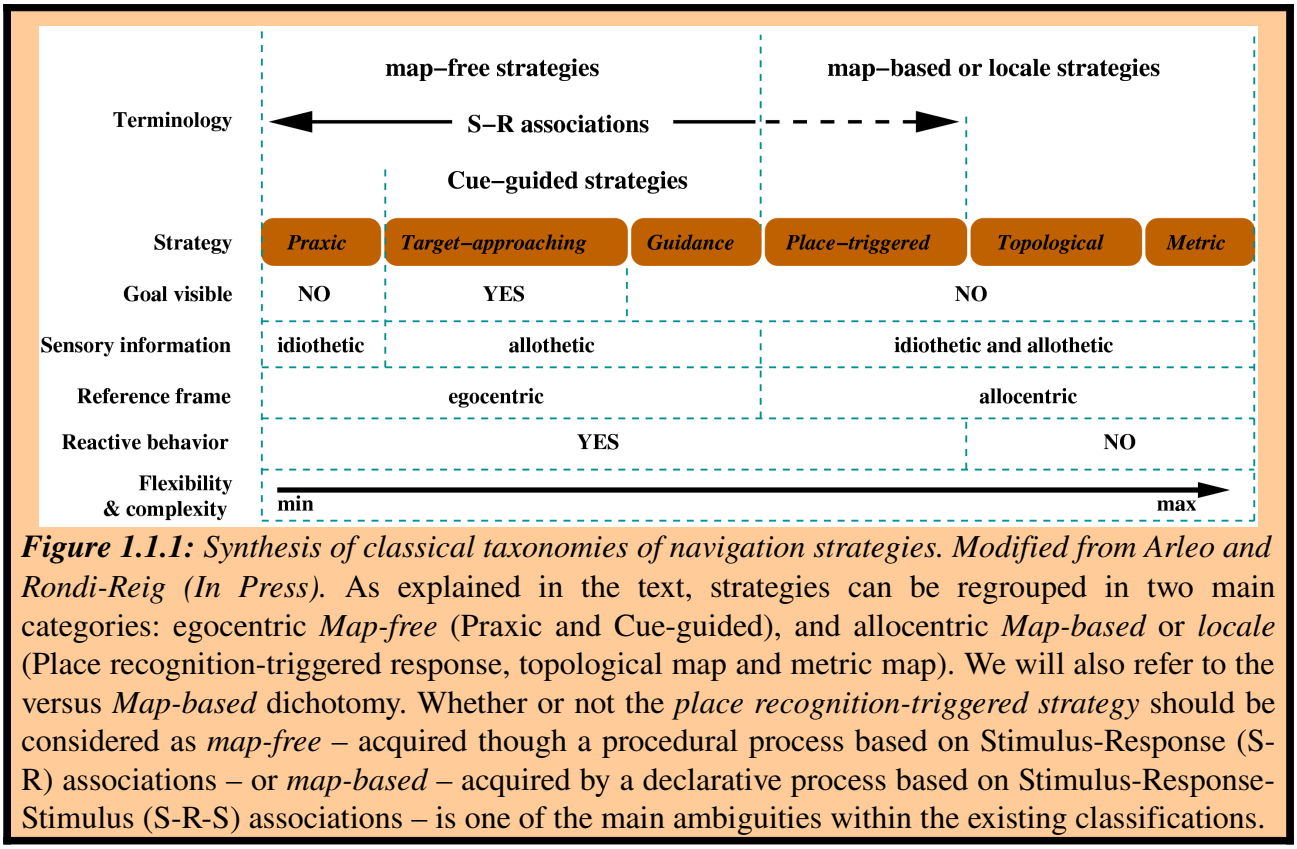


Figure 1.1.1: *Synthesis of classical taxonomies of navigation strategies. Modified from Arleo and Rondi-Reig (In Press).* As explained in the text, strategies can be regrouped in two main categories: egocentric *Map-free* (Praxic and Cue-guided), and allocentric *Map-based* or *locale* (Place recognition-triggered response, topological map and metric map). We will also refer to the versus *Map-based* dichotomy. Whether or not the *place recognition-triggered strategy* should be considered as *map-free* – acquired though a procedural process based on Stimulus-Response (S-R) associations – or *map-based* – acquired by a declarative process based on Stimulus-Response-Stimulus (S-R-S) associations – is one of the main ambiguities within the existing classifications.

These criteria lead to the following simplified overview of existing categories of strategies – which will be more precisely defined below (figure 1.1.1):

1. *Cue-guided* strategies, where a reactive action selection process depends on an external stimulus such as a visual cue. This category includes *target-approaching*, *guidance*, *taxon* navigation, and can be further elaborated in the form of a sequence or chaining of Stimulus-Response (S-R) associations when new cues result from the previous displacement.
2. *Praxic* strategies, where the animal executes a fixed motor program (example: « go straight for a certain distance, then turn right... »). These strategies can also be viewed as S-R associations.
3. *Map-based* or *locale* strategies, which rely on a spatial localization process, and can be either reactive behaviors depending on place recognition (e.g. *place recognition-triggered* response), or can imply a *topological* or *metric* map of the environment – the term *map* being defined by Gallistel (1990) as « *a record in the central nervous system of macroscopic geometric relations among surfaces in the environment used to plan movements through the environment* ».

The next sections provide a more detailed description at the behavioral level of each strategy.

1.2 Cue-guided strategies

Within the framework of the behaviorist theory, the animal's behavior is considered as limited to stereotyped Stimulus-Response (S-R) associations (Thorndike, 1911; Watson, 1913). In the case of navigation, this can be the case when the goal place is visible, or when it is signalled by a single prominent cue, sometimes named a *beacon* in the literature (Leonard and McNaughton, 1990). In such a case, the Stimulus-Response type of association performed by the animal is referred to as ***target-approaching*** or ***beacon-approaching*** (Trullier, 1998). Some authors also refer to it as ***taxon navigation*** which consists in identifying a cue and moving towards it (Morris, 1981; Redish, 1999).

Biegler and Morris (1993) showed that rats are able to perform this kind of *S-R strategy* by learning to discriminate between relevant and irrelevant landmarks in a given environment. They further showed that this type of discrimination required landmark stability, stressing the lack of flexibility of S-R strategies.

Maintaining « *a certain egocentric relationship [with respect to a] particular landmark or object* » is what O'Keefe and Nadel (1978) call ***guidance***, sometimes named *view-based navigation* (Steck and Mallot, 2000). It is a more elaborate situation of S-R association that is considered when the goal is neither visible nor signalled by a beacon. In this case, the animal can use the spatial distribution of landmarks, that is, a configuration of landmarks, relatively to its proper orientation. At the goal, the animal memorizes the spatial relationship between itself and the landmark configuration. Later on, it will attempt to return so as to replicate this view.

As Trullier and colleagues (1997) stressed, « *the memorization of a specific spatial relationship with respect to a landmark-configuration does not necessarily require high-level information such as the identities of landmarks, their positions or the distances to them.* ». In other words, this navigation strategy does not require the processing of an internal spatial representation, nor the use of declarative memory. Indeed, the animal can memorize the raw sensory information associated to the landmark distribution, and later on, can select an appropriate behavior in order to minimize the mismatch between the perceived configuration of landmark and the memorized one.

Target-approach, beacon approach, taxon navigation and guidance can be considered as *Cue-based strategies*. They are considered by authors as S-R associations since the selected response is not based on a representation of the consequence of the action, but rather triggered by a stimulus (Yin and Knowlton, 2006). They are generally described as slow to acquire, that is, rats need several trials before getting a good performance in a task that requires such strategies (O'Keefe and Nadel, 1978; Packard and McGaugh, 1992; Redish, 1999; Yin and Knowlton, 2006).

1.3 Praxic or response strategies

The praxic strategy refers to the case where the animal always executes the same chaining of movements. Some authors refer to this strategy as a *response* behavior (Ragozzino et al., 2002; Yin and Knowlton, 2006). For instance, as shown by Packard and McGaugh (1996), animals perform a praxic strategy in a plus-maze by consistently executing the same body turn (i.e. 90° left) at the center of the maze (figure 1.1.2). This type of response is adapted when the spatial relationship between the departure point and the goal is constant[1]. As a consequence, the praxic strategy is not considered as a flexible strategy but rather exemplifies automatic or habitual behaviors (Chang and Gold, 2003). While some authors assume that the *praxic* strategy requires many trials for its acquisition (Honzik, 1936: O'Keefe and Nadel, 1978; Packard and McGaugh, 1996; Redish, 1999), several authors have reported rapidly learned *praxic* strategies (Pych et al., 2005; see Willingham, 1998; Hartley and Burgess, 2005 for reviews including rodent data).

1 However, Wiener and Schenk (2005) have shown that, if the departure points are few, rats are able to memorize the direction and distance of the goal from each of these points.

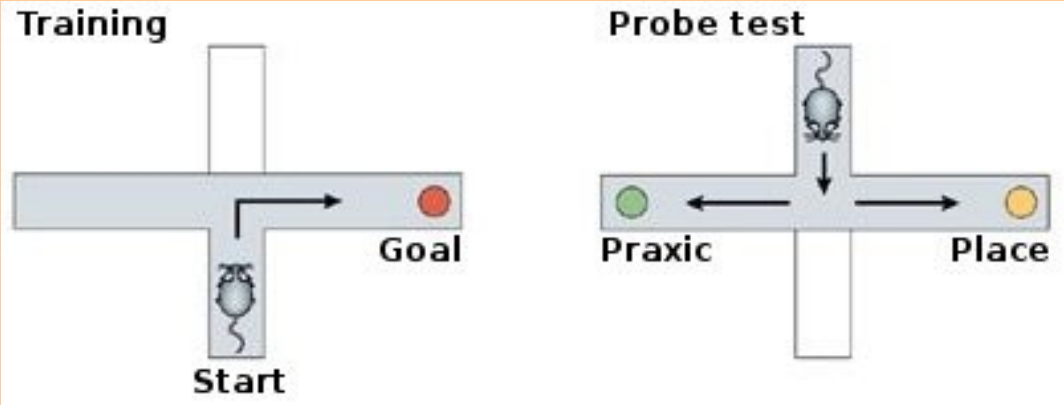


Figure 1.1.2: *Plus-maze setup representing a classical test to discriminate a praxic strategy from a locale strategy (Tolman, 1948; Packard and McGaugh, 1996; Chang and Gold, 2003). Adapted from Yin and Knowlton (2006). Left:* Training setup. Both the starting position (south) and the baited arm (east) remain fixed. *Right:* Testing setup. During the training phase, access to the arm opposite to the start location remains blocked (white arm) to form a T-maze. Animals are trained to enter a consistently baited arm – here, the right arm. Then the configuration is rotated by 180°, the starting point is changed to the north arm and the access to the south arm is now blocked. Animals expressing a praxic strategy perform the same body turn than during training at the intersection of the maze: a right turn which results in entering the western arm. In contrast, animals entering the east arm are considered to have memorized the east location in an allocentric representation. As a consequence, they are considered to be performing a place response as described in paragraph 1.4.

1.4 Map-based or locale strategies

Navigation strategies requiring a localization process can be regrouped into a single category named *map-based strategies* (Arleo and Rondi-Reig, in press) or *locale strategies* (Redish, 1999; Chavarriaga, 2005). They rely on the use of place information, distinguishable from map-free information in the plus maze mentioned above (figure 1.1.2). They are generally assumed to be faster acquired than cue-based or praxic strategies (O'Keefe and Nadel, 1978; Packard and McGaugh, 1992, 1996; Redish, 1999; Yin and Knowlton, 2006) – when a quick exploration of the environment enables animals to build a spatial representation based on latent learning (Blodget, 1929). However, it is important to expose the different strategies constituting this category since they are grounded on different computational principles, are characterized with different levels of complexity and flexibility, and are supposed to differentially involve the prefronto-striatal system, as we will see later on.

Moreover, there is an ambiguity between different usages of the term *locale*. Some authors employ this term to refer to the whole category of map-based strategies (O'Keefe, 1990; Prescott, 1996; Redish, 1999), whereas more and more computational models consider that *locale* navigation refers to a subset where the decision of the behavioral response to perform is based on local spatial information (e.g. a *place recognition triggered* response, Trullier and Meyer, 1997; Arleo and Gerstner, 2000).

Thus we will briefly present each of the so-called map-based strategies in this section.

1.4.1 The place recognition-triggered response strategy

The *place recognition-triggered response* strategy is the process of choosing an action based on the

recognition of places in the environment. Instead of guidance (view-based), place recognition is independent from the observer's orientation and viewing direction (Poucet, 1993). This recognition can be based on *allothetic* cues – external information provided by the environment such as visual, auditory, olfactive or haptic cues – or on *idiothetic* cues – the animal's internal information such as vestibular, proprioceptive, kinesthesic cues or efferent copies that enable an animal to perform *path integration*.

Experiments in the Morris water maze have demonstrated rodents' ability to localize themselves based on allothetic information (Morris, 1981). The maze, a circular pool filled with opaque water (figure 1.1.3), is situated in a room with several extramaze landmarks. To escape, the animal has to find a hidden platform immersed in the water. Animals can learn to take a direct path towards the hidden platform location even when starting from several random departure points, preventing the use of a unique trajectory that could have been memorized based on self-body movements (idiothetic information). The animal is rather presumed to exploit invariant information in the environment as a compass – preferentially using distal rather than proximal cues.

Figure 1.1.3: *The Morris water maze (Morris, 1981). Adapted from Burguière (2006).* In this device, a circular pool is filled of opaque water. Animals have to find a particular location in space, materialized by a platform which enables them to rest without needing to swim. *Left:* Place strategy: The platform is emerged under the water. Animals have to find it using extramaze cues (not represented) while starting from different locations from trial to trial. *Right:* Taxon strategy: The platform is signalled by a visible cue (beacon). Animals have to reach it without being provided with extramaze cues.

Because of its allocentric reference frame, and because it is also considered as more flexible than view-based navigation – probably due to the sparse and coarse information provided by the decomposition of the environment in several places, (Arleo, 2000) –, this strategy is considered by some authors as belonging to the *map-based* category (O'Keefe and Nadel, 1978; Redish, 1999; Arleo and Rondi-Reig, In press; Yin and Knowlton, 2006). However, some other authors consider it as *map-free*, since it does not require the geometrical relationships between memorized locations in the environment that characterize a map (Trullier et al., 1997; Franz and Mallot, 2000). Consequently, learning processes involved are assumed to be different: Stimulus-Stimulus associations (and particularly, Place-Place associations) for map-based and S-R associations for map-free (Balkenius, 1994; Trullier, 1998), and thus respectively fast and slow to acquire. So this strategy appears more difficult to classify than others, and Trullier et al. (1997) « *question the necessity [for distinguishing a] difference between guidance and place recognition-triggered response* ».

1.4.2 The route strategy

O'Keefe and Nadel (1978) call *route strategy* a chain of Stimulus-Response-Stimulus associations. Some authors refer to the general case of chaining sequences of visually-guided, praxic or place recognition-triggered substrategies (Redish, 1999; Arleo and Rondi-Reig, In press). However, this strategy is classified within the map-based category when it is applied to the case where the considered stimuli represent places, making some author view the *route strategy* as a combination or alternation between *place recognition-triggered* and *guidance* strategies (Trullier et al., 1997; Wiener and Schenk, 2005). Redish (1999) mainly applies the route strategy to cases requiring a localization process, and defines it as « an association between positions and vectors (directions of intended motion) ».

Figure 1.1.4 describes the difference between the *route strategy* and topological mapping with a schema. While performing a *route strategy* from a stimulus S1 to another stimulus S2, an animal starts by selecting a response associated to S1. This response is also related to the stimulus S2 that the animal is supposed to reach. As a consequence, the animal can adapt its trajectory before reaching S2, thanks to the *guidance strategy* applied to the approach of S2. However, this process does not provide a bidirectional link between stimuli S1 and S2, and routes S1-R1-S2 and S2-R2-S1 are considered as different and independent. Moreover, this strategy does not take into account the fact that two different routes may pass through the same places, and thus does not imply a topological representation.

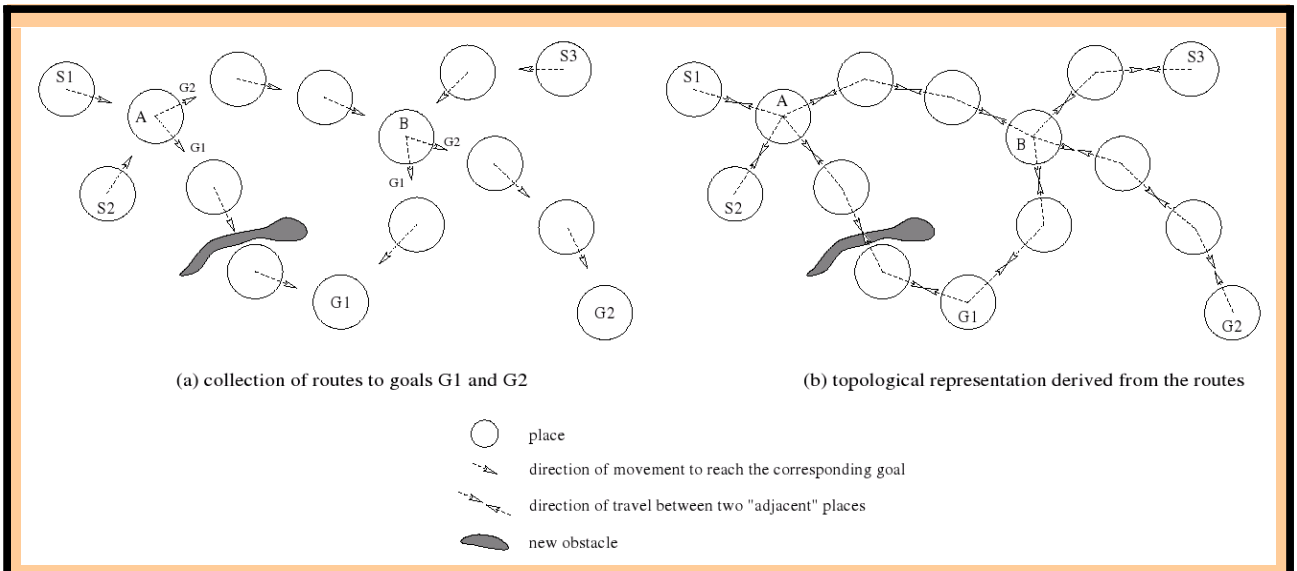


Figure 1.1.4: Comparison of *route* and *topological map-based strategies. Adapted from Trullier et al., 1997. (a)* With the place recognition-triggered response strategy there can be an ensemble of intersecting routes. The animal is able to go from S1 to G1, from S2 to G2, and from S3 to G1. However, if there is a new obstacle on the way from S1 to G1, as on this figure, the animal is lost because the route from S1 to G1 is unique. (b) In contrast, if the animal merges its representations of routes into a topological representation, then it can go back to place A, take the sub-route between places A and B, and take the sub-route from place B to goal G1. The resulting path is the concatenation of three sub-sequences, derived from three different routes.

1.4.3 Topological mapping strategy

A topological representation can be expressed in mathematical terms as a graph, where nodes represent places and edges represent adjacency, or direct connectivity. Then, two nodes are linked if there is a previously visited direct path which leads from one corresponding place to the other

corresponding place, without going through a third intermediate known place.

A topological representation of the environment can be obtained during exploration by merging place-action-place associations derived from a collection of routes. Such a topological map provides a goal-independent and structured representation of places. Because this process provides a bidirectional link between places, it is more flexible than the *route strategy* (figure 1.1.4): when an obstacle is encountered, alternative intersecting paths can be taken.

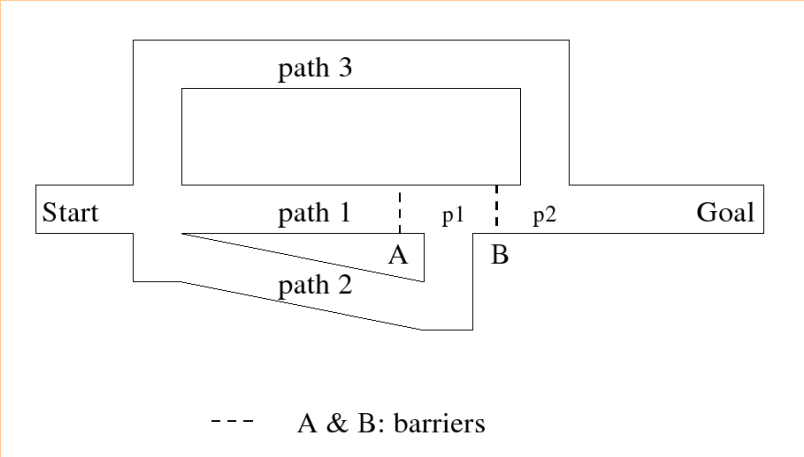


Figure 1.1.5: *Tolman and Honzik's detour problem. Adapted from Tolman, 1948.* After exploration of the entire maze, the path usually taken by the rats is path 1, but when the barrier A is put in place, the rat shifts its choice to path 2, shorter than path 3. If barrier B is put in place instead of barrier A while the rat is performing path 1, then the rat reacts by choosing path 3, without trying path 2.

Behavioral experiments have provided evidence that a strategy based on a topological representation of the environment can be employed by rodents (Tolman, 1948; Thinus-Blanc, 1996; Poucet and Hermann, 2001) or cats (Poucet, 1984). Tolman and Honzik's detour problem is such a case (figure 1.1.5). In this experiment, a rat is required to select one of three paths leading to a reward. It first learns to use the shortest one, that is, path 1. When path 1 is blocked with a barrier, after several trials, the rat chooses path 2, which is the second shortest path. However, if a barrier is put at the end of path 1 while the rat is performing this path (barrier B on figure 1.1.5), then the rat shifts its choice to path 3 without trying path 2. The authors' interpretation was that the rat has the « insight » that both path 1 and path 2 are blocked by barrier B. Such an « insight » does not necessarily require a metric representation of the environment because it can be solved by simply suppressing the link between places p1 and p2 in a topological representation of the experimental setup. Moreover, taking the shortest available path (for instance taking path 2 when path 1 is obstructed), can be explained using a topological map without metric representation. Indeed, the number of consecutive places or *nodes* required to encode path 2 within the map is supposed to be smaller than for path 3.

1.4.3 Strategies based on a metric map

As explained in the previous paragraph, in some cases, a topological map can provide some distance information without using any metric representation. However, this is possible only for known paths and cannot be applied for planning detours and shortcuts in paths never explored before.

Figure 1.1.6 illustrates two situations that cannot be solved with a topological map. In the first example, the animal starts from position A and finds an obstacle B on the path it already experienced to reach E. In such a case, the animal has to make a detour through an unknown region. Choosing the shortest inexperienced detour requires an estimation of the size of the unknown region within an incomplete map of the environment. In the second example, the animal is traversing a

familiar path from A to C. It is assumed that C cannot be perceived from B because there is a forest between them. Knowing that a path goes round the forest, the animal can deduce the direction of a shortcut though the forest towards point C.

Several experiments report the ability of animals to rely on metric information for navigation, such as execution of paths in the dark (Collett et al., 1986), shortcuts (Gallistel, 1990; Roberts et al., 2007), or the planning of paths from unexplored areas in a Morris water maze (Matthews et al., 1999). However, it is not always clear whether rodents perform metric navigation using a computational procedure that subsumes topological mapping as proposed by (Trullier et al., 1997), which view has been criticized by some authors on the ground that animals can solve certain tasks requiring limited metric information by simply using a simple praxic strategy (Foster et al., 2000).

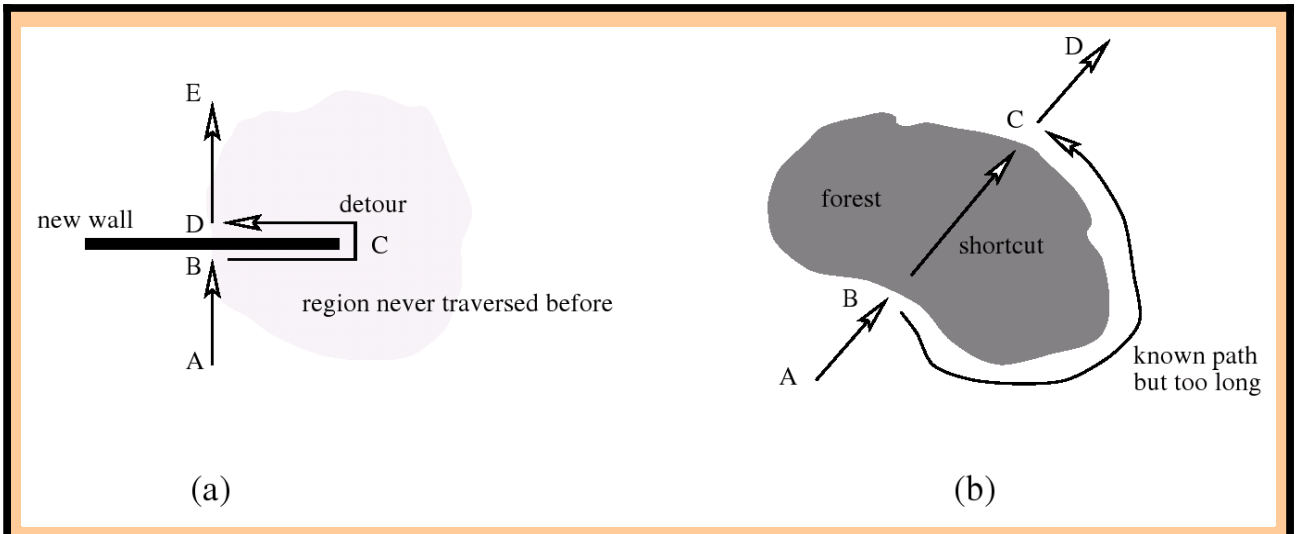


Figure 1.1.6: *Adapted from Trullier et al., 1997. (a):* Metric detour and *(b):* Metric shortcut behaviors provided by a strategy based on a metric map. In both cases, the animal takes a path never experienced before, without being able to use familiar landmarks (the new wall is assumed to be tall and the forest is assumed to be dense).

1.5 Discussion of the classifications

As we tried to point out, there are some inconsistencies between existing classifications of navigation strategies. These reveal some ambiguities in the terminology adopted and on the distinctions between categories.

Indeed, it appears to some authors that these classifications lend too much importance to the issue of involvement of a spatial localization process for the categorization of navigation strategies (Trullier et al., 1997; Sutherland and Hamilton, 2004). Flexible, rapidly acquired, declarative and, as we will see later, *hippocampus-dependent* strategies, have often been assimilated with spatial (allocentric), map-based strategies. In constrast, inflexible, slowly acquired, procedural and, as we will see later, *striatum-dependent* strategies like the praxic and cue-guided strategies, have been regrouped in map-free Stimulus-Response strategies.

However, as we have seen above, on the one hand, certain strategies relying on allocentric representations of space such as the place recognition-triggered response do not require a map and are inflexible, while on the other hand, there are cases where a praxic or a cue-guided strategy can be rapidly acquired. The latter case has been extensively described in the field of instrumental conditioning, where an animal introduced in a novel environment, can quickly learn to associate responses to external cues (such as a light or a tone), and can remain in a flexible behavior – called

goal-directed, in opposition to *habitual behavior* – until extensive training has been undertaken (Dickinson and Balleine, 1994; see Cardinal et al., 2002; Yin and Knowlton, 2006 for reviews). This type of flexible cue-guided behaviors have been recently described as relying on a ***world model***, not necessarily a *map* since this term has an allocentric connotation, but still using a structured representation of transitions between task *states* (Sutton and Barto, 1998; Doya, 1999; Kawato, 1999; Daw et al., 2005; Samejima and Doya, 2007). This model of the environment can be viewed as echoing the term "cognitive graph" (Muller et al., 1991; Trullier, 1998). The latter was proposed to counterbalance the "cognitive map" term by getting rid of the assumption of existence of a neural metric representation, which too strongly resembles the "map in the head" assumption deplored by some authors (Kuipers, 1982).

Then a distinction between ***model-free*** and ***model-based*** behavioral strategies appears to be interesting for disambiguating certain navigation strategies. Thus, in the next section, we will first explain the difference between *model-based* and *model-free* behaviors (or strategies), using an example taken from an instrumental conditioning task. Then we will attempt to characterize the navigation strategies described above within this framework. Yet, there was not have enough time in the presently described work to extensively discuss the possible contribution of this attempt. We will rather propose an attempt to reconcile some of the inconsistencies described above, while over-simplifying other aspects of these classifications. Further investigations will be indeed required to evaluate this proposition (for example by proposing a behavioral protocol where the model-free/model-based dichotomy might be more appropriate than previous navigation strategies to describe the mode of acquisition of animals' behaviors). However, as stated at the beginning of this chapter, it is still a proposition which, in the framework of this thesis, will help us make the link between navigation strategies, neurophysiological data and computational models.

1.6 *Model-based* versus *model-free* strategies

These terms, coming from the Computational Modeling community, refers to models implementing learning processes that employ a *world model*, that is, a representation of the *transition* from one state to another that results from a behavioral response (Sutton and Barto, 1998; Doya, 1999; Kawato, 1999; Daw et al., 2005, 2006; Doya, 2007). In other words, this transition information provides Action-Outcome (A-O) associations (Dickinson and Balleine, 1994). This representation of the estimated consequences of actions can be used in the action selection process, making it more flexible than *model-free* behaviors. This *world model* can either implement allocentric positions within the environment or, more generally, states of a given task.

The *model-based* versus *model-free* dichotomy has recently been applied successfully to replicating rats' ability to alternate between a flexible visually-guided behavior and a reactive visually-guided behavior. In the field of instrumental conditioning, each of them refer to distinct learning processes, named *goal-directed learning* and *habit learning* (Dickinson, 1980; Dayan and Balleine, 2002; Daw et al., 2005, 2006). According to Colwill and Rescorla (1986) and Dickinson (1980), the former is controlled by the anticipation of the outcome and its performance is flexible since it is sensitive to reward devaluation, whereas the latter is controlled by antecedent stimuli, its performance being inflexible because insensitive to the manipulation of the outcome.

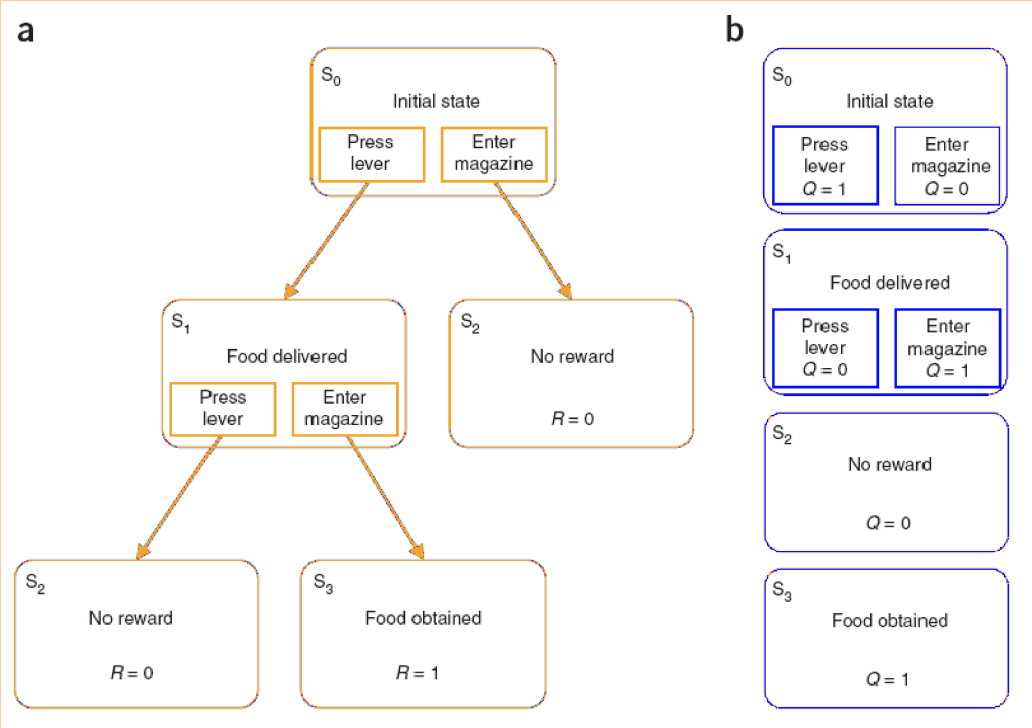


Figure 1.1.7: Example of Model-free / Model-based system. Adapted from Daw et al. (2005). **a:** model-based controller, **b:** model-free controller. The former has a representation of expected consequences of actions, and use it for action selection, whereas the latter has not. In the model-based controller, *S0* and *S1* are two different states that do however correspond to a unique location in the environment.

In the task employed by Daw et al. (2005), rats have to learn to press a lever in response to an external cue, and then to enter a magazine in order to get a food reward. Figure 1.1.7 describes the task in a schema where different *states* (e.g. possible situations within the task) are represented by leaves in a tree-graph, whereas arrows represent possible transitions from one state to another. After training, when rats have learned the task, a reinforcer devaluation is imposed to rats. This can be done, for example, by feeding the animal until satiation, or by pairing the food reward with illness to induce aversion (see Dickinson and Balleine, 2002 for a review). After that, animals are tested to see whether they will continue to perform the actions previously associated with the newly devalued outcome. Strikingly, while after moderate pretraining rats stop performing the task that leads to food reward, after extensive pretraining rats persist in pressing the lever even if the outcome had been devaluated. In the former case, the animal's behavioral is said to be sensitive to outcome devaluation (e.g. *goal-directed*), whereas in the latter case, the extensive training has built a *habit*, insensitive to devaluation.

Daw et al. (2005) could reproduce these two situations by implementing two different models: one using a representation of the consequences of actions – a *world model* or *tree*; the other learning simple Stimulus-Response associations (figure 1.1.7). The former is called a *model-based* controller, requires more computations and memory – for A-O associations – is quickly learned and remains flexible in order to adapt to new environments or to changing tasks. The latter is called a *model-free* controller, is simpler and less computationally expensive. Because of the absence of representation of A-O associations, it is slower to acquire – hence requiring extensive training –, and is less flexible to task changes. The precise computational reason for this will be explained in the *modeling section* at the end of this chapter.

The *model-free/model-based* dichotomy strongly resembles the one previously defined by authors

between flexible *map-based* strategies and automatic *map-free* ones – such as cue-based and praxic strategies. However, in this dichotomy, the main difference between behavioral strategies does not rely on spatial versus non spatial information, but rather on the type of learned associations, respectively A-O and S-R, which result in providing different degrees of flexibility. Indeed, Daw et al. (2005)'s *model-based* controller, the *world model*, contains un-necessarily allocentric states, as shown on figure 1.1.7. In their graph, states S0 and S1 correspond to two different states within the task – before and after lever-pressing – but to the same position in space.

Thus, A-O associations can also be learned within an egocentric framework. This gives an argument for the existence of rapid, flexible and "declarative" (because relying on A-O associations) *cue-guided* or *praxic* strategies, whose action selection mechanism corresponds to a similar graph than the one displayed on figure 1.1.7 (left part), representing a subject's estimated states in prediction of the performance of a sequence of egocentric movements. This assumption is indeed to be checked with experiments in which an outcome devaluation procedure would be imposed to animals.

Symmetrically, the previously mentioned *place recognition-triggered* strategy, which used to be included in *map-based strategies* but is considered by some authors as relying on S-R associations only, would have the same action selection mechanisms as the *model-free* part of Daw et al. (2005)'s system (right part of figure 1.1.7).

Extending this dichotomy to navigation, we will consider two main categories: *model-based* versus *model-free* strategies. Within these two main groups, strategy differentiation relies on the *dimension*, defined by their reference frame and modality of processed stimuli:

- egocentric reference frame, relying on idiothetic (praxic), or allothetic (cue-guided) stimuli;
- allocentric reference frame, relying idiothetic and/or allothetic stimuli (place).

Thus we will adopt the following notation for the rest of the manuscript:

Model-free strategies: *Praxic model-free* (idiothetic egocentric S-R), *cue-guided model-free* (allothetic egocentric S-R), and *place model-free* (place allocentric S-R) respectively correspond to praxic, cue-guided and place recognition-triggered (PTR) strategies in the previous classification.

As mentioned in the introduction of this chapter, assuming that the *place recognition-triggered strategy* is « model-free » does not mean that no model is used at the level of place recognition processes. It only considers that the action selection process is reactive and relies on S-R associations.

Model-based strategies: *Praxic model-based* strategy (idiothetic egocentric A-O), *cue-guided model-based* strategy (allothetic egocentric A-O), together with strategies based on a spatial topological map that will be noted *place model-based* strategies (place allocentric A-O).

Figure 1.1.8 summarizes the resulting taxonomy.

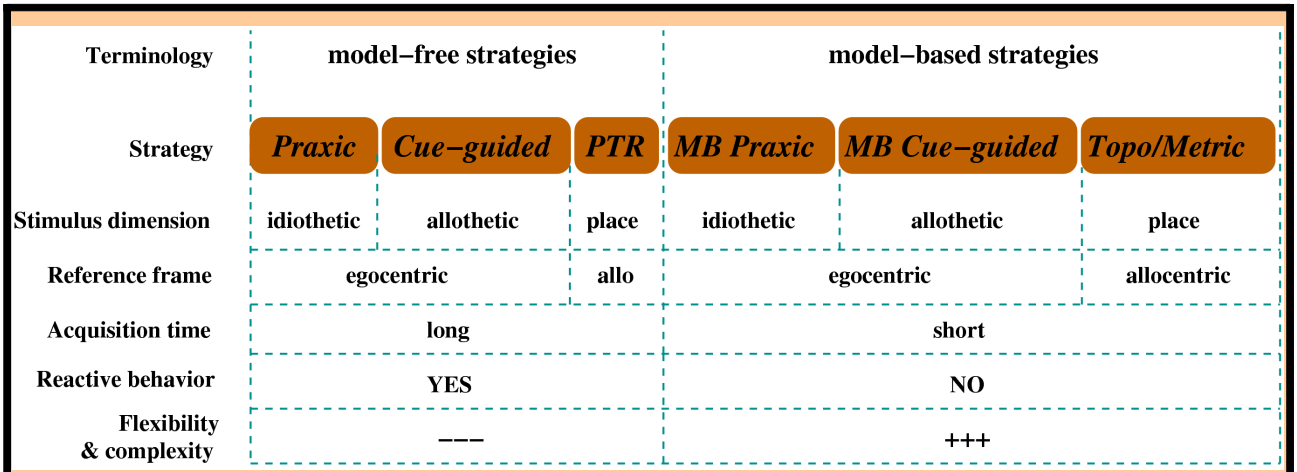


Figure 1.1.8: *Model-free / Model-based* taxonomy applied to navigation strategies. Inside the two main groups, strategy differentiation relies on the dimension, referring to the sensory modality of processed stimuli (idiothetic, allothetic and both: place) and on the reference frame (egocentric versus allocentric). MB: model-based; Topo: Topological. Model-free are considered as slower to acquire and less flexible than model-based strategies. PTR: Place recognition-triggered response.

It would be interesting to study if this way to classify navigation strategies, despite being very simplistic and schematic, can help explain some contradictory results found in the literature. For instance, some authors have reported a more rapid acquisition of the *praxic* strategy than the *locale* strategy (Pych et al., 2005). These results appear to be in contradiction with previous observations that *praxic* strategies should be slower to acquire than *locale* strategies (Packard and McGaugh, 1996). It could be the case that rats in the experiment of Pych et al. (2005) indeed were using a *praxic model-based strategy*, which is assumed to be more flexible than the *place recognition-triggered response* strategy in the model-based/model-free classification.

Furthermore, it would be interesting to see whether postulating that different brain regions subserve *model-based visually-guided* versus *model-free visually-guided* strategies can help explain the differential impairments of visually-guided behaviors resulting from lesions of different brain areas. Indeed, without getting into much details on the neurobiology here (see next sections), it is worthy of note here that lesions of the dorsal striatum are found to impair the navigation towards a visible platform in the Morris water maze (Packard and McGaugh, 1992), whereas after extensive training, lesions of the same brain region only impaired the flexibility of visually-guided behaviors, while still enabling rats to find the visible platform (McDonald and White, 1994). Indeed, it could be that lesions of the dorsal striatum only impaired one of the two visually-guided strategies postulated in the model-free/model-based dichotomy, while sparing the other one, and thus still enabling some visually-guided behaviors.
However, much more work is needed to rigorously analyse the above mentioned experiments in the light of the model-based/model-free dichotomy, to see whether it can or cannot bring complementary contributions to the previous classifications of navigation strategies.

As we will see, this dichotomy between *model-free* and *model-based* strategies will help us bring together neurobiological and computational data on the rat prefrontal cortex and striatum reported by different scientific disciplines. The classification into different *dimensions* will have a direct implication on the consideration of behavioral shifting between navigation strategies, as described below.

1.7 Strategy shifts

Rats' ability to shift from one navigation strategy to another has been strongly supported by the seminal work of Krech (1932). In Krech's experiments, rats were trained in a maze that had four choice points. The experimenter changed the layout of the maze after each trial, and varied the stimuli that were relevant (left-right, light-dark), so that the problem was objectively unsolvable. Krech discovered that his rats did not respond randomly, but instead responded systematically first to one set of stimuli for a few trials, then to another, and so on. These results were taken to suggest that the rats were "trying out hypotheses", and that their learning was guided by confirmation or rejection of strategies, rather than by kinesthetic stimuli.

Pursuing the investigation, Krech argued that the rat attends to only one dimension of the discrimination problem at a time – e.g. spatial position (left or right) and not brightness of the goal box (light or dark) –, instead of gradually learning how to solve the task. In this view, the rat would try different hypotheses, and only learn about the value of left over right when it hit upon the correct hypothesis. Thus, Krech's theory considered learning to be noncontinuous and insightful – a distinct shift in attention from one dimension to another.

In this manuscript, the simplification that we adopt considers two different conditions for shifting, and two types of shifts.

The two conditions are: 1) stability of the task; versus 2) a change in the task.

Within the case of a task change, the two types of shifts considered are: intradimensional shifts (within the same modality); versus extradimensional shifts (e.g. praxis/cue-guided, or cue-guided/place, ...).

1.7.1 Two conditions for shifting

Separating strategy shifts in response to a change in the task from strategy shifts in a situation of task stability was originally justified by Sherry and Schacter (1987)'s suggestion that different brain systems should subserve these two conditions. According to them, « *preservation of variance across episodes* » and « *detection and preservation of invariances across episodes* » are two mutually incompatible encoding processes.

The first condition considered (task stability) refer to the case when the animal is getting familiar with an unchanging task. In such a case, the animal can progressively abandon a flexible but cognitively expensive strategy, and rather shift to a more reactive model-free strategy, which is less flexible, but to which the environment's invariance let the time to be learned. This kind of shift precisely refers to the shift from a *goal-directed behavior* to a *habit* described above. So this shift will be simplistically considered as a shift from a *model-based strategy* to a *model-free strategy*, as modeled in the case of a lever-press task (Daw et al., 2005).

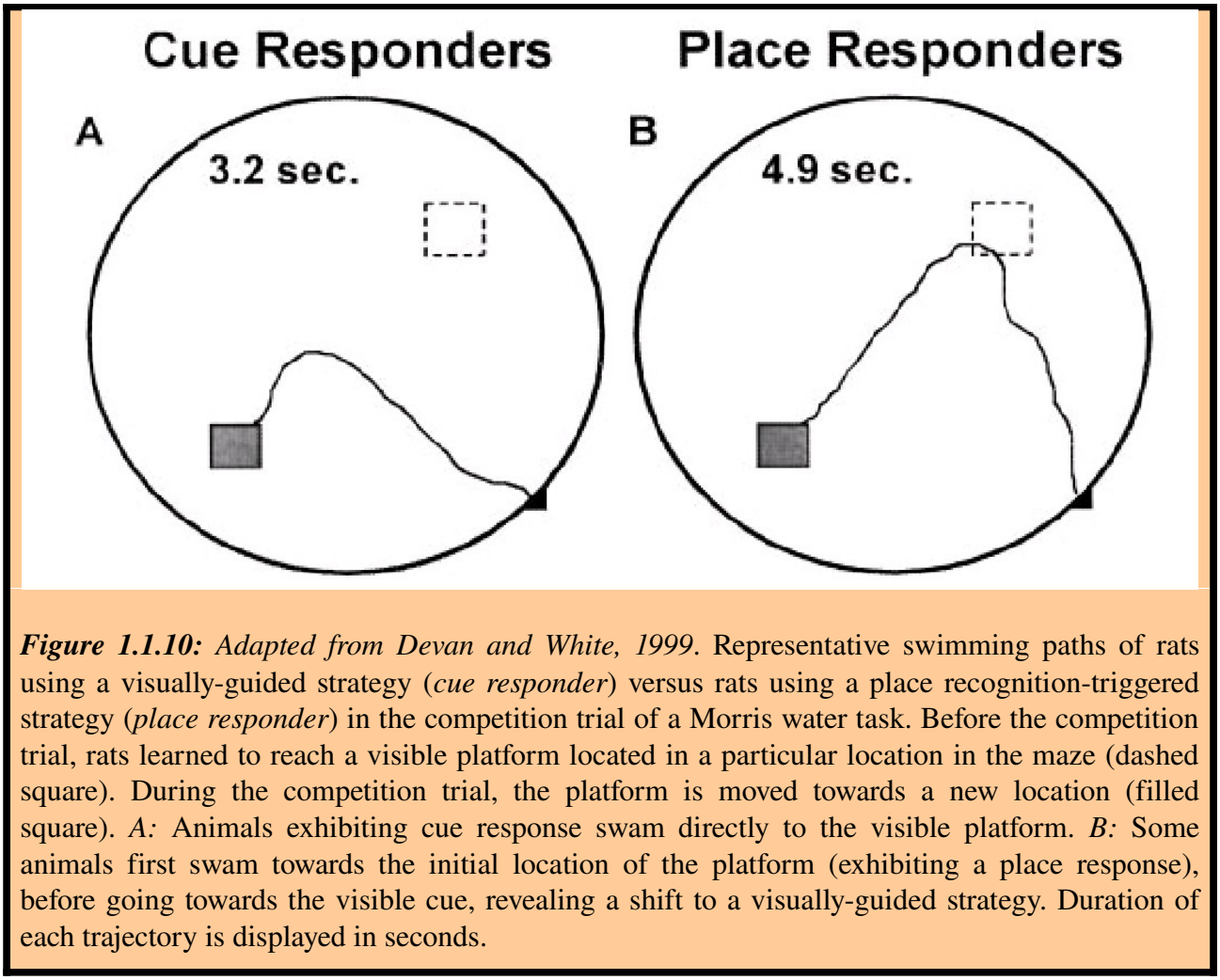


Figure 1.1.10: *Adapted from Devan and White, 1999*. Representative swimming paths of rats using a visually-guided strategy (*cue responder*) versus rats using a place recognition-triggered strategy (*place responder*) in the competition trial of a Morris water task. Before the competition trial, rats learned to reach a visible platform located in a particular location in the maze (dashed square). During the competition trial, the platform is moved towards a new location (filled square). *A:* Animals exhibiting cue response swam directly to the visible platform. *B:* Some animals first swam towards the initial location of the platform (exhibiting a place response), before going towards the visible cue, revealing a shift to a visually-guided strategy. Duration of each trajectory is displayed in seconds.

Several studies have reported a progressive shift from one strategy to another after extensive training in an unchanging task (Dickinson, 1980; Packard and McGaugh, 1996; Pearce et al., 1998; Chang and Gold, 2003). In most navigation tasks, the observed shift is from a *locale* (i.e. model-based place strategy) strategy to a model-free *visually-guided* or *praxic* strategy. For instance, Packard and McGaugh (1996) trained rats in the plus-maze task displayed on figure 1.1.2. During this training phase, rats started from the south arm and had to learn to go to the east arm (turn right to find a reward. After sixteen trials, the maze was rotated so that the animals now started from the north arm. Rats had to spontaneously make a single choice and predominantly chose to turn left, that is go to the same east location than during training. Then, the maze was rotated for a second time and sixteen more trials of training were given to the animals. After a final location, starting from the north arm, rats predominantly turned right, that is, they made the same body turn than during training. The generally accepted interpretation is a shift from *locale* (model-based) to *praxic* (model-free) under stable task conditions, rats having an initial preference for a spatial strategy (Gallup and Diamon, 1960).

The second condition considered (change in the task) can either be a change in the reward position – for instance if the experimenter translates the hidden platform towards a new location in the Morris water maze; or it can take the shape of a change in the landmark cues that signal the presence of the reward (for example if a green tree used to indicate the reward location while now, the leaves fall let the tree brown). It can also be the disappearance of a food source in a familiar area; or it can be the appearance of an obstacle across a familiar path. In these cases, an animal needs to shift its navigation strategy in order to further explore, or in order to build a behavior associated to another cue present near the reward location, or even so as to invoke its mental model

and plan a path that can replace the usual reactive model-free behavior that the rat was relying on.

Observations of rats ability to shift their navigation strategy in response to a change in the task have been previously described (Packard and McGaugh, 1996; Packard, 1999; McIntyre et al., 2003; Hamilton et al., 2004). Figure 1.1.9 displays an example of such a strategy shift observed in a Morris water maze task (Devan and White, 1999). In this task, rats first learned to reach a visible platform located in a particular location in the maze (dashed square). Then, a competition trial is imposed where the platform is moved towards a new location (filled square). Animals exhibiting a visually-guided strategy swam directly to the visible platform. Some animals first swam towards the initial location of the platform (exhibiting a place strategy), before going towards the visible cue, revealing a shift to a visually-guided strategy.

Moreover, several studies show an enhanced learning of one strategy in response to a task change produced by the inactivation of another strategy (McDonald and White, 1995; Matthews et al., 1999; Ferbinteanu and McDonald, 2001; Chang and Gold, 2003; Poldrack and Packard, 2003; Canal et al., 2005). This suggests that such kind of shifts can result from a competition between different brain pathways mediating alternative navigation strategies.

1.7.2 Two types of shifts in response to task change

We will distinguish here between **two types of shifts** – extradimensional (ED) and intradimensional (ID) shifts. Considering an initial condition where an animal has learned the association between a stimulus S1 and a behavioral response (S1-R), an ID shift refers to a shift to an association S2-R where S2 shares the same dimension than S1 – e.g. both are visual, or both are places in the environment, etc... – whereas an ED shift implies an association S2'-R where S2' has a different dimension than S1 (e.g. from cue-guided to place).

ID shifts were found to be easier to learn for rats than ED shifts (Trobalon et al., 2003; Block et al., 2007).

It is very important to note that here the envisaged decomposition of conditions and types of shift is simplified and limited. Indeed, there are other possible conditions for shifting, and there are many factors that are supposed to influence the recrutement of one strategy or another, such as physiological states, characteristics of experimental settings, training stage, individual preference, sex differences (d'Hooge and Dedeyn, 2001). However, the different cases of shifts considered here already provide a richness and variety of behaviors. Application to robotics of such a system of learning and shifting different navigation strategies could provide robots with interesting flexible behaviors and abilities to adapt to unexpected changes in the environment. There already exist such enterprises both in biomimetic robotics (Guazelli et al., 1998; Gaussier et al., 2000; Banquet et al., 2005; Chavarriaga et al., 2005b; Girard et al., 2005; Doya and Uchibe, 2005) and in classical robotics.

The next section, titled « Neural systems involved in learning and shifting among navigation strategies », will present the neurophysiological background.

2. Neural systems involved in learning and shifting among navigation strategies

Two of the principal brain structures examined here, the medial prefrontal cortex (mPFC) and the striatum, are globally considered to be involved in action selection and decision-making, including in the spatial domain (Pennartz et al., 1994; Fuster, 1997; Graybiel, 1998; Redgrave et al. 1999a; Granon and Poucet, 2000; Cardinal et al., 2002; Berthoz, 2003a; Wiener et al., 2003; Kesner and Rogers, 2004; Balleine et al., 2007; Samejima and Doya, 2007; Prescott and Humphries, 2007). On the one hand, the striatum, and the basal ganglia in general – a set of subcortical nuclei whose main entry point is the striatum – are considered to be globally involved in reactive, automatic and habitual action selection (Mink, 1996; Prescott et al., 1999; Redgrave et al., 1999a; see Greenberg, 2001 for a review), and in learning to adapt this action selection based on reward (Graybiel and Kimura, 1995; Houk et al., 1995; Wickens and Rötter, 1995; Kelley et al., 1997). On the other hand, the rat mPFC, having functional homologies with the primate dorsolateral PFC (Kolb, 1990; Uylings et al., 2003; Voorn et al., 2004; Vertes, 2006), is considered to have a role in high-level cognitive processes, usually referred to as executive functions, that is "complex cognitive processes required to perform flexible and voluntary goal-directed behaviors based on stored information in accordance with the context" (Granon and Poucet, 2000).

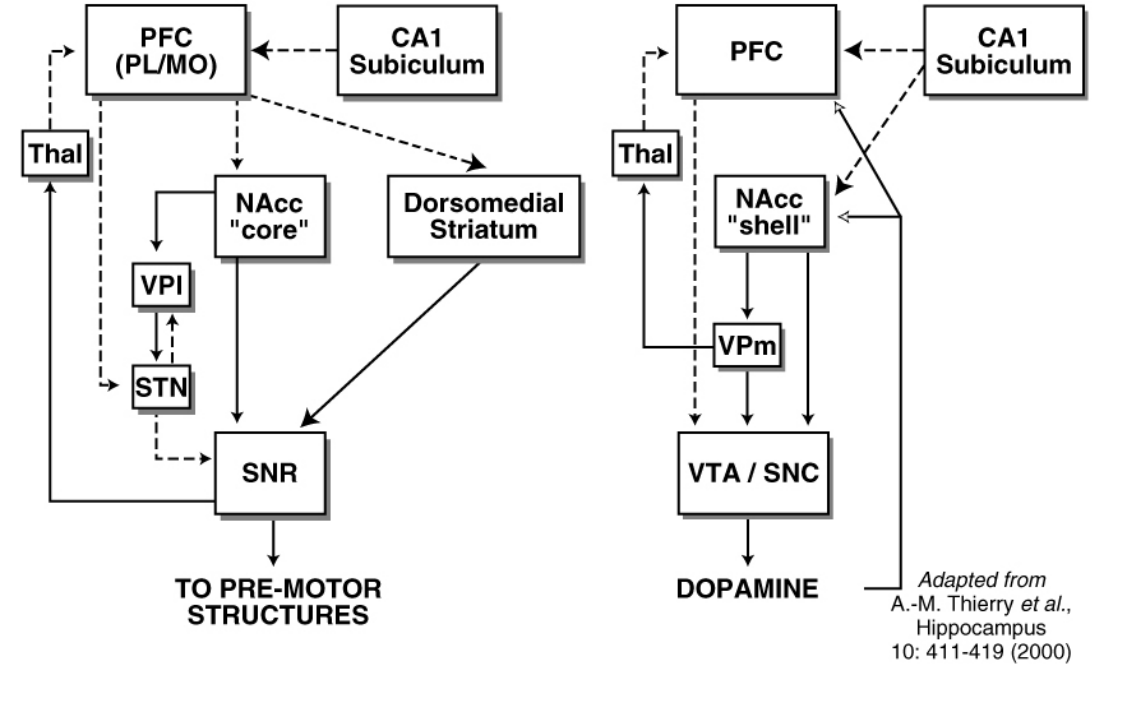


Figure 1.2.1: *Schematic representation of circuits mediating hippocampal spatial and other contextual information through the prefrontal cortex and striatum. Adapted from (Thierry et al., 2000).* Dashed lines represent excitatory projections. Solid lines with filled arrows represent inhibitory projections. Solid lines with empty arrows represent dopaminergic neuromodulations. CA1 – Hippocampus; Nacc – Nucleus Accumbens; Thal – Mediodorsal thalamic n.; PFC – Prefrontal Cortex; SN(C&R) – Substantia Nigra (pars compacta & reticulata); STN – Subthalamic nucleus; VP – Ventral pallidum (lateral & medial).

Both mPFC and the striatum are strongly interconnected with a system emitting a neuromodulator

called *dopamine* which can play a role in reward-based learning (Robbins and Everitt, 1992; Schultz et al., 1997; Berridge and Robinson, 1998), and thus can participate in the adaptation of action selection. However, the precise interaction and complementarity of mPFC and the striatum in learning and shifting particular navigation strategies is not yet precisely understood.

The prefrontal cortex and striatum are anatomically organized in parallel loops receiving different information (figure 1.2.3), and some of which are innervated by key structures in different types of navigation: the *hippocampal system*, as well as sensorimotor and parietal cortices (Thierry et al., 2000; Tierney et al., 2004). The hippocampus is of particular interest for understanding the neural basis of cognitive function since it is involved in the elaboration of abstract cue-invariant representations of the environment (Wiener, 1996). The hippocampal system is considered to play an important role in the elaboration of spatial information that are required for learning certain navigation strategies (O'Keefe and Nadel, 1978; Poucet and Hermann, 1990; Muller et al., 1999; Poucet et al., 2003); Indeed, the present studies are part of a long term research program (for review, see Wiener et al., 2003) of how hippocampal representations are exploited for behavior, and the striatum and prefrontal cortex were selected since they are of its principal output destinations that are in turn connected to premotor systems (figure 1.2.1) (Pennartz et al., 1994; Thierry et al., 2000; Battaglia et al., 2004b; Voorn et al., 2004).

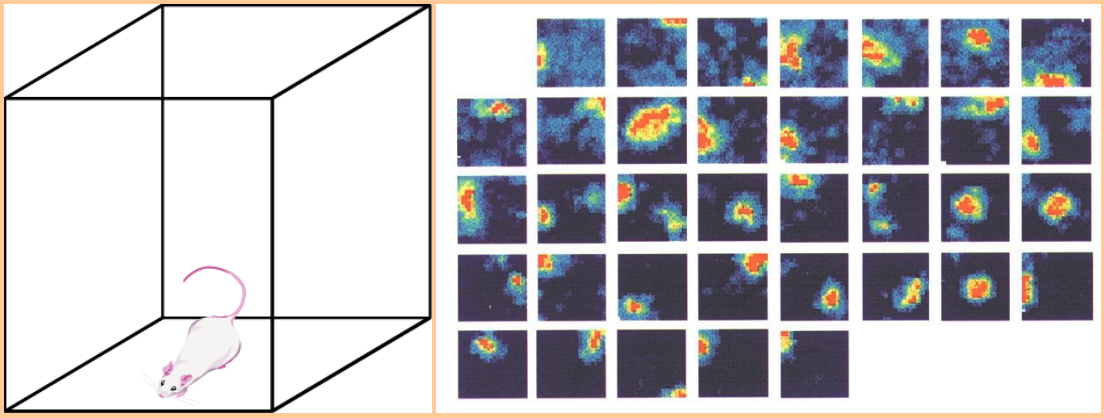
So this chapter will first briefly present the hippocampus, then examine the anatomical loops characterizing the prefronto-striatal system. We will present the basis for a theoretical framework wherein the striatum is involved in learning of several navigation strategies, and describe how dopamine signals can participate in these learning processes. Finally, we will see the foundations in the literature for the hypothesis that the prefrontal cortex could be involved in shifting among strategies.

2.1 The hippocampus and the elaboration of spatial information

A key finding in the rodent hippocampus is the so-called *place cells*. In a freely moving animal, the electrophysiological response of these hippocampal pyramidal neurons show a remarkable correlation to the location of the animal (O'Keefe and Dostrovsky, 1971; Battaglia et al., 2004a). Each of these neurons responds when the animal occupies a particular place in the environment, and at the level of the neuronal population, the entire surface of an experimental surface is represented, as shown in figure 1.2.2 adapted from Jung, Wiener and McNaughton (1994). These results led to the theory that the hippocampus participates in the storage of an allocentric spatial map for navigation (O'Keefe and Nadel, 1978), the *cognitive map* whose existence in the brain was postulated by Tolman (1948).

Moreover, lesions of the hippocampus impair learning of *locale* navigation strategies while sparing *taxon, praxic* and *guidance* strategies (Morris, 1981; Devan and White, 1999; Pearce et al., 1998). Translated into the terminology adopted in the previous chapter, this means that hippocampal lesions impair navigation strategies based on the *place* dimension, while sparing strategies based on single *visual* or *idiothetic* dimensions. This suggests that the hippocampus is crucial for the acquisition of place dimension strategies. However, it is generally admitted that the hippocampus does not participate in the control of such navigation strategies, but rather sends spatial information to other brain structures involved in decision-making, such as the prefrontal cortex and striatum (Pennartz et al., 1994; Devan and White, 1999; Thierry et al., 2000; Voorn et al., 2004).
Apparently inconsistent with this view are the findings of several correlates of hippocampal neurons with decision-making parameters. These include behavioral correlates (Wiener et al., 1989), movement correlates (Fukuda et al., 1997; Yeshenko et al., 2004), reward correlates (Dayawansa et al., 2006) and goal correlates (Hok et al., 2007). However, the consequence of these results on the interpretation of hippocampal function will not be discussed here. For such topic, we invite the

reader to refer to some review articles (Wiener, 1996; Mizumori et al., 2004; Poucet et al., 2004).
As a last point concerning the elaboration of spatial information in the hippocampal neural system, it is important to mention the existence of *head-direction (HD) cells* and *grid cells*. The former are neurons that we had the occasion to study as an initiation to electrophysiological techniques at the beginning of the PhD period (see Zugaro et al., 2004; Arleo et al., 2004 in appendix). Characteristically, the activity of these neurons reflects the animal's current head direction, independent of its position in the environment (Ranck, 1984). HD cells have a single *preferred direction* at which they fire maximally, and their firing rates decrease monotonically as the animals' orientation moves progressively farther away from the preferred direction. Because a cell's preferred direction does not change over the space of an environment (Taube et al., 1990a,1990b), the cell cannot be encoding egocentric bearing to a landmark; it must be encoding allocentric bearing to a *reference direction*. HD cells were found in a number of structures tightly interconnected with the hippocampal system, such as the postsubiculum (Ranck, 1984), the anterodorsal thalamic nucleus (Blair and Sharp, 1995; Knierim et al., 1995; Taube, 1995; Zugaro et al., 2001), entorhinal cortex and even a small population in the hippocampus itself (Leutgeb et al., 2000). HD cells are required for hippocampal place cell stability (Calton et al., 2003; Degris et al., 2004) and thus could participate in navigation strategies requiring an allocentric orientation process.



(a) (b)

Figure 1.2.2: *Coding of an ensemble of locations in a given environment by a population of hippocampal places cells. Adapted from (Jung et al., 1994).* (a) The rat was introduced into a cubic arena. (b) Electrophysiological response of a set of hippocampal neurons as a function of place. The squares correspond to responses of individual dorsal hippocampal neurons in overhead views of the arena. Dark blue zones correspond to no activity, while the neurons fired maximally in red zones (generally about 1-40 impulses per second depending upon the neuron).

A second recently discovered neural substrate for spatial navigation are the *grid cells* in a part of the hippocampal system named the *entorhinal cortex* (Fyhn et al., 2004; Hafting et al., 2005; Sargolini et al., 2006). These cells are active when the rat occupies a set of regularly spaced places, tesselating the environment in a hexagonal pattern. This activity can be interpreted as a basis for a self-motion or path integration based map of the spatial environment (McNaughton et al., 2005), and are likely to be essential for the elaboration of the hippocampal spatial responses (Hafting et al., 2005). Consistent with this hypothesis are behavioral results showing that lesions of the entorhinal cortex impair spatial navigation based on distal cues in a Morris water maze (Parron et al., 2004), distal cues being crucial for the control of place cells activity (Cressant et al., 1997), and thus for navigation based on a cognitive map (Pearce et al., 1998).

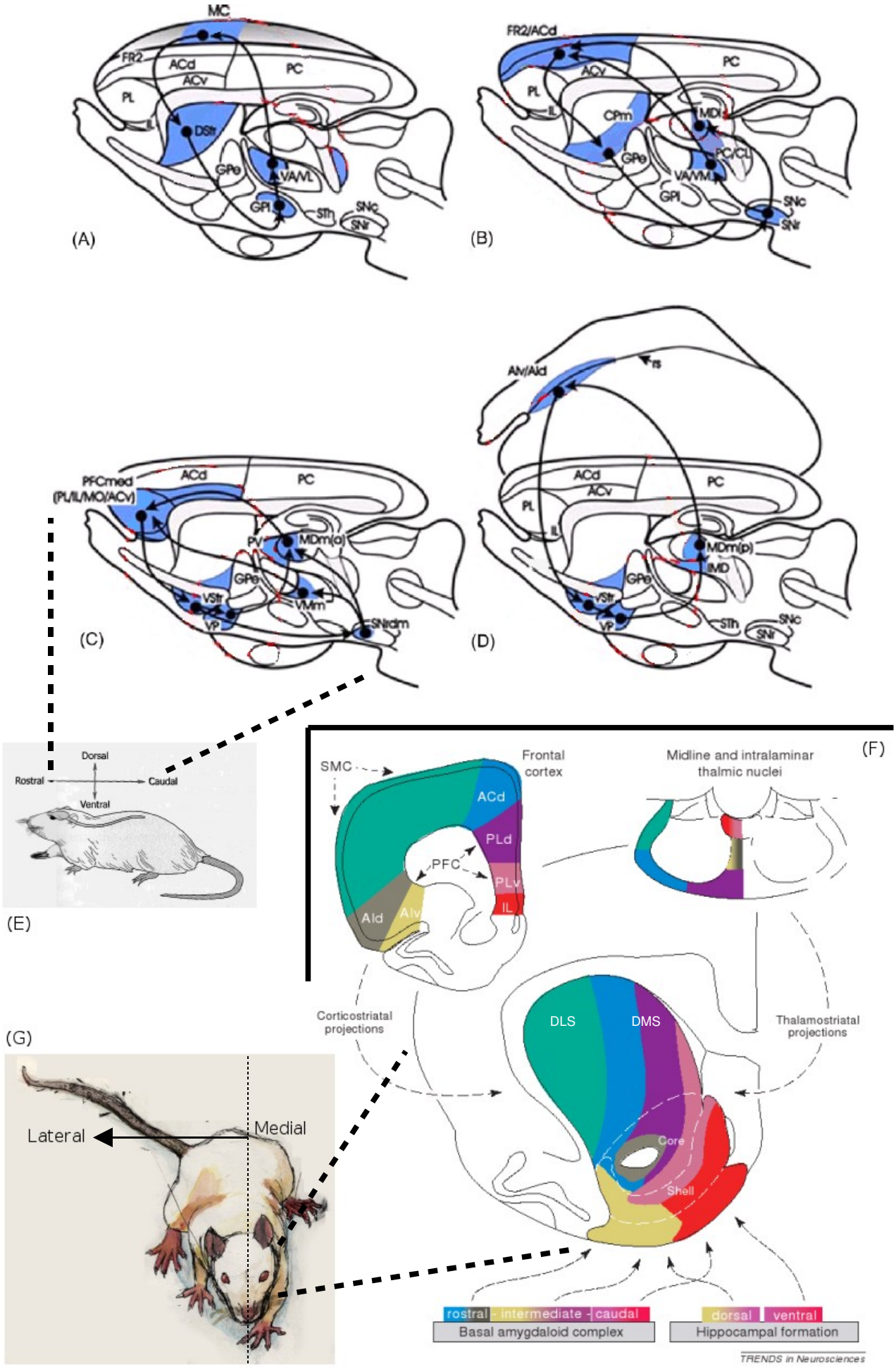

MC
FR2
ACd
ACV
PC
PL
IL
DStr
GPe
VA/VL
GPi
STh
SNc
SNr
(A)
FR2/ACd
ACV
PC
PL
IL
CPm
MDl
GPe
PC/CL
VA/VM/L
GPi
SNc
SNr
(B)
PFCmed
(PL/IL/MO/ACv)
ACd
PC
PV
MDm(o)
GPe
VStr
VMm
VP
SNrdm
(C)
AIv/AId
rs
ACd
ACV
PC
PL
IL
MDm(p)
IMD
GPe
VStr
VP
STh
SNc
SNr
(D)
Dorsal
Rostral
Caudal
Ventral
(E)
(F)
SMC
Frontal
cortex
ACd
PLd
PFC
PLv
IL
AId
AIv
Midline and intralaminar
thalmic nuclei
Corticostriatal
projections
DLS
DMS
Thalamostriatal
projections
Core
Shell
rostral - intermediate - caudal
Basal amygdaloid complex
dorsal
ventral
Hippocampal formation
TRENDS in Neurosciences
(G)
Lateral
Medial

Figure 1.2.3 (previous page): *Prefronto-striatal loops presented in two different schemas.* A motor loop, B associative loop, C loop limbic1 core, D loop limbic2 shell, adapted from Uylings et al. (2003); E reference for orientation in A-D; F parallel cortico-striatal loops (with midline nuclei) in the rat are indicated by color code adapted from Voorn et al. (2004); G reference for orientation in F, adapted from Popolon (2007). List of abbreviations ... ACd, dorsal anterior cingulate area; ACv, ventral anterior cingulate area; AId, dorsal agranular insular area; AIv, ventral agranular insular area; DLS, dorsolateral striatum; DMS, dorsomedial striatum; DStr, dorsal striatum; FR2, frontal cortical area 2; GPe, globus pallidus, external segment; GPi, globus pallidus, internal segment; IL, infralimbic cortical area; IMD, intermediodorsal thalamic nucleus; MC, motor cortex; MDl, mediodorsal thalamic nucleus, lateral segment; MDm, mediodorsal thalamic nucleus, medial segment; MDm(a), anterior part of MDm; MDm(p), posterior part of MDm; MO, medial orbital cortical area; PC, paracentral thalamic nucleus; PFC, prefrontal cortex; PFCmed, medial prefrontal cortex; PL, prelimbic cortical area; PLd, dorsal PL; PLv, ventral PL; rs, rhinal sulcus; SMC, sensorimotor cortex; SNc, substantia nigra pars compacta; SNr, substantia nigra reticulata; SNrdm, dorsomedial part of SNr; STh, subthalamic nucleus; VA, ventral anterior thalamic nucleus; VL, ventral lateral thalamic nucleus; VM, ventral medial thalamic nucleus; VMm, medial part of VM; VP, ventral pallidum; VStr, ventral striatum.

2.2 Prefronto-striatal anatomical loops

In mammals, the frontal cortex and striatum are anatomically organized in parallel loops engaging different cognitive functions such as motor, associative, limbic and oculomotor (Alexander and Crutcher, 1990; Alexander et al., 1990). Within these loops, cortical information enter the basal ganglia via the striatum. Information processed within the basal ganglia by a disinhibitory process is then sent back to cortical areas in the frontal lobe – which include prefrontal, premotor and motor cortices – through the mediodorsal thalamic nucleus (Chevalier and Deniau, 1990).
In the rat, four principal loops can be distinguished, which correspond to different territories of the striatum as shown on figure 1.2.3 (Uylings et al., 2003). To broadly summarize, all of the neuron groups in a given territory can be characterized by the regions they receive input from:
A) In the motor loop, the dorsolateral striatum (DLS) is related to the sensorimotor cortex;
B) In the associative loop, the dorsomedial striatum (DMS) is linked to the dorsomedial prefrontal cortex – including the prelimbic area (PL) – and the premotor cortex;
C) In loop limbic 1, the accumbens "core" – belonging to the ventral striatum – is related to the dorsomedial prefrontal cortex – including PL – and the amygdala;
D) In loop limbic 2, the accumbens "shell" – belonging to the ventral striatum – is connected with the hippocampus, the amygdala, the ventromedial prefrontal cortex – including PL and IL (infralimbic area) –, the orbitofrontal cortex and the agranular insular cortex.

2.2.1 Feature 1: similar anatomical organization between loops

A first important feature of these loops is the similar anatomical organization from one loop to the other – for this reason they are referred to as parallel. Indeed, as shown in figures 1.2.3F and 1.2.4, within each loop, a given cortical subterritory projects to an associated striatal territory, which in turn sends projections through a series of basal ganglia nuclei that are similarly organized in all loops (Mink, 1996; Wickens, 1997; Maurice et al., 1997, 1999); these nuclei will not be described here. Then from an output nucleus of the basal ganglia – either the Substantia Nigra Reticulata (SNr) or the Entopeduncular nucleus (EP) –, projections are sent back to the Cortex via the mediodorsal thalamic nucleus (Deniau et al., 1994).
The loops are similarly characterized by a set of patterns that can be roughly enumerated as:
1) the existence of a *direct/indirect* dissociation between pathways through the basal ganglia (Albin

et al., 1989);
2) the existence of a dissociation between striatal neurons' *D1/D2* receptors (Parent and Hazrati, 1995a,b), which makes them differently sensible to the neuromodulator *dopamine*;
3) the existence, in each loop, of subdivisions of striatal territories into the *Striosomes* which project towards dopaminergic neurons and thus have an influence on dopamine release, and *Matrisomes* which do not (Gerfen, 1984,1992; Gerfen and Wilson, 1996; Desban and Kemel, 1993; Graybiel and Kimura, 1995). However, the shell differs from this, as explained in paragraph 2.2.2
As a consequence of this first feature, **computationally similar signal processing can be subserved by each of these loops** (Houk et al., 1995), which is also true in the monkey (Hikosaka et al., 1999).

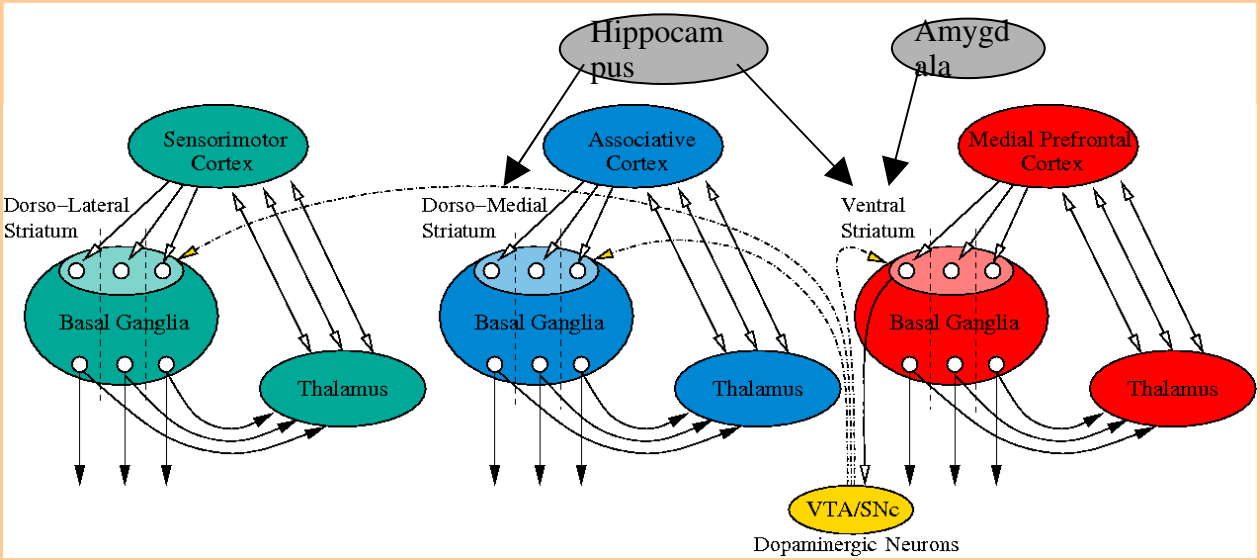


Figure 1.2.4: *Prefronto-striatal loops*. This scheme represent three major cortico-basal ganglia loops, involving different striatal territories: respectively the dorso-lateral striatum, the dorso-medial striatum and the ventral striatum. Solid lines with filled arrows represent inhibitory connections. Dashed lines represent dopaminergic neuromodulation. SNc, substantia nigra pars compacta; VTA, ventral tegmental area.

2.2.2 Feature 2: interaction between loops through the dopaminergic system

The loop associated with the *shell* zone of the ventral striatum differs substantially from the above mentioned stereotyped parallel organization. Indeed, the shell, and particularly the medial shell where we have recorded neurons (see chapter 2), is endowed with some unique anatomical and neurophysiological characteristics which distinguish it functionally (Voorn et al., 2004). The principal difference to be evoked here is a stronger influence on the dopaminergic system than is exerted by *striosomes* that are located in other striatal regions (Groenewegen et al., 1996; Joel and Weiner, 2000; Thierry et al., 2000; J.-M. Deniau et al., unpublished but cited in Voorn et al., 2004). This puts the shell in control of dopaminergic input to other loops, evoking the *spiral* dopaminergic modulation of the cortico-striatal loops reported in primates (Haber et al., 2000). The spiral refers to the crossover in the parallel loops for both MD projections to cortex and for striatal projections to the dopaminergic VTA and SNc wherein there is overlapping from the limbic loop to associational zones and from the associational loop to sensorimotor zones. As we will see later, **this puts the shell at the top of a hierarchy where it can potentially modulate learning processes in the other loops**.

2.2.3 Feature 3: diverse input for each loop

A third important feature is the difference in the input information received by each loop. It is informative to contrast the dorsolateral striatum (motor loop) which primarily receives sensorimotor

inputs (McGeorge and Faull, 1989), the dorsomedial striatum (associative loop), which also receives hippocampal system inputs (Groenewegen et al., 1987; McGeorge and Faull, 1989), and has allocentric spatial responses in the form of head direction cells (Wiener, 1993), and their counterpart the nucleus accumbens (limbic loops) which receives hippocampal, prefrontal, amygdalar and entorhinal inputs. The latter permit access to signals concerning place, motivation, reward signals, head direction and path integration information (Pennartz et al., 1994; Groenewegen et al., 1996; Pennartz et al., 2000).

We can already see that based on these anatomical data, **different loops might be engaged in learning of action selection based on different stimulus information, while the loop associated to the shell would exert an overall dopaminergic influence on other loops, and while striosomes within each loop could participate in the modulation of dopamine release**.

In the next section, we review lesion and electrophysiological results which relate each loop with a particular navigation strategy.

2.3 Different striatal regions involved in different navigation strategies

2.3.1 Lesion studies

Initially, the striatum was considered as globally involved in egocentric navigation strategies, in contrast to the hippocampus which was assumed to participate in allocentric strategies. Indeed, whereas lesions of the *hippocampus* impaired *locale* strategies (Morris, 1981; Devan and White, 1999), lesions of the striatum were found to impair both *praxic* (Potegal, 1969; Cook and Kesner, 1988; Colombo et al., 1989; Kesner et al., 1993) and *taxon* strategies while sparing *locale* strategies (Whishaw and Mittleman, 1986; Packard et al., 1989; Brasted et al., 1997; DeCoteau and Kesner, 2000; Adams et al., 2001; Ragozzino and Kesner, 2001; Packard and Knowlton, 2002).

However, some recent studies of lesions restricted to striatal territories corresponding to a single loop reveal their specific roles in particular navigation strategies. In the variation of the Morris water maze task presented on Figure 1.1.10 (Devan and White, 1999), after learning to reach a visible platform at a particular position, rats were exposed to a competition trial were the platform was visible but moved. On the one hand, rats with DLS lesions moved towards the uncued first location, thus expressing a *spatial* strategy. On the other hand, rats with DMS lesions preferred the visible platform at the new location, thus expressing a *cue-based* strategy. Devan and White (1999) interpret these results as revealing an involvement of DMS in place learning. These results are consistent with the anatomical organization reported in the previous section that DMS had afferents from the hippocampal system (Groenewegen et al., 1987; McGeorge and Faull, 1989).

In line with this DMS/DLS dichotomy, in a lever-press task, DLS lesions impair procedural S-R learning based on a visual stimulus (Yin et al., 2004), whereas lesions of DMS do not affect rats' performance in a T-maze task requiring a praxic strategy, but rather alters choice behavior based on the flexible use of place cues (Yin and Knowlton, 2004). Furthermore, lesion of DMS affect flexible place reversal learning – change in the place associated with reward – in a plus-maze (Ragozzino and Choi, 2003).

This suggests that DLS can be involved in cue-based and praxic navigation strategies, whereas DMS can subserve place strategies.

DMS also appears to be involved in goal-directed behaviors, since lesions of the posterior part of DMS impair learning and expression of the contingency between instrumental actions and their outcomes (Yin et al., 2005a,b), which, as we have seen in the first chapter, is one of the necessary memory components for model-based strategies.

This suggests that DMS could also participate in the acquisition and expression of goal-directed behaviors (Balleine, 2005), thus playing a role in model-based strategies.

The precise role of the ventral striatum (VS) in particular navigation strategies is less clear. Lesions of VS impair spatial learning, thus suggesting its involvement in place strategies (Sutherland and

Rodriguez, 1989; Ploeger et al., 1994; Setlow and McGaugh, 1998; Albertin et al., 2000). For instance, lesions of the rat accumbens medial shell – corresponding to the region we have recorded – impair the rat in learning and recalling sites providing larger rewards (Albertin et al., 2000), which conveys an alteration of the reward-place associations.
More recent studies even reveal that **VS function is not restricted to strategies based on the place dimension** but can also participate in others. For instance, DeLeonibus et al. (2005) report that lesions of VS impair the acquisition of both allocentric and egocentric strategies in a task requiring the detection of a spatial change in the configuration of four objects placed in an arena.

Furthermore, it appears that different subdivisions of VS may subserve different behavioral functions and thus can be considered separately. In this manuscript, we will distinguish the accumbens « core » and accumbens « shell » (Zahm and Brog, 1992). Shell lesions and pharmacological manipulations within the shell impair various forms of instrumental conditioning (Corbit et al., 2001; Fenu et al., 2001; Phillips et al., 2003), thus suggesting a role of the shell in reward-based learning of S-R associations.
Moreover, the shell appears not to be required for knowledge of the contingency between instrumental actions and their outcomes (Balleine and Killcross, 1994; Dickinson and Balleine, 1994; Corbit et al., 2001; see Cardinal et al., 2002 for a review), which, as we have seen in the first chapter, is one of the necessary memory components for model-based strategies.
However, it should be clear that the core/shell segregation of VS is oversimplified, since certain results suggest a finer subdivision (Heimer et al., 1997; Ikemoto, 2002), and other results reveal overlapping behavioral functions, thus stressing a continuum between core and shell (see Voorn et al., 2004 for a review).
So, following the terminology that we have adopted in the section concerning navigation strategies, it seems that **the shell could possibly be important for learning model-free strategies** in any reference frame (egocentric or allocentric), and thus for any stimulus type (place, simple allothetic or idiothetic), whereas storage and expression of these model-free strategies would require motor and associative loops.

In contrast, accumbens core lesions do not impair conditioned reinforcement (Parkinson et al., 1999; Hall et al., 2001; see Cardinal et al., 2002 for a review), but rather impair the animal's sensitivity to outcome devaluation (Corbit et al., 2001), and the acquisition of action-outcome contingencies (Kelley et al., 1997). **Thus, the core could be assumed not to be involved in learning of model-free strategies, but rather could be important in goal-directed behaviors (Dayan and Balleine, 2002), thus playing a role in model-based strategies**.

However, these hypotheses are simplified, and a rigorous investigation of the functional roles of the core and shell should data from the fields of classifical conditioning (see Cardinal et al., 2002), drug addiction (Robbins and Everitt, 1992; Berridge and Robinson, 1998), and lesion studies concerning their role in unlearned behaviors (Kelley, 1999) and motivational processes (Kelley et al., 1997; Aberman and Salamone, 1999; Cardinal et al., 2001).

2.3.2 Electrophysiological studies

Consistent with lesion studies, electrophysiological recordings show that each of the parameters required for storage and learning of the respective navigation strategies (i.e. stimuli, behaviors, space, rewards) are encoded in zones of the rodent striatum.
Cues and movements correlates: More precisely, in the dorsal striatum (without distinction of DMS/DLS subdivisions), neurons were found which respond to movements, turning movements, grooming movements, head direction, auditory cues, visual cues and olfactory cues (Gardiner and Kitai, 1992; Callaway and Henriksen, 1992; Wiener, 1993; Lavoie and Mizumori, 1994; Carelli et

al., 1997; Aldridge and Berridge, 1998; Jog et al., 1999; Ragozzino et al., 2001; Daw et al., 2002; Setlow et al., 2003; Nicola et al., 2004; Yeshenko et al., 2004; Barnes et al., 2005; Wilson and Bowman, 2005). For example, among the results of Setlow et al. (2003), neurons were found to encode specific combinations of cues and associated motor responses in a « go / no-go » olfactory discrimination learning and reversal task. In this task, rats had first to learn to associate an odor with a positive reward (sucrose) and another one with an aversive gustatory stimulus (quinine). Then rats were exposed to reversal learning where the odor-outcome contingencies were changed.

Several studies have reported the specificity of ventral striatal responses to cues and movements, showing that neurons that are responsive during a task would not be responsive outside the task (Gardiner and Kitai, 1992; Carelli et al., 1997; Aldridge and Berridge, 1998). **These results suggest that the dorsal striatum can store part of learned S-R associations.**

Similar encoding of stimulus and movement information are found in the monkey striatum (Rolls et al., 1983; Kimura, 1986,1990,1995; Kermadi et al., 1993; Kermadi and Joseph, 1995; Miyachi et al., 1997; Hikosaka et al., 1998; Kawagoe et al., 1998; Shidara et al., 1998; Shidara and Richmond, 2004; Ravel et al., 1999,2003; Hikosaka et al., 2000; Lauwereyns et al., 2002a,b; Itoh et al., 2003; Watanabe et al., 2007). Several studies have also reported spatial correlates in monkey ventral striatal and caudate (equivalent to rat DMS) neurons (Hassani et al., 2001; Takikawa et al., 2002; Cromwell and Schultz, 2003; Ravel et al., 2006). However, the latter spatial aspect is not strictly comparable to place encoding in the rat, since it corresponds to selectivity to areas on a screen displaying stimuli.

Spatial correlates: Interestingly, in rodents, neurons with spatial correlates were found both in DMS and VS (Wiener, 1993; Lavoie and Mizumori, 1994; Martin and Ono, 2000; Shibata et al., 2001; Chang et al., 2002; Mulder et al., 2004; Schmitzer-Torbert and Redish, 2004; Yeshenko et al., 2004). To the best of our knowledge, none or few neurons from DLS are selective to spatial positions. Synchronized ensemble activity between hippocampus and VS (including core and shell) during behavior (Tabuchi et al., 2000) and during memory consolidation phase in sleep were also reported (Pennartz et al., 2004). This supports the hypothesis that the hippocampus and ventral striatum interact with each other in relation with learning spatial tasks.

These findings of spatial modulation in DMS and VS neurons activity suggest that the limbic and associative loops can store place navigation strategies.

Learning correlates: Change in striatal neurons activity during learning were also reported in rodents (Graybiel, 1995; Jog et al., 1999; Setlow et al., 2003; Barnes et al., 2005), as well as in the monkey (Aosaki et al., 1994a,b; Tremblay et al., 1998; Pasupathy and Miller, 2005). **These results confirm that the striatum can be instrumental for learning of navigation strategies.**

Reward correlates: Finally, reward information is signalled in ventral striatal neuronal activity. VS neurons respond to reinforcements including food, drink, drugs of abuse, and intracranial electrical stimulations (in the rat: Lavoie and Mizumori, 1994; Bowman et al., 1996; Carelli and Deadwyler, 1997; Chang et al., 1997; Martin and Ono, 2000; Shibata et al., 2001; Carelli, 2002; Janak et al., 2004; Nicola et al., 2004; Wilson and Bowman, 2005; in the monkey: Hikosaka, 1989; Apicella et al., 1991a,b, 1992, 1998; Schultz et al., 1992; Hollerman et al., 2000; Cromwell and Schultz, 2003).

Several studies report reward expectations responses in monkey VS (Hollerman et al., 1998; Hassani et al., 2001; Cromwell and Schultz, 2003; Kawagoe et al., 1998,2003) and combinations of reward and action information in monkey caudate nucleus (Samejima et al., 2005). In the rat, strict reward expectations are less clearly discriminated, mainly because in freely-moving rats, many experimental designs fail to dissociate reward information from other behavioral components. Nicola et al. (2004) show VS neurons encoding the motivational significance of stimuli predicting rewards. VS neurons also encode predictive information concerning the type of reward (food vs. drink) that the animal receives (Miyazaki et al., 1998,2004; Daw et al., 2002) or concerning the reward value (positive or aversive) (Setlow et al., 2003). However, none of these studies report purely behavior-independent reward expectation, distinguishing between behaviors leading up to

rewards, and the rewards proper. The aim of the experiment presented in chapter 2 is to clarify this ambiguity and discriminate reward anticipations from other behavioral parameters. **In summary, these reward correlates suggest that VS can subserve reward-based learning of navigation strategies.**

Overall, these data are consistent with a role of the striatum in learning model-free navigation strategies. However, it should also be noted that a few studies cite the striatum's involvement in shifts in task rules. A set of striatal neurons change their activity in response to a change in the task rule – from place to visual, place to praxic, and vice versa (Shibata et al., 2001; Eschenko and Mizumori, 2007). Shifts from a praxic task to a visual task are impaired by lesions of either the accumbens core (Floresco et al., 2006) or DMS (Ragozzino et al., 2002). Lesions of DMS also impair reversal learning (Shirakawa and Ichitani, 2004). However, the medial prefrontal (presented in section 2.5) rather than the striatum (which receives prefrontal inputs) is generally considered as playing a key role in strategy selection. Above, we alluded to a hypothetical mechanism subserving reward-based learning of navigation strategies within the striatum. This mechanism is generally considered to rely on dopamine. The particular patterns of dopamine release have strong computational consequences for the models elaborated in this PhD thesis (chapter 2).

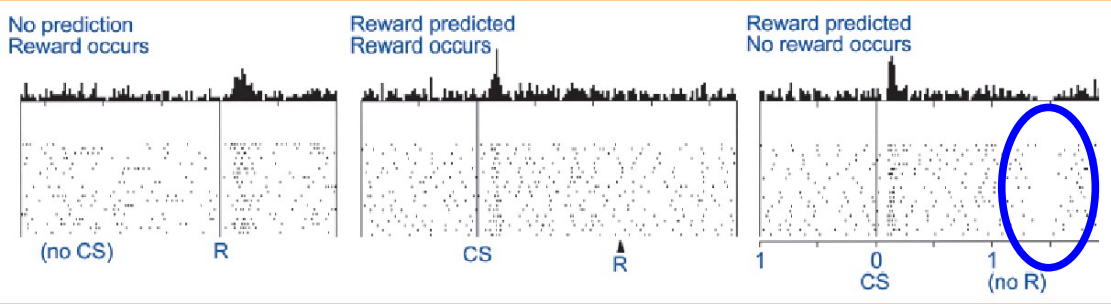


Figure 1.2.6 : *Monkey dopaminergic neurons activity during three task conditions. Adapted from (Schultz, 2001).* Black dots represent action potentials of measured neurons. These action potentials are plotted across successive trials (y-axis) and synchronised with the occurrence of certain task events – CS: Conditioned stimulus (a tone), R: Reward. Top histograms cumulated action potentials across trials.

2.4 Dopamine mediated reward-based learning of navigation strategies in the striatum

A possible mechanism underlying learning of navigation strategies within the striatum could be the reinforcement of stimulus-response associations that lead to reward, i.e. instrumental conditioning. In the framework where different striatal territories store S-R associations based on different types of stimuli and responses, such a learning mechanism would require the appropriate release, in these respective territories, of reinforcement signals depending on the behavioral context and occurrence or absence of rewards.

Indeed, such reinforcement could involve dopaminergic (DAergic) signals (Robbins and Everitt, 1992; Schultz et al., 1997; Berridge and Robinson, 1998; Satoh et al., 2003; Nakahara et al., 2004; Cheng and Feenstra, 2006). Dopamine (DA) is a neuromodulator emitted by a set of *dopaminergic neurons.* Of particular interest here are two DAergic brainstem nuclei: the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). Schultz and colleagues (1992, 1995, 1998) performed electrophysiological recordings of DAergic neurons during a task where monkeys learned

to respond to a stimulus (such as a tone) in order to earn a juice reward. They found a set of dopaminergic neurons which respond to unexpected rewards, i.e. prior to learning the Stimulus-Reward association (figure 1.2.6). This activity vanishes as the reward becomes predictable, roughly tracking improved performance (Mirenowicz and Schultz, 1994; Hollerman and Schultz, 1998; Fiorillo et al., 2003). Meanwhile, the activity of the same dopaminergic neurons gradually begins to respond to stimuli predictive of reward – the latter becoming a *conditioned stimuli (CS)*. Finally, when a reward predicted by a CS fails to arrive, a number of DA neurons exhibit a momentary pause in their background firing, timed to the moment the reward was expected. These findings support the idea that **DAergic neurons signal errors in reward prediction, these signals being crucial for reward-based learning** (Houk et al., 1995; Schultz et al., 1997). These responses are summarized in figure 1.2.6.
VTA and SNc are known to send projections to the prefrontal cortex and the striatum (Joel and Wiener, 2000; Thierry et al., 2000). Long-term modifications – in the form of Long Term Potentiation (LTP) or Long Term Depression (LTD) – have been observed at corticostriatal synapses after exposure to dopamine (Centonze et al., 2001; Reynolds et al., 2001). This supports the hypothesis that these signals are implicated in learning processes of S-R associations taking place in the striatum (Houk et al., 1995). Moreover, whereas all territories receive DAergic inputs, the accumbens shell's singular status as a major source of afferences to VTA/SNc (Joel and Wiener, 2000; Thierry et al., 2000) makes it a good candidate for influencing dopamine release within other striatal territories, and thus for driving dopamine-based reinforcement learning in the striatum (Dayan, 2001; Daw et al., 2006; Deniau et al., 2007). Consistently, the incidence of reward-responsive cells is greater in the accumbens than in the dorsal striatum (Apicella et al., 1991a; Schultz et al., 1992; Williams et al., 1993; Lavoie and Mizumori, 1994; Carelli and Deadwyler, 1994; see Pennartz et al., 2000 for a discussion of this point). Finally, the ratio of DA concentrations in monkey striatum / amygdala / premotor cortex / hippocampus was estimated to be 411/9.4/2.7/1 (Brown et al., 1979; see Pennartz, 1996 for a discussion of this point), supporting the view that the striatum is a main targets of DAergic reinforcement signals.

However, the theory of dopamine as a prediction error signal is criticized by some authors, and several points challenging this theory can be listed:
1) **Latency of DA responses.** Redgrave et al. (1999b) observe that DA neurons respond to a visual event well before a visual saccade to it, and therefore, identification of the reward-predicting properties of the stimulus or assessment of reward itself. A visual saccade has a latency of 180-200 ms or 80-110 ms for express saccades (Moschovakis et al., 1996) whereas the latency of dopaminergic neurons' responses reported by Schultz and colleagues is around 70-100 ms in overtrained animals.
Redgrave et al. (1999b; Redgrave and Gurney, 2006) suggest instead that dopamine signals are elicited by projections from the superior colliculus (SC) – whose functions include orientation of ocular saccades towards stimuli capturing the animal's attention. According to Redgrave and Gurney (2006), SC is the most likely source of visual input to DA neurons (Coizet et al., 2003; Comoli et al., 2003; Overton et al., 2005). In contrast, a study in monkey states that early visual responses in the striatum occur about the same time or after phasic DA signalling (Hikosaka et al., 1989). Thus, Redgrave et al. (1999b) propose the alternative hypothesis that short latency DA, triggered by SC, signals salient events that cause a shift in animals' behavior. In line with this view is Horvitz (2000)'s attentional hypothesis of dopamine. Moreover, several studies report DAergic excitation to novel neutral stimuli (Ljungberg et al., 1992; Horvitz et al., 1997).
2) **Influence on synaptic plasticity.** Several studies appear to contradict the possibility that DA can exert an influence on corticostriatal synaptic plasticity (see Pennartz, 1996 for a review). For instance, Pennartz et al. (1993) report an absence of DAergic modulation on LTP for the prefrontal-ventral striatal loop in vitro, as indicated by a lack of effect of both DA and of a mixture of D1 and

D2 antagonists in intra- and extra-cellular recordings.
3) **DA interference with pre- and post-synaptic activity.** Finally, DA release in the striatum exerts some immediate effects on signal transmission which are not expected by the reward prediction error theory, and which could interfere with expected long term learning effects (see Pennartz, 1996 for a review). Instead, the immediate DA effect on the striatum can be interpreted as an influence on the control of action selection and movement initiation (Gurney et al., 2001a,b; Humphries, 2003).
These discrepancies lead some authors to propose an alternative hypothesis of reinforcement learning in the cortico-striatal loops, which does not involve dopamine but rather relies on a glutamatergic signal (Pennartz, 1997; Pennartz et al., 2000).
In parallel, other work has been undertaken to examine possible resolutions to these discrepancies.
For instance, Kakade and Dayan (2002) show that models of DA's role in reinforcement learning based on reward prediction errors can account for DAergic neurons' responses to novelty. Moreover, these models envision positive errors to cues predicting reward only probabilistically (see Daw, 2003 for a discussion of this issue), which would explain the possibility of short latency prediction error signals in response to yet unidentified task-related stimuli.
In addition, certain components of dopamine responses relative to uncertainty-based attentional processes (Fiorillo et al., 2003) can be modeled as emerging features of DA's involvement in reward prediction error (Niv et al., 2005).
A final intriguing element is the existence of several different DA signals within the striatum. Grace (1991) distinguishes a *tonic* dopamine signal – persistent and long-lasting –, from a *phasic* DA signal – i.e. transient. These two signals may have different effects on corticostriatal plasticity and on corticostriatal neurotransmission, with different roles in different loops, and thus could subserve different functions, leaving the field free for several theories of the functional role of DA. For instance, Wickens et al. (2007a,b) propose that these interloop variations can be understood in terms of the temporal structure of activity in the inputs sent to different striatal territories, and the requirements of different learning operations. In this perspective, they suggest that DLS may be subject to “brief, precisely timed pulsed of dopamine [corresponding to] reinforcement of habits”, whereas ventromedial striatal regions integrate “dopamine signals over a longer time course [corresponding to] incentive processes that are sensitive to the value of expected rewards”. Some recent models have captured the differential effects of these different DA signals on reinforcement learning processes and modulation of action selection (McClure et al., 2003; Niv et al., 2007), whereas another model proposes an integrative theory of tonic and phasic dopamine signals' effects on decision making (Dreher and Burnod, 2002).
Interestingly, the interzone dissociation of DA effects finds some echo in the drug addiction literature. Di Chiara (2002) reviews differences in effects between the shell and core regions of the ventral striatum. He notes that repetitive, non-decremental stimulation of DA transmission by drugs in the shell abnormally strengthens stimulus-drug associations, while stimulation of DA transmission in the core appears to have an effect on instrumental performance.
In conclusion, the precise relation between DA release and reward prediction error is still unresolved. Further investigations will be necessary to determine whether short latency DA signals facilitate reinforcement learning, whether DA's influence on synaptic plasticity is consistent with learning related processes, and whether behavioral effects of DA manipulations are consistent with the reward prediction error theory.

2.5 Summary of the hypothetized roles of the striatum in learning

We have seen in this chapter the striatum and the cortex are organized in four principal anatomical loops, associating corresponding territories.
Within each loop, the involved striatal and cortical territories are proposed to interact and work together to perform action selection (Redgrave et al., 1999a; Gurney et al., 2001a,b) and DA-mediated reward-based reinforcement learning (Schultz et al., 1997; Satoh et al., 2003). Besides, the

hippocampus is assumed to elaborate and sends contextual and allocentrically based spatial information, to the associative and limbic2 loops, respectively via DMS and the shell.
A number of lesion studies and electrophysiological data, taken together, suggest a respective role within these loops of:

1. DLS in the storage and expression of praxic and cue-guided model-free strategies;
2. DMS in the storage and expression of place model-free and model-based strategies;
3. the shell in the learning of model-free strategies via the dopaminergic system (VTA/SNc);
4. the core in the storage and expression of model-based strategies.

Figure 1.2.7 summarizes the functional architecture resulting from these reports. One of the key pending questions within this architecture is the following: Which brain structure can subserve the role of « strategy shifter » presented in this figure ? That is, which part of the central nervous system can detect when current behavior is not adequate, and can create new rules, or select, among existing strategies, the one to perform ? In the next section, we review anatomical, lesion and electrophysiological data supporting the medial prefrontal cortex (mPFC) as subserving such a function. The mPFC is considered to play an important role in flexible executive functions, it is strongly interconnected with the accumbens core, and is implicated in goal-directed behaviors. We will thus present the data suggesting its interaction with the core in model-based strategy, and its possible role in strategy shifting.

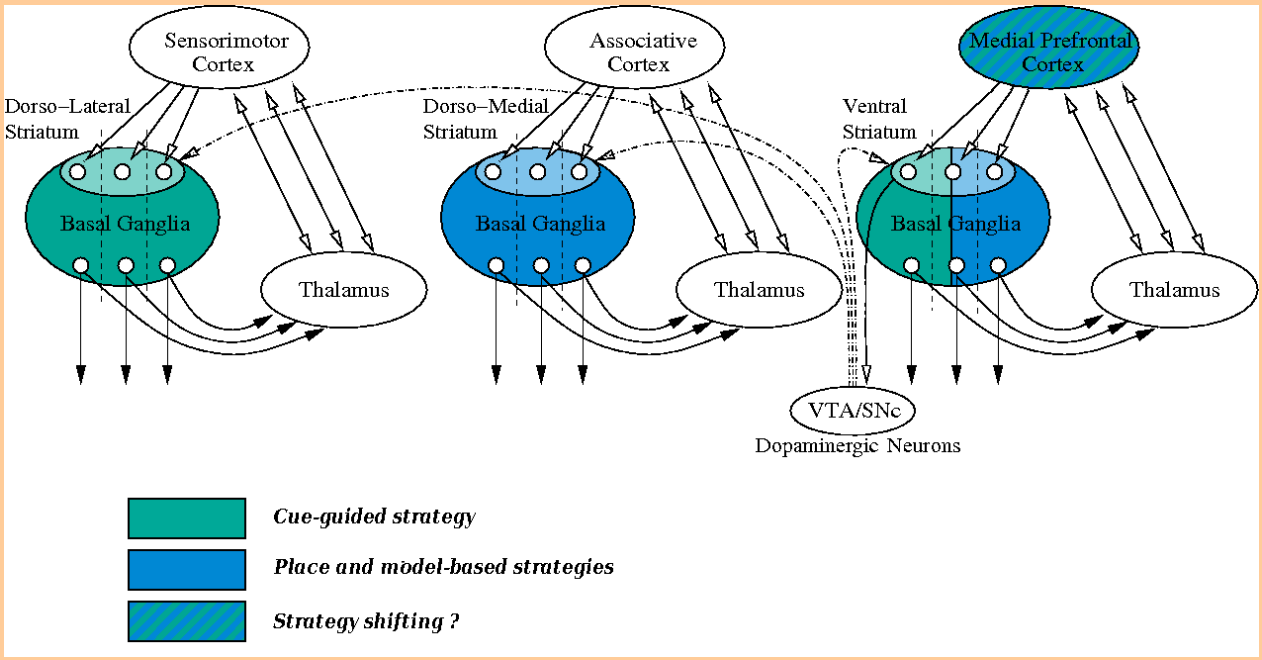


Figure 1.2.7 : *Possible functional architecture of the striatum where different territories of the striatum subserve different navigation strategies.* One of the key pending questions within this architecture is the following: Which brain structure can subserve the role of « strategy shifter » presented in this figure ? That is, which part of the central nervous system can detect when current behavior is not adequate, and can create new rules, or select, among existing strategies, the one to perform ? VS : Ventral Striatum ; DMS : Dorsomedial Striatum ; DLS : Dorsolateral Striatum ; VTA : Ventral Tegmental Area ; SNc : Substantia Nigra pars compacta.

2.6 The prefrontal cortex and flexible strategy shifting

Throughout evolution, the cortical mantle is the neural structure that has developed the most in mammals in comparison to other brain components (Northcutt and Kaas, 1995). The prefrontal cortex (PFC) was originally defined in humans and non human primates (NHPs) as the most rostral

portion of the frontal lobe lying rostrally adjacent to the premotor cortex and motor cortices (figure 1.2.8). With respect to all other cortical areas, it is particularly developed in primates and in humans and has come to represent nearly a third of the cortex (Brodmann, 1895).
Functionally, PFC is considered as a critical component of the "generator of planned behavior", according to Dickinson (1980). In humans, it is considered as a key structure for cognitive control, that is to say, the ability to coordinate thoughts and actions in relation with internal goals (Miller and Cohen, 2001; Koechlin et al., 2003). There exist different concurrent theories of the primate prefrontal cortex emphasizing respectively (see Koechlin and Summerfield, 2007 for a review): working-memory – characterized by the temporary storage of information required for its internal manipulation – (Goldman-Rakic, 1987; Petrides et al., 1993; D'Esposito et al., 1995; Petrides, 1996; Dreher et al., 2002; Guigon et al., 2002), representation of events of varying durations (Grafman, 2002), inhibition of irrelevant information (Fuster, 1997), attentional control (Shallice, 1988; Desimone and Duncan, 1995), executive processes (Shallice, 1996), voluntary action selection based on reward (Shima and Tanji, 1998a,b), control of the balance between planning and automaticity (Posner and Snyder, 1975; Shiffrin and Schneider, 1984), or control of the balance between cognition and emotion (Bechara et al., 2000).
Generally, it is admitted that the prefrontal cortex plays an important role in flexible behavior planning. Patients with prefrontal cortex damage show impaired performance in rule-shifting tasks and tend to persist in applying the previously relevant rule even after it becomes inappropriate (Milner, 1963; Drewe, 1974; Goldstein et al., 2004). Moreover, prefrontal cortical neurons show correlates with a set of parameters required for action sequencing, such as correlates with the relevant rule of a given task (Sakagami and Niki, 1994), with action-reward combinations (Matsumoto et al., 2003), and with the temporal organization of action sequences (Tanji and Shima, 1994; Carpenter et al., 1999; Procyk et al., 2000; Tanji and Hoshi, 2001; Mushiake et al., 2006). Besides, other PFC neurons show correlates with learning-related parameters such as reward expectation (Watanabe, 1996; Schultz et al., 1998) and error detection (Amiez et al., 2005). Finally, lateral prefrontal neurons encode context-dependent switching behaviors (Konishi et al., 1998; Nakahara et al., 2002; Amemori and Sawaguchi, 2006). These functions are found to be distributed over distinct regions of the prefrontal cortex, namely dorsolateral, anterior cingulate, medial and orbitofrontal regions (Fuster, 1997).

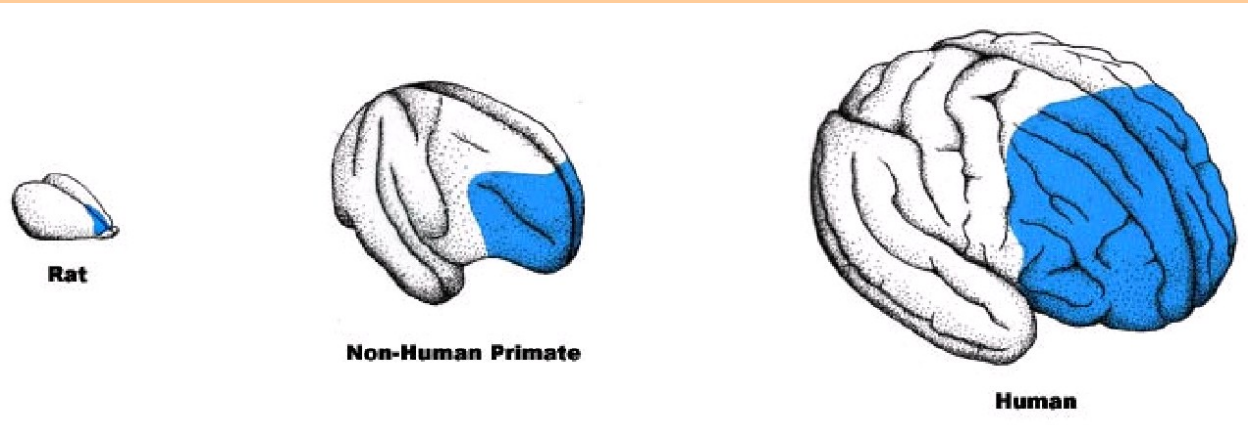


Figure 1.2.8 : *Adapted from Tierney (2006).* The PFC in the rat, NHP and human occupies the rostral portion of the frontal lobe.

In rodents, the prefrontal cortex is much less differentiated in terms of anatomy and function, and there are discrepancies between anatomical and functional homologies with regions of the primate PFC (Preuss, 1995; Granon and Poucet, 2000; Uylings et al., 2003). However, the rat prefrontal cortex can also be divided in the medial prefrontal cortex, which itself can be divided in the

ventromedial prefrontal cortex – comprising the Prelimbic and Infralimbic regions (PL/IL) and the medial orbitofrontal (MO) –, **the dorsomedial prefrontal cortex** – comprising the frontal area 2 (Fr2) and the dorsal anterior cingulate (ACd) –, **the agranular insular** (AI), and the **lateral orbitofrontal** areas (Uylings et al., 2003; Vertes, 2006).
Of particular interest here is the medial prefrontal cortex, and more specificially the prelimbic area, which shows strong functional homologies with the primate dorsolateral cortex, that is the region that is mostly implicated in flexible and attentional behavior planning and shifting (Granon and Poucet, 2000; Uylings et al., 2003).

2.6.1 Lesion studies

2.6.1.1 Rat mPFC is not a pure working-memory system

Early behavioral experiments suggested that the rat mPFC is involved in working-memory (see Kolb, 1990 for a review). Originally defined in humans, the concept of working memory combines, within a single model:
(a) a system for temporary storage and
(b) a mechanism for online manipulation of information that occurs during a wide variety of cognitive activities (Baddeley, 1996).
In lower vertebrates (rodents and birds), working memory was originally defined in a similar way (Honig, 1978; Olton et al., 1979) but was rapidly restricted to refer to a memory buffer that maintains information on-line in order to perform the task correctly.
Recent lesion studies suggest that mPFC is not involved in the on-line maintenance of information, and thus is not a pure working-memory system (see Granon and Poucet, 2000; Gisquet-Verrier and Delatour, 2006 for reviews). More precisely, whereas some studies report that mPFC damage produce a delay-dependent memory deficits in spatial delayed alternation tasks in a Y-maze or a T-maze (Van Haaren et al., 1985; Brito and Brito, 1990; de Brabander et al., 1991; Delatour and Gisquet-Verrier, 1999), in a « go / no-go » task, Delatour and Gisquet-Verrier (1996) reported no detrimental effects of increasing the delay in rats with PL lesions.
Granon and Poucet (2000) propose an explanation of these apparently inconsistent results by noting that the latter experiment requiring a simple runaway from the animal, it might engage less effortful processing than spatial delayed alternation tasks. Thus, they propose that **working-memory processes should be affected by mPFC lesions only when combined with other factors** such as:

- the difficulty of the task – which, for example, is increased when selection of the correct response must operate on a greater number of alternatives;
- the requirement for attentional mechanisms;
- the requirement for flexible behavior.

In line with this view, whereas lesions of mPFC do not impair the performance in a pure attentional task where rats have to detect spatial changes in their environment (Granon et al., 1996), and mPFC damage does not impair the performance in a task requiring the rats to pay attention to only two possible positions (Granon et al., 1998), lesions of mPFC impair the performance in a task where rats have to pay attention to a brief visual stimulus (a light) that could randomly occur in one of five possible positions (Muir et al., 1996).
Moreover, several studies show that mPFC lesions lead to attentional deficits (Birrell and Brown, 2000; Delatour and Gisquet-Verrier, 2000), to behavioral inflexibility (Burns et al., 1996; Delatour and Gisquet-Verrier, 2000; Dias and Aggleton, 2000), to impaired retrieval processes (Botreau et al., 2004)

2.6.1.2 PL is not a pure spatial system but appears instead to be involved in model-based (goal-directed) behavior

Moreover, within working memory, PL function appears to be better characterized by its involvement in a certain type of information processing (e.g., the type of associations: Stimulus-Response (S-R) or Action-Outcome (A-O) stored and manipulated) than by its involvement in

processing certain types of information (e.g. spatial vs. non spatial information) (see Granon and Poucet, 2000 for a review of this issue). Indeed, PL neurons fail to show spatial responses similar to hippocampal place neurons (Poucet, 1997; Jung et al., 1998; Pratt and Mizumori, 2001; Battaglia et al., In press). Consistent with this, rats with PL lesions are neither impaired in short-term memory of a spatial movement (Poucet, 1989; Granon et al., 1996), nor in place navigation (de Bruin et al., 1994; Granon and Poucet, 1995; Delatour and Gisquet-Verrier, 2000), nor in spatial discrimination (Ragozzino et al., 1999a,b; Hannesson et al., 2004).
Furthermore, Balleine and Dickinson (1998) found that prelimbic lesions impair action-outcome contingencies while sparing learning of S-R associations. In their task, rats were trained to perform two actions concurrently for two different food rewards. In addition, one of those reinforcers was delivered non-contingently with respect to the animal's behavior, thus resulting in a selective A-O contingency degradation. PL lesions rendered the rats insensitive to this contingency manipulation, suggesting that such rats might truly be “creatures of habit” (see Cardinal et al., 2002; Dalley et al., 2004 for reviews).
Other lesion studies confirm an involvement of PL in goal-directed behaviors (Corbit and Balleine, 2003; Killcross and Coutureau, 2003; Dalley et al., 2004; Ostlund and Balleine, 2005; see Cardinal et al., 2002 for a review). In contrast, IL lesions appear to affect habitual behavior following overtraining (Quirk et al., 2000; Morgan et al., 2003; Dalley et al., 2004; Coutureau and Killcross, 2003).
Interestingly, in the Y maze experiment of Delatour and Gisquet-Verrier (1996), rats with PL lesions were initially impaired during acquisition but eventually recovered with extensive training. These results are consistent with the hypothesis that PL lesions impair the goal-directed behavior system while sparing the habit system. The extensive training in this study may have enabled the latter to eventually mediate learning of the task. Similarly, Fritts et al. (1998) found that the impairment in a radial arm maze task induced by PL lesions was mainly due to a difficulty during the acquisition phase, and did not last more than eight days.

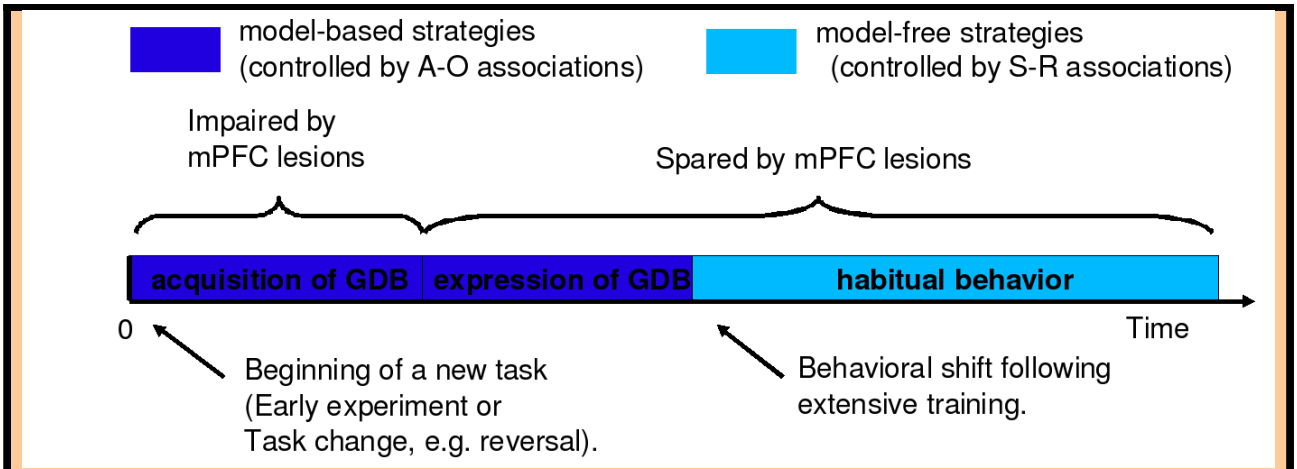


Figure 1.2.9 : *A possible role of the mPFC in the acquisition of goal-directed behaviors (GDB, i.e. model-based strategies), but not in the expression of GDB*. The schema sketches three different stages of learning, starting from the beginning of a new task. Extensive training within the same task – which remains constant – enables shifting to habitual behavior (*model-free strategies*), as described in section 1.7. As mentioned in the previous section, the expression of GDB could involve DMS or the core (Yin et al., 2005), while habitual behavior could be controlled by DLS (Yin et al., 2004).

Following our terminology, these results suggest that PL lesions could impair model-based strategies while leaving intact model-free strategies.
However, a recent study reports that only pre-training lesions of PL impair the animal's sensitivity to

outcome devaluation, while post-training lesions spare it (Ostlund and Balleine, 2005). The authors interpret these results as suggesting that the rat PL is more crucial for the acquisition of goal-directed behavior rather than for its expression.
In other words, PL seems to be important at the early stage of learning at the beginning of a task, or after a reversal, when a change in the task rule requires the animal to flexibly shift its behavior (Salazar et al., 2004). Figure 1.2.9 summarizes this hypothesis.

2.6.1.3 PL as a detector of task rule changes

Indeed, it seems that the rat PL plays a important role in attention focusing on the detection of external events indicating when the learning rule countradicts either spontaneously engaged or previously learned strategies. Rodents with mPFC lesions are unable to learn new task contingencies and continue applying the previously learned rule despite no longer being consistently rewarded for it (Delatour and Gisquet-Verrier, 2000; Dias and Aggleton, 2000). Attentional set-shifting or rule shifting depend on the mPFC (de Bruin et al., 1994; Birrell and Brown, 2000; Colacicco et al., 2002; McAlonan and Brown, 2003; Lapiz and Morilak, 2006). Moreover, PL damage-induced impairment is significantly increased when the task requires shifting from one strategy to another, whether the initial strategy has been learned (Granon and Poucet, 1995; Ragozzino et al., 1999a,b) or is spontaneously used by the animal (Granon et al., 1994).
More precisely, a particular subset of strategy shifts are impaired by PL lesions, refering to the different types of shifts we defined in the first section. Whereas lesions of the orbitofrontal cortex are found to impair intradimensional shifts (Kim and Ragozzino, 2005), lesions of PL-IL impair extradimensional shifts but intradimensional shifts are spared (Joel et al., 1997; Birrell and Brown, 2000; Ragozzino et al., 2003).

2.6.4 Electrophysiological data on mPFC

Electrophysiological studies in the rat confirm that the medial prefrontal cortex can integrate movement, motivational, reward and spatial information required for flexible model-based strategies. Cells recorded in mPFC have correlates with movement (Poucet, 1997), with reward, sometimes in an anticipatory manner (Pratt and Mizumori, 2001; Bouret and Sara, 2004), are selective to a lesser extent than VS for the type of reward the animal receives (Miyazaki et al., 2004). Activity in mPFC show a working-memory component (Baeg et al., 2003), correlates with spatial goals (Hok et al., 2005) and with action-outcome contingencies (Mulder et al., 2003 ; Kargo et al., 2007). Medial prefrontal neurons also show encoding of some spatial information. Even if spatial selectivity is less important than in the hippocampus (Poucet, 1997; Jung et al., 1998; Pratt and Mizumori, 2001), some mPFC neurons show spatial correlates (Pratt and Mizumori, 2001; Hok et al., 2005) or correlates with combined movement and location (Jung et al., 1998).
More recently, a study showed that mPFC neural activity could react to a behavioral shift. Notably, the functional connectivity between neurons within the mPFC was found to be the highest at the early stage of a new learning phase following a reversal (Baeg et al., 2007).
These results suggest that mPFC could play a important role in detecting a need to shift behavior after a change in the environment or in the task contingencies. One could predict from this evidence that neurons could be found in mPFC detecting changes in the task rule, for example by showing transitions of activity in response to such changes. Another prediction, in addition to the detection of task rule changes, is that mPFC's possible participation in the selection of the new strategy to perform after the change could take the form of neurons being selective to the ongoing strategy spontaneously performed by the animal.

3. Neuromimetic models of rodent navigation

In this closing-section of the introduction, we review some computational models of strategy learning and of strategy shifting. We restrict here to neuromimetic models, and more precisely to models involving the prefrontal cortex or the striatum, or including one or several prefronto-striatal loops. As a consequence, the models presented here are based on neural networks.
These networks correspond to control architectures for animats whose function is to deal with the coordination of captors and actuators in order to efficiently reach resources within a given environment.

In the review of computational models presenting here, we will restrict to a particular situation where an animal or an artificial agent has to learn to perform a sequence of actions leading to reward (figure 1.3.0). The main questions that these models will help us to solve are: how to choose which actions to reinforce when getting a reward ? And when to reinforce these actions ? For example, in figure 1.3.0, action 2 was inappropriate for reaching reward, and thus shall not be positively reinforced. As we will try to highlight in this section, one of the main difference between considered groups of models relies in the type of representation that they use. On the one hand, *model-based* systems memorize the whole sequence of actions and use a representation of the respective consequence of each action within the sequence. As a consequence, the system can quickly update the sequence when the environment changes because it can estimate the respective contribution of each action to the reward. However, such *model-based* systems are computationally expensive.
On the other hand, *model-free* systems reinforce each action individually, as soon as one action has been performed, and without taking into account the global sequence of actions. As a consequence, each action is learned independently from preceding and succeeding actions. Thus, *model-free* systems are computationally very simple. However, they are much slower to learn.

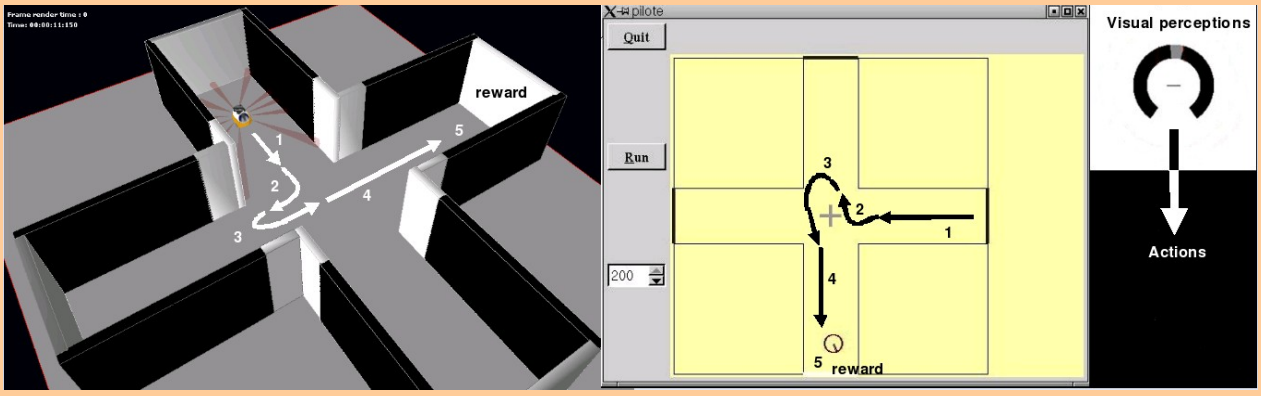


Figure 1.3.0 : *Navigation paradigm considered in this review of computational models.* In this example environment, a virtual agent is located in a plus-maze, and perceives visual information from this environment. The agent has to learn to associate actions to the visual information that it perceives at particular moment. We suppose here that the agent has performed a sequence of 5 consecutive actions and eventually got a reward. The issue at stake is to decide which actions were appropriate and should be reinforced.

3.0 Overview

A biomimetic navigation model aims at improving the autonomy of animats in the processing of available sensory information (allothetic, idiothetic, both), and the selection of the ones that are

relevant according to the context; in the association of these informations with behavioral responses that will enable the animat to maximize reward achieving by reaching the resources; in the evaluation of the efficiency of the animat's responses in order to change them as soon as they become inappropriate.
For this purpose, several types of learning techniques can be used:

1. Unsupervised learning, a correlation-based form of learning that does not consider the result of the output, i.e. which does not need any explicit target outputs or environmental evaluations associated with each input (Dayan, 1999);
2. reinforcement learning, which processes sensorimotor associations by trial-and-error with a scalar reward feedback from its environment. Here the feedback signal is only evaluative, not instructive;
3. and supervised learning, which differs from reinforcement learning in that the feedback provides the motor output solution is instructive, and must be given by some external ”teacher”.

Interestingly, this difference between computational learning mechanisms finds some echo within the brain, since the cerebral cortex, the basal ganglia and the cerebellum appear to respectively rely on unsupervised, reinforcement and supervised learning mechanisms (Doya, 1999, 2000b).
Here, we will not address supervised learning, but we will rather focus on two kinds of **reinforcement learning:** model-free and model-based. As we will see, the former is used in most systems implementing *navigation strategies* in which the *basal ganglia* - at the core of the computational model we have designed during this thesis - are involved. The latter, or algorithms mathematically equivalent to the latter, are mostly used to implement *behavioral* or *navigation strategies* involving the prefrontal cortex.
We will also describe some examples of **unsupervised learning**, which can be used to learn the *structure of the environment* on the basis of allothetic and/or idiothetic inputs, or to build graphs of possible movement transitions used in *model-based navigation*. In our case, we used unsupervised learning to categorize visual perceptions with a method called *self-organizing maps* (Kohonen, 1995), and employed it for the coordination of several model-free reinforcement learning modules (see chapter 2).

3.1 Basic notions on neural networks

Within the neural networks used by the reviewed models, a formal neuron has an associated stored vector of real values, representing a memory of the strength (or *weight*) of « synaptic » connections with afferent neurons. Each neuron is also provided with a fixed threshold – defining how much input activity is required to trigger a response from the neuron –, and a filtering function defining how the neuron output is affected by its inputs (Churchland and Sejnowski, 1995). Most often, neurons within the models presented in this section are *rate coding neurons*, that is neurons whose activity represents a rate averaged over time, in contrary to *spiking neurons* where spike timing information is represented (Gerstner and Kistler, 2002; Dayan and Abbott, 2005).
Finally, learning within these models is represented by a modification of the synaptic weights of concerned neurons, thus changing the way information is processed through the networks, and altering the way in which perceptions of an artificial agent or *animat* are associated with behavior.

3.2 Unsupervised Learning in neural networks

In the late 1940s, Donald Hebb made one of the first hypotheses for a mechanism of neural plasticity (i.e. learning), *Hebbian Learning* (Hebb, 1949), that states that the connection between two neurons is strengthened if the neurons fire simultaneously, or within a time interval. Hebbian learning is considered to be a 'typical' unsupervised learning rule and it (and variants of it) was an early model for *Long-Term Potentiation* and *Long-Term Depression* (resp. increase and decrease of synaptic efficiencies) observed in biology (Ito and Kano, 1982; Bear et al., 1987).
Unsupervised Hebbian learning can be employed to allow an agent to build its allocentric spatial

map of the environment based on its own experience (Burgess et al., 1994; Trullier and Meyer, 1997; Gaussier et al., 1998; Arleo and Gerstner, 2000). For exemple, in the model of Gaussier et al. (1998), the synaptic weight $\omega_{i,j}$ of the link between two successively visited places *j* and *i* is increased by Hebbian learning. Thus, $\omega_{i,j}=0$ when there is no path from *j* to *i*, whereas $0<\omega_{i,j}\leq 1$ when *i* is directly reachable from *j*. A scheme illustrating the architecture of this model, and more information about it are provided in section 3.8.1.

Another possible use of unsupervised Hebbian learning is for the implementation of *self-organizing maps* (Kohonen, 1995), which we employed for the coordination of Stimulus-Response learning modules in our model (see chapter 2, section 4). Despite the word « map », this algorithm does not necessarily apply to the building of an allocentric map of the environment. It is an artificial neural network trained to produce low dimensional representation of the training samples – e.g. a set of visual inputs perceived by an agent in an environment – while preserving the topological properties of the input space (Ritter et al., 1992). For exemple, an agent introduced into a continuous environment containing two different cues – *cue1* and *cue2* – will receive an ensemble of continuous visual inputs between « perceiving *cue1* only », « perceiving *cue1* and *cue2* », « perceiving *cue2* only », and a set of intermediary visual perceptions. In this case, the goal of learning in the self-organizing can be to associate certain neurons (or *nodes*) in the map with certain input patterns in order to build a discrete approximation of the distribution of visual perceptions. Learning will lead to bring closer and to cluster neurons in the map which respond similarly to a given visual input, while moving away and separating neurons which respond differently. Thus, after learning, such a map will « represent » approximate categories within the visual input space. This is partly motivated by how visual, auditory and other sensory information is handled in separate parts of the cerebral cortex in the brain (Haykin and Simon, 1999).

3.3 Reinforcement Learning in neural networks

While methods presented in the previous section can provide algorithms to build representations of an agent's perceptions, reinforcement learning provide a tool to adapt the agent's actions to the environment.

3.3.1 Markovian decision processes

Within a community with intuitions from animal learning theory, researchers have provided a theoretical framework for *Reinforcement Learning* in order to have an agent learn by trial-and-error to adapt its actions in a given environment so as to maximize some notion of long-term reward (see Sutton and Barto, 1998). This theoretical framework is grounded on *Markov Decision Processes (MDP)*, in which it is assumed that the agent state at a given moment only depends on two factors: its state at the previous instant, and the action it has just performed (Bellman, 1957). This provides a deterministic framework in which one can prove conditions for learning convergence.

The agent's behavior is identified as its *policy*, a function $\Pi: S\times A\rightarrow \Pi(A)$ which indicates, for each state $s \in S$, the probability distribution that the agent chooses each action $a \in A$ it this state. The agent's state can refer to various parameters such as its perceptions, internal metabolic state, or location within the environment. Usually, the probability to perform action *a* in state *s* is noted as $\Pi(s,a)$. Once the action is chosen, a certain transition function *T* determines for each (state,action) couple the probability distribution that the agent can reach each possible state based on the action it has just performed in the former state.

In *model-free* reinforcement learning, the transition function is unknown and cannot be learned by the agent (Sutton et al., 1992).

In *model-based* reinforcement learning, the agent can learn and use the transition function. Thus the agent is provided with a *world model* enabling it to choose its actions based on an estimation of their consequences (i.e. in which final state they will lead) (e.g., Sutton, 1990; Barto et al., 1995,

Kaelbling et al., 1996).

3.3.2 Learning based on reward

In a given environment, there can be a state *s* where performance of action *a* by the agent can lead to a reward: $R(s,a)$. The main objective for the agent is to adopt a policy which enables it to maximise the frequency and the value of these rewards. Thus, the agent shall proceed with a certain learning to adapt its policy.

Formally, in reinforcement learning, in order to evaluate the agent's policy π, a value function $V_{\Pi}(s)$ associates to each state *s* an estimation of the cumulated reward the agent will get if it performs this policy starting from state *s*. The cumulated reward consists in the sum of all future reinforcement signals and can be written as:

$$R_{\Pi}(t) = r_t + \gamma . r_{t+1} + \gamma^2 . r_{t+2} + \dots \quad \text{(E.1)}$$

where $0 < \gamma < 1$ is a *discount factor* which limits the capacity to take into account rewards in the far future and prevents this sum from being infinite. Equation E.1 can be written as:

$$R_{\Pi}(t) = \sum_{i=1}^{\infty} \gamma^{(i-t)} . r_i \quad \text{(E.2)}$$

Then, the value function is defined as the expected (or *predicted*) reward return starting from state S and following policy π, and can be written as :

$$V_{\Pi}(s) = \sum_{a \in S} \Pi(s,a) \left[R(s,a) + \gamma \sum_{s' \in S} T(s,a)(s') V_{\Pi}(s') \right] \quad \text{(E.3)}$$

where $T(s,a)(s')$ denotes the probability, based on the transition function *T*, to reach state *s'* after performing action *a* in state *s*. Equation E.3 is called Bellman equation for policy π. This equation plays a fundamental role at the core of optimisation methods which permit to define reinforcement learning algorithms.

There exist three main classes of algorithms permitting to an agent facing an MDP to discover an optimal policy – a policy by which the agent can get a maximal cumulated long-term reward:

1. *Monte Carlo* methods, which do not require any a priori knowledge of the environment (model-free), and have the lack not to rely on an incremental estimation of an optimal policy;
2. *Temporal Difference* methods, which are also model-free, and rely on an incremental estimation of an optimal policy;
3. *Dynamic Programming* algorithms, which are used when the agent is provided with a world model (model-based), that is when both transitions and reward functions are known. Examples of such algorithms are *Dyna-Q, prospective planning* or *Tree-search*;

In this manuscript, we will not explain the first class which is reviewed in (Sutton and Barto, 1998; Cornuéjols and Miclet, 2002; Sigaud, 2004).

We will first focus on *Temporal Difference methods*, which are widely used in the field of Reinforcement Learning because they gather interesting properties from the two other classes: like Dynamic Programming algorithms, they are incremental (the estimated value *V(s)* in the current state *s* is updated based on the estimated value *V(s')* in the forthcoming state *s'*); like Monte Carlo methods, they do not require any model of the environment.

However, Temporal Difference methods are slow to learn and suffer from inflexibility. Thus we will also present a few examples of model-based *Dynamic Programming* methods, which have a high computational cost and degree of complexity due to the manipulation of the world model, but which are more flexible than model-free algorithms.

3.3.3 The model-free Temporal Difference (TD) algorithm

The method consist in comparing two consecutive reward estimations (or predictions) $V_{t-1}(s)$ and $V_t(s')$, the agent having performed an action *a* between state *s* and *s'*. For simplicity, we note these reward estimations V_{t-1} and V_t , following the demonstration of Barto (1995):

$$V_{t-1}=r_t+\gamma.r_{t+1}+\gamma^2.r_{t+2}+\dots \quad (E.4)$$

$$V_t=r_{t+1}+\gamma.r_{t+2}+\gamma^2.r_{t+3}+\dots \quad (E.5)$$

Notice that equation E.4 can be reformulated as:

$$V_{t-1}=r_t+\gamma.\left(r_{t+1}+\gamma.r_{t+2}+\dots\right) \quad (E.6)$$

Which, combined with equation E.5, gives:

$$V_{t-1}=r_t+\gamma.V_t \quad (E.7)$$

This is the consistency condition that is satisfied by the correct predictions. The error by which any two adjacent predictions fail to satisfy this condition is called the *temporal difference error* (TD error) by Sutton (1988) and is computed as: $r_t+\gamma.V_t-V_{t-1}$.

Then learning does not consist in waiting for a long term reinforcement signal, but rather in modifying at each timestep the value function $V_\Pi(s)$ as a function of the TD error between two consecutive reward estimations:

$$V_\Pi(s)\leftarrow V_\Pi(s)+\eta.\overbrace{\left[r_t+\gamma.V_t-V_{t-1}\right]}^{\text{TD error}} \quad (E.8)$$

where $\eta>0$ is the learning rate. In the same manner, the policy of the agent can be updated by modifying the probability to perform again the same action *a* in the same state *s*:

$$\Pi(s,a)\leftarrow\Pi(s,a)+\eta.\overbrace{\left[r_t+\gamma.V_t-V_{t-1}\right]}^{\text{TD error}} \quad (E.9)$$

This learning procedure leads to progressively translate reinforcement signals from the time of reward occurrence to environmental contexts (i.e. *states*) that precede the reward, and further to *states* preceding *states* preceding reward, ... as described with the example in BOX1.

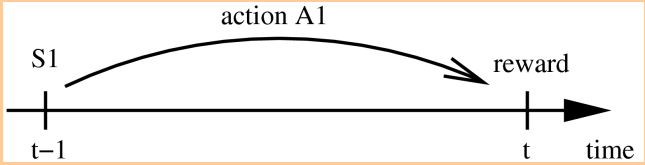
In this example, a stimulus *S1* preceding reward has its value reinforced. Then, the perception of *S1* becomes reinforcing, and enables to increase the value of a stimulus *S2* preceding S1. Barto describes this method with the expression: « It is [...] like the blind being led by the slightly less blind » (Barto, 1995). Dayan and Sejnowski (1994) proved that this method converges with probability 1. This method can enable an agent to find an optimal behavior within a discretized environment (see Sutton and Barto, 1998 for a review) or a continuous one (Doya, 1999, 2000a; Tani et al., 2007).

A widely used architecture for the implementation of the Temporal Difference learning method is the **Actor-Critic architecture** as described by Barto (1995) and displayed on figure 1.3.1. On the one hand, the Actor is the memory zone which stores the agent policy and performs action selection depending on the agent's state within the environment. On the other hand, the Critic evaluates the value function at each timestep. To do so, it makes a reward prediction. Then at the next timestep, it computes its reward prediction error based on actual feedback from the environment (the primary reward which can be positive, negative or null), and according to the TD error. If the primary reward is better than expected (respectively worst than expected), the Critic sends a positive (respectively negative) reinforcement to the Actor, thus permitting the Actor to adapt action selection. Besides, the same reinforcement signal is also used by the Critic in order to precise its reward predictions.

In the field of model-free reinforcement learning, other methods have derived from Temporal Difference Learning. We briefly mention methods such as *SARSA* and *Q-learning*, which similarly to the TD algorithm, are employed in some models of rodent navigation or in models of reinforcement learning within the cortico-striatal system. In contrast to the TD method which estimates a value function $V_\Pi(s)$ of a state *s* (i.e. a reward prediction computed in state *s*), the SARSA algorithm works with the *quality* of the (state,action) couple – also called the *action value function* –, written as $Q(s,a)$. As a consequence, SARSA is required to predict one step further which action *a'* the agent will perform at the next timestep. SARSA is updated according to the following equation adapted from E.8:

$$Q(s,a) \leftarrow Q(s,a) + \eta . \left[r_t + \gamma . Q(s',a') - Q(s,a) \right] \quad \text{(E.10)}$$

BOX1: Example of an agent learning a sequence of stimulus-action associations.



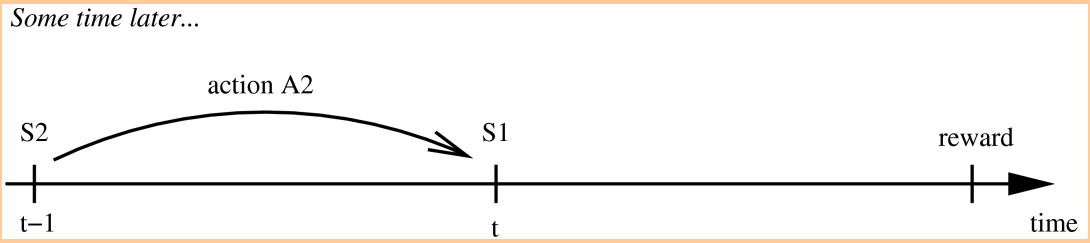
Let's reduce the state to the information concerning a single salient stimulus S1 in the environment. At time t-1, the agent performs action A1 in response to stimulus S1. This leads the agent to reach, at time t, a reward. Thus, based on equation E.8, the reward value associated to stimulus S1 is increased, building a stimulus-reward association ***S1-reward****:*

$$V_\Pi(S1) \leftarrow V_\Pi(S1) + \overbrace{\eta . \left[1 + \gamma . 0 - 0 \right]}^{>0}$$

On the same occasion, the policy associated to stimulus S1 is modified, thus building a stimulus-action association ***S1-A1****:*

$$\Pi(S1,A1) \leftarrow \Pi(S1,A1) + \overbrace{\eta . \left[1 + \gamma . 0 - 0 \right]}^{>0}$$

Then, let's consider that some time later, the agent experiences a new stimulus S2, selects randomly an action A2 in response to S2, which results in putting the agent in front of the known stimulus S1.



Stimulus S1 having been associated with reward, can itself become a source of reinforcement thanks to TD learning, thus increasing the value of stimulus S2 according to the following equation:

$$V_\Pi(S2) \leftarrow V_\Pi(S2) + \overbrace{\eta . \left[0 + \gamma . V_\Pi(S1) - 0 \right]}^{>0}$$

On the same occasion, the stimulus-action association ***S1-A1*** is increased*:*

$$\Pi(S2,A2) \leftarrow \Pi(S2,A2) + \overbrace{\eta . \left[0 + \gamma . V_\Pi(S1) - 0 \right]}^{>0}$$

As a consequence, the agent has learned a sequence S2-A2 – S1-A1 to reach the reward.

Improving the SARSA algorithm, Q-learning does not need to predict the action performed at the next timestep, since it updates the quality function based on the estimated optimal action at the next timestep:

$$Q(s,a) \leftarrow Q(s,a) + \eta . \left[r_t + \gamma . max_a Q(s',a) - Q(s,a) \right] \quad \text{(E.11)}$$

The term $Q(s',a')$ in equation E.10 has been replaced by $max_a Q(s',a)$ in equation E.11. This could appear equivalent when the agent always chooses the action that maximizes reward (in this case, $a'=argmax_a Q(s',a)$). However, the necessity to realize an exploration/exploitation trade-off makes this equality generally false. Thus, it appears that the SARSA algorithm processes its updating as a function of actions actually performed, whereas Q-learning processes it updating as a function of optimal actions, which makes it easier and more efficient (Watkins, 1989). Formal proofs of convergence of the Q-learning algorithm have been produced (Watkins and Dayan, 1992).

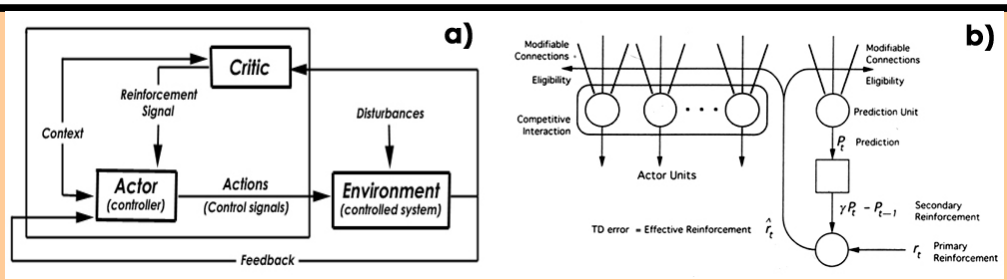


Figure 1.3.1 : *Actor-Critic architecture. Adapted from Barto (1995).* a) Global schema describing interactions between the Actor, the Critic and the environment. At each new timestep, the Critic sends to the Actor a reinforcement signal (which can be null) computed out of the TD error. b) Barto's canonical neural network implementation of the Actor (left) and the Critic (right). Each possible action is represented by a neural unit in the Actor part. A neural unit computes reward predictions in the Critic. Consecutive reward predictions are combined with a primary reinforcement (the actual reward received at time *t*) within a neural unit computing the TD error. The resulting effective reinforcement learning is used to reinforce synaptic weights of both the Actor and the Critic.

It is important to notice that, due to the property of TD learning to incrementally transferring reward information from reward itself to stimuli that precede it, **this model-free method is slow to learn and suffers from inflexibility** (Samejima and Doya, 2007).
It is slow to learn, because if the agent randomly performs a sequence of actions A1-A2-A3 which lead to reward, only action A3 is reinforced the first time. Then the agent has to perform the sequence again so that action A2 is reinforced, and so on. It is inflexible, because when the condition of the reward is changed (e.g. in case of a *reversal*), the agent has to experience many failures to depreciate action A3, then action A2, then action A1. It is only when action A1 has been depreciated that the agent can perform a new action sequence A4-A5-A6.

3.3.4 Model-based reinforcement learning algorithms

In order to improve the learning speed and the flexibility, researchers developed model-based reinforcement learning (Sutton and Barto, 1998), in which the consequences of actions is learned and used under the form of the transition function:

$$T: s \times s' \times a \rightarrow Proba\langle s'|s,a\rangle \qquad \text{(E.12)}$$

This function gives the probability to reach state s' after having performed action a in state s.
As explained in the first chapter of the introduction, if the states represent positions in a given environment, then this function can consist in an allocentric graph linking places together and providing transitions from one place to another. If the states represent visual stimuli, or any other external cues, then this function can consist in a graph which is not necessarily allocentric. In any case, building the model of transitions does not require reward information, and thus can be learned through unsupervised learning, simply by associating temporally consecutive states and actions.
For example, in the model of Gaussier et al. (2002), while building an association in the map between two successively visited places *j* and *i*, the model can also learn to associate the action *a* that was performed to enable the agent to reach place *i* starting from place *j*.
Then, the reinforcement learning part of the method consists in associating reward or *goals* with certain states in the model (e.g. associating certain places with the presence of a reward, Gaussier et al., 2002). Provided with such information, the action selection process can take into account estimated consequences of action and can be done with several algorithms. An example is *prospective planning* or *tree-search*, or *look-ahead planning*, in which the agent can anticipate several actions in advance without actually performing them, and thus can predict a hypothetical outcoming state (Baldassarre, 2002b,2003; Gaussier, 2002; Butz et al., 2003; Daw et al., 2005;

Hasselmo, 2005; Degris et al., 2006).
Model-based reinforcement learning can also be used to disambiguate noisy or missing sensory inputs, in a form called « belief states » (Samejima and Doya, 2007). For instance, when an animat navigating in the environment experiences two visually similar states, by combining the *world model* and history of the past actions stored in working-memory, the animat can estimate which of the two possible states is the most reliable. An efficient method to combine noisy observation and dynamic prediction based on a world model is the framework of Bayesian inference (Doya et al., 2007).
Model-based reinforcement learning algorithms have the advantage to provide rapid learning since the *world model* permits to "try" actions and evaluate their consequence without actually performing them, rather simply by simulating the performance of these actions within the *world model* (Sutton and Barto, 1998; Coulom, 2002; Degris, 2007). Moreover, these algorithms provide more flexibility than model-free methods, since the devaluation of a reward or the change in the reward position can be taken into account by changing the states with which reward is associated. Then, planning within the latter adapted *world model* can provide a one-trial adaptation of the behavior to perform by the agent (Daw et al., 2005).
However, model-free methods such as TD learning have the advantage not to require the storage of such a *world model* (only stimulus-reward and stimulus-action associations are stored, respectively by the value function and the policy function). Moreover, action selection does not rely on a complex and computationally expensive exploration of a *world model*. Rather, the perception of a stimulus is enough to trigger an action.
Thus, model-free and model-based learning methods appear to have complementary advantages that are appropriate in different situations (Sutton and Barto, 1998; Uchibe and Doya, 2004). Model-based methods are particularly suited when facing novelty (e.g. a task rule change), whereas model-free methods are adapted when the task is stable.
Note that certain authors have proposed to combine model-based methods with TD-learning (Sutton and Barto, 1998; Doya et al., 2002).
In the next section, we present the *analogy between model-free / model-based methods and some neurobiological data*. It turns out that the TD learning algorithm and the Actor-Critic architecture within which it is anchored show a strong resemblance with the way dopamine is released within the basal ganglia.
Besides, models of transition within the environment, planning and belief states were found to accurately describe prefrontal activity during certain tasks.

3.4 Analogy between the TD error and dopamine signals within the basal ganglia

As we have seen in section 2.4, electrophysiological data recorded in the monkey suggest that dopaminergic neurons respond to unexpected rewards, or to conditioned stimuli predicting reward, whereas they do not respond to predicted rewards (Schultz, 1998; Schultz, 2001). This pattern of response is very similar to the Temporal Difference error, and thus to the reinforcement signal described in the previous section.
Indeed, when considering the reinforcement based on the TD error equation: $\hat{r}_t = r_t + \gamma . P_t - P_{t-1}$ where $\hat{r}_t$ is the effective reinforcement at time *t* used by Barto (1995) as shown on figure 1.3.1, assuming the usual case where γ is close to 1 (e.g. $\gamma = 0{,}98$ such as in the model of Suri and Schultz, 1998,2001), we can see that:

1. At the occurrence of an unexpected reward, we have $r_t = +1, P_t = 0, P_{t-1} = 0$, thus $\hat{r}_t = +1$
2. At the time of a reward predicting stimulus *S*,we have $r_t = 0, P_t \approx +1, P_{t-1} = 0$, thus $\hat{r}_t \approx +1$
3. At the time of the reward predicted by *S*, we have $r_t = +1, P_t = 0, P_{t-1} \approx +1$, thus $\hat{r}_t \approx 0$
4. When an expected reward does not occur, we have $r_t = 0, P_t = 0, P_{t-1} \approx +1$, thus $\hat{r}_t \approx -1$,

which results in an interruption of activity in « dopaminergic neurons » having a baseline activity within models (Suri and Schultz, 1998).

These various cases describe quite accurately the different situations where dopaminergic neurons were recorded as shown on figure 1.3.6. This report gave birth to the hypothesis that dopamine could encode the temporal difference error (Barto, 1995; Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). Moreover, the two-compartmental structure of the Actor-Critic architecture implementing TD-learning was found to reflect the anatomical dissociation between striatal territories which project to dopaminergic neurons – thus able to drive reinforcement learning –, and striatal territories which do not – only subject to RL. As described in section 3.2, this dichotomy can be either considered between striosomes (Critic) and matrisomes (Actor) (Gerfen, 1984,1992) or between the shell (Critic) versus other parts of the striatum (Thierry et al., 2000; Voorn et al., 2004).

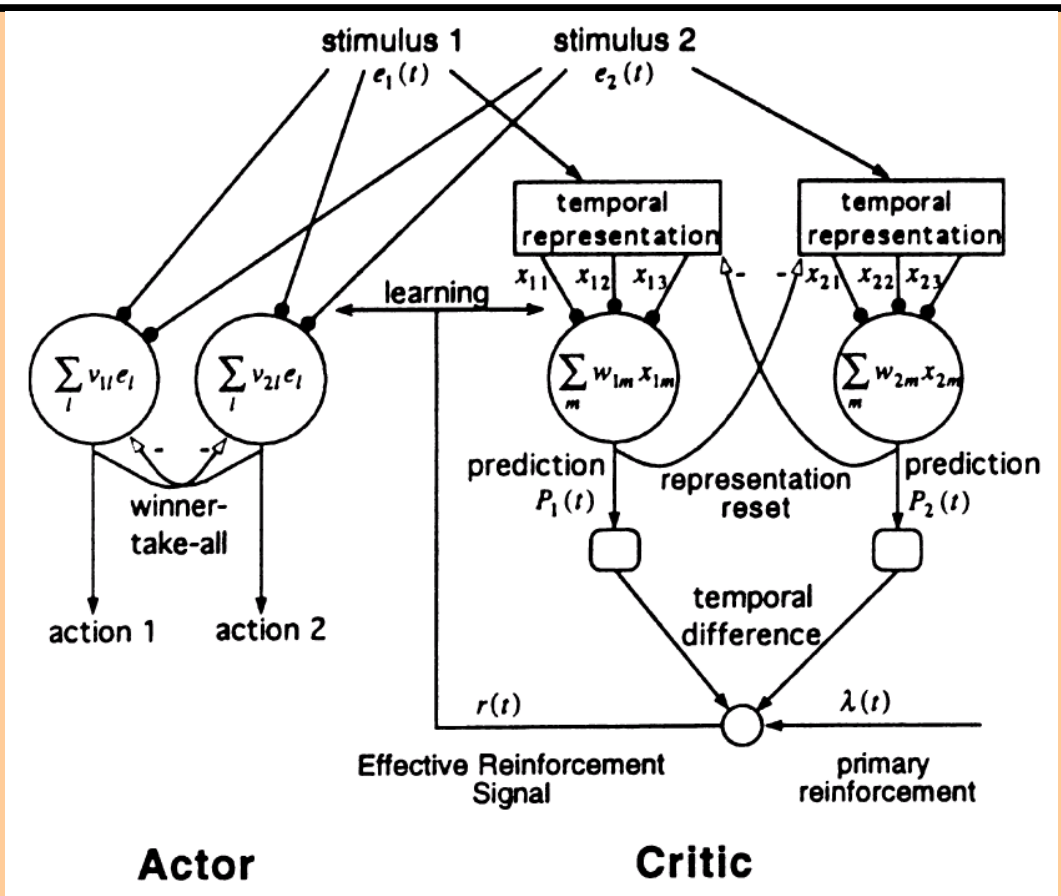


Figure 1.3.2 : *Actor-Critic model using a temporal representation of stimuli. Adapted from Suri and Schultz (1999).* The model consists of an Actor component (left) and a Critic component (right). Actor and Critic receive input stimuli 1 and 2 which are coded as functions of time, $e_1(t)$ and $e_2(t)$, respectively. The Critic computes the Effective Reinforcement Signal *r(t)* which serves to modify the weights v_{nl} of the Actor and the weights w_{lm} of the Critic at the adaptive synapses (heavy dots). Within the Actor, a winner-takes-all rule prevents the Actor from performing two actions at the same time. Within the Critic, every stimuus *l* is represented as a series of components x_{lm} of different durations. Each of these components influences the reward prediction signal according to its own adaptive weight w_{lm}. This form of temporal stimulus representation allows the Critic to learn the correct duration each stimulus-reward interval. Computation of the Temporal Difference error and adaptation of synaptic weights are subserved in a similar manner as described in figure 1.3.1.

3.5 Computational model-free systems in the basal ganglia

Starting from the resemblance between TD-learning and dopaminergic neurons activity, a number of Actor-Critic computational models were developed to represent the functional role of the basal ganglia in action selection and reinforcement learning (see Joel et al., 2002; Gurney et al., 2004 for reviews; and Khamassi et al., 2005 for a more recent comparison). Until recently, these models individually focused on the modelling of reinforcement learning of a single behavioral or navigational strategy, starting from the hypothesis that the basal ganglia were dedicated to habitual S-R learning (Graybiel and Kimura, 1995; Houk et al., 1995). In most cases, the modeled strategy was one of the two following: a model-free allothetic dimension or a model-free place dimension strategy.

3.5.1 Models associating cues with actions

Indeed, an important subset of these models focused on the reproduction of the precise temporal patterns of response of dopaminergic neurons in Schultz's task involving an association between an external cue (tone or light) and a reward, thus able to provide cue-based learning strategies (Houk et al., 1995; Montague et al., 1996; Suri and Schultz, 1998, 1999, 2001; Suri, 2002; Perez-Uribe, 2001; Sporns and Alexander, 2002).

For example, in the model of Suri and Schultz (1999), displayed in figure 1.3.2, the Actor-Critic employs a temporal representation of stimuli providing precise durations of stimulus-reward intervals. This temporal representation is called a “complete serial compound stimulus” (Montague et al., 1996) and varying, time after time, the values $x_{i,j}$ representing stimuli. For instance, if the Actor-Critic model gets an input $x_{1,2}=1$, it means that *stimulus1* was perceived 2 ms ago. If it gets an input $x_{2,5}=1$, it means that *stimulus2* was perceived 5 ms ago. Using such a representation, the Actor-Critic model is able to predict the precise moment when the reward usually occurs. If a predicted reward does not occur, the model is able to produce a negative signal similar to dopaminergic neurons' response in the same situation (Schultz, 1998).

In the article presented in chapter 2, section 2.2, we adapted this component to reproduce reward anticipatory activities recorded in the ventral striatum.

Moreover, Suri and Schultz (1999)'s model employs an ad hoc association between different stimuli and different Actor-Critic modules (see figure 1.3.2), thus preventing any interference the modules' responses to different stimuli. The latter feature is not satisfying for navigation in autonomous animats facing changing environments, since the experimenter cannot manually add a new Actor-Critic module each time the animat is facing a new stimulus.

Some Actor-Critic models were design to solve this issue by implementing a *mixture of Actor-Critic* experts (Baldassarre, 2002; Doya et al., 2002). The mixture of experts algorithm was proposed as a formal architecture to coordinate different *experts* competing and learning a given task (Jacobs et al., 1991). It mathematically parametrizes how *experts* “share” learning signals. Moreover, both Baldassarre (2002) and Doya et al. (2002) combine the *mixture of experts* with a certain rule controlling the latter parameters, that is, deciding which experts should but trained at each given moment. In both models, each expert has a component which learns to predict future states. Then, the expert which has the best performance in computing accurate predictions in a given state will be trained. As a consequence, each expert becomes specialized in a particular subset of a given task.

Note that in Doya et al. (2002)'s model, the components predicting future states are model-based systems. However, these components are used to coordinate reinforcement learning modules which subserve a model-free action selection (without engaging a planning procedure). This is why we mentioned this model in this section. These combinations of TD-learning and mixture of experts are particularly suitable in robotics reinforcement learning tasks with an important number of states (Baldassarre, 2002; Doya and Uchibe, 2005).

Finally, let's mention that other groups of Actor-Critic models have been proposed, notably focusing on the basal ganglia's ability to generate sequences of actions in response to sequences of external cues, these sequences being either immediately performed by an agent or stored in a cortical

working memory buffer (Berns and Sejnowski, 1996,1998; Beiser and Houk, 1998; Doya, 1999,2000; Hikosaka et al., 1999; Frank et al., 2001; Nakahara et al., 2001; Brown et al., 2004; O'Reilly and Frank, 2006).
A final group of models focused on the way basal ganglia anatomy could play a Critic-like role in generating temporal difference errors (Brown et al., 1999; Contreras-Vidal and Schultz, 1999; Doya, 1999,2000; Bar-Gad et al., 2000; Suri et al., 2001; see Daw and Doya, 2006 for a review).

3.5.2 Models associating places with actions

Another group of Actor-Critic models were more specifically dedicated to navigation and studied how hippocampal spatial input to the basal ganglia combined with a TD-learning algorithm could subserve a place recognition-triggered strategy (Brown and Sharp, 1995; Arleo and Gerstner, 2000; Foster et al., 2000).
As we did not implement a model for learning *locale* navigation, we will not precisely describe these models here. However, the important thing to note is that, in contrast to the model of Suri and Schultz (1999) presented in the previous section, models of *locale* navigation roughly consist in replacing input stimuli sent the Actor-Critic system by input places computed by localization system. Then, both reinforcement learning and action selection methods are similar to the models presented in the previous section.
It is interesting to mention that several models implementing reinforcement learning within the basal ganglia have replaced the TD-learning system by a Q-learning system (Strösslin and Gerstner, 2003; Chavarriaga et al., 2005a,b; Daw et al., 2005; Sheynikhovich et al., 2005; Hadj-Bouziane et al., 2006; Haruno and Kawato, 2006). As mentioned earlier, whereas TD-learning separately learns the value function and the policy, respectively within the Critic and the Actor, Q-learning systems combine the two in a *quality function* (or *action-value function*).
Whereas section 3.4 presented evidence that the dopaminergic neurons can encode a TD-learning signal, Samejima et al. (2005) have shown that part of the striatum could subserve Q-learning. They indeed showed in a free choice task in monkey that many striatal neurons represent action-specific reward prediction, which can be related to the *quality* function used in the Q-learning algorithm.

3.6 Computational model-based systems in the rat prefrontal cortex

During this PhD thesis, we did not systematically review computational models implementing decision making within the prefrontal cortex, such as the one proposed by Daw et al. (2005). However, in order to prepare the section concerning models of the prefronto-striatal system's role in navigation, we will describe the model of Hasselmo (2005) which is particulartly dedicated to the *model-based locale* strategy.
Hasselmo proposed a model of cortical organisation based on minicolumns that he tested on a very simple navigation task in a discrete environment. The model creates a topological representation of the environment with three types of minicolumns: state, action and goal columns. State columns are associated to the possible states of the world. Action columns are associated to the action that can be performed in this world. Goal columns are associated to the reward the agent can get.
Before the task starts, all minicolumns that might be necessary are created and connections between them are initialised. All these minicolumns have the same architecture (Fig. 1.3.3): they are composed of 5 populations of cells in which each cell is a representation of a minicolumn of the network (hence there are as many cells in a population as the total number of minicolumns). Two of these populations, g_i and g_o, form the upper division of a minicolumn and two others, c_i and c_o form the lower division. A fifth population a codes for inputs that represent the agent situation (only the column corresponding to the current situation will have its a population activated). Figure 1.3.3 shows how a network of columns can use an algorithm close to activation-diffusion planning in order to trigger an action.

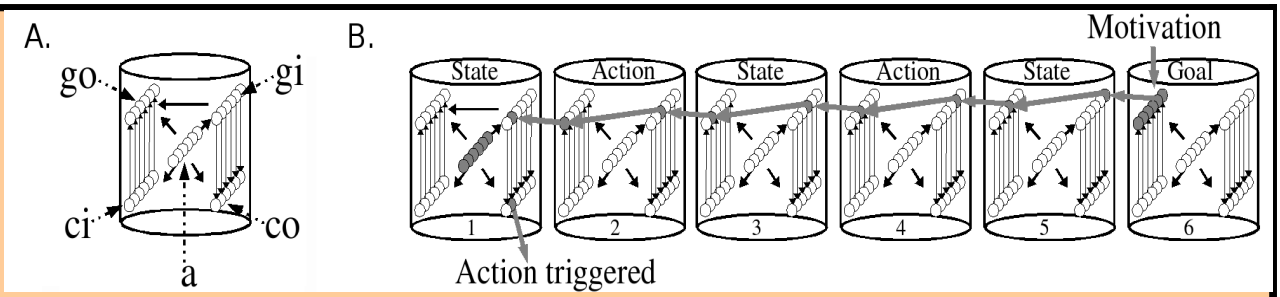


Figure 1.3.3: *A model of planning, adapted from Hasselmo (2005).* **A.** Populations of cells in Hasselmo's minicolumn. go, goal output cells; gi, goal input cells; co, state output cells; ci, state input cells; a, action cells. **B.** Functioning of a minicolumn network in Hasselmo's model. The upper divisions perform back-propagation and the lower division triggers an action. This model builds sequences of action-outcome associations. Then, the system can propagate both information concerning the position of the agent (here state 1) and it's the goal (i.e. its motivation) to deduce a path from the current position to the goal.

This phase is called the " retrieval phase ": a motivational signal is fed to the goal minicolumn (number 6 in the figure) and is back-propagated in the network until it reaches the minicolumn associated with the current state which a population is active (number 1 in the figure), an action is then triggered. It is important to notice that the signal goes through inter and intra-column connections. Inter-column connections learn the temporal relationships between two situations, and intra-column connections learn a short temporal sequence indicating which situations can precede and follow the current one (e.g. in minicolumn 3, there is a connection between cell 2 of g_o and cell 4 of g_i, it indicates that minicolumns 2,3 and 4 where activated sequentially).
These connections are reinforced during an encoding phase. At each time step, the network performs a retrieval phase to determine its next movement (decided randomly when none is triggered) and an encoding phase to refresh its connections.
Such a system can perform planning by propagating both information concerning the position of the agent and it's the goal to deduce a path from the current position to the goal. When the goal is changed, the system can plan a path towards the new goal without needing to re-learn action-outcome associations encoded in the model. Thus, this model enable flexible behaviors, similarly to other model-based system inspired by the prefrontal cortex (Dehaene and Changeux, 1997, 2000; Daw et al., 2005).

3.7 Analogy between model-based decision making and prefrontal activity

As mentioned in section 2.6, the prefrontal cortex shows an activity that can be related to goal-directed behaviors in general, and flexible action planning, both in the monkey (Sakagami and Niki, 1994; Tanji and Shima, 1994; Watanabe, 1996; Schultz et al., 1998; Carpenter et al., 1999; Procyk et al., 2000; Tanji and Hoshi, 2001; Matsumoto et al., 2003), and in rodents (Baeg et al., 2003; Mulder et al., 2003; Hok et al., 2005; Kargo et al., 2007). More specifically some parameters of model-based learning and decision making could be encoded in the prefrontal cortex, such as planned future actions (Mushiake et al., 2006), stored sequences of actions (Averbeck and Lee, 2007), goals (Hok et al., 2005; Genovesio et al., 2006), action-outcome associations (Matsumoto et al., 2003 ; Mulder et al., 2003; Kargo et al., 2007), and working-memory components (Baeg et al., 2003).
Based on these data, some authors have recently postulated that the prefrontal cortex can realize model-based reinforcement learning (Samejima and Doya, 2007). However, the precise algorithmic mechanisms that could be processed within the prefrontal cortex are not yet clear. Still, several models have been proposed to represent how the prefrontal cortex could learn a world model and plan actions based on expected consequences.
In the next section, we review several models implementing at least two navigation strategies in an

architecture inspired by the prefronto-striatal system. In mots cases, one of the navigation strategy is assumed by its authors as a model-based strategy and relies on the prefrontal cortical network, whereas the other strategy is model-free and relies on the striatum. These models coordinate strategies in two different manners: a set of models implement **strategy fusion**, where decision of the action to be performed by the agent is the result of a sum of candidate actions proposed by each strategy. Thus, these models do not require any strategy shifting mechanism since strategies are cooperating in decision-making.

The other set of models reviewed here implement a **strategy competition** mechanism to decide which strategy controls the agent's actions at any given moment. The latter models enable strategy shifting, as described below.

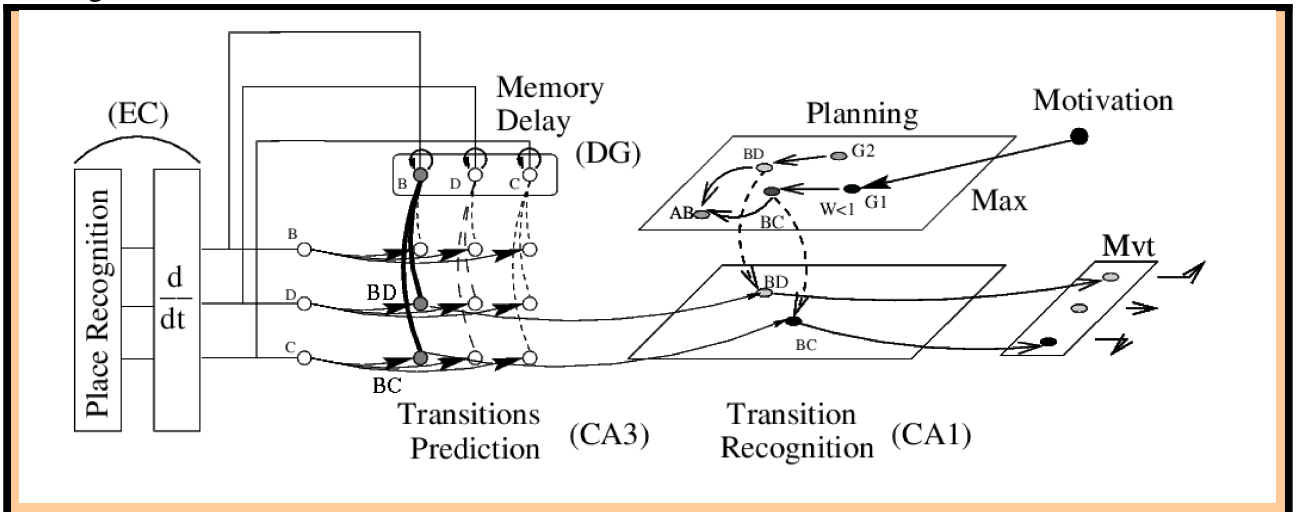


Figure 1.3.5 : *Model of Banquet et al. (2005).* The left part of the figure represent the hippocampal system implementing place recognition and associating places visited successively through *transition prediction*. CA1-CA3, brain regions forming the hippocampus proper; DG, dentate gyrus; EC, entorhinal cortex. The right part displays the decision making system. Top: cortical map employed for planning. Bottom: striatal movement selector. After recognizing that the agent can go from place *B* to place *D*, and from place *B* to place *C*, and after learning that place *C* is associated by goal *G1* satisfying the current motivation of the agent, the planning system selects transition *BC* and triggers a "turn right" movement. When the cortical planning map is disabled, transition recognition can directly trigger movements if a certain reinforcement learning process has been performed.

3.8 Computational models of navigation strategies in the cortico-striatal system

In this section, we review several recent models implementing different navigation strategies within different subparts of the cortico-striatal system. These models will capture our interest since they distinguish several navigation strategies, since they propose different roles of the cortico-striatal loops in the learning of these strategies, and since they incorporate mechanisms than enable ED strategy shifts mostly between place and cue-guided dimensions.

3.8.1 Models of strategy fusion

Gaussier et al. (1998, 2002), Banquet et al. (2005)

In this model, the implemented strategies are: 1) a place model-based strategy (« planning »); 2) a place model-free strategy.

The former involves a brain network including the hippocampal system and the prefrontal cortex. Within the hippocampal system, places are recognized and are associated though unsupervised Hebbian learning in order to build representations of transitions from places to places (figure 1.3.5). Then a cortical map enables the model to perform *planning* based on the agent's motivation.

Besides, the model-free strategy is computed within the striatum. The system does not implement place-recognition triggered responses but rather transition recognition-triggered responses.
The method for coordinating the two strategies is classified as “strategy fusion” since the planning network directly influences the model-free system. Movements performed by the agent are based on the sum of expected rewards computed by each strategy. Moreover, the system is able to deal with several motivations (hunger, thirst). This model was successfully applied to navigation within a T-maze and in robotics survival tasks.
Note that whereas the model implement a detailed architecture inspired by the hippocampal system, the basal ganglia part (including the striatum) is simplified.

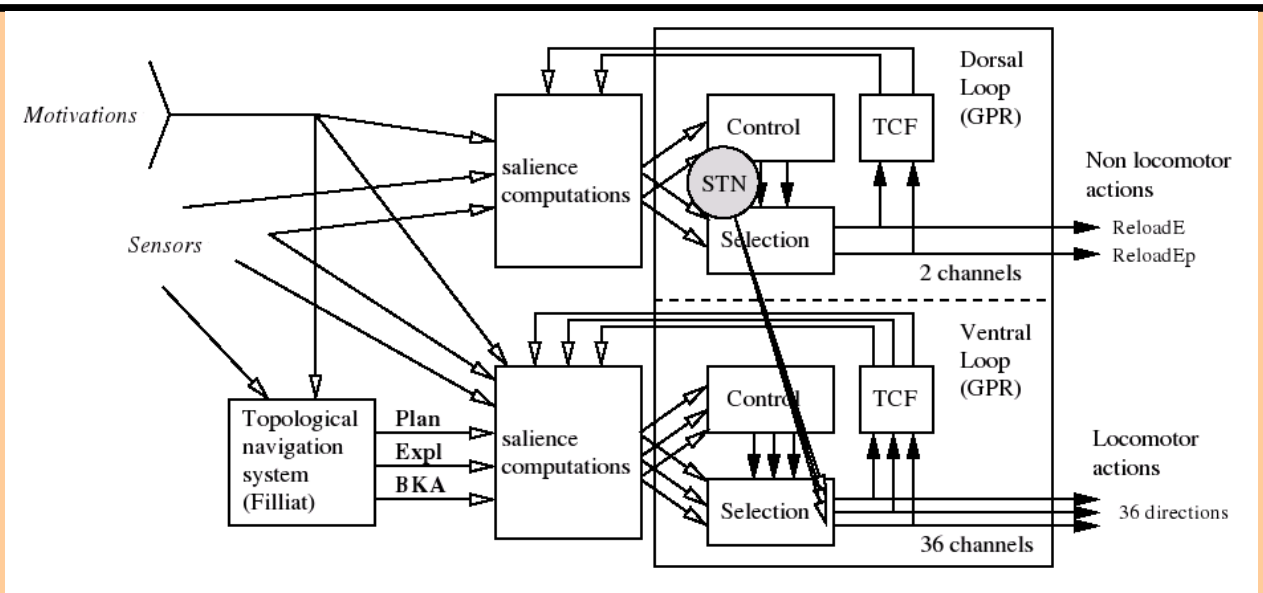


Figure 1.3.6 : *Model with two cortico-striatal loops. Adapted from Girard et al. (2005).* The model implements two basal ganglia loops, each one based on a biologically plausible model of the basal ganglia (i.e. the GPR model, Gurney et al., 2001b). The “dorsal loop” is responsible for triggering consumatory behaviors (ReloadE for reloading energy, ReloadEp for reloading potential energy). The “ventral loop” selects locomotor actions based on propositions subserved by two different navigation strategies. Both loops include several computational modules such as a selection module, a control module, and a thalamo-cortical feedback (TCF). The first strategy implemented is a hippocampo-prefrontal topological planning system. It systematically propose propositions of action based on three different motivations: plan a path towards a given goal; explore; back to a known area (BKA). The second strategy is a “target approach” equivalent to a visual model-free system, and is subserved by direct sensory and motivational inputs to the “salience computations” module. Decisions made by each strategy system are merged within the ventral loop to subserve strategy fusion. Loops coordination relies on the subthalamic nucleus (STN) which prevents from selecting locomotor actions when the agent is consuming a reward.

Girard (2003); Girard et al. (2004, 2005)
In this model, the implemented strategies are: 1) a place model-based place strategy (« topological navigation »); 2) a visual model-free strategy (“target approach”). Similarly to Banquet et al. (2005)'s model, the former strategy is assumed to rely on the hippocampo-prefrontal system, whereas the former is implemented in the striatum (more precisely in the ventral striatum). In contrast with the previous model, Girard et al. (2005) implemented a detailed biologically plausible model of the basal ganglia (Gurney et al., 2001a,b; Humphries and Gurney, 2002; Girard et al., 2002, 2003), whereas the hippocampo-prefrontal system is much less detailed.
Interestingly, different cortico-striatal loops implemented in the model do not represent different

navigation strategies, but rather distinguish locomotor actions (resulting from all navigation strategies) from non-locomotor actions which enable the robot to stop at resources and to consume rewards (figure 1.3.6).
Similarly to the previous model, the system implement a strategy fusion mechanism. The model was successfully applied to a T-maze task, to reproduce opportunistic behavior, danger avoidance, and to a robotics survival task.
The system can also deal with several motivations. However, no learning mechanism is implemented at the level of striatal action selection. Rather, hand-tuned synaptic weights determine how stimuli influence action selection. One of the goals of the modelling work presented in chapter 2 section 3 is to solve this issue by using Temporal-Difference Learning to autonomously adapt the cortico-striatal synaptic weights.

Guazelli et al. (1998)
In this model, implemented strategies are: 1) a place model-based strategy (« world graph »); 2) a visual model-free strategy ("taxon navigation"). The former involves the parietal and prefrontal cortices. The latter involves the basal ganglia.
In contrast to the two previous models, decisions taken by prefrontal model-based system are not executed by the striatum. Rather, they are combined within the premotor cortex with decisions taken by the basal ganglia model-free system (figure 1.3.7). Action selection is based on the sum of expected rewards computed by each strategy, thus also implementing strategy fusion.
The model was used to simulate gradual/sudden choice changes in a T-maze reversal task in fornix lesion experiments (O'Keefe, 1983).

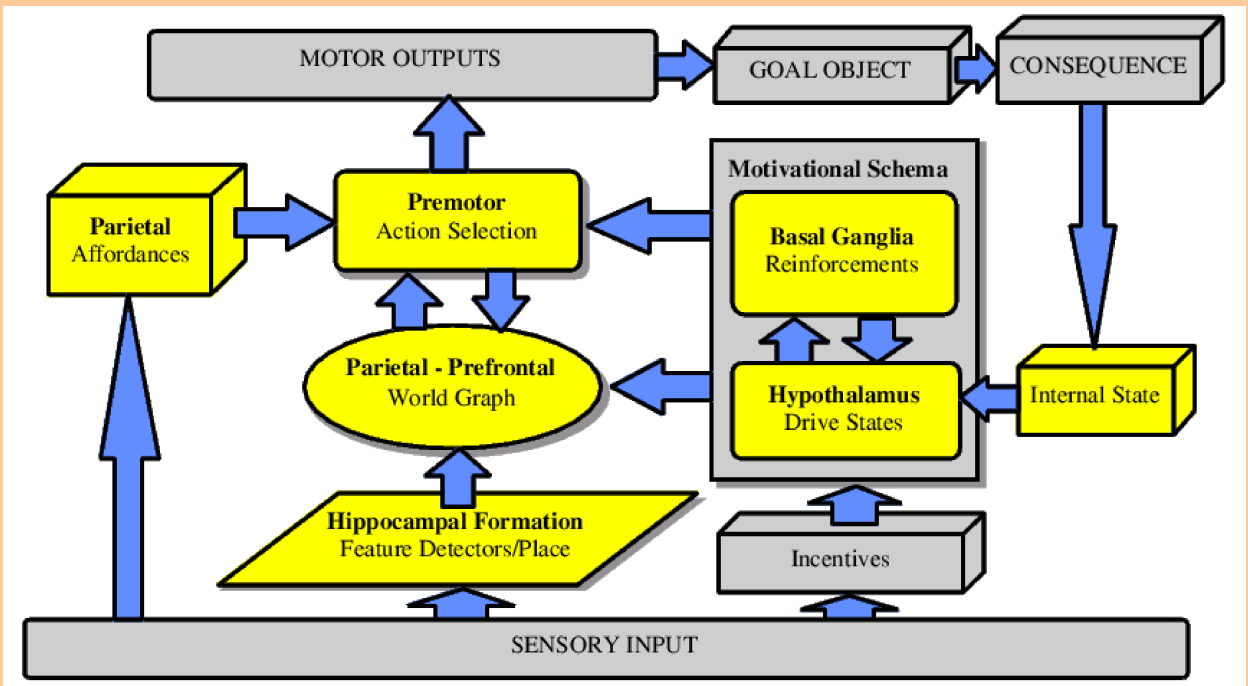


Figure 1.3.7 : *Model combining a world graph (model-based) strategy and a visual model-free strategy. Adapted from Guazelli et al. (1998).* The yellow boxes represent neural networks inspired by several brain structures. The hippocampo-parieto-prefrontal system implements model-based navigation by planning within a "world graph". The basal ganglia implements model-free reinforcement learning. Action selection is subserved within the premotor cortex based on a fusion of decisions taken by different strategies.

3.8.2 Models of strategy competition

Chavarriaga (2005); Chavarriaga et al. (2005a,b)

In this model, the implemented strategies are: 1) a place model-free strategy (« locale »); 2) a visual model-free strategy ("taxon"). The former involves the hippocampo-ventral striatal system, whereas the latter is subserved by the dorsolateral striatum (figure 1.3.8).

In contrast to the previous models, this system implement a strategy selection mechanism, that is, at any given moment, decisions relies on only one navigation strategy. The system learns to associate a strategy to a given context based on the reward prediction by each strategy. As a consequence, the shifting mechanism depends on the animat's perceptions and on reward expectations. A persistence mechanism decreases the probability to perform two strategy shifts in less than 100 trials.

Interestingly, among the models reviewed here, it is the only one where strategies employ the same learning mechanism based on dopaminergic TD-learning.

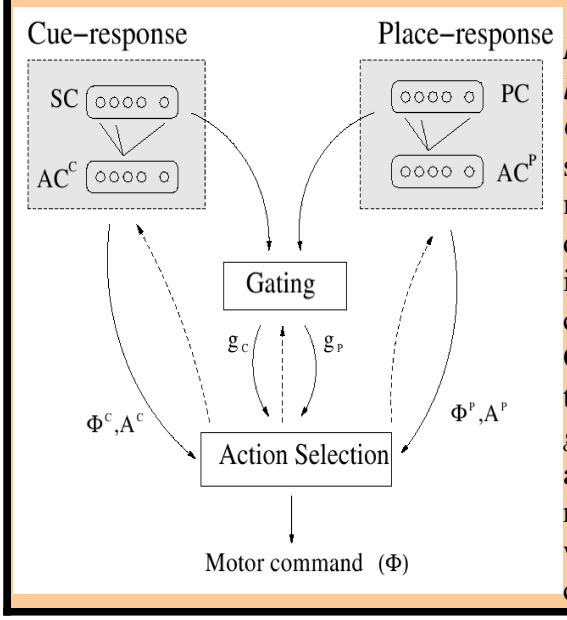


Figure 1.3.8 : *Model involving two cortico-striatal loops for* locale *and* taxon *navigation. Adapted from Chavarriaga et al. (2005a).* The two loops have the same structure and involve the same learning mechanism: a dopamine-based TD-learning. The difference between the two loops relies on the input information triggering actions. On the left, visual cues constitute the input, thus performing *taxon* navigation. On the right, *place* information constitute the input, thus performing *locale* navigation. In the center, a *gating network* decides which strategy controls the agent's behavior at any given moment based on the reward predicted by each strategy, and on a memory of which strategy was previously associated to the current context.

The model manages to produce intra-trial strategy shifts in response to a task change in a Morris water maze task (Devan and White, 1999), which task was described in section 1.7, as well as progressive shift from place to visual within a constant task in another Morris water maze task (Pearce et al., 1998). It also well reproduces rats' bias towards one strategy when either the hippocampus or DLS are lesioned (Pearce et al., 1998; Packard and Knowlton, 1992).

Daw et al. (2005)

As we already described this model in section 1.6, we will only mention here important features that contrast with models above. Daw et al. (2005)'s model also implements strategy selection. However, in contrast to Chavarriaga et al. (2005a), one of the two strategies implemented is a model-based one. It implements: 1) a visual model-based strategy (« tree-search »); 2) a visual model-free strategy. The former involves the prefrontal cortex, the accumbens core (ventral striatum) and the dorsomedial striatum (DMS). The latter involves the dorsolateral striatum (DLS) and the accumbens shell (ventral striatum).

The system implements a strategy shifting mechanism: the most reliable strategy (based on a measure of Bayesian uncertainty) selects actions to be performed by the agent. The model successfully reproduced devaluation effects on extinction in a lever-pressing task.

3.8.3 Conclusion

The models described above all consider the prefronto-striatal system and implement different

behavioral strategies. Interestingly, they have in common to involve part of the striatum in model-free strategies. Moreover, most of them involve the prefrontal cortex in high-level decision making relying on model-based learning processes. They differ on the role of the ventral striatum which is in some case responsible for strategy fusion (Banquet et al., 2005; Girard et al., 2005), in another case for model-free navigation (Chavarriaga et al., 2005a), and in a last one for subserving model-free reinforcement learning (Guazelli et al., 1998; Daw et al., 2005). None involve an explicit prefrontal mechanism for strategy shifting.
Finally, let's briefly mention that several other models were proposed which involve different prefronto-striatal loops in different behavioral strategies, not necessarily for navigation, yet in primates. Some of these models were proposed for saccade generation (Dominey and Arbib, 1992; Arbib and Dominey, 1995; Brown et al., 2004), for the generation of visuo-motor sequences (Nakahara et al., 2001), or for advantageous stimulus-action-reward performance such as button-push in response to visual stimuli (Haruno and Kawato, 2006).

The experimental work that will be described in the following chapters will contribute to:
1) (Chapter 2) clarifying the role of the ventral striatum in model-free learning by:
(a) (section 2) analysing electrophysiological data recorded in the VS of rats performing a reward-seeking task in a plus-maze (Khamassi et al., paper submitted to *J Neurophysiol*, in revision);
(b) (sections 3 and 4) designing an Actor-Critic model of S-R learning where VS is the Critic which drives learning, whereas DLS is the Actor which memorizes S-R associations. This model is applied to robotics simulations, and compared with existing models in a virtual plus-maze; (Khamassi et al. 2005 *Adaptive Behavior*, 2006 *SAB06*);
2) (Chapter 3) studying the role of mPFC in strategy shifting by means of electrophysiological recordings in the mPFC of rat performing a task requiring such kind of shifts. (Khamassi et al., paper in preparation)
Following hypotheses emerging from the neurobiological literature, we aim at finding reward anticipatory activity in VS, that could confirm the involvement of VS in a *Critic* subserving TD-learning for model-free strategies. Moreover, we expect to find neurons in the mPFC detecting task changes and other neurons showing correlates with the current strategy performed by the animal. These neurons could participate in a strategy shifting mechanism.

CHAPTER 2 : ROLE OF THE VENTRAL STRIATUM IN LEARNING CUE-GUIDED MODEL-FREE STRATEGIES

1. Introduction

The objectives of the experimental work presented in this chapter are: 1) to better understand neuronal activity of the ventral striatum in reward-based learning; 2) to study the efficiency of models of reinforcement learning inspired by the striatum in a simulated robotics task; 3) to benefit from the pluridisciplinary approach of the subject. That is, on the one hand, to improve models based on our neurophysiological data, on the other hand, to make predictions on functional mechanisms to biology based on our models' simulations.

Three different studies are presented here:

- The first is an electrophysiological study of the rat ventral striatum in a reward-seeking task in a plus-maze. The aim of the study is to test if neuronal activity in the rat ventral striatum demonstrates reward anticipations compatible with the Actor-Critic theory for learning a cue-guided navigation strategy (corresponding to a cue-guided *model-free* strategy explained in the previous chapter).
- The second compares the efficiency of computational principles extracted from several Actor-Critic models in a simulated robotics version of the plus-maze task. On the one hand, the aim is to reproduce rats behavioral performance in solving the task. On the other hand, the study analyses how these principles can integrate within a biologically plausible model of the basal ganglia.
- The last proposes a new method to improve the performance of Actor-Critic models in simulated robotics that consist in combining *self-organizing maps* with a *mixture of experts* in order to automatically adapt several Actor-Critic submodules, each module being an expert trained in a particular subset of the task.

Each of these three works will be presented in the form of articles that are published or submitted, and will be preceded with a short presentation and summary of methods and results:

- Khamassi, M., Mulder, A.B., Tabuchi, E., Douchamps, V., and Wiener, S.I. Actor-Critic models of reward prediction signals in the rat ventral striatum require multiple input modules. *Submitted to Journal of Neurophysiology, in revision.*
- Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., and Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior, Special Issue Towards Artificial Rodents*, 13(2):131-148.
- Khamassi, M., Martinet, L.E., and Guillot, A. (2006). Combining self-organizing maps with mixture of experts: Application to an Actor-Critic model of reinforcement learning in the basal ganglia. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.A., Miglino, O., Parisi, D. (Eds), Proceeding of the Ninth International Conference on the Simulation of Adaptive Behavior, SAB06, Lecture Notes in Artificial Intelligence, pp. 394-405, Springer-Verlag.

Related works done during this thesis (in appendix) consist in: a preliminary comparison of Actor-Critic models in a simulated robotics task (Khamassi et al., 2004); two articles concerning the Psikharpax project, the artificial rat whose control architecture integrates our Actor-Critic models with other models (Filliat et al., 2004; Meyer et al., 2005); a poster presenting a model coordinating

two competing navigation strategies controlled by two cortico-striatal loops in a water maze task (Dollé et al., 2006)

2. Critic-like reward anticipation in the rat VS

Khamassi, Mulder, Tabuchi, Douchamps and Wiener (submitted to Journal of Neurophysiology).

2.1 Summary of objectives

The aim of this study is to test if neuronal activity in the rat ventral striatum could be involved in reward anticipation activity compatible with the Actor-Critic theory. As mentioned in the previous chapter, there is now strong evidence in primates that the ventral striatum can subserve the role of a *Critic* by anticipating reward, and using reward expectations to modulate the release of dopamine reinforcement signal in the striatum (Cromwell and Schultz, 2003; O'Doherty et al., 2004). This mechanism is assumed to rely on the *Temporal Difference Learning* algorithm, in which temporally consecutive reward predictions are compared in order to enable learning based on reward in the far future (Schultz et al., 1997). Figure 2.2.1 summarizes this hypothesis.

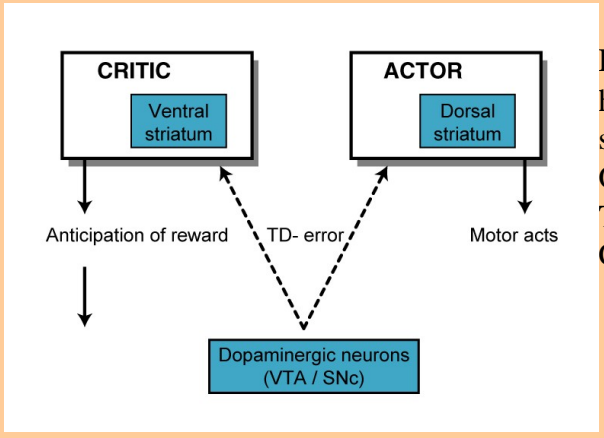


Figure 2.2.1 : Schematic representation of the hypothesized role of the dorsal and ventral striatum in reinforcement learning, based on O'Doherty et al. (2004)'s data. VTA: Ventral Tegmental Area; SNc: Substantia Nigra Pars Compacta.

However, in the rat, it is less clear if the ventral striatal activity can be similar to a Critic since, in previous protocols, anticipatory responses are difficult to discriminate from spatial or behavioral correlates (Lavoie and Mizumori, 1994; Chang et al., 1997; Miyazaki et al., 1998; Martin and Ono, 2000; Setlow et al., 2003; Nicola et al., 2004; Wilson and Bowman, 2005). So we used a novel experimental design that can dissociate these different types of activity.

2.2 Summary of methods

Seven rats implanted with electrodes in the ventral striatum (ventromedial caudate, medial accumbens core and dorsomedial accumbens shell) were recorded while searching for different volumes of reward at the respective arms of a plus-maze. Multiple rewards were provided at 1 s intervals while the rat remained immobile. Neuronal responses were analysed during a window starting 1 second before deliverey of the first droplet reward until 1 second after the last.

2.3 Summary of results

We found neurons discharging phasically prior to each droplet of water, both when the rat approached or was immobile at the goal, demonstrating that this activity is predictive. Strikingly, this activity often reappeared after the final droplet was delivered, while the rat was still immobile, as if in anticipation of yet another reward.

We replicated these neuronal activities by simulating a multiple-module Actor-Critic model whose originality resides in providing different input information to each module. Four modules independently process the same TD-learning algorithm based upon a different mix of spatial and temporal inputs. The spatial information corresponds to the position (i.e., location of the respective maze arms relative to one another and the room) and the available sensory cues, such as cue lights at reward sites. The temporal information corresponds to the addition or suppression of the *'complete serial compound stimulus'* component proposed by Montague et al. (1996).

The co-existence of cells' responses corresponding to different modules in the model suggests a multiple-input model of the rat ventral striatum.

Interestingly, whereas Schultz and colleagues reported similar stereotyped activity in all dopaminergic neurons recorded (Schultz, 1998), one of the predictions of our model is that different sub-groups of brainstem dopaminergic neurons would be anatomically connected to respective TD-learning modules, and would, in the same plus-maze task, exhibit differential responses to reward (see figure 9 in the paper): a set of dopamine neurons responses to reward should vanish as in the seminal study of Schultz et al. (1997); another set of dopamine neurons related to the TD-learning module which erroneously anticipates an additional droplet of water in our model should have negative responses. In the latter case, the model's prediction constitutes an interesting situation where the negative response to reward could have been intuitively interpreted as contradictory to the TD-learning theory, whereas indeed, this response is predicted by the slightly different version of the TD-learning model proposed here.

2.4 Discussion

We found phasic responses that anticipates a sequence of consecutive water rewards, independently from the rat behavior. **These results agree with the hypothesis that neurons of the rat ventral striatum could participate in the role of a Critic in the framework of the TD-learning theory**. Furthermore, the regular timing of these anticipatory reward responses in the absence of any explicit trigger stimulus suggests that these neurons have access to some kind of internal clock signals. However, our experimental design was not conceived to precisely study this timing mechanism.

Interestingly, the reward expectation information reported here could be provided to the ventral striatum by the orbitofrontal cortex (OPFC). The latter has been suggested to code the motivational value of environmental stimuli, and OPFC neurons were recently found to code reward expectancy, regardless of reward magnitudes, in an olfactory discrimination « go / no-go » task (van Duuren et al., 2007).

Another interesting point to note is that anticipatory responses to reward are not the only type of responses that we found in the rat ventral striatum. Previous results recorded in the same experiment show goal-approach neurons (Khamassi, 2003; Mulder et al., 2004). The latter constitute a population of cells whose activity cut the behavioral sequence performed by the rat in subparts such as « from departure to maze center », « from maze center to goal », « from half of departure arm to goal », ...etc. Moreover, 25% of these neurons were spatially modulated, which means for example that one neuron would respond from departure to maze center only when the rat started from arm #2 or #3.

The coexistence within the ventral striatum of *goal-approach neurons* and *reward anticipation neurons* would be consistent with previous results from the literature (see section 1.2.3). As discussed before, the former is consistent with the hypothesis that the core participates in goal-directed learning (Dayan and Balleine, 2002; Cardinal et al., 2002) – corresponding to *model-based* strategies explained in the previous chapter; whereas the latter is consistent with the hypothesis that the shell mediates learning of reactive and procedural navigation strategies (Dayan, 2001; Corbit et

al., 2001) – corresponding to *model-free* strategies. However, we did not find consistent differences between shell and core neurons. Both kinds of correlates (goal-approach and reward anticipation) were found in shell and in the core. Thus this is rather consistent with data stressing an anatomical and functional continuum between core and shell (Heimer et al., 1997; Ikemoto, 2002; see Voorn et al., 2004 for a review)

Finally, our task was not designed to distinguish different navigation strategies. Indeed, in our task, rats had to perform a mix between visually-guided and spatial strategies: they had to memorize different volumes of reward located in space, and simultaneously, they had to recall that only rewards signalled by a light cue were available. Thus, we cannot study whether different groups of reward anticipation neurons – like those we reported – are responsible for different navigation strategies within the ventral striatum. Further investigations will be required to answer this question.

Title: Actor-Critic models of reward prediction signals in the rat ventral striatum require multiple input modules

Preprint, submitted to Journal of Neurophysiology (June 2007)

Authors and author addresses:
Mehdi Khamassi[1,2,§], Antonius B. Mulder[1,§], Eiichi Tabuchi[1], Vincent Douchamps[1] and Sidney I. Wiener[1]
[1]CNRS-Collège de France Laboratoire de Physiologie de la Perception et de l'Action
UMR-C7152, 11 pl. Marcelin Berthelot, 75231 Paris Cedex 05 France.
[2]Université Pierre et Marie Curie – Paris 6, CNRS FRE 2507, ISIR, Paris, France
[§] MK and ABM contributed equally to this work
Running head: Striatal cells predict rewards like TD learning Critics

Corresponding author: Sidney I Wiener, CNRS-Collège de France LPPA, 11 pl. Marcelin Berthelot, 75231 Paris Cedex 05, France. Telephone 33-1-44271621; Fax 33-1-44271382; Electronic mail: sidney.wiener@college-de-france.fr
9 Figures, no tables, 2 Supplementary Figures
35 text pages

Abstract

The striatum is proposed to play a vital role in learning to select appropriate actions optimizing rewards according to the principles of 'Actor-Critic' models of trial-and-error learning. The ventral striatum (VS), as Critic, would employ a Temporal-Difference (TD) learning algorithm to predict rewards and drive dopaminergic brainstem neurons. In previous studies reporting anticipatory responses in the rat VS, the experimental protocols did not control for possible confounds with spatial or behavioral correlates; thus these data fail to provide strong support for Actor-Critic models. Hence here we used a novel experimental design where, in rats searching for different volumes of reward at the respective arms of a plus-maze, multiple rewards were provided at 1 s intervals while the rat remained immobile. Neurons discharged phasically prior to each droplet of water, both when the rat approached or was immobile at the goal, demonstrating that this activity is predictive. In different neurons, the anticipatory activity commenced from 800-200 msec prior to rewards and this activity could be greater for early, middle or late droplets in the sequence. Strikingly, this activity often reappeared after the final droplet was delivered, as if in anticipation of yet another reward. Basic TD learning models cannot replicate this rich variety of anticipatory responses. Thus we developed a new model with multiple modules, the originality of which resides in modules processing different mixes of input information with different 'discount factors' (accounting for future rewards). This TD learning variant thus constitutes a more biologically plausible neural substrate for reinforcement learning.

Keywords: Accumbens, TD learning, caudate, reinforcement learning, dopamine

Introduction

The prefrontal cortex-basal ganglia loop has been identified as instrumental for orchestrating behavior by linking past events and anticipating future events (Fuster, 1997; Otani, 2004). It is proposed to enable learning mechanisms for goal-directed behaviors, particularly those requiring chaining of sequences of behaviors orders of magnitude greater than the time scale of postsynaptic events (Joel et al. 2002).

For example the striatum is hypothesized to organize action sequences leading to habit formation (Graybiel, 1998). Indeed some striatal neurons are selectively active in the successive actions comprising goal directed behaviors (e.g., Kawagoe et al.. 1998; Itoh et al. 2003; Mulder et al. 2004; Schmitzer-Torbert and Redish 2004), and yet others fire in relation to reinforcements including food, drink, habit-forming drugs, and intracranial electrical stimulation (Hikosaka et al. 1989; Schultz et al. 1992; Wiener 1993; Lavoie and Mizumori 1994; Miyazaki et al. 1998; Martin and Ono 2000; Shibata et al. 2001; Daw et al. 2002; Cromwell and Schultz. 2003; Takikawa et al. 2002; Nicola et al. 2004; Wilson and Bowman 2005). By virtue of their projections to brainstem dopaminergic (DA) neurons (Houk et al. 1995; Schultz et al. 1997) the latter signals could help resolve the classic 'credit assignment problem', namely, how to strengthen connections which were active thousands of milliseconds prior to when the reward was received and thus outside the time window of conventional synaptic plasticity mechanisms.

Temporal difference (TD) learning (Sutton and Barto, 1998) models of these reinforcement learning mechanisms engage an 'Actor' and a 'Critic'. As Sutton (1997) explains: "Actor-critic methods are TD methods that have a separate memory structure to explicitly represent the policy independent of the value function. The policy structure is known as the *actor*, because it is used to select actions, and the estimated value function is known as the *critic*, because it criticizes the actions made by the actor." Numerous studies have attempted to identify the brain areas corresponding to these roles. Whereas neurons coding actions, for example in the dorsal striatum, would play the role of the Actor, the identity of the Critic is controversial (Joel et al. 2002). Candidates include dorsal striatal striosomes, ventral striatum, and prefrontal cortex, all with neurons with apparently anticipatory activity and sending projections to DA neurons, which would then transmit the error prediction signal.

Although there are now numerous data in monkey (Schultz) and in human (O'Doherty et al. 2004) supporting the hypothesis that the ventral striatum show Critic-like reward anticipation activity, there is a lack of evidence in the rat striatal recording literature since, in the experimental designs employed until now, anticipatory neural responses could be confounded with activity associated with reward-directed behaviors. Thus we recorded ventral striatal neurons in rats as they approached goals and also rested immobile awaiting successive rewards presented at 1 s intervals. This revealed anticipatory responses, some selective early, middle, late parts of the reward sequence. Strikingly, these neurons also discharged in anticipation of yet another reward after the final one. We demonstrate that these responses are indeed compatible with the TD learning model but only when endowed with a novel capacity to engage multiple modules which process temporal or spatial inputs with differing weights.

Materials and methods

Animals and apparatus

Seven Long-Evans male adult rats (220 to 240 g) were purchased (from the Centre d'Elevage René Janvier, Le Genest-St-Isle, France) and kept in clear plastic cages bedded with wood shavings. The rats were housed in pairs while habituating to the animal facility environment. They were weighed and handled each work day. Prior to training they were placed in separate cages and access to water was restricted to maintain body weight at not less than 85% of normal values (as calculated for animals of the same age provided *ad libitum* food and water). The rats were examined daily for their state of health and were rehydrated at the end of each work week. This level of dehydration was necessary to motivate performance in the behavioral tasks, and the rats showed

neither obvious signs of distress (excessive or insufficient grooming, hyper- or hypo-activity, aggressiveness) nor health problems. The rats were kept in an approved (City of Paris Veterinary Services) animal care facility in accordance with institutional (CNRS Comité Opérationnel pour l'Ethique dans les Sciences de la Vie), national (French Ministère de l'Agriculture, de la Pêche et de l'Alimentation No. 7186) and international (US National Institutes of Health) guidelines. A 12 hr/12 hr light/dark cycle was applied.

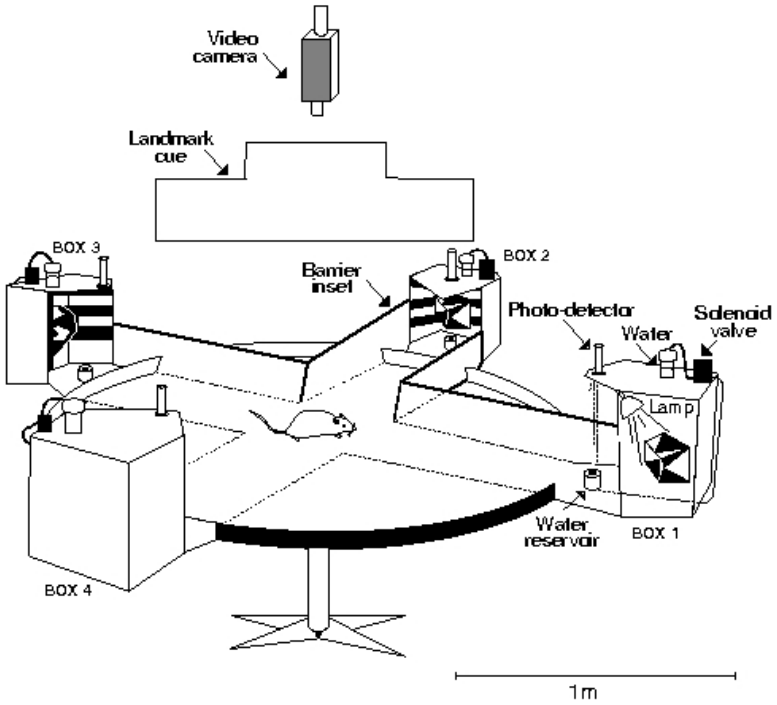


Figure 1. *The experimental apparatus. The rat performed the behavioral task on a 180 cm diameter platform with a low border. Barriers placed on the platform (dashed lines) restricted the movements of the rats to four alleys. Four reward boxes (30 x 30 x 30 cm) were attached to the edge of the platform and were equally spaced and oriented toward the corners of the experimental room. Each box contained identical, highly contrasted polyhedrons suspended in front of a striped background. Each reward box could be illuminated independently under computer control. The main sources of illumination in the experimental room were the lamps directed towards the salient cues in the reward boxes, the overhead lamp and two miniature lamps on the headstage of the rat. (Adapted from Tabuchi et al., 2000).*

Training and experiments took place in a four arm 'plus' maze. The arms were 70 cm long and 30 cm wide with 40 cm high sloped black walls while the center was a 30 x 30 cm square. This was placed in a darkened square room (3 x 3 m) bordered by opaque black curtains (Figure 1). At the end of each of the four arms was an alcove (30 x 30 x 30 cm) containing a water reservoir and a large highly contrasted, three-dimensional visual cue. The cues were identical in each of the boxes but could be illuminated independently. Room cues included a wide inverted-T shaped white poster board (185 x 60 cm) as well as a white rectangular box (56 x 25 cm), each mounted 70 cm from the platform on walls respectively opposite or adjacent to the entrance of the curtained area. The poster board was spotlit by a ceiling-mounted incandescent lamp (60 W) during both training and recording sessions.

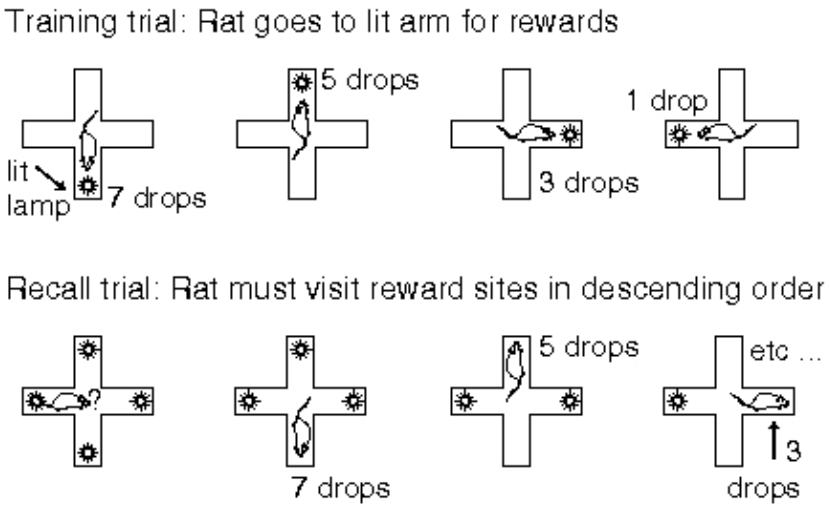


Figure 2. *The experimental task. First the rats performed a series of training trials where the correct choice was guided by the lit cue lamp in the appropriate reward box. Each trial comprised a sequence of visits to the four reward boxes providing 7, 5, 3 and 1 droplets of water. During recall trials all cue lamps were lit, then were turned off one by one as the rat visited the reward boxes in the same order of descending reward value. Reward values were then re-assigned for the second half of the session, and were also changed daily. (Adapted from Tabuchi, et al., 2000).*

Each reward box was equipped with automated water delivery and infrared photo-emitter/detector systems. At the entry of each reward alcove stood a short (3 cm high) cylindrical block (the 'water reservoir'). Tubing transported water from elevated bottles to computer-controlled solenoid valves that in turn led to each water reservoir. When the rat arrived at the water trough and blocked the photobeam, the computer triggered release of the water reward(s) there. The volume of the water droplets was calibrated to 30 µl by regulating the time that the solenoid valves remained open. Multiple droplets of water were provided at 1 s intervals. The solenoid valves made an audible click when opening and closing. The times of the photobeam occlusions as well as solenoid valve openings were recorded as event flags in the data file. Photodetectors also registered when the rat arrived at the center of the maze.

The differentially rewarded plus maze task (Figure 2)

Details of the task and training protocols may be found in Tabuchi et al (2000, 2003). In each session the rats were exposed to a novel distribution of different reward volumes at the four respective arms of the maze and then were required to recall the sequence in order of decreasing volume. After this, the reward distribution was changed and a second series of trials were run while recording the same cells.

In the *training phase*, reward availability was signaled by cue lamps in the reward boxes. The rat was thus cued to go in order to the respective boxes that provided 7, 5, 3, or 1 droplets of water. For the multiple rewards, the successive droplets of water were delivered at 1 s intervals while the cue lamp remained lit. After the rat consumed the water it returned to the center of the maze and the lamp on the next arm was then lit automatically.

In the *recall phase,* all reward alcoves were illuminated, and turned off successively as the rats visited them in order of descending reward value. The task design exploited the tendency for rats to prefer locations with greater rewards (e.g., Albertin et al. 2000). If the rat entered an arm out of sequence, all cue lamps were turned off and the same lamps were lit again when the rat returned to the maze center. The rats only very rarely continued to the end of the arm in these cases, and thus there was insufficient data to analyse error trials.

Electrode implantation and recordings

Electrodes were surgically implanted ater the performance level exceeded 70% correct

(rewarded) visits (usually after 4 to 6 weeks of training). The rat was returned to ad lib water, tranquillized with 0.1 ml of 2% xylazine (i.m.) and anesthetized with 40 mg/kg pentobarbital intraperitoneally. Two bundles of eight 25 μm formvar-insulated nichrome wires with gold plated tips (impedance 200-500 kΩ) were stereotaxically implanted. Each bundle was installed in a guide tube (a 30 gauge stainless steel cannula) and mounted on one of two independently advanceable assemblies on a single headstage (Wiener, 1993). A ground screw was installed in the cranial bone. One group of electrodes was placed above either the ventrolateral shell region of Acb (AP 10.7 to 11.2, ML 1.7 to 2.2), or the medial shell of Acb (AP 11.2 to 11.6, ML 0.7 to 0.9). The second bundle was placed above the hippocampus (data reported in Tabuchi et al. 2000, 2003). About one week later, after complete recovery from the surgery, water restriction and training were resumed. The screws of the advanceable electrode drivers were gradually rotated daily until neurons were isolated (the drivers advanced 400 μm for each full rotation); then multiple single units were recorded as the rat performed the tasks. The electrodes were advanced at least 3 hr prior to recording sessions to promote stability.

Electrode signals passed through FETs (field effect transistors), then were differentially amplified (10,000 x) and filtered (300 Hz to 5 kHz, notch at 50 Hz). Single unit activity was discriminated post-hoc with DataWave software, where single unit isolation was performed using 8 waveform parameters (positive, negative and entire spike amplitude, spike duration, amplitude windows immediately prior to and after the initial negative-going peak, and time until maximums of positive and negative peaks) on the filtered waveform signals. Isolation was confirmed in interspike interval histograms which had, on average, only 0.3% occupancy of the first 3 ms bins corresponding to the refractory period. Waveforms are presented in Supplementary Figure 1 and as insets to raster and histogram figures. Putative fiber responses were identified by extremely short spike durations (on the order of 0.1 ms) and by distinctive waveform characteristics - these were discarded from analyses.

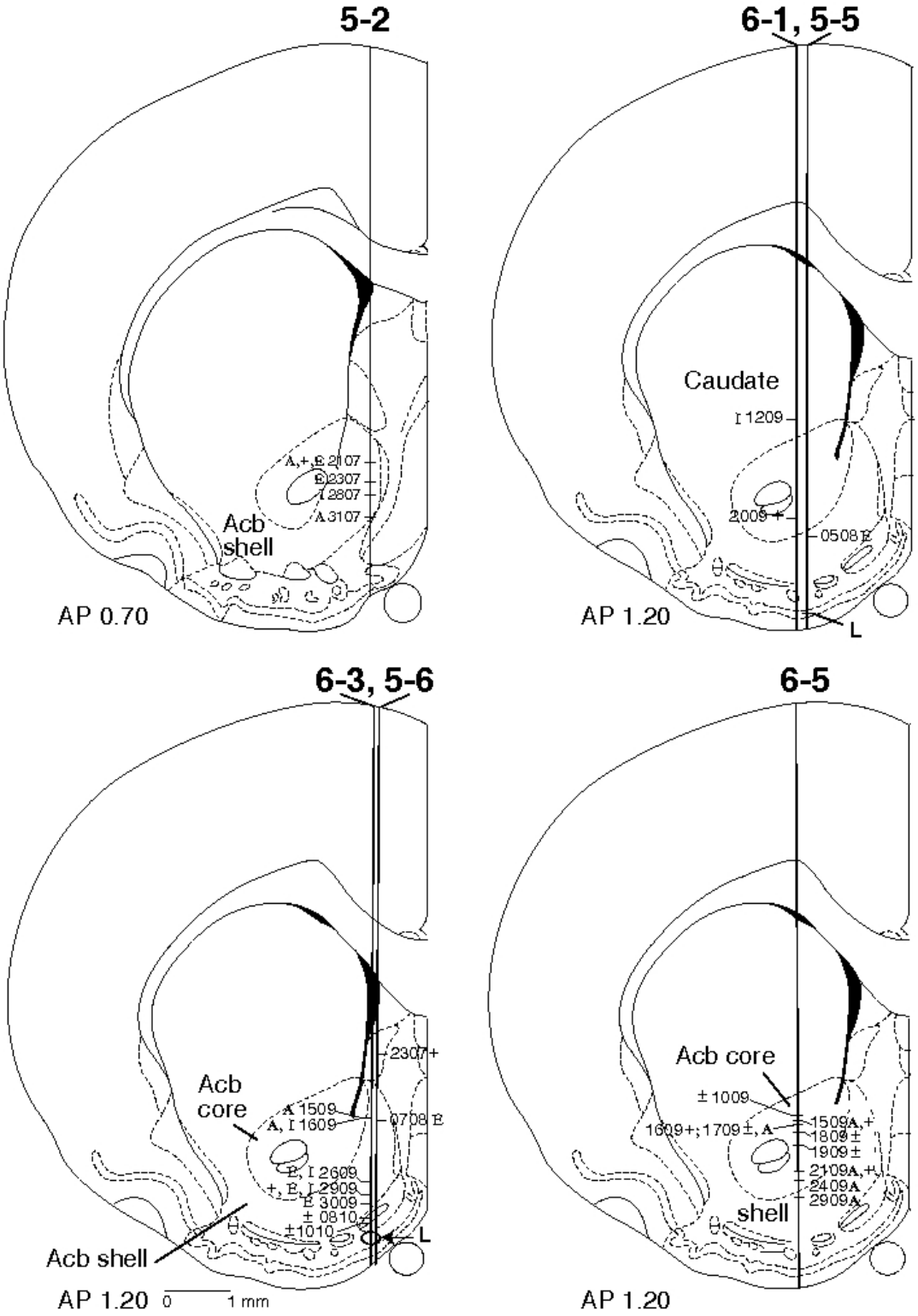
Two small lamps (10 cm separation) were mounted 10 cm above the headstage. Reflectors were attached to the rostral lamp to aid the tracking system in distinguishing it from the caudal lamp. The two lamps were detected with a video camera mounted above the platform and transmitted to a video tracking system (DataWave, Longmont, CO, USA) and a video monitor. All of the action potential (digitized waveforms and timing) and behavioral (position of the animal, photobeam crossings, water delivery) data were simultaneously acquired on a personal computer with software operating under DOS (DataWave, Longmont, CO, USA).

In preparation for recording sessions, the rat was placed in a cage with transparent plastic walls (and no wood shavings) then brought into the experimental room. The recording cable was attached to the headstage and the rat was placed in a cubic cardboard box (with sides ~40 cm). Then the electrode recording channels were examined for signs of discriminable neuronal activity. If this was successful, the data acquisition system was initialized and the lamp assembly was attached. The rat was then placed in the experimental apparatus where the lamp at the first reward box was already lit. No attempts were made to disorient the rat, and the lengthy training period assured that the environment was familiar. The rats always immediately started performing the task.

The neuronal discharge data are described strictly in terms of their synchronization with reward deliveries. Thus no error trials are included in analyses since no rewards were delivered then. Sessions usually lasted about 20 minutes.

Figure 3 (next page). *Reconstruction of recording sites on the basis of histological preparations. Animal identification numbers appear above respective electrode tracks. Recording sites are marked by cross bars and numbers. Neurons are identified according to the following code: A - anticipatory responses for individual droplets of water, E - uniform increase in firing rate during drinking, I - inhibition during drinking, + - excitatory response for first droplet only, ± - Excitation and inhibition during first droplet, L – lesion site. Multiple single neurons recorded at the same site are separated by commas. Histological analyses showed tracks in animal 6-2 were indeed in*

ventral striatum but sites could not be reconstructed with precision (data not shown). (Figure templates adapted from Paxinos and Watson 1998 with permission).



Data analysis

Data from all recorded neurons with average firing rates greater than 0.1 Hz during the experiment were submitted to statistical analyses. The synchronization point for analyses of cell activity was selected as the instant that the computer triggered the first droplet of water after the tip of the rat's muzzle blocked the photobeam at the reward boxes. In this experimental design, analysis of variance (ANOVA) was selected for determining the correlations of spatial position, behavior, reward and task phase with the firing rate of the neurons. In order to better approximate a gaussian distribution, spike count data were first transformed to the sum of the square root of the count summed with the square root of the count incremented by one (Winer, 1971). ANOVA has been shown to be robust even in cases where the underlying distribution is not perfectly gaussian (Lindman, 1974).

Two different analyses of variance (ANOVAs) tested for the first two or all of the following three factors: 1) behavioral correlates - comparisons of firing rates during reward site approach, arrival and water consumption (two 0.5 s periods prior to and after delivery of the first droplet of water); 2) position correlates - differences in firing rate when the rat occupied the different maze arms, and 3) comparisons between phases of the experiment (training versus recall phases and after changes in the reward distribution). Data were also recombined from recordings on different arms that provided the same reward volume during the course of a session (e.g., as shown in Figure 4). Statistical results were considered significant at $p<0.05$. The Student-Newman-Keuls test was employed for *post-hoc* analyses. ANOVAs and post-hoc tests were performed with Statistica® (Statsoft, Tulsa, OK, USA) and other tests performed with Microsoft Excel®.

Histology

After experiments were completed the rat was rehydrated for at least a day, and then deeply anesthetized with pentobarbital. A small electrolytic lesion was made by passing DC current (20 μA, 10 s) through one of the recording electrodes to mark the location of the electrode tip. Intracardial perfusion with saline was followed by 10% formalin in 0.1M phosphate buffer (pH 7.4). Serial frozen sections (50 μm thickness) were stained with cresyl violet. Recording sites were reconstructed by detecting the small electrolytic lesion and the track left by the guide tube, then taking into account the distance that the microelectrode driver had been advanced from the point of stereotaxic placement of the electrodes. The recording sites were calculated by interpolation along the electrode track between the lesion site and the implantation site.

Results

Task performance levels

These data were recorded in 35 experimental sessions in 8 rats. In all cases performance was nearly perfect on light-cued training trials. Consistent with our goal of studying the neural bases of Actor-Critic modes of learning by trial-and-error, in recall trials rats sometimes incorrectly entered maze arms that did not provide the greatest of the remaining rewards. The mean percentage of correct visits was 79±8% and the range was from 60 to 92%. The number of completely correct trials, that is, four visits in sequence of descending reward quantity, was 37 ± 14% (standard error of the mean) and ranged from 0 to 83% in individual sessions. (Note that the probability of correctly performing a complete trial by chance is less than 4%, that is 0.25x0.33x0.50).

Cell localization

Electrode placements were intentionally made in different parts of the ventral striatum in order to explore diverse sub-regions for possible reward-associated responses. Figure 3 shows that recording sites were distributed in the core of the nucleus accumbens, the medial shell of the nucleus accumbens, and the ventromedial part of the caudate nucleus. There was no anatomical segregation of different response types (chi-square, $p>0.05$).

Cell activity profiles

The ANOVAs revealed significant behavioral correlates in about 75% of the neurons recorded in the nucleus accumbens core (33 of 43), accumbens shell (60 of 81), and ventromedial

part of the caudate nucleus (53 of 68). The present study focuses on those cells that showed significant changes in firing rate when rewards were delivered (n=46; other neurons are reported in Mulder et al. 2004).

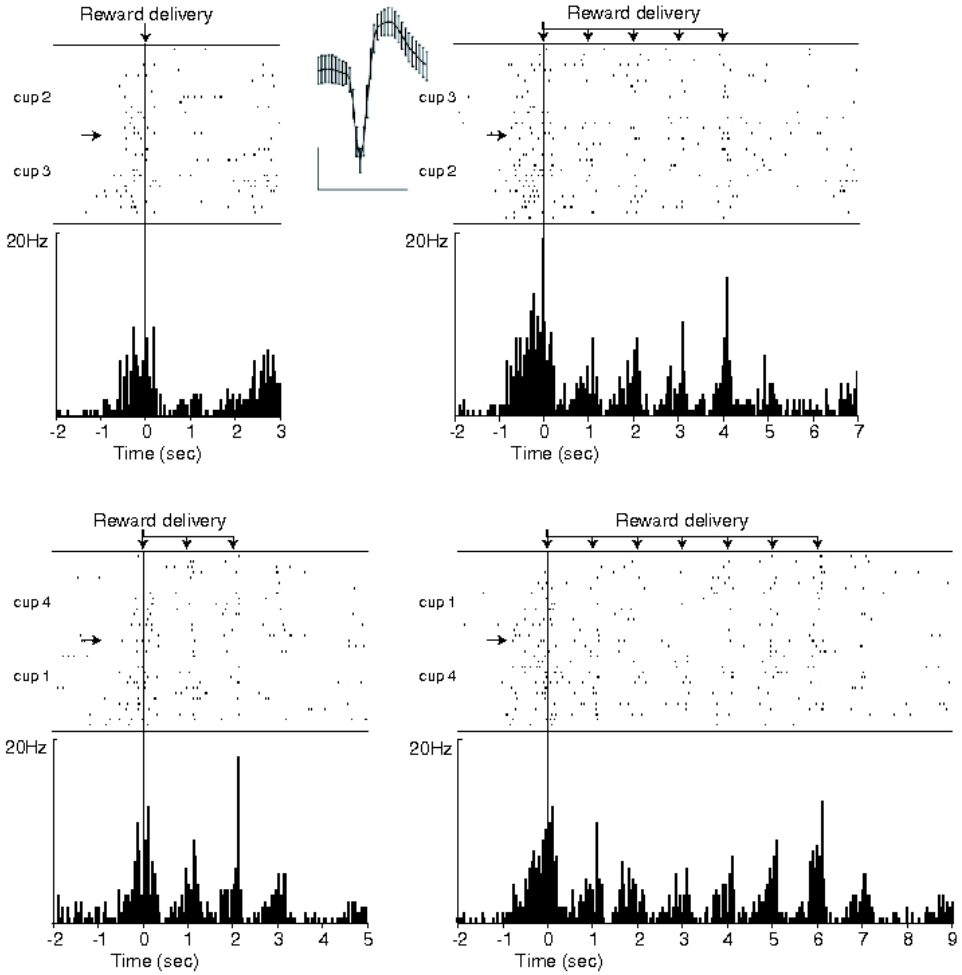


Figure 4. *Phasic excitatory activity predicting reward delivery in a nucleus accumbens core neuron. Raster displays and corresponding histograms (50 ms binwidth) are synchronized with the onset of reward delivery (arrows above). Since the reward value distribution was changed in the middle of the session, data have been regrouped from the cups to combine data corresponding to 1, 3, 5 and 7 droplets of water respectively. Arrows at the left in the raster displays separate data acquired at the respective cups. The discharge activity began as early as 800 msec prior to the reward delivery. Note in the lower left panel, there is a fourth peak in the histogram at time 3 s, even though no fourth reward was delivered then. The same inaccurate predictive activity also appears in the right panels corresponding to 5 and 7 droplets. Activity at the right border of the panels corresponds to arrivals at the next reward site. Waveform average is displayed in inset above (scales 50 V, 1 msec). (rat 6-5, session 2409, unit 0-1).*

Among these 46 cells showing reward related activity, we distinguish phasic neurons and tonically active neurons (TANs). As in previous work (Mulder et al. 2005) TANs were identified principally by *1*) the absence of 'silent' periods (when the firing rate went below 1 imp/s) of 2 s or longer along the course of a trial, and *2*) a significant decrease or increase firing (relative to baseline) during a task event. In contrast, phasic neurons had silent periods interspersed with brief bouts of behaviorally correlated activity. This pattern of phasic activity superimposed upon negligible background activity is consistent with identification as a medium spiny principal neuron (see Mulder et al. 2005). While only 14 of the total 66 (21%) phasic neurons with significant behavioral correlates fired during reward delivery, 32 of the 80 statistically significant tonic neurons

(41%) had these properties. Other neurons had average firing rates of less than 0.1 imp/s (n=11) - since such data are unsuitable for the statistical analyses planned in the experimental design they were not considered further.

Overview of cell response types

Three principal categories of reward-related responses were distinguished to classify the cell activity profiles that we found. First, phasic firing rate increases prior to and during delivery of the successive droplets of water (n=14). These anticipatory neurons had the striking property of discharging after the final reward was delivered and the light in the reward arm was turned off. The second group showed a firing rate increase (n=14) or mixed excitation and inhibition (n=7) during delivery of only the first droplet of water. These responses do not anticipate later rewards at the same site and thus a more closely correlated with reward approach behaviors. Finally, a group with tonic firing rate increases (n=5) or decreases (n=6) throughout the period when multiple droplets of water were delivered. These cells correspond most closely to well-documented tonically active neurons (TANs) and some show very regular spike timing that could provide a possible mechanism for the elaboration of the regular anticipation times of the first group of neurons. Note that these groups could easily be confounded with one another in experimental protocols providing only single rewards. Examples of each of these response types will be presented first. Then their relevance to validation of the TD learning algorithm will be evaluated and a novel modified Actor-Critic model will be presented accounting for observed inconsistencies.

Reward anticipatory responses. Figure 4 shows data from a nucleus accumbens core neuron that started to discharge above baseline about 600-800 msec prior to each reward release, with peak activity on average 100 msec before each droplet. The greatest responses occur for the first and last drops of water. Although the activity preceding the first drop of water could be associated with sensory or motor events (the looming image of the lit cue in the reward box, deceleration, assuming an immobile stance and initial licking), this is not plausible for the subsequent responses for the subsequent droplets since the rats invariably remained stably positioned at the water trough. In this cell the activity precedes the subsequent rewards (indicated by arrows above the rasters) by 300 to 700 msec with a peak in the interval 200 msec prior to and following the reward trigger.

Interestingly, this same activity occurred in the same time window one second after the final droplet was delivered. This is consistent with prediction of a final reward that was never provided. This anticipatory activity occurred on both visually-guided and memory-guided trials (data from the entire session are shown in the Figures.) This activity is surprising since it occurred after the lamp signalling cue availability had been turned off. Recall that in training trials the rats reliably used these same lights to locate the baited reward site. Note that in the present case this 'erroneously predictive' activity occurred on less than half of the trials, yielding smaller histogram peaks than observed for the preceding rewards.

Thus there was no clear correlation between the appearance of this activity on a given trial and whether there were errors on that trial. There was also no relation between the overall performance level of the rat and the incidence of erroneously predictive activity – the latter appeared in sessions where the rat made 90% correct visits. Furthermore, this activity always occurred while the animal still blocked the photobeam at the reward trough and only irregularly coincided with licking. Thus it is parsimonious to consider this activity to be associated with the episodic anticipation of another droplet of water rather than motor preparation of the subsequent departure (since movement timing was the same on trials with and without the predictive activity). Activity in these neurons was not correlated with departures (not shown).

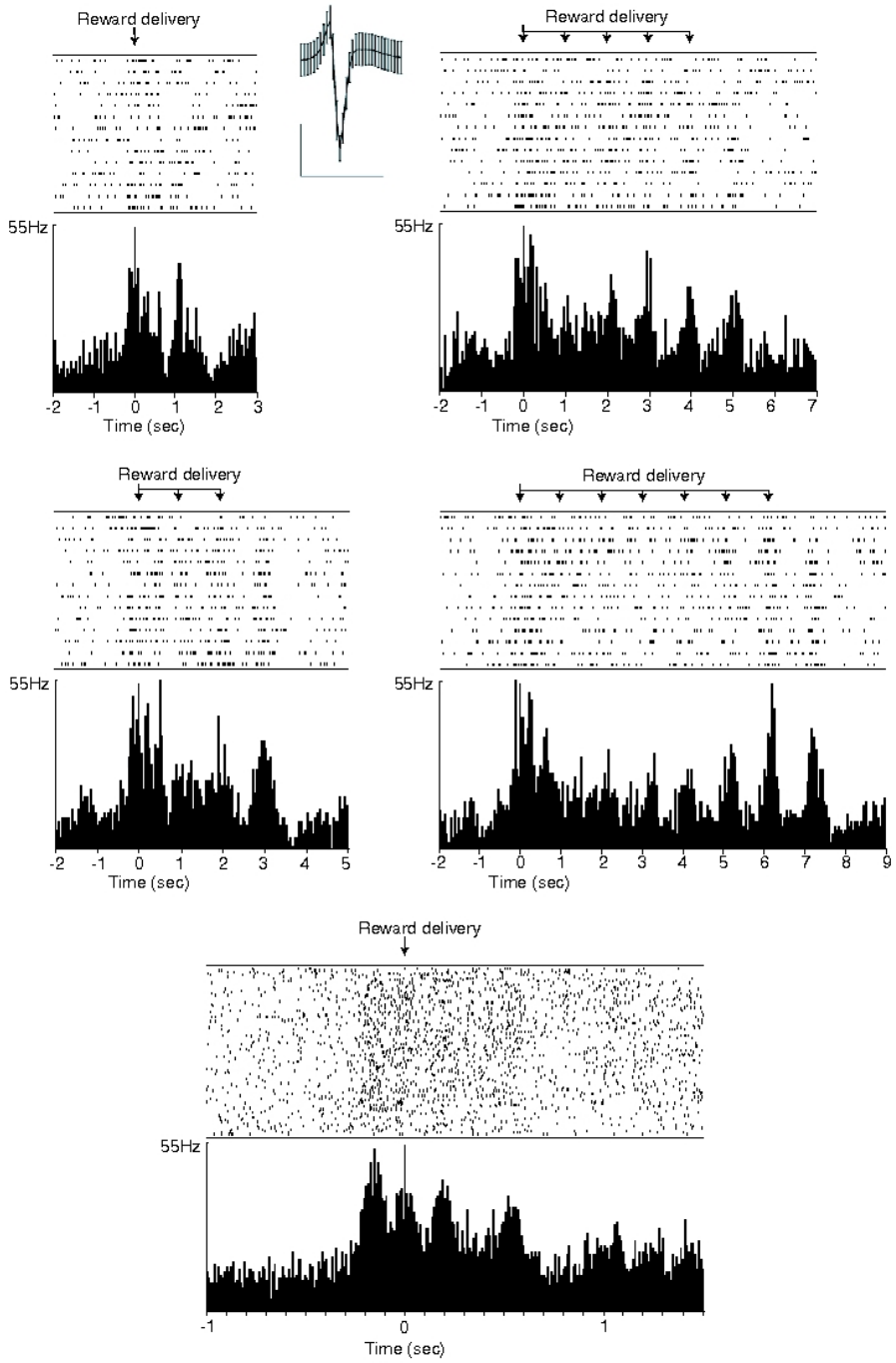


Figure 5. *A tonically active ventromedial caudate neuron with phasic excitatory activity predicting and following rewards. The response is greatest for the first droplet of water, lower for the final droplets, while the weakest responses are found for intermediate droplets. Moderately high activity also appears one second after the final droplet was delivered. The neuron discharged from about 200 ms prior to reward trigger until 300 ms afterwards. Only data from the first half of the session are shown here – the remaining data show similar properties. Average waveform is displayed at the top center. Scales are the same as Figure 4.*
Lower panel) Data from all reward sites for the entire session are displayed at an expanded time scale to demonstrate the fine structure of the activity during delivery of the first droplet of water. Four peaks appear centered on –180, 0, 200 and 520 msec relative to the release of the first droplet of water. In contrast, such fine structure for later droplets of water was not discernible in data in the upper panels where the activity is a broad peak centered about the water delivery. Waveform average is shown at top, same scale as Figure 4. (rat 6-2, session 1609, unit 2-2).

Figure 5 demonstrates another variation of this type of response in a ventromedial caudate neuron with a higher firing rate, a tonically active neuron. This neuron started to fire above background rate at 200 ms prior to the first reward trigger and continued until 300-500 ms afterwards. Similar to Figure 4, maximal responses occur at the first reward but, in contrast, there is also a second major peak for the erroneous reward prediction at the end. The anticipatory activity also is more robust here, occurring on virtually all trials for all droplets as well as the erroneous prediction. The similarity of this final response to the others demonstrates that the persistence of this activity in the 300-500 msec following the reward is independent of the presence or absence of reward. In the histograms the later peaks appear to be narrower and clearly defined, with a trough of reduced activity prior to the reward-predictive increases in activity.
All fourteen neurons with anticipatory activity also showed this 'erroneously predictive' activity (see Supplementary Figure 2 for more examples). Of these, eight neurons were tonically active (like in Figure 5), while the remaining six were phasic (as shown in Figure 4). These neurons were found with similar incidence in the accumbens core (n=5) and shell (n=9; p=0.28, chi-square test).

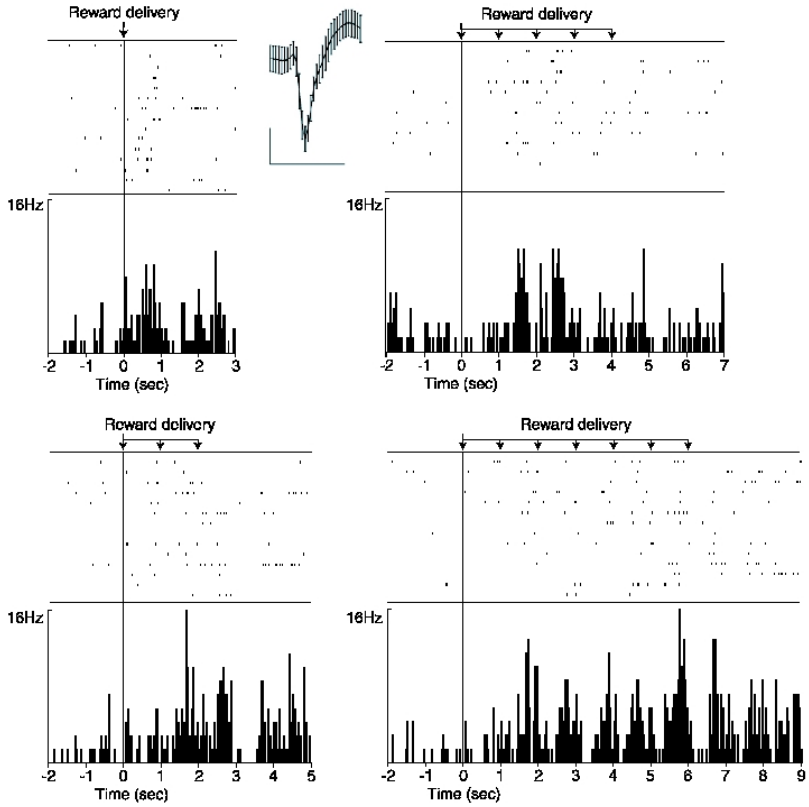


Figure 6. *Phasic excitatory activity anticipating reward delivery in a nucleus accumbens core neuron. This neuron discharged little for the first two droplets of water. The activity was greatest for the final droplet of water and for the corresponding period one second after the final droplet (corresponding to inappropriate anticipation of another reward). This neuron was distinguished from others in this group by a rather low firing rate. Discharges started during the 800 ms preceding water rewards. Waveform average is shown above, scales as in Figure 4. (Rat 6-5, session 1709, unit 0-1)*

Neurons in this group had particular preferential selectivities for the order of presentation of water droplets: early, midway or late in the sequence. For example, the neuron of Figure 4 had larger responses for the first and last rewards. Figure 6 is an example of a phasic neuron in this group that had only minor responses for the first and second droplets of water, but peak activity for the final reward (except when there was only one water droplet). This variability demonstrates an uncoupling between presumed level of anticipation or expectation and the activity of the individual neurons. The first drop of water should have been anticipated with a very high degree of certainty, yet there is

little such anticipatory activity in the neuron of Fig. 6. Yet other neurons had different order preferences, for example two neurons fired maximally prior to and during delivery of the fourth droplet of water (rows 4 and 5 in Supplemental Figure 2) exceeding the responses for the first or last droplets. As detailed in the computational modeling section below, these characteristics will require an adaptation of Actor-Critic models for TD learning so that they are constituted of multiple modules, with the particularity that each module processes a different information concerning the task.

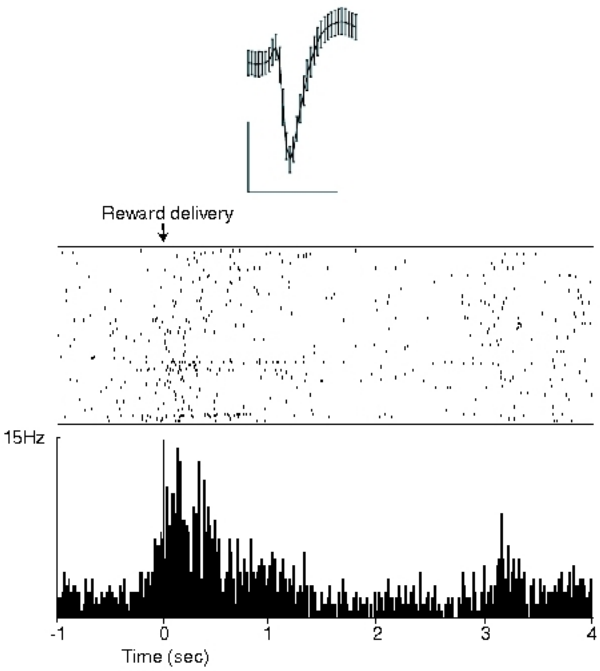


Figure 7. *This ventromedial caudate neuron discharged prior to and after delivery of the first droplet of water, but had no response to any other successive droplets. At least two peaks are discernible here centered on 100 and 350 ms following reward delivery.* This neuron was exceptional in that it showed an increase in firing rate to about 8 Hz prior to and during departures from the water troughs (shown *below). Average waveform is shown on top; scales are same as Figure 4. Bin width = 20 ms. (Rat 6-2, session 0809, unit 1-1.)*

Activity increase during release of only the first droplet of water. In the neurons of Figure 7 (top), the firing rate started to increase at 100 msec prior to when the rat blocked the photodetector at the water trough and the activity peaked at about 150 msec afterwards. No further activity was observed for the following droplets of water at the same site (data are shown for all trials of the session). While neurons in this group varied in the onset time (from 1 s prior to arrival until slightly after arrival) and the offset time, the activity was only observed for the first droplet of water. These neurons thus would not provide a reliable signal for reward anticipation.

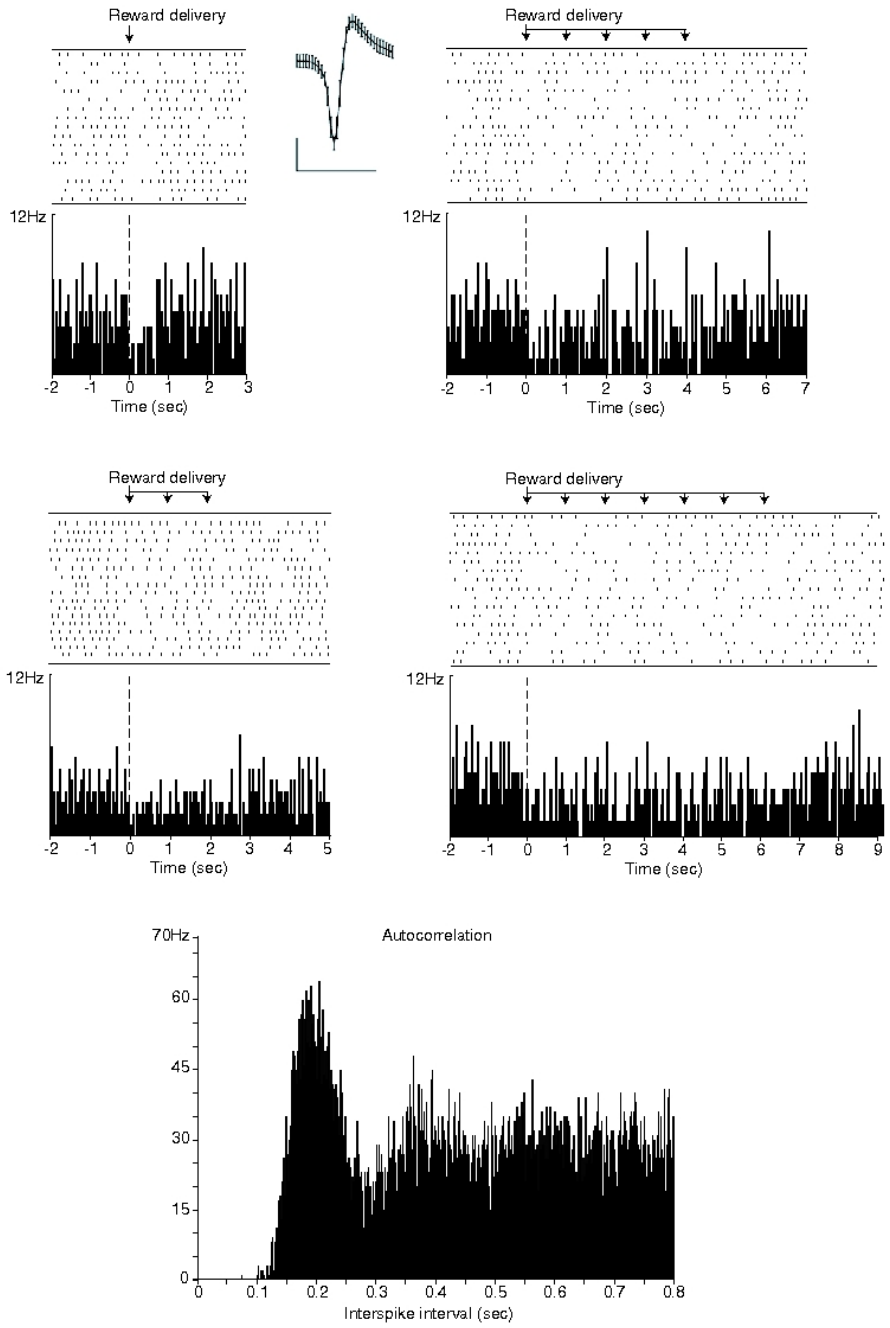


Figure 8. *Inhibition during rewards in a tonically active neuron. The tonic activity at about 8 Hz diminishes to about 3 Hz while the rat is at the reward trough consuming and waiting for more water droplets. Unlike the neurons described above, the activity resumed during the second following the final droplet of water and there was no prolongation for an additional second. Below, the interspike interval histogram has a peak at 0.2 sec, corresponding to regular firing at 5 Hz. Waveform average appears at top, scales as in Figure 4. (Rat 6-1,session 1209, unit 0-1).*

Uniform increase or decrease of firing rate while multiple droplets of water were delivered. Figure 8 is taken from a tonically active neuron with inhibition during the period that the rat consumed water rewards. This response profile strikingly resembles tonically active neurons reported in the monkey striatum (see e.g., Apicella et al. 1991a). In contrast with neurons of the first group above, here inhibition persisted during only 1 s after the final droplet was delivered and did not continue for an additional 'erroneous' second. While this suggests that the response is correlated with the actual presence of reward, it must be noted that the onset of the inhibition began the instant the reward

delivery was triggered, immediately prior to when reward would have entered the rat's mouth.

The autocorrelation analysis of this neuron's activity at the bottom of Figure 8 demonstrates a strikingly regular timing. Note that the principal peak occurs at 0.2 s, corresponding to a frequency of 5 Hz. While all of the neurons in this group were tonically active, several instead had irregularly timed and bursty activity as shown in Figure 5. Tonically active neurons with reward responses were found in the medial shell of the nucleus accumbens, the ventromedial caudate and, in one case, at the junction of medial shell and ventral pallidum.

Ventral striatal neurons as Critic in TD-learning: the need for multiple input modules

This study aimed to determine if rat ventral striatal activity is compatible with the role of this structure in the temporal difference (TD) learning model developed by Sutton and Barto (1998; see Annex I) as a means for reward signals to reinforce neural circuits mediating goal directed behavior. While existing models were effective for cases of single rewards, multiple rewards are more challenging since in the models the striatal reward prediction signal drops to zero the instant the first reward arrives (Barto 1995; Foster et al. 2000; Baldassarre 2003). The few models that were tested with a temporally prolonged, but single, reward (Montague et al. 1996; Suri and Schultz. 2001) hold that reward prediction signals should decrease while consuming successive rewards in or and finally disappearing at the final reward (similar to the black trace in Figure 9A). These models cannot account for the variety of neural responses observed here, such as, reward anticipation signals that were greater for either early, middle or later rewards, and the anticipation of an extra droplet of water which was never provided or variations in the timing of the predictive activity. We were inspired by previous approaches employing multiple modules TD-learning (MMTD) models where each module produces a particular response in a given task (Baldassarre. 2002; Doya et al. 2002; Khamassi et al. 2006). In these models, the differences among modules lies in the 'responsibility' signal that gates their output: the modules share the same input signal, but vary with respect to the task component for which they are responsible. Here, we extended the MMTD framework to enable each module to process different *information inputs*, thus permitting the model to better emulate the neurophysiological data.

The present TD-learning model has four Actor-Critic modules. Each module independently processes the same TD-learning algorithm based upon a different mix of spatial and temporal inputs. The spatial information here, that is, the state *S* of the animal, consists of spatial position (i.e., location of the respective maze arms relative to one another and the room) and sensory cues such as cue lights at reward sites. For temporal information, we tested the different responses of the modules by adding or suppressing the 'complete serial compound stimulus' component proposed by Montague et al. (1996). This component gives temporal information about a given stimulus, and enables this model to "count" the number of droplets of reward already received at a given moment. All model variants had full access to signals concerning position along the path between the maze center and the ends of the arms. The modules of the model also varied in the value of the discount factor γ, which indicates the capacity to take future rewards into account (cf. Annex I). The four TD learning modules of the model could be considered as embodied in four different zones within the striatum each receiving different information about the state of the animal, perhaps due to local variations in the composition of convergent populations of afferent cortical and hippocampal inputs as well as locally specialized signal processing.

We simulated each module on 25 trials where the rats visited each of the four maze arms. For each trial, the TD-learning algorithm was computed once every 250 ms, starting 5 seconds before the first droplet reward and ending 2 seconds after the last. We did not study how the Actor part of the model should learn to build appropriate behavior for task resolution, since this was done in a previous robotics simulation (Khamassi et al. 2006). The goal here was only to study if and how the Critic could learn to anticipate rewards in a manner similar to ventral striatal neurons, in conditions like those faced by the rats in our task: facing the reservoir and waiting for successive rewards while a light stimulus was maintained on until the last droplet of water. Since this happened only during correct trials in the real experiment (error trials were aborted to enforce the trial-and-

error learning), in the simulations, the Actor part of the model had a fixed repetitive behavior.

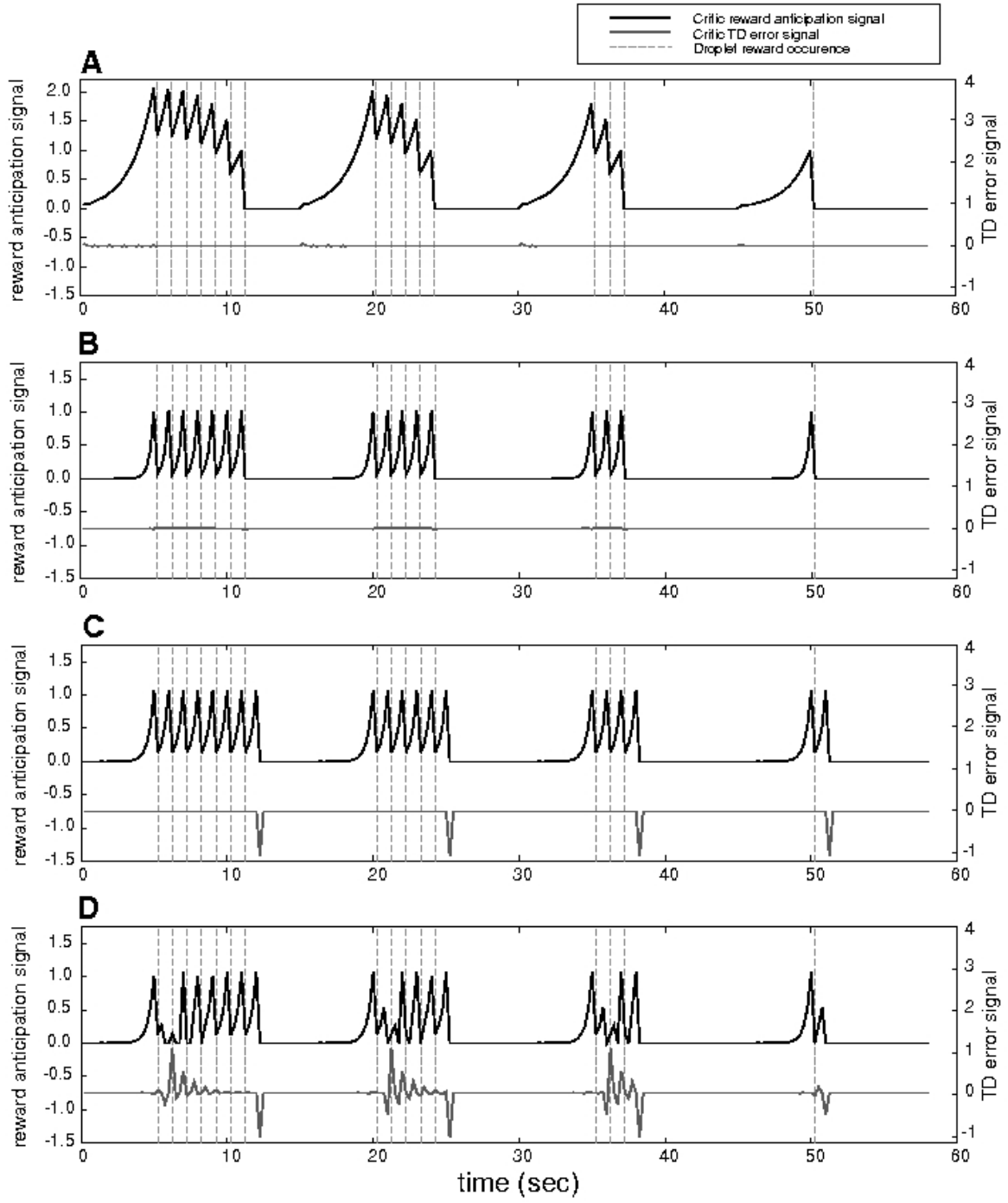


Figure 9. *Simulations of cell activity in four compartments of the TD learning model with varied inputs concerning state (spatial and sensory information), temporal inputs, as well as discount factor (related to how far in the future predictions are made). The ordinate indicates average firing rate and the abscissa is time. The vertical dashed gray lines indicate the onset of rewards and the displays show successive visits to reward sites on the four arms in order of descending reward volume. A) These parameters permit the model to replicate the results of Suri and Schultz (2001). B, C, D) Reducing the discount factor and changing state and temporal inputs reproduces several of the activation patterns recorded in ventral striatal neurons.*

Figure 9A shows the results from the model's first module which has explicit and precise inputs concerning state S as well as temporal information and a discount factor γ of 0.85. This result

resembles that of Suri and Schultz (2001). The long lead in initial reward predictive activity is due to the elevated discount factor and the gradual reduction of response strength results from the accurate temporal input signals. The model's second module (Figure 9B) has no temporal inputs but has precise state information. The discount factor is 0.40 here. While the initial onset of predictive activity starts later, there is still no prediction of a droplet of water after the final one. The latter only emerges in the third module (Figure 9C), where there is highly ambiguous state information (for example, this module would not receive information about the light going off in the reward box) as well as poor temporal information. The discount factor here is 0.50. (Interestingly, in our recordings, 'post-final droplet' anticipatory activity only occurred on a fraction of trials suggesting that such inputs could be subject to gating or other modulation). Finally Figure 9D demonstrates how variations in selective activity during early, middle or late droplets can appear by varying inputs. In this module, there is partially ambiguous state information but no temporal inputs (again discount factor $\gamma = 0.50$).

Discussion

Here rats received experimental multiple rewards at 1 s intervals on the respective arms of a plus-maze. This experimental design aimed to disambiguate activity associated with reward-directed behaviors from actual anticipatory activity predicted by Actor-Critic models of TD learning. We found the latter in the form of phasic increases in firing rate anticipating and accompanying delivery of individual droplets of water, a novel finding in the rat striatum. This contrasted with other responses more likely associated with reward site approach behaviors and associated sensations, which took the form of phasic increases (sometimes coupled with decreases) in firing rate for the first droplet of water only.

The anticipatory lag varied among individual neurons, commencing from 800 to 200 msec prior to the reward. Previous studies have generally shown accumbens responses that begin immediately after reward delivery (Lavoie and Mizumori 1994; Miyazaki et al. 1998; Martin and Ono 2000; Wilson and Bowman 2004), but in some cases precede rewards by 300 to 500 msec (Nicola et al. 2004; Taha and Fields 2005), and even as much as 1-2 s (Tremblay et al. 1998; Schultz et al. 1992; Shibata et al. 2001; Janak et al. 2004). However since only single rewards were provided in those experiments it is not clear whether this activity in rats might be associated with sensory cues or behaviors preceding reward acquisition or rather are actually associated with reward anticipation. In the immobile awake monkey preparation (Cromwell and Schultz 2003) and in humans (O'Doherty et al. 2004), however, it has been easier to reduce the risk of such confounds.

The regular timing of these anticipatory reward responses in the absence of any explicit trigger stimulus suggests that these neurons have access to some kind of internal clock signals. One possible source for this would be TANs such as the one shown in Figure 8. The highly regular 5 Hz discharges could provide a reliable basis for such timing. Although these neurons fired at a lower firing rate during the reward period, this appeared to be due to spikes dropping out while the remaining activity maintained the regular timing. Interestingly, three of the peaks observed for the first droplet of water in the neuron of Figure 5 (bottom) also had 200 msec intervals (5 Hz) between them.

Implications for models of reinforcement learning. The present results bear on recent theories and models of mechanisms of goal-directed rewarded learning engaging basal ganglia activity (Schultz et al. 1997; Graybiel, 1998). The TD learning algorithm (Sutton and Barto 1998) has been successfully employed in Actor-Critic architectures to endow robots with reinforcement learning capacities (see Khamassi et al. 2005 for a review). In the original formulation, the striatum makes successive predictions of reward, whose accuracy is used to compute an error prediction signal at the level of striatal-afferent dopaminergic neurons (Houk et al. 1995). This prediction error, combined with signals of the presence or absence of reward would then enable dopaminergic neurons to emit reinforcement signals that in turn modify cortico-striatal synaptic plasticity. Such

modifications would lead to learning by increasing the probability of selecting an action that previously led to a reward. Modification of behavior following TD-learning rules has already been observed in rats (see Daw et al. 2005 for a review) and monkeys (Samejima et al. 2005) during reward-based habit-learning tasks. The present results extend this by demonstrating that the diversity of striatal responses anticipating multiple consecutive rewards is coherent with TD-learning. The striatal responses 'erroneously' predicting another droplet of water after the last one can be accounted for in the simulations as reflecting weak levels of temporal information input while state information varies from somewhat to highly ambiguous. When state information is more precise, such activity ceases. As a consequence, ventral striatal activity is consistent with parallel TD-learning systems processing varying input signals, analogous to the multiple module approach recently employed to model spatial navigation in rodents (Chavarriaga et al. 2005a). The notion that different neurons receive different mixes of input information of varying levels of accuracy is consistent with known patterns of input projections to the ventral striatum. Moreover, discount factors of 0.4 to 0.5 provided the anticipatory activity on the time scales observed here. Higher discount factors gave longer lead times for the anticipatory activity corresponding to more gradual buildup of activity prior to rewards, as found in other neurons here and elsewhere (Suri and Schultz 2001). Interestingly in a recent brain imaging study of humans performing a reward motivated task, different striatal subregions were selectively active according to the discount factor that best modeled the subjects' strategy concerning short or long term gain (Tanaka et al. 2004). Recent imaging studies in humans by O'Doherty et al. 2004) are also consistent with the ventral striatum being engaged in Critic-like functions.

The present simulations only concern activity in the striatum prior to and during rewards while the animals were immobile at the end of maze arms, and thus can be interpreted as a reinforcement signal. Furthermore, the areas where these neurons were recorded send projections to brainstem dopaminergic areas: the substantia nigra pars compacta (SNpc) is principally afferented by the dorsal striatum and accumbens core, while the ventral tegmental area (VTA) is more influenced by the accumbens shell (Haber et al. 2000; Ikemoto 2002). These zones then send dopaminergic projections to respective striatal areas which would then modulate learning processes specific to their functional modalities (such as motor sequencing, habit, or goal-directed behaviors). A prediction of our model is that sub-groups of brainstem dopaminergic neurons would be associated with different TD-learning modules, and would, in the same plus-maze task, exhibit differential responses to reward (see figure 9): some dopamine neurons responses to reward should vanish as in the seminal study of Schultz et al. (1997) Other dopamine neurons related to the TD-learning module which erroroneously anticipates an additional droplet of water should then have negative responses.

Daw et al. (2005) have recently argued that anticipatory activity for motivated behavior in rats cannot be completely explained with TD-learning models. Thus their model employs a TD module to drive habitual behavior, and this competes with a higher level tree-search module dedicated to goal-directed behavior. The present work shows that a TD-learning based mechanism is computationally sufficient to model the diverse anticipatory responses. However, other ventral striatal neurons recorded in the present protocols (reported in Mulder et al. 2004) which were active from initiation to completion of goal approach behaviors could be an embodiment of the tree-search model since they 'chunk' (see Graybiel 1998) the behavioral sequence until the outcome. The dichotomy in reward anticipation and goal approach correlates is consistent with the hypothesis that functionally distinct groups of the rat nucleus accumbens could be differentially involved in TD-learning or in goal-directed behavior (Dayan 2001).

The reward-related activity observed here could serve as a Critic signal to help establish functional circuits (by a loop through VTA) for sequencing the activity of the goal appproach neurons (Mulder et al. 2004), first orchestrating then automatizing the sequence of successive steps to satisfy task exigencies. The neurons selective for goal-directed behavior would also affect

dopaminergic neurons which would then transmit Critic signals to more dorsal striatal regions implicated in habit learning. Selection among alternative goal choices or even among cognitive strategies would thus be carried out in associative and limbic regions situated more ventrally in the striatum. This could lead to a hierarchy of behavioral control which might lead to cognitive correlates, for example, context or reward-dependence in the more dorsal basal ganglia responses (Hikosaka et al. 1989).

Acknowledgements: Thanks to Alain Berthoz for indispensable support throughout all stages of this project, Dr. M. Zugaro for assisting with the recordings and comments on the manuscript, F. Maloumian for figure preparation, S. Doutremer for histology, S. Lemarchand for animal care, Drs. A.-M. Thierry, J.-M. Deniau, A. Guillot and B. Girard for helpful discussions. Portions of this work have been presented as posters at meetings. The present address for ABM is: Cognitive Neurophysiology-CNCR, Dept. of Anatomy and Neurosciences, VU University Medical Center, van de Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands and for ET is: Department of Analysis of Brain Function, Faculty of Food Nutrition, Toyama College, 444 Gankaiji, Toyama 930-0193, Japan

Grants: Human Frontiers Fellowship and the French Embassy in the Netherlands for support for (to A.B.M.); Cogniseine; the European Community Human Capital and Mobility program Groupement d'Intérêts Scientifiques, ACI, European Community Integrated Project ICEA, European Community Integrated Project BACS.

References

Albertin SV, Mulder AB, Tabuchi E, Zugaro MB, Wiener SI. Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behav Brain Res* 117: 173-183, 2000.

Aosaki T, Kimura M, Graybiel AM. Temporal and spatial characteristics of tonically active neurons of the primate's striatum. *J Neurophys* 73: 1234-1252, 1995.

Aosaki T, Tsubokawa H, Ishida A, Watanabe K, Graybiel AM, Kimura M. Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *J Neurosci* 14: 3969-3984, 1994.

Apicella P, Scarnati E, Schultz W. Tonically discharging neurons of monkey striatum respond to preparatory and rewarding stimuli. *Exp Brain Res* 84: 672-675, 1991.

Apicella P, Legallet E, Trouche E. Responses of tonically discharging neurons in monkey striatum to visual stimuli presented under passive conditions and during task performance. *Neurosci Lett* 203: 147-150, 1996.

Baldassarre G. Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In: *Adaptive behavior in anticipatory learning system*s, edited by Butz M, Sigaud O, Gerard P, pp. 179-200. Berlin: Springer Verlag, 2003.

Barnes TD, Kubota Y, Hu D, Jin DZ, Graybiel AM. Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* 437: 1158-1161, 2005.

Barto A. Adaptive critics and the basal ganglia. In: *Models of Information Processing in the Basal Ganglia*, edited by Houk JC, Davis JL, Beiser DG, 1995, pp 215-232.Cambridge, MA: MIT.

Chavarriaga R, Strösslin T, Sheynikhovich D, Gerstner W. A computational model of parallel navigation systems in rodents. *Neuroinformatics* 3: 223-242, 2005.

Cromwell HC, Schultz W. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J Neurophys* 89: 2823-2838, 2003.

Daw ND, Touretzky DS, Skaggs WE. Representation of reward type and action choice in ventral and dorsal striatum in the rat. *Soc Neurosci Abstr* 28: 765.11, 2002.

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8: 1704-1711, 2005.

Dayan, P. Motivated reinforcement learning. In: *Advances in Neural Information Processing Systems,* edited by Dietterich TG, Becker S, Ghahramani Z, 2001, pp 11-18. Cambridge, MA: MIT Press, 2001.

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

Foster D, Morris R, Dayan P. Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus* 10: 1-16, 2000.

Fuster JM. *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe, (3rd ed.),* Philadelphia: Lippincott-Raven, 1997.

Gardiner TW, Kitai ST. Single unit activity in the globus pallidus and neostriatum of the rat during performance of a trained head movement. *Exp Brain Res* 88: 517-530, 1992.

Graybiel AM. The basal ganglia and chunking of action repertoires. *Neurobiol Learn Mem* 70: 119-136, 1998.

Groenewegen HJ, Wright CI, Beijer AV. The nucleus accumbens: gateway for limbic structure to reach the motor system? *Prog Brain Res* 107: 485-511, 1996.

Haber SN, Fudge JL, McFarland NR. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci* 20: 2369-2382, 2000.

Hikosaka O, Sakamoto M, Usui S. Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J Neurophys* 61: 814-832, 1989.

Houk JC, Adams JL, Barto AG. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: *Models of information processing in the basal ganglia,* edited by Houk JC, Davis JL, Beiser D, pp 249-270, Cambridge, MA: MIT Press, 1995.

Ikemoto S. Ventral striatal anatomy of locomotor activity induced by cocaine, (D)-amphetamine, dopamine and D1/D2 agonists. *Neurosc*i 113: 939-955, 2002.

Itoh H, Nakahara H, Hikosaka O, Kawagoe R, Takikawa Y, Aihara K. Correlation of primate caudate neural activity and saccade parameters in reward-oriented behavior. *J Neurophysiol* 89: 1774-1783, 2003.

Janak PH, Chen MT, Caulder T. Dynamics of neural coding in the accumbens during extinction and reinstatement of rewarded behavior. *Behav Brain Res* 154: 125-135, 2004.

Joel D, Niv Y, Ruppin E. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw* 15(4-6): 535-547, 2002.

Kawagoe R, Takikawa Y, Hikosaka O. Expectation of reward modulates cognitive signals in the basal ganglia. *Nat Neurosci* 1: 411-416, 1998.

Khamassi M, Lachèze L, Girard B, Berthoz A, Guillot A. Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adapt Behav*, *Spec Issue Towards Artificial Rodents* 13(2): 131-148, 2005.

Khamassi M, Martinet, L-E, Guillot, A. Combining self-organizing maps with mixture of experts: Application to an Actor-Critic model of reinforcement learning in the basal ganglia. In: *From Animals to Animats 9. Proceedings of the Ninth International Conference on Simulation of Adaptive Behavior* edited by Nolfi S, Baldassare G, Calabretta R, Hallam J, Marocco D, Meyer J-A, Miglino O, Parisi D, pp 394-405. Springer - Lecture Notes in Artificial Intelligence 4095, 2006

Kimura M, Rajkowski J, Evarts E. Tonically discharging putamen neurons exhibit set-dependent responses. USA 81: *Proc Nat Acad Sci*: 4998-5001, 1984.

Kimura M. The role of primate putamen neurons in the association of sensory stimuli with movement. *Neurosci* Res 3: 436-443, 1986.

Lavoie AM, Mizumori SJ. Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Res* 638: 157-168, 1994.

Lindman HR. *Analysis of Variance in Complex Experimental Designs*. San Francisco: W. H.

Freeman and Co, 1974.

Martin PD, Ono T. 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. *Behav Brain Res* 116: 23-38, 2000.

Matsumoto N, Minamimoto T, Graybiel AM, Kimura M. Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *J Neurophysiol* 85: 960-976, 2001.

Miyazaki K, Mogi E, Araki N, Matsumoto G. Reward-quality dependent anticipation in rat nucleus accumbens. *Neuroreport* 9: 3943-3948, 1998.

Mogenson GJ, Jones DL, Yim CY. From motivation to action: Functional interface between the limbic system and the motor system. *Prog Neurobiol* 14: 69-97, 1980.

Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 6(5): 1936-1947, 1996.

Mulder AB, Gijsberti Hodenpijl M, Lopes da Silva FH. Electrophysiology of the hippocampal and amygdaloid projections to the nucleus accumbens of the rat: convergence, segregation and interaction of inputs. *J Neurosci* 18: 5095-5102, 1998.

Mulder AB, Shibata R, Trullier O, Wiener SI. Spatially selective reward site responses in tonically active neurons of the nucleus accumbens in behaving rats. *Exp Brain Res* 163: 32-43, 2005.

Mulder AB, Tabuchi E, Wiener SI. Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur J Neurosci* 19 : 1923-1932, 2004.

Nicola SM, Yun IA, Wakabayashi KT, Fields HL. Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task. J *Neurophysiol* 91: 1840-1865, 2004.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304(5669): 452-454, 2004.

Otani S. *Prefrontal Cortex: From Synaptic Plasticity to Cognition.* Boston: Kluwer Acadamic, 2004.

Paxinos G, Watson C. *The Rat Brain in Stereotaxic Coordinates (CD-ROM version)*. NY: Acad Press, 1998.

Ravel S, Legallet E, Apicella P. Responses of tonically active neurons in the monkey striatum discriminate between motivationally opposing stimuli. *J Neurosci* 23: 8489-8497, 2003.

Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. *Science* 310: 1337-1340, 2005.

Schmitzer-Torbert N, Redish AD. Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J Neurophysiol* 91: 2259-2272, 2004.

Schoenbaum G, Setlow B, Saddoris MP, Gallagher M. Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron* 39: 855-867, 2003.

Schultz W, Apicella P, Scarnati E, Ljungberg T. Neuronal activity in monkey ventral striatum related to the expectation of reward. *J Neurosci* 12: 4595-4610, 1992.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science* 275: 1593-1599, 1997.

Shibata R, Mulder AB, Trullier O, Wiener SI. Position sensitivity in phasically discharging nucleus accumbens neurons of rats alternating between tasks requiring complementary types of spatial cues. *Neurosci* 108: 391-411, 2001.

Suri RE, Schultz W. Temporal difference model reproduces anticipatory neural activity.

Neural Comput 13: 841-862, 2001.

Sutton RS http://www.cs.ualberta.ca/~sutton/book/6/node7.html, 1997.

Sutton RS, Barto AG. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press, 1998.

Tabuchi ET, Mulder AB, Wiener SI. Position and behavioral modulation of synchronization of hippocampal and accumbens neuronal discharges in freely moving rats. *Hippocampus* 10: 717-728, 2000.

Tabuchi E, Mulder AB, Wiener SI. Reward value invariant place responses and reward site associated activity in hippocampal neurons of behaving rats. *Hippocampus* 13: 117-132, 2003.

Taha SA, Fields HL. Encoding of palatability and appetitive behaviors by distinct neuronal populations in the nucleus accumbens. *J Neurosci* 25: 1193-1202, 2005.

Takikawa Y, Kawagoe R, Hikosaka O. Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. *J Neurophysiol* 87: 508-515, 2002.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci* 7: 887-893, 2004.

Thierry AM, Gioanni Y, Degenetais E, Glowinski J. Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus* 10: 411-419, 2000.

Tremblay L, Hollerman JR, Schultz W. Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J Neu*rophys 80: 964-977, 1998.

Wiener SI. Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J Neurosci* 13: 3802-3817, 1993.

Wilson DI, Bowman EM. Nucleus accumbens neurons in the rat exhibit differential activity to conditioned reinforcers and primary reinforcers within a second-order schedule of saccharin reinforcement. *Eur J Neurosci* 20: 2777-2788, 2004.

Wilson DI, Bowman EM. Rat nucleus accumbens neurons predominantly respond to the outcome-related properties of conditioned stimuli rather than their behavioral-switching properties. *J Neurophysiol* 94: 49-61, 2005.

Winer, BJ. *Statistical Principles in Experimental Design. 2nd ed*. New York: McGraw-Hill, 1971.

Yang CR, Mogenson GJ. Electrophysiological responses of neurones in the accumbens nucleus to hippocampal stimulation and the attenuation of the excitatory responses by the mesolimbic dopaminergic system. *Brain Res* 324: 69-84, 1984.

3. Comparison of Actor-Critic models in simulated robotics

Khamassi, Lachèze, Girard, Berthoz, Guillot (2005). *Adaptive Behavior*.

3.1 Summary of objectives

The objective of the work presented here is to integrate an efficient Actor-Critic model with an existing biologically plausible model of action selection inspired by the rat basal ganglia, and to simulate the system in a visual cue-guided reward-seeking task in a virtual plus-maze.

As described in section 4.4, numerous Actor-Critic models inspired by the basal ganglia have been developed since 1995. These models were tested on different tasks, and it is therefore difficult to compare their performance. Moreover, whereas a few models were tested in complex conditions incorporating a continuous environment (Doya et al., 2002; Strösslin and Gerstner, 2003), most were simulated in tasks involving a small number of finite states (Houk et al., 1995; Montague et al., 1996; Berns and Sejnowski, 1998; Suri and Schultz, 1999; Suri et al., 2001; Frank et al., 2001; Brown et al., 2004).

Besides, if an important proportion of these models focused on timing mechanisms, it is because inspiring studies in monkeys employed fixed temporal bins (generally 2 seconds) between the stimulus and the reward, thus allowing temporally calibrated responses of dopaminergic neurons (Montague et al., 1996; Suri and Schultz 2001). However, in more natural situations where a rodent or an animat needs to find reward for its survival, temporal characteristics of the tasks are rarely fixed, but rather depend on the agent's behavior and on change in the environment.

To deal with more complex tasks, several authors proposed to coordinate several Actor-Critic modules within a *mixture of experts* architecture (Baldassarre, 2002; Doya et al., 2002). The mixture of experts architecture was proposed by Jacobs et al. (1991, see previous chapter) and, for each subset of a given task, consist in specializing the module that gives the best performance. Both Baldassarre (2002)'s model and Doya et al. (2002)'s model used the experts' performance in predicting future states of the animat in the environment for this specialization process. The former model was combined successfully with Actor-Critic models in a task where a simulated animat learns to navigate towards three goals (Baldassarre, 2002). The latter model was applied successfully in swinging up an inverted pendulum (Doya et al., 2002).

Finally, in their review, Joel et al. (2002) report that Actor-Critic models of reinforcement learning do not take into account known anatomy and physiology of the basal ganglia. For instance, they usually implement a simple winner-takes-all mechanism for action selection (the best action is selected regardless of the value of other actions), whereas evidence suggest that an interaction of selection and control pathways (depending on other actions' value) within the basal ganglia subserve action selection (Gurney et al., 2001a). The latter authors proposed a model named *GPR* solving this issue, and showing interesting energy saving properties for robotics (Montez-Gonzalez et al., 2000; Gurney et al., 2001b; Girard et al., 2002,2003). However, the model was not yet provided with reinforcement learning capabilities, and « strengths » of stimulus-response association employed in the model were hand-tuned.

The following section will present our work consisting in comparing several principles taken from previous Actor-Critic models within a common architecture using the GPR model, and on a common continuous state-space version of the plus-maze task. Four Actor-Critic frameworks were compared : a single-component Critic; several Critic modules controlled by a gating network within a *mixture of experts* architecture; several Critic modules a priori associated with different subparts of the task and connected to a single Actor ; a similar combination of several Critic modules, but

implementing several Actor components.

3.2 Summary of methods

Two virtual versions of the plus-maze task are simulated: one in a simple 2D environment; the other in a 3D simulator working in real time and implementing physical dynamics. The task employed mimics the training phase of the article presented in the previous section of this chapter. At each trial, one of the maze's arm is randomly chosen to deliver reward. The associated wall is colored in white whereas walls at the three other extremities are dark gray. The animat has to learn that selecting the action *drinking* when it is near the white wall (distance < 30 cm) and faces it (angle < 45 degrees) gives it a reward. Here we assume that reward = 1 for n iterations (n = 2), without considering how the hedonic value of this reward is determined.

The animat is equipped with a linear panoramic camera providing the color of the nearest perceived segment every 10°. This results in a 36 color table that constitute the animat's visual perception. At each timestep, this 36 color vector is sent to a primitive visual system that estimates the importance of each color on the agent's "retina", the angle between each color and the center of the "retina". This provides a 13 dimension state space (including 9 continuous variables and 4 binary variables) that constitutes the input information to the Actor-Critic model.

We expect the animat to learn a sequence of context-specific behaviors, so that it can reach the reward site from any starting point in the maze:

- When the white wall is not visible, orient towards the center of the maze and move forward.
- Upon arriving at the center (the white wall is visible), turn towards the white stimulus.
- Move forward until close enough to reward location.
- Drink.

The trial ends when the reward is consumed: the color of the wall at reward location is changed to dark gray, and a new arm extremity is randomly chosen to provide the next reward. The animat then has to perform the learned behavioral sequence again. Note that there is no delay between two consecutive trials: trials follow each other successively.
The more efficiently and fluidly the animat performs the above-described behavioral sequence, the less time it will take to reach the reward. As a consequence, the criterion chosen to validate the models is the time to goal, plotted along the course of the experiment as the learning curve.

3.3 Summary of results

We find that a multiple modules Actor-Critic system is required to solve the task. Moreover, the classical method used to coordinate Actor-Critic modules – e.g. the mixture of experts proposed by Jacobs et al. (1991, see previous chapter) which, for each subset of a given task, specializes the module that gives the best performance – does not provide a satisfying specialization of modules in our task: in some cases the specialization is unstable, in others only one module is trained. We propose a new way to coordinate modules independently from their performances, in partitioning the environment into several sub-regions in which each expert is at work. This method gives good results in the two simulated environment.

3.4 Discussion

We find that a biologically plausible implementation of an Actor-Critic model can provide good results in a simulated robotics task involving a visual cue-guided strategy. We show that multiple modules are necessary to solve a continuous state space task like this one. However, existing methods to coordinate modules did not achieve good results in our task. We propose a new method

which is more adapted to the task and yields better performance. However, this method lacks autonomy as well as generalization abilities. Moreover, we did not study the precise influence of our biologically detailed architecture on the specialization process used in this method.
In the fourth part of this chapter, we present an improvement of the autonomy of our method by combining it with self-organizing maps (Kohonen, 1995).

Actor-Critic Models of Reinforcement Learning in the Basal Ganglia: From Natural to Artificial Rats

Preprint : accepted for publication in *Adaptive Behavior 13(2):131-148, Special Issue Towards Artificial Rodents, 2005.*

Mehdi Khamassi[1,2], Loïc Lachèze[1], Benoît Girard[1,2], Alain Berthoz[2] and Agnès Guillot[1]

[1]AnimatLab, LIP6, 8 rue du capitaine Scott, 75015 Paris, France

[2] LPPA, Collège de France, 11 place Marcellin Berthelot, 75005 Paris, France

Since 1995, numerous Actor-Critic architectures for reinforcement learning have been proposed as models of dopamine-like reinforcement learning mechanisms in the rat's basal ganglia. However, these models were usually tested in different tasks, and it is then difficult to compare their efficiency for an autonomous animat. We present here the comparison of four architectures in an animat as it performs the same reward-seeking task. This will illustrate the consequences of different hypotheses about the management of different Actor sub-modules and Critic units, and their more or less autonomously determined coordination. We show that the classical method of coordination of modules by mixture of experts, depending on each module's performance, did not allow solving the task. Then we address the question of which principle should be applied to efficiently combine these units. Improvements for Critic modeling and accuracy of Actor-critic models for a natural task are finally discussed in the perspective of our Psikharpax project – an artificial rat having to survive autonomously in unpredictable environments.

Keywords animat approach - TD learning - Actor-Critic model - S-R task - taxon navigation

1. Introduction

This work aims at adding learning capabilities in the architecture of action selection introduced by Girard *et al.* in this issue. This architecture will be implemented in the artificial rat Psikharpax, a robot that will exhibit at least some of the capacities of autonomy and adaptation that characterize its natural counterpart (Filliat *et al.*, 2004). This learning process capitalizes on Actor-Critic architectures, which have been proposed as models of dopamine-like reinforcement learning mechanisms in the rat's basal ganglia (Houk *et al.*, 1995). In such models, an Actor network learns to select actions in order to maximize the weighted sum of future rewards, as computed on line by another network, a Critic. The Critic predicts this sum by comparing its estimation of the reward with the actual one by means of a Temporal Difference (TD) learning rule, in which the error between two successive predictions is used to update the synaptic weights (Sutton and Barto, 1998). A recent review of numerous computational models, built on this principle since 1995, highlighted several issues raised by the inconsistency of the detailed implementation of Actor and Critic modules with known basal ganglia anatomy and physiology (Joel *et al.*, 2002). In the first section of this paper, we will consider some of the main issues, updated with anatomical and neurophysiological knowledge. In the second section, we will illustrate the consequences of alternative hypotheses concerning the various Actor-Critic designs by comparing animats that perform the same classical instrumental learning (S-R task). During the test, the animat freely moves in a plus-maze with a reward placed at the end of one arm. The reward site is chosen randomly at the beginning of each trial and it refers to site-specific local stimuli. The animat has to autonomously learn to associate continuous sensory information with certain values of reward and to select sequences of behaviors that enable it to reach the goal from any place in the maze. This experiment is more realistic than others used to validate Actor-Critic models, often characterized by an a priori fixed temporal interval between a stimulus and a reward (e.g., Suri and Schultz, 1998), by an unchanged reward location over trials (e.g., Strösslin, 2004), or by a discrete state space (e.g., Baldassarre, 2002).

We will compare, in this task, four different principles inspired by Actor-Critic models trying to tackle the issues evocated in the first section. The first one is the seminal model proposed by Houk *et al.* (1995), which uses one Actor and a single prediction unit (*Model AC* – one Actor, one Critic), which is supposed to induce learning in the whole environment. The second principle implements one Actor with several Critics *(Model AMC1* – one Actor, Multiple Critics). The Critics are combined by a mixture of experts where a gating

network is used to decide which expert – which Critic – is used in each region of the environment, depending on its performance in that region. The principle of mixture of experts is inspired from several existing models (Jacobs *et al.*, 1991; Baldassarre, 2002; Doya *et al.*, 2002). The third one is inspired by Suri and Schultz (2001) and uses also one Actor with several Critic experts. However, the decision of which expert should work in each sub-zone of the environment is independent from the experts' performances, but rather depends on a partition of the sensory space perceived by the animat (*Model AMC2* – one Actor, Multiple Critics). The fourth one (*Model MAMC2* – Multiple Actors, Multiple Critics) proposes the same principle as the previous Critic, combined with several Actors, which latter principle is one of the features of Doya *et al.*'s model (2002), particularly designed for continuous tasks, and is also a feature of Baldassarre's model (2002). Here we will implement these principles in four models using the same design for each Actor component. Their comparison will be made on the learning speed and on their ability to extend learning to the whole experimental environment.
The last section of the paper will discuss the results on the basis of acquired knowledge in reinforcement learning tasks in artificial and natural rodents.

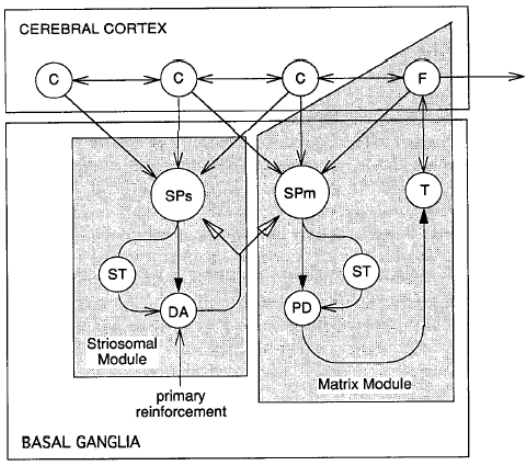


Figure 1 Schematic illustration of the correspondence between the modular organization of the basal ganglia including both striosomes and matrix modules and the Actor-Critic architecture in the model proposed by Houk et al, (1995). F, columns in the frontal cortex; C, other cortical columns; SPs, spiny neurons striosomal compartments of the striatum; SPm, spiny neurons in matrix modules; ST, subthalamic sideloop; DA, dopamine neurons in the substantia nigra pars compacta; PD, pallidal neurons; T, thalamic neurons. (adapted from Houk et al., 1995).

2. Actor-Critic designs: the issues

The two main principles of Actor-Critic models that lead to consider them as a good representation of the role of the basal ganglia in reinforcement learning of motor behaviors are (i): the implementation of a Temporal Difference (TD) learning rule which leads to translate progressively reinforcement signals from the time of reward occurrence to environmental contexts that precede the reward.; (ii): the separation of the model in two distinct parts, one for the selection of motor behaviors (actions) depending on the current sensory inputs (the Actor), and the other for the driving of the learning process via dopamine signals (the Critic).

Schultz's work on the electrophysiology of dopamine neurons in monkeys showed that dopamine patterns of release are similar to the TD learning rule (see Schultz, 1998 for a review). Besides, the basal ganglia are a major input to dopamine neurons, and are also a privileged target of reinforcement signals sent by these neurons (Gerfen *et al.*, 1987). Moreover, the basal ganglia appears to be constituted of two distinct sub-systems, related to two different parts of the striatum – the major input nucleus of the basal ganglia –, one projecting to motor areas in the thalamus, the other projecting to dopamine neurons, influencing the firing patterns of these neurons at least to some extent (Joel and Weiner, 2000).
These properties lead the first Actor-Critic model of the basal ganglia to propose the matrisomes of the striatum to constitute the Actor, and the striosomes of this very structure to be the Critic (Houk *et al.*, 1995, figure 1). The classical segregation of 'direct' and 'indirect' pathways from the striatum to the dopaminergic

system (SNc, substantia nigra pars compacta, and VTA, ventral tegmental area; Albin *et al.,* 1989) was used in the model to explain the timing characteristics of dopamine neurons' discharges.
Numerous models were proposed to improve and complete the model of Houk *et al*. However, most of these computational models have neurobiological inconsistencies and lacks concerning recent anatomical hypotheses on the basal ganglia (Joel *et al.*, 2002).
An important drawback is that the Actor part of these models is often simplistic compared to the known anatomy of the basal ganglia and does not take into account important anatomical and physiological characteristics of the striatum. For example, recent works showed a distinction between neurons in the striatum having different dopamine receptors (D1-receptors or D2-receptors; Aizman *et al.*, 2000). This implies at least two different pathways in the Actor, on which tonic dopamine has opposite effects, going beyond the classical functional segregation of 'direct' and 'indirect' pathways in the striatum (Gurney *et al.*, 2001).
Likewise, some constraints deriving from striatal anatomy restrict the possible architectures for the Critic network. In particular, the striatum is constituted of only one layer of medium spiny neurons – completed with 5% of interneurons (Houk *et al.*, 1995). As a consequence, Critic models cannot be constituted of complex multilayer networks for reward prediction computation. This anatomical constraint lead several authors to model the Critic as a single-neuron (Houk *et al.*, 1995; Montague *et al.*, 1996), which works well in relatively simple tasks. For more complicated tasks, several models assign one single Critic neuron to each subpart of the task. These models differ in the computational mechanism used to coordinate these neurons. Baldassarre (2002) and Doya et al. (2002) propose to coordinate Critic modules with a mixture of experts method: the module that has the best performance at a certain time during the task becomes expert in the learning process of this subpart of the task. Another model proposes an affectation of experts to subparts of the task (such as stimuli or events) in an a priori manner, independently from each expert's performance (Suri and Schultz, 2001). It remains to assess the efficiency of each principle, as they have been at work in heterogeneous tasks (*e.g.* Wisconsin Card Sorting Test, Discrete Navigation Task, Instrumental Conditioning).
These models also question the functional segregation of the basal ganglia in 'direct' and 'indirect' pathways (see Joel *et al.*, 2002 for a review). These objections are built on electrophysiological data (for review see Bunney *et al.*, 1991) and anatomical data (Joel and Weiner, 2000) which show that these two pathways are unable to produce the temporal dynamics necessary to explain dopamine neurons patterns of discharge. These findings lead to question the localization of the Critic in the striosomes of the dorsal striatum, and several models capitalized on its implementation in the ventral striatum (Brown *et al.*, 1999; Daw, 2003). These works are supported by recent fMRI data in humans, showing a functional dissociation between dorsal striatum as the Actor and ventral striatum as the Critic (O'Doherty *et al*., 2004), but they may be controversial for the rat, as electrophysiological data (Thierry *et al.*, 2000) showed that an important part of the ventral striatum (the nucleus accumbens core) does not project extensively to the dopamine system in the rat brain.
We can conclude that the precise implementation of the Critic remains an open question, if one takes also into account a recent model assuming that a new functional distinction of striosomes in the dorsal striatum – based on differential projections to GABA-A and GABA-B receptors in dopamine neurons – can explain the temporal dynamics expected (Frank *et al.*, 2001).
Besides these neurobiological inconsistencies, some computational requirements on which numerous Actor-Critic models have focused seem unnecessary for a natural reward-seeking task. For example, as Houk *et al*.'s model could not account for temporal characteristics of dopamine neurons firing patterns, most of the alternative models focused on the simulation of the depression of dopamine at the precise time where the reward is expected when it eventually does not occur. To this purpose, they concentrated on the implementation of a temporal component for stimulus description – which is computed outside of the model and is sent as an input to the model via cortical projections (Montague *et al.*, 1996; Schultz *et al.*, 1997). These models were tested in the same tasks chosen by Schultz *et al.* (1993) to record dopamine neurons in the monkey, using a fixed temporal bin between a stimulus and a reward. However, in natural situations where a rodent needs to find food or any other type of reward, temporal characteristics of the task are rarely fixed but rather depend on the animal's behavior and on the environment's changes/evolution.

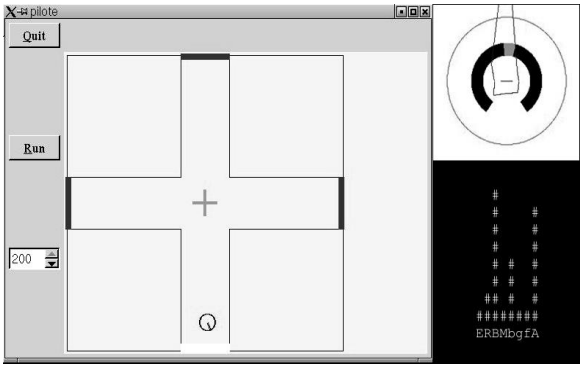


Figure 2 Left: the robot in the plus maze environment. A white arm extremity indicates the reward location. Other arm extremities do not deliver any reward and are shown in black. Upper right: the robot's visual perceptions. Lower right: activation level of different channels in the model.

3. Method

The objective of this work is to evaluate the efficiency of the main principles on which current Actor-Critic models inspired by the basal ganglia are designed, when they are implemented in the same autonomous artificial system. The main addressed issues are:

5. The implementation of a detailed Actor, whose structure would be closer to the anatomy of the dorsal striatum, assessing whether reinforcement learning is still possible within this architecture.
6. The comparison of the function of one Critic unit, versus several alternative ways to coordinate different Critic modules for solving a complex task where a single-neuron is not enough.
7. The test of the models in a natural task involving taxon navigation where events are not predetermined by fixed temporal bins. Instead, the animat perceives a continuous sensory flow during its movements, and has to reactively switch its actions so as to reach a reward.

3.1 The simulated environment and task

Figure 2 shows the experimental setup simulated, consisting in a simple 2D plus-maze. The dimensions are equivalent to a 5m * 5m environment with 1m large corridors. In this environment, walls are made of segments colored on a 256 grayscale. The effects of lighting conditions are not simulated. Every wall of the maze is colored in black (luminance = 0), except walls at the end of each arm and at the center of the maze, which are represented by specific colors: the cross at the center is gray (191), three of the arm extremities' walls are dark gray (127) and the fourth is white (255), indicating the reward location (equivalent to a water trough delivering two drops – non instantaneous reward – not a priori known by the animat).
The plus-maze task mimics the neurobiological and behavioral studies that will serve as future validation for the model (Albertin *et al.*, 2000). In this task, at the beginning of each trial, one arm extremity is randomly chosen to deliver reward. The associated wall is colored in white whereas walls at the three other extremities are dark gray. The animat has to learn that selecting the action *drinking* when it is near the white wall (distance < 30 cm) and faces it (angle < 45 degrees) gives it a reward. Here we assume that reward = 1 for n iterations (n = 2), without considering how the hedonic value of this reward is determined.
We expect the animat to learn a sequence of context-specific behaviors, so that it can reach the reward site from any starting point in the maze:

4. When not seeing the white wall, face the center of the maze and move forward.
5. As soon as arriving at the center (the animat can see the white wall), turn to the white stimulus.
6. Move forward until being close enough to reward location.
7. Drink.

The trial ends when reward is consumed: the color of the wall at reward location is changed to dark gray, and a new arm extremity is chosen randomly to deliver reward. The animat has then to perform again the learned behavioral sequence. Note that there is no break between two consecutive trials: trials follow each other successively.
The more efficiently and fluently the animat performs the above described behavioral sequence, the less time it will take to reach the reward. As a consequence, the criterion chosen to validate the models is the time to

goal, plotted along the experiment as the learning curve of the model.

3.2 The animat

The animat is represented by a circle (30 cm diameter). Its translation and rotation speeds are 40 $cm.s^{-1}$ and $10°.s^{-1}$. Its simulated sensors are:

- An omnidirectional linear camera providing every 10° the color of the nearest perceived segment. This results in a 36 colors table that constitute the animat's visual perception (see figure 2),
- Eight sonars with a 5m range, an incertitude of ±5 degrees concerning the pointed direction and an additional ±10 cm measurement error,

The sonars are used by a low level obstacle avoidance reflex which overrides any decision taken by the Actor-Critic model when the animat comes too close to obstacles.
The animat is provided with a visual system that computes 12 input variables $(\forall i \in [1;12], 0 < \mathrm{var}_i < 1)$ out of the 36 colors table at each time step. These sensory variables constitute the state space of the Actor-Critic and so will be taken as input to both the Actor and the Critic parts of the model (figure 3). Variables are computed as following:

1. *seeWhite*(resp. *seeGray*, *seeDarkGray*) = 1 if the color table contains the value 255 (resp. 191, 127), else 0.
2. *angleWhite*, *angleGray*, *angleDarkGray* = (number of boxes in the color table between the animat's head direction and the desired color) / 18.
3. *distanceWhite*,*distanceGray, distanceDarkGray* = (maximum number of consecutive boxes in the color table containing the desired color) / 18.
4. *nearWhite* (resp. *nearGray*, *nearDarkGray*) = 1 – *distanceWhite* (resp. *distanceGray, distanceDarkGray*).

Representing the environment with such continuous variables will imply for the model to permanently receive a flow of sensory information and having to learn autonomously the events (sensory contexts) that can be relevant for the task resolution.
The animat has a repertoire of 6 actions: *drinking*, *moving forward*, *turning to white perception*, *turning to gray perception*, *turning to dark gray perception*, and *waiting*. These actions constitute the output of the Actor model (described below) and the input to a low-level model that translates it into appropriate orders to the animat's engines.

3.3 The model: description of the Actor part

The Actor-Critic model is inspired by the rat basal ganglia. As mentioned in section 2, the Actor can be hypothesized as implemented in the matrix part of the basal ganglia, while striosomes in the dorsal striatum are considered as the anatomical counterpart for the Critic. The Critic produces dopamine-like reinforcement signals that help it learn to predict reward during the task, and that make the Actor learn to select appropriate behaviors in every sensory context experienced during the task.
The architecture implemented in the Actor is a recent model proposed by Gurney, Prescott and Redgrave (2001a,b) – henceforth called GPR model - that replaces the simple winner-takes-all which usually constitutes Actor models and is supposed to be more biologically plausible.
Like other Actors, the GPR is constituted of a series of parallel channels, each one representing an action (in our implementation, we used 6 channels corresponding to the 6 actions used for the task). This architecture constitutes an alternative view to the prevailing functional segregation of the basal ganglia into 'direct' and 'indirect' pathways discussed in section 1 (Gurney *et al.*, 2001). All these channels are composed by two different circuits through dorsal striatum: the first is the 'selection' pathway, implementing action selection properly via a feed-forward off-center on-surround network, and mediated by cells in the dorsal striatum with D1-type receptors. The second is the 'control' pathway, mediated by cells with D2-type receptors in the same area. Its role is to regulate the selection by enhancing the selectivity inter-channels, and to control the global activity within the Actor. Moreover, a cortex-basal ganglia-thalamus loop in the model allows it to take into account each channel's persistence in the process of selection (see Gurney *et al.*, 2001, for detailed description and mathematical implementation of the model). The latter characteristic showed some

interesting properties that prevented a robot from performing behavioral oscillations (Montes-Gonzalez *et al.*, 2000; Girard *et al.*, 2003).
In our implementation, the input values of the Actor model are saliences – i.e. the strength of a given action – that are computed out of the 12 sensory variables, a constant implementing a bias, and a persistence factor – equal to 1 for the action that was selected at previous timestep (figure 3). At each timestep t (timesteps being separated by a 1 sec bin in our simulations), the action that has the highest salience is selected to be performed by the animat, the salience of action i being:

$$sal_i(t) = \left[\sum_{j-1}^{13} \text{var}_j(t) \cdot w_{i,j}(t) \right] + persist_i(t) \cdot w_{i,14}(t) \qquad (1)$$

where $\text{var}_{13}(t) = 1, \ \forall t$, and the $w_{i,j}(t)$ are the synaptic weights representing, for each action *i*, the association strength with input variable *j*. These weights are initiated randomly $(\forall i,j, -0.02 < w_{i,j}(t=0) < 0.02)$ and the objective of the learning process will be to find a set of weights allowing the animat to perform the task efficiently.
An exploration function is added that would allow the animat to try an action in a given context even if the weights of the Actor do not give a sufficient tendency to perform this action in the considered context.
To do so, we introduce a clock that triggers exploration in two different cases:

1. When the animat has been stuck for a large number of timesteps (*time* superior to a fixed threshold α) in a situation that is evaluated negative by the model (when the prediction *P*(t) of reward computed by the Critic is inferior to a fixed threshold).
2. When the animat has remained for a long time in a situation where *P*(t) is high but this prediction doesn't increase that much ($|P(t+n) - P(t)| < \varepsilon$) and no reward occurs.

If one of these two conditions is true, exploration is triggered: one of the 6 actions is chosen randomly. Its salience is being set to 1 (Note that: when exploration = false, $sal_i(t) < 1, \ \forall i,t, w_{i,j}(t)$) and is being maintained to 1 for a duration of 15 timesteps (time necessary for the animat to make a 180° turn or to run from the center of the maze until the end of one arm).

3.4 The model: description of the Critic part

For the Critic part of the model, different principles based on existing techniques are tested. The idea is to test the hypothesis of one single Critic unit first, but also to provide the Critic with enough computational capacities so that it can correctly estimate the value function over the whole environment of the task. In other words, the Critic will have to deal with several different sensory contexts – corridors, maze center, extremity of arms, etc. equivalent to different stimuli –, and will have to associate a correct reward prediction to these contexts.
One obvious possibility would be a multilayer perceptron with several hidden layers but, as mentioned before in section 2, there are anatomical constraints which prevent us from adopting this choice: our Critic is supposed to be situated in the striosomes of dorsal striatum, which structure is constituted of only one layer of medium spiny neurons (Houk *et al.*, 1995). Thus we need a more general method that combines several Critic modules, each one being constituted of a single neuron and dealing with a particular part of the problem space.
The method adopted here is the mixture of experts, which was proposed to divide a non-linearly separable problem into a set of linearly separable problems, and to affect a different expert to each considered sub-problem (Jacobs *et al.*, 1991).
The Critics tested in this work differ mainly in two following manners:

- The first (*Model AMC1*) implements a mixture of experts in which a gating network is used to decide which expert is used in each region.
- The second (*Model AMC2*) implements a mixture of experts in which a hand-determined partition of the environment based on a categorization of visual perceptions is used to decide which expert works in each sub-zone.

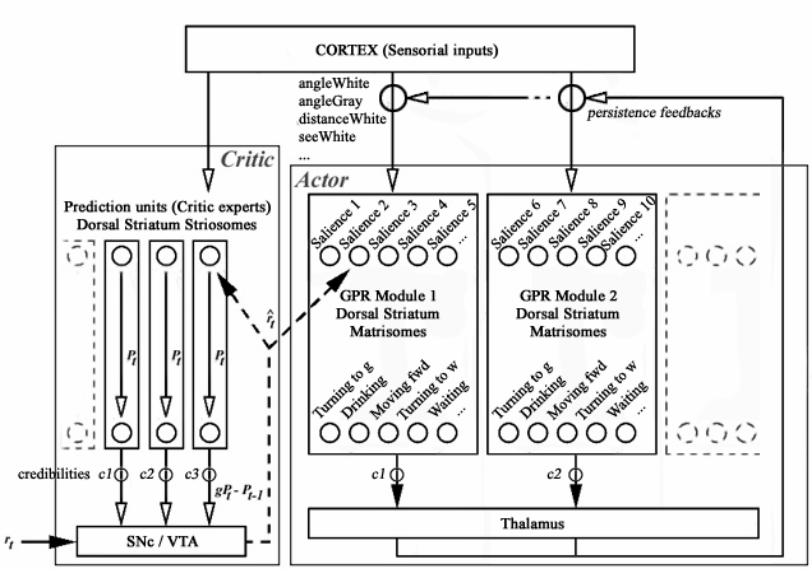


Figure 3 General scheme of the models tested in this work. The Actor is a group of GPR modules with saliences as inputs and actions as outputs. The Critic (involving striosomes in the dorsal striatum, and the substantia nigra compacta (SNc)) propagates towards the Actor an estimate ř of the instantaneous reinforcement triggered by the selected action. The particularity of this scheme is to combine several modules for both Actor and Critic, and to weight the Critic experts' predictions and the Actor modules' decisions with credibilities. These credibilities can be either computed by a gating network (*Model AMC1*) or in a context-dependent manner (*Models AMC2 and MAMC2*).

Moreover, since the animat has to solve a task in continuous state space, there could be interferences between reinforcement signals sent by different Critic experts to the same single Actor. In this way, whereas one model will employ only one Actor (*Model AMC2*), another one will use one Actor module associated to each expert (*Model MAMC2*). Figure 3 shows the general scheme with different modules employed as suggested by the models presented here.
Performances of *Models AMC1*, *AMC2* and *MAMC2* will be compared, together with the one of the seminal Actor-Critic model inspired by the basal ganglia, proposed by Houk, Adams and Barto (1995), and using a single cell Critic with a single Actor (*Model AC*).
We will start by the description of the simplest Critic, the one belonging to *Model AC*.

3.4.1 Model AC

In this model, at each timestep, the Critic is a single linear cell that computes a prediction of reward based on the same input variables than the Actor, except the persistence variable:

$$P(t)=\sum_{j=1}^{13} \mathrm{var}_j(t)\cdot w'_j(t) \qquad (2)$$

where $w'_j(t)$ are the synaptic weights of the Critic.
This prediction is then used to calculate the reinforcement signal by means of the TD-rule:

$$\hat{r}(t)=r(t)+gP(t)-P(t-1) \qquad (3)$$

where $r(t)$ is the actual reward received by the animat, and g is the discount factor ($0 < g < 1$) which determines how far in the future expected rewards are taken into account in the sum of future rewards.
Finally, this reinforcement signal is used to update both Actor's and Critic's synaptic weights according to the following equations respectively:

$$w_{i,j}(t)\leftarrow w_{i,j}(t-1)+\eta\cdot\hat{r}(t)\cdot \mathrm{var}_j(t-1) \qquad (4)$$

$$w'_j(t)\leftarrow w'_j(t-1)+\eta\cdot\hat{r}(t)\cdot \mathrm{var}_j(t-1) \qquad (5)$$

where $\eta>0$ is the learning rate.

3.4.2 Model AMC1

As this Critic implements N experts, each expert k computes its own prediction of reward at timestep t:

$$p_k(t)=\sum_{j=1}^{13} w'_{k,j}(t)\cdot \mathrm{var}_j(t) \quad (6)$$

where the $w'_{k,j}(t)$ are the synaptic weights of expert *k*.

Then the global prediction of the Critic is a weighted sum of experts' predictions:

$$P(t)=\sum_{k=1}^{N} cred_k(t)\cdot p_k(t) \quad (7)$$

where $cred_k(t)$ is the credibility of expert *k* at timestep *t*. These credibilities are computed by a gating network which learns to associate, in each sensory context, the best credibility to the expert that makes the smaller prediction error. Following Baldassarre's description (2002), the gating network is constituted of N linear cells which receive the same input variables than the experts and compute an output function out of it:

$$o_k(t)=\sum_{j=1}^{13} w''_{k,j}(t)\cdot \mathrm{var}_j(t) \quad (8)$$

where $w''_{k,j}(t)$ are the synaptic weights of gating cell *k*.

The credibility of expert *k* is then computed as the softmax activation function of the outputs $o_f(t)$:

$$cred_k(t)=\frac{o_k(t)}{\sum_{f=1}^{N} o_f(t)} \quad (9)$$

Concerning learning rules, whereas equation (3) is used to determine the global reinforcement signal sent to the Actor, each Critic's expert has a specific reinforcement signal based on its own prediction error:

$$\hat{r}_k(t)=r(t)+gP(t)-p_k(t-1) \quad (10)$$

The synaptic weights of each expert *k* are updated according to the following formula:

$$w''_{k,j}(t)\leftarrow w''_{k,j}(t-1)+\eta\cdot\hat{r}_k(t)\cdot \mathrm{var}_j(t-1)\cdot h_k(t) \quad (11)$$

where $h_k(t)$ is the contribution of expert *k* to the global prediction error of the Critic, and is defined as:

$$h_k(t)=\frac{cred_k(t-1)\cdot corr_k(t)}{\sum_{f=1}^{N} cred_f(t-1)\cdot corr_f(t)} \quad (12)$$

where $corr_k(t)$ is a measure of the « correctness » of the expert *k* defined as:

$$corr_k(t)=\exp\left(\frac{-\hat{r}_k(t)^2}{2\sigma^2}\right) \quad (13)$$

where σ is a scaling parameter depending on the average error of the experts (see parameters table in the appendix section).

Finally, to update the weights of the gating network, we use the following equation:

$$w''_{k,j}(t)\leftarrow w''_{k,j}(t-1)+m\cdot diff(t)\cdot \mathrm{var}_j(t-1) \quad (14)$$

with $diff(t)=h_k(t)-cred_k(t-1)$ (15)

where *m* is a learning rate specific to the gating network.

So the credibility of expert *k* in a given sensory context depends on its performance in this context.

3.4.3 Model AMC2

The Critic also implements N experts. However, it differs from *Model AMC1* in the way the credibility of each expert is computed.

The principle we wanted to bring about here is to dissociate credibilities of experts from their performance. Instead, experts would be assigned to different subregions of the environment – these regions being computed as windows in the perceptual space –, would remain enchained to their associate region forever, and would progressively learn to accurate their performance along the experiment. This principle is declined

from Houk et al. (1995) for the improvement of their model, assuming that different striosomes may be specialized in dealing with different behavioral tasks. This proposition was implemented by Suri and Schultz (2001) in using several TD models, each one computing predictions for only one event (stimulus or reward) that occurs in the simulated paradigm.
To test this principle, we replaced the gating network by a hand-determined partition of the environment (e.g. a coarse representation of the sensory space): At timestep *t*, the current zone β depends on the 12 sensory variables computed by the visual system. *Example: if (seeWhite = 1 and angleWhite < 0.2 and distanceWhite > 0.8) then zone = 4 (e.g. β=4).*Then $cred_\beta(t)=1$, $cred_k(t)=0$ for all other experts, and expert β has then to compute a prediction of reward out of the 12 continuous sensory variables. Predictions and reinforcement signals of the experts are determined by the same equations than Critic of *Model AMC1*.
This was done as a first step in the test of the considered principle. Indeed, we assume that another brain region such as the parietal cortex or the hippocampus would determine the zone (sensory configuration) depending on the current sensory perception (McNaughton, 1989; Burgess *et al.*, 1999), and would send it to the Actor-Critic model of the basal ganglia. Here, the environment was partitioned into N=30 zones, an expert being associated to each zone. The main difference between this scheme and the one used by Suri and Schultz is that, in their work, training of experts in each sub-zone was done in separated sessions, and the global model was tested on the whole task only after training of all experts. Here, experts will be trained simultaneously in a single experiment.
Finally, one should note that this method is different from applying a coarse coding of the state space that constitutes the input to the Actor and the Critic (Arleo and Gerstner, 2000). Here, we implemented a *coarse coding of the credibility space* so as to determine which expert is the most credible in a given sensory configuration, and kept the 12 continuous sensory variables, plus a constant described above, as the state space for the reinforcement learning process. This means that within a given zone, the concerned expert has to learn to approximate a continuous reward value function, based on the varying input variables.

3.4.4 Model MAMC2

The Critic of this Model is the same as in *Model AMC2* and only differs from its associated Actor.
Instead of using one single Actor, we implemented N different Actor modules. Each Actor module has the same structure than the simple Actor described in section 3.4 and is constituted of 6 channels representing the 6 possible actions for the task. The difference resides in the fact that only actions of the Actor associated with the zone in which the animat is currently are competing to determine the animat's current action.
As a consequence, if the animat was in zone β at time *t* and performed action *i*, the reinforcement signal $\hat{r}(t+1)$ computed by the Critic at next timestep will be used to update only weights of action *i* from the Actor β according to the following equation:

$$w_{k,i,j}(t) \leftarrow w_{k,i,j}(t-1)+\eta\cdot\hat{r}(t)\cdot\mathrm{var}_j(t-1) \qquad (16)$$

Other equations are the same than those used for Critic of *Model AMC2*. As mentioned above, this principle – using a specific controller or a specific Actor for each module of the Actor-Critic model – is inspired by the work of Doya *et al.*, (2002).

4. Results

In order to compare the learning curves of the four simulated models, and so as to evaluate which models manage to solve the task efficiently, we adopt the following criterion: after 50 trials of training (out of 100 for each experiments), the animat has to achieve an equivalent performance to a hand-crafted model that can already solve the task (Table 1). To do so, we simulated the GPR action selection model with appropriate hand-determined synaptic weights and without any learning process, so that the animat can solve the task as if it had already learned it. With this model, the animat performed a 50 trials experiment with an average performance of 142 iterations per trial. Since each iteration lasted approximately 1 sec, as mentioned above, it took a little bit more than 2 min per trials to this hand-craft animat to reach the reward.

Table 1. Performances of each model.

Model	GPR	AC	AMC1	AMC2	MAMC2
Performance	142	587	623	3240	97

Table 1 shows the performance of each model, measured as the average number of iterations per trial after trial #50. Figure 4 illustrates results to the four experiments performed in the 2D environment, one per model. The x-axis represents the successive trials along the experiments. For each trial, y-axis shows the number of iterations needed for the animat to reach the reward and consume it. Figure 4.a shows the learning curve of *Model AC*. We can first notice that the model increased rapidly its performance until trial 7, and stabilized it at trial 25. However, after trial 50, the average duration of a trial is still 587 iterations, which is nearly 4 times higher than the chosen criterion. We can explain this limitation by the fact that *Model AC* is constituted of only one single neuron in the Critic, which can only solve linearly separable problems. As a consequence, the model could learn only a part of the task – in the area near the reward location –, but it was unable to extend learning to the rest of the maze. So the animat has learned to select appropriate behaviors in the reward area, but it still performs random behaviors in the rest of the environment.

Model AMC1 is designed to mitigate the computational limitations of *Model AC*, as it implies several Critic units controlled by a gating network. Figure 4.b shows its learning curve after simulation in the plus-maze task. The model has also managed to decrease its running time per trial at the beginning of the experiment. However, we can notice that the learning process is more unstable than the previous one. Furthermore, after the 50th trial, the model has a performance of 623 iterations, which is not better than *Model AC*. Indeed, the model couldn't extend learning to the whole maze either. We can explain this failure by the fact that the gating network did not manage to specialize different experts in different subparts of the task. As an example, figure 5 shows the reward prediction computed by each Critic's expert during the last trial of the experiment. It can be noticed that the first expert (dark curve) has the highest prediction throughout the whole trial. This is due to the fact that it is the only one the gating network has learned to consider as credible – its credibility remains above 90% during the whole experiment. As a consequence, only one expert is involved in the learning process and the model becomes computationally equivalent to *Model AC*: it cannot extend learning to the whole maze, which is confirmed by the absence of any reward prediction before the perception of the reward site (stimulus occurrence) in Figure 5.

Figure 4.c shows the learning curve of *Model AMC2* which implements another principle for experts coordination. This model cannot suffer from the same limitations than *Model AMC1*, since each expert was a priori assigned to a specific area of the environment. As a consequence, it quickly managed to extend learning to the whole maze. However, the consequence of this process is to produce interferences in the Actor's computations: the same Actor receives all experts' teaching signals, and it remains unable to switch properly between reinforced behaviors. For example, when the action '*drinking*' is reinforced, the Actor starts selecting this action permanently, even when the animat is far from reward location. These interferences explain the very bad performances obtained with *Model AMC2*.

The last simulated model (*Model MAMC2*) performed best. Its learning curve is shown on figure 4.d. This model implements several Actor modules (an Actor module connected to each Critic expert). As a consequence, it avoids interferences in the learning process and rapidly converged to a performance of 97 iterations per trial. This good performance cannot been reached with the multi-Actor only, since we tried to combined several Actor modules to model *AMC1* and got a performance of 576 iterations per trial. So the achievement of the task implies the combination of a multi-Actor and a good specialization of experts.

For checking the ability of *Model MAMC2* to learn the same task in more realistic conditions, we simulated it a 3D environment, working in real time and implementing physical dynamics (Figure 7). This experiment constituted an intermediary step favoring the implementation into an actual Pekee robot (Wany Robotics). The animat is still able to learn the task in this environment and gets good performances after 35 trials (Figure 6; corresponding average performance of the animat between trials 35 and 65: 284 iterations per trial).

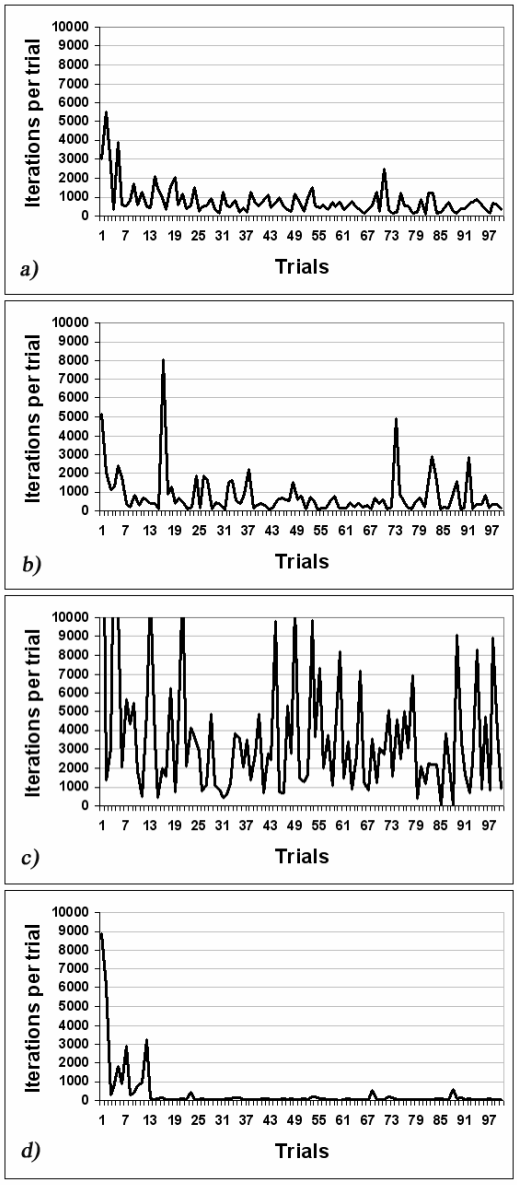


Figure 4 Learning curves of the four models simulated in the 2D plus-maze task over 100 trials experiments. X-axis: trials. Y-axis: number of iterations per trial (truncated to 10000 it. for better readability). **a)** *Model AC*. **b)** *Model AMC1*. **c)** *Model AMC2*. **d)** *Model MAMC2*.

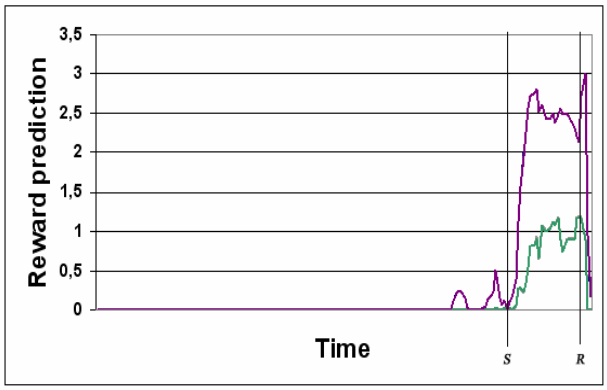


Figure 5 Reward prediction computed by each Critic's expert of Model AMC1 during trial #100 of the experiment. Time 0 indicates the beginning of the trial. S: perception of the stimulus (the white wall) by the animat. R: beginning of reward delivery. The dark curve represents the prediction of expert 1. The other experts' predictions are melted into the light curve or equal to 0.

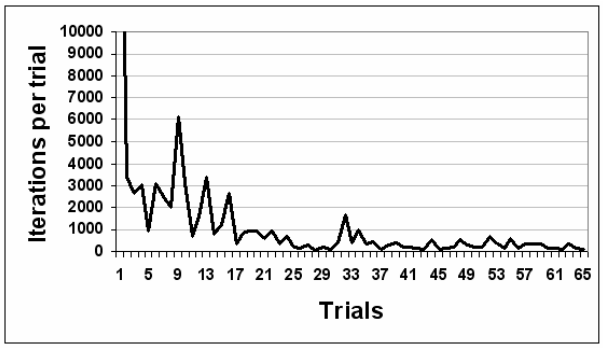


Figure 6 Learning curve in the 3D environment. X-axis: trials. Y-axis: number of iterations per trial.

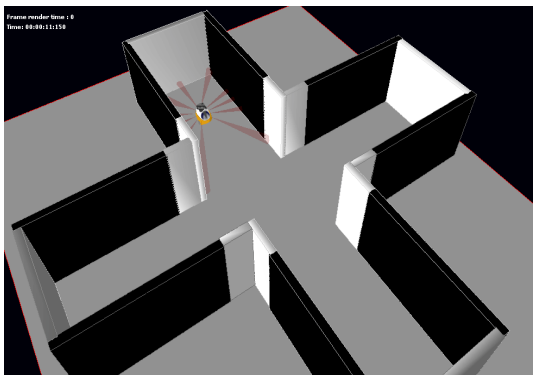


Figure 7 Simulation of the plus-maze task in a 3D environment. Like the 2D environment, one random arm extremity is white and delivers reward. The animat has to perform taxon navigation so as to find and consume this reward. Gray

stripes arising from the animat's body represent its sonar sensors used by its low level obstacle avoidance reflex.

5. Discussion and future work

In this work, we compared learning capabilities on a S-R task of several Actor-Critic models of the basal ganglia based on distinct principles. Results of simulations with *models AC, AMC1, AMC2* and *MAMC2* demonstrated that:

- A single-component Critic cannot solve the task (*Model AC*);
- Several Critic modules controlled by a gating network (*Model AMC1*) cannot provide good specialization, and the task remains unsolved.
- Several Critic modules a priori associated with different subparts of the task (*Model AMC2*) and connected to a single Actor (an Actor component being composed of a 6 channels GPR) allow learning to extend to areas that are distant from reward location, but still suffer from interferences between signals sent by the different Critic to the same single Actor.

Model MAMC2, combining several Critic modules with the principle of *Model AMC2*, and implementing several Actor components produces better results in the task at matter, spreading learning in the whole maze and reducing the learning duration. However, there are a few questions that have to be raised concerning the biological plausibility and the generalization ability of this model.

5.1 Biological plausibility of the proposed model

When using a single GPR Actor, each action is represented in only one channel – an Actor module being constituted of one channel per action (Gurney *et al.*, 2001) – and the structural credit assignment problem – which action to reinforce when getting a reward – can be simply solved: the action that has the highest salience inhibits its neighbors via local recurrent inhibitory circuits within D1 striatum (Brown and Sharp, 1995). As a consequence, only one channel in the Actor will have enough pre- and post-synaptic activity to be eligible for reinforcement.

When using several Actor modules, this property is not true anymore: even if only one channel per Actor module may be activated at a given time, each Actor module will have its own activated channel, and several concurring synapses would be eligible for reinforcement within the global Actor. To solve this problem, we considered in our work that only one channel in the entire Actor is eligible at a given time. However, this implies for the basal ganglia to have one of the two following characteristics: it should either exist non-local inhibition between Actor modules within the striatum, or there should be some kind of selectivity in the dopamine reinforcement signals so that even if several channels are activated, only those located in the target module receives dopamine signals.

To the best of our knowledge, these characteristics were not found in the basal ganglia, and a few studies tend to refute the dopamine selectivity (Pennartz, 1996).

5.2 Computational issues

Several computational issues need also to be addressed. First, the results presented here show that the learning process was not perturbed by the fact to use an Actor detailing the action selection process in the basal ganglia. This Actor has the property to take into account some persistence provided by the cortex-basal ganglia-thalamus-cortex loops. The way this persistence precisely influence the learning process in the different principles compared in this work was not thoroughly studied here. However we suspect that persistence could probably challenge the way different Actors interact with Critic's experts, as switching between actions does not exactly follow switches in sensorimotor contexts with this model. This issue should be examined in a future work.

Generalization ability of the multi-module Actor: Another issue that needs to be addressed here is the generalization ability of the multi-module Actor model used in this experiment. Indeed, Model MAMC2 avoids interferences in the Actor because hand-determined subzones of the maze are absolutely disjoint. In other words, learned stimulus-response associations in a given zone cannot be performed in another zone, and do not interfere with the learning process is this second zone even if visual contexts associated to each of

them are very similar. However, this leads also to an inability to generalize from one zone to the other: even if the distinction we made between two zones seemed relevant for the plus-maze task, if these two zones are similar and would imply similar motor responses in another task, the animat would have to learn twice the same sensorimotor association – one time in each zone. As a consequence, the partition we set in this work is task-dependent.
Instead, the model would need a partitioning method that autonomously classifies sensory contexts independently from the task, can detect similarities between two different contexts and can generalize learned behaviors in the first experienced context to the second one.

About the precise time of reward delivery:
In the work presented here, the time of reward delivery depends exclusively on the animat's behavior, which differs from several other S-R tasks used to validate Actor-Critic models of the basal ganglia. In these tasks, there is a constant duration between a stimulus and a reward, and several Actor-Critic models were designed so as to describe the precise temporal dynamics of dopaminergic neurons in this type of task (Montague *et al.*, 1996). As a consequence, numerous Actor-Critic models focused on the implementation of a time component for stimulus representation, and several works capitalized on this temporal representation for the application of Actor-Critic models of reinforcement learning in the basal ganglia to robotics (Perez-Uribe, 2001; Sporns and Alexander, 2002). Will we need to add such a component to our model to be able to apply it to certain type of natural tasks, or survival tasks?
In the experiments presented here, we didn't need such a temporal representation of stimuli because there was sufficient information in the continuous sensory flow perceived by the animat during its moves, so that the model can dynamically adapt its reward predictions, as observed also in another work (Baldassarre and Parisi, 2000). For example, when the animat is at the center of the maze, perceives the white wall (stimulus predicting reward) and moves towards reward location, the latter stimulus becomes bigger in the visual field of the animat, and the model can learn to increase its reward prediction, as shown in figure 8. We didn't aim at explaining the depression of dopamine neurons' firing rates when a reward doesn't occur, nevertheless we were able to observe this phenomenon in cases where the animat was approaching the reward site, was about to consume it, but finally turned away from it (R events in figure 8).

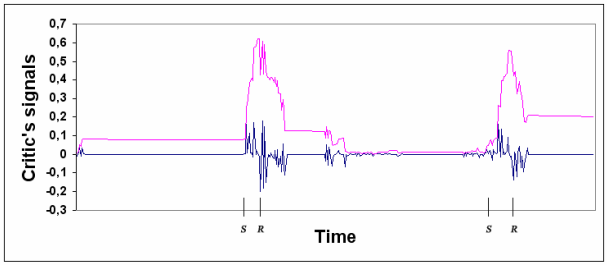


Figure 8 Reward prediction (light curve) and dopamine reinforcement signal (dark curve) computed by Critic of Model *MAMC2* in the 3D environment. X-axis: time. Y-axis: Critic's signals. S : perception of the stimulus (white wall) by the animat; R: Reward missed by the animat.

Using Critics dependent or independent from the performance: In our experiments, *Model AMC1*, implementing a gating network for experts' credibilities computation, did not solve the task. We saw in section 2 that, during the simulations, one expert became rapidly the most credible, which forced the model to use only one neuron to solve the task. The use of gating networks in the frame of mixture of experts methods has already being criticized (Tang *et al.*, 2002). According to these authors, this approach works well on problems composed of disjoint regions but does not generalize well, suffering from effects on boundaries of regions.
In our case, we explain the failure in the experts' specialization with *Model AMC1* by the observation that until the model has started to learn the task, and so can propagate teaching signals to the rest of the maze, only reward location has a value. As a consequence, it is the only area where the gating network tries to train an expert, and the latter rapidly reaches a high credibility. Then, as reward value starts to be extended to a new zone, this same expert still has the best credibility while getting bad performances. Other experts do not have significantly better performances – since they were not trained yet and since the new area and the first one are not disjoint. As a consequence, they remain non credible and the model starts having bad performances.

In his work, Baldassarre managed to obtain a good specialization of experts (Baldassarre, 2002). This may be partly explained by the fact that his task involved three different rewards located in three different sensory contexts. The simulated robot had to visit all rewards alternatively since the very beginning of the task. This may have helped the gating network to attribute good credibilities to several experts. However, reward locations in Baldassarre's task are not perfectly disjoint, which result in a difficult specialization: one of the experts is the most credible for two of the three rewards (see Baldassarre, 2002).
Another model (Tani and Nolfi, 1999) proposes a different mixture of experts where the gating network is replaced with a dynamical computation of experts' credibilities. Their model managed to categorize the sensori-motor flow perceived by a simulated robot during its movements. However, their method does not use any memory of associations between experts' credibilities and different contexts experienced during the task. As a consequence, experts' specialization is even more dependent to each expert's performance than Baldassarre's gating network, and suffers from the same limitation when applied to reinforcement learning in our plus-maze task - as we experimented in unpublished work.

Combining self-organizing maps with mixture of expert: To test the principle of dissociating the experts credibility from their performance, we partitioned the environment into several sub-regions. Yet, this method is ad hoc, lacks autonomy, and suffers generalization abilities if the environment is changed or becomes more complex. We are currently implementing Self-Organizing Maps (SOM) as a method of autonomous clustering of the different sensory contexts will be used to determine these zones. Note that this proposition differs from the traditional use of SOM to cluster the state space input to experts or to Actor-Critic models (Smith, 2002; Lee *et al.*, 2003). It is rather a clustering of the credibility space, which was recently proposed by Tang *et al.* (2002). We also would like to compare the use of SOM with the use of place cells. Indeed models of hippocampal place cells have already been used for coarse coding of the input state space to the Actor and the Critic (Arleo and Gerstner, 2000; Foster *et al.*, 2000; Strösslin, 2004) but, in our case, we would like to use place cells to determine experts' credibilities.

As often mentioned in the literature, and as confirmed in this work, the application of Actor-Critic architectures to continuous tasks is more difficult than their use in discrete tasks. Several other works have been done on the subject (Doya, 2000). However, these architectures still have to be improved so as to decrease their learning time:
Particularly, the learning performance of our animat seems still far from the learning speed that real rat can reach in the same task (Albertin *et al.*, 2000), even if the high time constant that we used in our model does not allow a rigorous comparison yet (cf. parameters table in the appendix). This could be at least partly explained by the fact that we implemented only S-R learning (or habit learning), whereas it has recently been known that rats are endowed with two distinct learning systems related to different cortex-basal ganglia-thalamus loops: a habit learning system and a goal-directed learning one (Ikemoto and Panksepp, 1999; Cardinal *et al.*, 2002). The latter would be fast, used at the early stages of learning, and implying an explicit representation of rewarding goals or an internal representation of action-outcome contingencies. The former would be very slow and takes advantage of the latter when the animat reaches good performances and becomes able to solve the task with a reactive strategy (S-R) (Killcross and Coutureau, 2003; Yin *et al.*, 2004).
Some theoretical work has already been started to extend Actor-Critic models to this functional distinction (Dayan, 2001). In the practical case of our artificial rat, both such systems could be useful in two different manners.
First, it could be useful to upgrade the exploration function. This function could have an explicit representation of different places of the environment, and particularly of the reward site. Then, when the animat gets reward for the first time, the exploration function would guide it trying behaviors that can allow it to reach the explicitly memorized reward location. The function could also remember which behaviors have already been tried unsuccessfully in the different areas, so that untried behaviors are selected instead of random behaviors in the case of exploration. This would strengthen the exploration process and is expected to increase the animat's learning speed.
The second possible use of a goal-directed behavior component is to represent the type of reward the animat

is working for. This can be useful when an animat has to deal with different rewards (food, drink) so as to satisfy different motivations (hunger, thirst). In this case, a component that chooses explicitly the current reward the animat takes as an objective can select sub-modules of the Actor that are dedicated to the sequence of behaviors that leads to the considered reward. This improvement would serve as a more realistic validation of the artificial rat Psikharpax when it has to survive in more natural environments, satisfying concurrent motivations.

Acknowledgments

This research has been granted by the LIP6 and the Project *Robotics and Artificial Entities* (ROBEA) of the Centre National de la Recherche Scientifique, France. Thanks for useful discussions to Drs. Angelo Arleo, Gianluca Baldassarre, Francesco Battaglia, Etienne Koechlin and Jun Tani.

References

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

Albertin, S. V., Mulder, A. B., Tabuchi, E., 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(1-2):173-83.

Albin, R. L., Young, A. B. & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neuroscience*, 12:366-375.

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

Baldassarre, G. (2002). A Modular Neural-Network Model of the Basal Ganglia's Role in Learning and Selecting Motor Behaviors. *Journal of Cognitive Systems Research*, 3(1):5-13.

Baldassarre, G. & Parisi, D. (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In Meyer *et al.* (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior,* Supplement Volume (pp.131-139). The MIT Press, Cambridge, MA.

Brown, J., Bullock, D. & Grossberg, S. (1999). How the Basal Ganglia Use Parallel Excitatory and Inhibitory Learning, or Incentive Salience ? *Brain Research Reviews*, 28:309-369.

Brown, L. & Sharp. F. (1995). Metabolic Mapping of Rat Striatum: Somatotopic Organization of Sensorimotor Activity. *Brain Research*, 686:207-222.

Bunney, B. S., Chiodo, L. A. & Grace, A. A. (1991). Midbrain Dopamine System Electrophysiological Functioning: A Review and New Hypothesis. *Synapse*, 9:79-84.

Burgess, N., Jeffery, K. J. & O'Keefe, J. (1999). Integrating Hippocampal and Parietal Functions: a Spatial Point of View. In Burgess, N. *et al.* (Eds), *The Hippocampal and Parietal Foundations of Spatial Cognition*, pp.3-29, Oxford University Press, UK.

Cardinal, R. N., Parkinson, J. A., Hall, J. & Everitt, B. J. (2002). Emotion and Motivation: The Role of the Amygdala, Ventral Striatum and Prefrontal Cortex. *Neuroscience Biobehavioral Reviews*, 26(3):321-352.

Dayan., P. (2001). Motivated Reinforcement Learning. *NIPS,* 14: 11-18. The MIT Press.

Daw, N. D. (2003). *Reinforcement Learning Models of the Dopamine System and Their Behavioral Implications*. PhD Thesis, Carnegie Mellon University, Pittsburgh, PA.

Doya, K. (2000). Reinforcement Learning in Continuous Time and Space. *Neural Computation*, 12:219-245.

Doya, K., Samejima, K., Katagiri, K. & Kawato, M. (2002) Multiple Model-based Reinforcement Learning. *Neural Computation*. 14(6):1347-69.

Filliat, D., Girard, B., Guillot, A., Khamassi, M., Lachèze, L., Meyer, J.-A. (2004). State of the artificial rat Psikharpax. In Schaal *et al.* (Eds), *From Animals to Animats 8: Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior,* pp.2-12. The MIT Press, Cambridge, MA.

Foster, D., Morris, R. & Dayan, P. (2000). Models of Hippocampally Dependent Navigation using the Temporal Difference Learning Rule. *Hippocampus*, 10:1-16.

Frank, M. J., 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(2):137-160.

Gerfen, C. R., Herkenham, M. & Thibault, J. (1987). The Neostriatal Mosaic. II. Patch- and Matrix- Directed Mesostriatal Dopaminergic and Non-Dopaminergic Systems. *Journal of Neuroscience*, 7:3915-3934.

Girard, B., Cuzin, V., Guillot, A., Gurney, K. & Prescott, T. (2003). A Basal Ganglia inspired Model of Action Selection Evaluated in a Robotic Survival Task. *Journal of Integrative Neuroscience,* 2(22), 179-200.

Gurney, K. N., Prescott, T. J. & Redgrave, P. (2001a). A Computational Model of Action Selection in the Basal Ganglia. I. A new functional anatomy. *Biological Cybernetics*. 84, 401-410.

Gurney, K. N., Prescott, T. J. & Redgrave, P. (2001b). A Computational Model of Action Selection in the Basal Ganglia. II. Analysis and simulation of behavior. *Biological Cybernetics*. 84, 411-423.

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 *et al.* (Eds), *Models of Information Processing in the Basal Ganglia.* The MIT Press, Cambridge, MA.

Ikemoto, S. & Panksepp., J. (1999). The Role of the 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. I., Nowlan, S. J. & Hinton, G.E. (1991). Adaptive Mixture of Local Experts. *Neural Computation*, 3:79-87.

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

Joel, D. & Weiner, I. (2000). The Connections of the Dopaminergic System with Striatum in Rats and Primates: An Analysis with respect to the Functional and Compartmental Organization of the Striatum. *Neuroscience*, 96:451-474.

Killcross, A. S. & Coutureau, E. (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex*, 13(4):400-408.

Lee, J. K. & Kim, I. H. (2003). Reinforcement Learning Control Using Self-Organizing Map and Multi-Layer Feed-Forward Neural Network. In *International Conference on Control Automation and Systems, ICCAS 2003*.

McNaughton, B. L. (1989). Neural Mechanisms for Spatial Computation and Information Storage. In Nadel *et al.* (Eds), *Neural Connections, Mental Computations*, chapter 9, pp.285-350, MIT Press, Cambridge, MA.

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

Montes-Gonzalez, F. Prescott, T. J., Gurney, K. N., Humphries, M. & Redgrave, P. (2000). An Embodied Model of Action Selection Mechanisms in the Vertebrate Brain. In Meyer *et al*. (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* (pp.157-166). The MIT Press, Cambridge, MA.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R. (2004). Dissociable Roles of Dorsal and Ventral Striatum in Instrumental Conditioning. *Science*, 304:452-454.

Pennartz, C. M. A. (1996). The Ascending Neuromodulatory Systems in Learning by Reinforcement: Comparing Computational Conjectures with Experimental Findings. *Brain Research Reviews*, 21:219-245.

Perez-Uribe, A. (2001). Using a Time-Delay Actor-Critic Neural Architecture with Dopamine-like Reinforcement Signal for Learning in Autonomous Robots. In Wermter *et al.* (Eds), *Emergent Neural Computational Architectures based on Neuroscience: A State-of-the-Art Survey* (pp. 522-533). Springer-Verlag, Berlin.

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

Schultz, W., Apicella, P. & Ljungberg, T. (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task. *Journal of Neuroscience*, 13(3):900-913.

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

Smith, A. J. (2002). Applications of the Self-Organizing Map to Reinforcement Learning. *Neural Networks*, 15(8-9):1107-1124.

Sporns, O. & Alexander, W. H. (2002). Neuromodulation and Plasticity in an Autonomous Robot. *Neural Networks*, 15:761-774.

Strösslin, T. (2004). *A Connectionnist Model of Spatial Learning in the Rat.* PhD thesis, EPFL, Swiss Federal

Institute of Technology, Swiss.
Suri, R. E. & Schultz, W. (2001). Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation*, 13:841-862.
Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning:An Introduction.* The MIT Press Cambridge, MA.
Tang, B., Heywood, M. I. & Shepherd, M. (2002). Input Partitioning to Mixture of Experts. In *International Joint Conference on Neural Networks*, pp. 227-232, Honolulu, Hawaii.
Tani, J. & Nolfi, S. (1999). Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-motor Systems. *Neural Networks*, 12(7-8):1131-1141.
Thierry, A.-M., Gioanni, Y., Dégénetais, E., and Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus*, 10:411-419.
Yin, H. H., Knowlton, B. J. & Balleine, B. W. (2004). Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning. *European Journal of Neuroscience*, 19(1):181-189.

Appendix : Parameters Table

Symbol	***Value***	***Description***
t	1 sec.	Time constant – time between two successive iterations of the model.
α	40 iter.	Time threshold to trigger the exploration function.
g	0.98	Discount factor of the Temporal Difference learning rule.
η	0.01	Learning rate of the Actor and Critic modules.
N	30	Number of experts in the Critic of Models *AMC1*, *AMC2* and *MAMC2*.
σ	2	Scaling parameter in the mixture of experts of Model *AMC1*.
m	0.1	Learning rate of the gating network in Model *AMC1*.

4. An Actor-Critic model for robotics combining SOM with mixtures of experts

Khamassi, Martinet, Guillot (2006). *SAB06*.

4.1 Summary of objectives

The last work of this chapter follows through on the previous one presented. In the latter, we presented a method to coordinate Actor-Critic modules which performed well in our simulated plus-maze task, but which lacks autonomy and generalization abilities.
Here we improved the method by combining self-organizing maps with a mixture of experts. This method aims at autonomously growing the number of Actor-Critic modules, and to automatically adapt the specialization between these latter modules. The method had been proposed in the field of data clustering (Tang et al., 2002), but yet had not been applied to reinforcement learning.

4.2 Summary of methods

The test environment was the same 2D virtual plus-maze described in the previous section. Three different kinds of self-organizing maps are connected to the mixture of Actor-Critic experts and tested during 11 simulations each. One is the classical Kohonen maps which has a fixed number of modules (Kohonen, 1995). The second is the Growing Neural Gas which adds a new module every 100 iterations depending on a global error in the map (Fritzke, 1995). The last is the Growing When Required which adds new modules more adaptively than the previous method, only depending on local errors in the map (Marsland et al., 2002).
In our simulations, we separated the exploration phase from the reinforcement learning phase. During the forme, the animat moves randomly in the maze while the map is being trained. During the latter, the map is stabilized and the animat learns the task by trial-and-error. Note that the so-called “map” corresponds to a set of independent categorizations of visual inputs without any transitional links between them.

4.3 Summary of results

The three methods give comparably good results in the task, respectively 548, 404 and 460 iterations per trial to reach the reward after learning. This is much better than the average of 30000 iterations per trial needed by a random agent, and also outperforms the classical mixture of experts discussed in the previous section which, in our task, reaches a performance of 3500 iterations per trial after learning.

4.4 Discussion

We proposed a new autonomous and adaptive method for the coordination of reinforcement learning modules. The model can solve the task and the performance obtained after learning is good. However, the performance is not yet as good as the hand-tuned method we used in the previous paper. The latter was proposed as a principle to coordinate experts. Indeed, this hand-tuned method gives a performance of 94 iterations per trial after learning, which is almost optimal: it approximately corresponds to the time needed for the agent to go straight from one arm extremity to the reward located at the next arm extremity.
The not yet optimal performance can be explained by the high variability produced by the self-organizing maps. Indeed this method could create some maps with very good performance (less

than 100 iterations per trial after learning), but also made some rather bad maps (around 1000 iterations per trial).
The variability of the method comes from the strong dependence of the map training phase on the way the agent has explored the environment: if the animat spends a great amount of time near the reward location during the exploration phase, then the map will specialize several experts dedicated to the reward location (this corresponds to the familiarization phase used with real rats in this task). The reward location being more crucial for the task than other places, this gives better performance during the reinforcement learning phase.
A proposition to improve our method, is to slightly adapt the map near the reward location after a first stage of reinforcement learning has been reached. This adaptation could be triggered by a threshold that blocks the reinforcement learning process in order to allow a new exploration phase. This post-training adaptation of the map coordinating Actor-Critic modules could be interestingly related to neural activity in the striatum. Jog et al. (1999) and Barnes et al. (2005) report a redistribution of spatial selectivity of striatal neurons among the maze with extensive training.
In our method, we chose to separate the exploration phase from the reinforcement learning phase. This prevents interferences between the adaptation of the map and reinforcement learning.

In order to enable the model to adapt to a more complex robotics task, the system can give the priority back to the map training when it detects that the map's performance drops. For example, if our agent is well habituated to the plus-maze, and if an experimenter opens a door, giving access to a new corridor, then the map's classification error should increase sharply, providing a signal to another neural system that transiently blocks the current reinforcement learning and launches a new phase of map adaptation.

Combining self-organizing maps with mixtures of experts: Application to an Actor-critic model of reinforcement learning in the Basal Ganglia

Mehdi Khamassi[1,2], Louis-Emmanuel Martinet[1], and Agnès Guillot[1]

[1] Université Pierre et Marie Curie - Paris 6, UMR7606, AnimatLab - LIP6, F-75005 Paris, France ; CNRS, UMR7606, F-75005 Paris, France

[2] Laboratoire de Physiologie de la Perception et de l'Action, UMR7152 CNRS, Collège de France, F-75005 Paris, France

{mehdi.khamassi, louis-emmanuel.martinet, agnes.guillot}@lip6.fr

http://animatlab.lip6.fr

Preprint accepted for publication in SAB06, Roma, Italy, 25-29 sept. 2006.

Abstract. In a reward-seeking task performed in a continuous environment, our previous work compared several Actor-Critic (AC) architectures implementing dopamine-like reinforcement learning mechanisms in the rat's basal ganglia. The task complexity imposes the coordination of several AC submodules, each module being an expert trained in a particular subset of the task. We showed that the classical method where the choice of the expert to train at a given time depends on each expert's performance suffered from strong limitations. We rather proposed to cluster the continuous state space by an *ad hoc* method that lacked autonomy and generalization abilities. In the present work we have combined the mixture of experts with self-organizing maps in order to cluster autonomously the experts' responsibility space. On the one hand, we find that classical *Kohonen maps* give very variable results: some task decompositions provide very good and stable reinforcement learning performances, whereas some others are unadapted to the task. Moreover, they require the number of experts to be set a priori. On the other hand, algorithms like *Growing Neural Gas* or *Growing When Required* have the property to choose autonomously and incrementally the number of experts to train. They lead to good performances, even if they are still weaker than our hand-tuned task decomposition and than the best Kohonen maps that we got. We finally discuss on propositions about what information to add to these algorithms, such as knowledge of current behavior, in order to make the task decomposition appropriate to the reinforcement learning process.

1 Introduction

In the frame of the Psikharpax project, which aims at building an artificial rat having to survive in complex and changing environments, and having to satisfy different needs and motivations [5][14], our work consists in providing a simulated robot with habit learning capabilities, in order to make it able to associate efficient behaviors to relevant stimuli located in an unknown environment.

The control architecture of Psikharpax is expected to be as close as possible to known anatomy and physiology of the rat brain, in order to unable comparison between functioning of the model with electrophysiological and behavioral recordings. As a consequence, our model of reinforcement learning is based on an Actor-Critic architecture inspired from basal ganglia circuits, following well established hypotheses asserting that this structure of the mammalian brain is responsible for driving action selection [16] and reinforcement learning of behaviors to select via substantia nigra dopaminergic neurons [17].

At this stage of the work, our model runs in 2D-simulation with a single need and a single motivation. However the issue at stake already has a certain complexity: it corresponds to a continuous state-space environment; the perceptions have non monotonic changes; an obstacle-avoidance reflex can interfere with actions selected by the model; the reward location provides a non instantaneous reward. In a previous paper [11], we demonstrated that this task complexity requires the use of multiple Actor-Critic modules, where each module is an expert trained in a particular subset of the environment. We compared different hypotheses concerning the management of such modules, concerning there more or less autonomously determined coordination, and found that the classical mixture of experts method - where the choice of the expert to train at a given time depends on each expert's performance [3][4] – cannot train more than one single expert in our reinforcement learning task. We rather proposed to cluster the continuous state space and to link each expert to a cluster by an ad hoc method that could indeed solve the task, but that lacked autonomy and generalization abilities.

The objective of the present work is to provide an autonomous categorization of the state space by combining the mixture of experts with self-organizing maps (SOM). This combination has already been implemented by Tang et al. [20] - these authors having criticized the undesirable effects of classical mixture of experts on boundaries of non disjoint regions. However, they did not test the method in a reinforcement learning task. When they were used in such tasks [18][13] - yet without mixture of experts –, SOM were applied to the discretization of the input space to the reinforcement learning model, which method suffers from generalization abilities. Moreover, the method has limited performance in high-dimensional spaces and remains to be tested robustly on delayed reward tasks.
In our case, we propose that the SOM algorithms have to produce a clustering of the responsibility space of the experts, in order to decide which Actor-Critic expert has to work in a given zone of the perceptual state space. In addition, the selected Actor-Critic expert of our model will receive the entire state space, in order to produce a non constant reward prediction inside the given zone.

After describing the task in the following section, we will report the test of three self-organizing maps combined with the mixture of Actor-Critic experts, for the comparison of their usefulness for a complex reinforcement learning task. It concerns the classical *Kohonen* algorithm [12], which requires the number of experts to be a priori set; the *Growing Neural Gas* algorithm [6], improved by [9], which adds a new expert when an existing expert has a important error of classification; and the *Growing When Required* algorithm [15], which creates a new expert when habituation of the map to visual inputs produces a too weak output signal when facing new visual data.
In the last section of the paper, we will discuss the possible modifications that could improve the performance of the model.

2 The task

Figure 1 shows the simulated experimental setup, a simple 2D plus-maze. The dimensions are equivalent to a 5m * 5m environment with 1m large corridors. In this environment, walls are made of segments colored on a 256 grayscale. The effects of lighting conditions are not simulated. Every wall of the maze is colored in black (luminance = 0), except walls at the end of each arm and at the center of the maze, which are represented by specific colors: the cross at the center is gray (191), three of the arm ends are dark gray (127) and the fourth is white (255), indicating the reward location equivalent to a water trough delivering two drops (non instantaneous reward) – not a priori known by the animat.

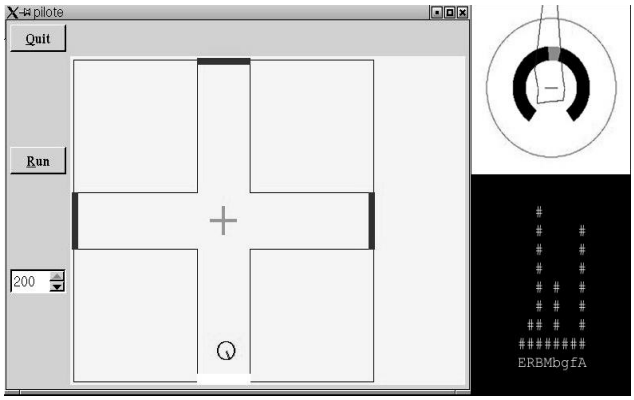


Fig. 1. Left: the robot in the plus-maze environment. Upper right: the robot's visual perceptions. Lower right: activation level of different channels in the model.

The plus-maze task reproduces the neurobiological and behavioral experiments that will serve as future validation for the model [1]. At the beginning of each trial, one arm end is randomly chosen to deliver reward. The associated wall becomes white whereas the other arm ends become dark gray. The animat has to learn that selecting the action *drinking* when it is near the white wall (distance < 30 cm) and faces it (angle < 45°) gives it two drops of water. Here we assume that reward = 1 for n iterations (n = 2) during which the action *drinking* is being executed. However, the robot's vision

does not change between these two moments, since the robot is then facing the white wall. As visual information is the only sensory modality that will constitute the input space of the Actor-Critic model, this makes the problem to solve a Partially Observable Markov Decision Process [19]. This characteristic was set in order to fit the multiple consecutive rewards that are given to rats in the neurobiological plus-maze, enabling comparison between our algorithm with the learning process that takes place in the rat brain during the experiments.
We expect the animat to learn a sequence of context-specific behaviors, so that it can reach the reward site from any starting point in the maze:
When not seeing the white wall, face the center of the maze and move forward
As soon as arriving at the center (the animat can see the white wall), turn to the white stimulus
Move forward until being close enough to reward location
Drink
The trial ends when reward is consumed: the color of the wall at reward location is changed to dark gray, and a new arm end is randomly chosen to deliver reward. The animat has then to perform another trial from the current location. The criterion chosen to validate the model is the time – number of iterations of the algorithm - to goal, plotted along the experiment as the learning curve of the model.

3 The animat

The animat is represented by a circle (30 cm diameter). Its translation and rotation speeds are 40 $cm.s^{-1}$ and $10°.s^{-1}$.
Its simulated sensors are:

4. Eight sonars with a 5m range, an incertitude of ±5 degrees concerning the pointed direction and an additional ±10 cm measurement error. The sonars are used by a low level obstacle avoidance reflex which overrides any decision taken by the Actor-Critic model when the animat comes too close to obstacles.
5. An omnidirectional linear camera providing every 10° the color of the nearest perceived segment. This results in a 36 colors table that constitute the animat's visual perception (see figure 1).

The animat is provided with a visual system that computes 12 input variables and a constant equal to 1 $(\forall i \in [1;13], 0 \leq \mathrm{var}_i \leq 1)$ out of the 36 colors table at each time step. These sensory variables constitute the state space of the Actor-Critic and so will be taken as input to both the Actor and the Critic parts of the model (figure 3). Variables are computed as following:

4. *seeWhite* (resp. *seeGray*, *seeDarkGray*) = 1 if the color table contains the value 255 (resp. 191, 127), else 0.
5. *angleWhite*, *angleGray*, *angleDarkGray* = (number of boxes in the color table between the animat's head direction and the desired color) / 18.
6. *distanceWhite*, *distanceGray, distanceDarkGray* = (maximum number of consecutive boxes in the color table containing the desired color) / 18.
7. *nearWhite* (resp. *nearGray*, *nearDarkGray*) = 1 – *distanceWhite* (resp. *distanceGray, distanceDarkGray*).

The model permanently receives a flow of sensory information and has to learn autonomously the sensory contexts that can be relevant for the task resolution.
The animat has a repertoire of 6 actions: *drinking*, *moving forward*, *turning to white perception*, *turning to gray perception*, *turning to dark gray perception*, and *waiting*. These actions constitute the output of the Actor model (described below) and the input to a low-level model that translates it into appropriate orders to the animat's engines.

4 The Model

4.1 The multi-module Actor-Critic

The model tested in this work has the same general scheme than described in [11]. It has two main components, an Actor which selects an action depending on the visual perceptions described above; and a Critic, having to compute predictions of reward based on these same perceptions (figure 2). Each of these two components is composed of N submodules or *experts*. At a given time, each submodule k $(k \in \{1\ ;N\})$ has a responsibility $c_k(t)$ that determines its weight in the output of the overall model. In the context of this work, we restrict to the case where only one expert *k* has its responsibility equal to 1 at a given moment, and $\forall\, j \neq k, \mathrm{c}_j(t) = 0$.

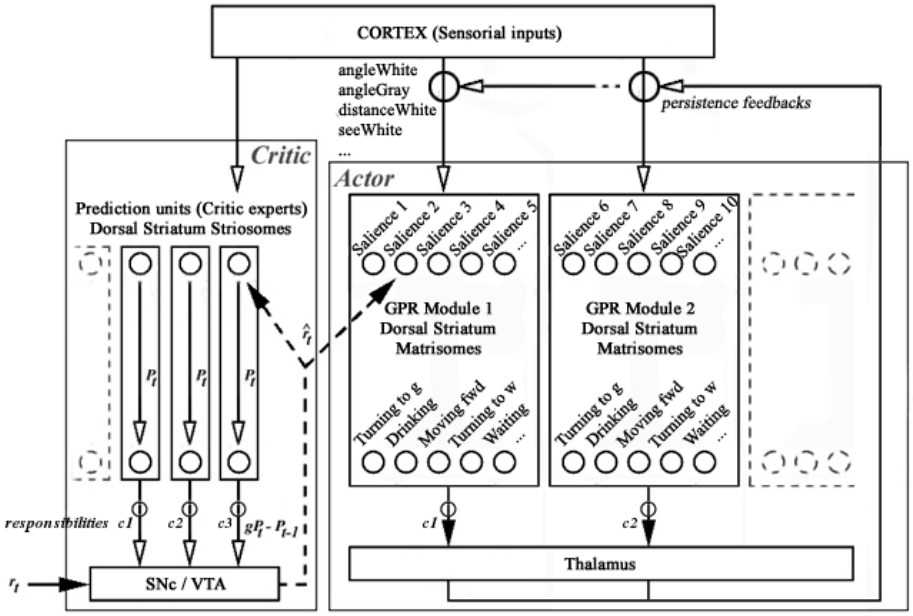


Fig. 2. General scheme of the model tested in this work. The Actor is a group of "GPR" modules [8] with saliences as inputs and actions as outputs. The Critic (involving striosomes in the dorsal striatum, and the substantia nigra compacta (SNc)) propagates towards the Actor an estimate ř of the instantaneous reinforcement triggered by the selected action. The particularity of this scheme is to combine several modules for both Actor and Critic, and to gate the Critic experts' predictions and the Actor modules' decisions with responsibility signals. These responsibilities can be either computed by a Kohonen, a GWR or a GNG map.

Inside the Critic component, each submodule is a single linear neuron that computes its own prediction of reward:

$$p_k(t) = \sum_{j=1}^{13} w'_{k,j}(t) \cdot \mathrm{var}_j(t) \qquad (1)$$

where $w'_{k,j}(t)$ is the synaptic weight of expert *k* representing the association strength with input variable j. Then the global prediction of the Critic is a weighted sum of experts' predictions:

$$P(t) = \sum_{k=1}^{N} c_k(t) \cdot p_k(t) \qquad (2)$$

Concerning the learning rule, derived from the Temporal-Difference Learning algorithm [19], each expert has a specific reinforcement signal based on its own prediction error:

$$\hat{r}_k(t) = r(t) + gP(t) - p_k(t-1) \qquad (3)$$

The synaptic weights of each expert *k* are updated according to the following formula:

$$w'_{k,j}(t) \leftarrow w'_{k,j}(t-1) + \eta \cdot \hat{r}_k(t) \cdot \mathrm{var}_j(t-1) \cdot c_k(t) \qquad (4)$$

Actor submodules also have synaptic weights $w_{i,j}(t)$ that determine, inside each submodule *k*, the salience – i.e. the strength – of each action *i* according to the following equation:

$$sal_i(t) = \left[\sum_{j-1}^{13} \mathrm{var}_j(t) \cdot w_{i,j}(t) \right] + persist_i(t) \cdot w_{i,14}(t) \qquad (5)$$

The action selected by the Actor to be performed by the animat corresponds to the strongest output of the submodule with responsibility 1. If a reinforcement signal occurs, the synaptic weights of the

latter submodule are updated following equation (4).

An exploration function is added that would allow the animat to try an action in a given context even if the weights of the Actor do not give a sufficient tendency to perform this action in the considered context.
To do so, we introduce a clock that triggers exploration in two different cases:

8. When the animat has been stuck for a large number of timesteps (*time* superior to a fixed threshold α) in a situation that is evaluated negative by the model (when the prediction $P(t)$ of reward computed by the Critic is inferior to a fixed threshold).
9. When the animat has remained for a long time in a situation where $P(t)$ is high but this prediction doesn't increase that much ($|P(t+n) - P(t)| < \varepsilon$) and no reward occurs.

If one of these two conditions is true, exploration is triggered: one of the 6 actions is chosen randomly.Its salience is being set to 1 (Note that: when exploration = false, $sal_i(t) < 1,\ \forall i,t,w_{i,j}(t)$) and is being maintained to 1 for a duration of 15 timesteps (time necessary for the animat to make a 180° turn or to run from the center of the maze until the end of one arm).

4.2 The self-organizing maps

In our previous work [11], we showed that the classical method used to determine the experts' responsibilities – a gating network, giving the highest responsibility to the expert that approximates the best the future reward value [3][4] – was not appropriate for the resolution of our reinforcement learning task. Indeed, we found that the method could only train one expert which would remain the more responsible in the entire state space without having a good performance. As our task is complex, we rather need the region of the state space where a given expert is the most responsible to be restricted, in order to have only limited information to learn there. As a consequence, we propose that the state space should be clustered independently from the performance of the model in learning the reward value function.

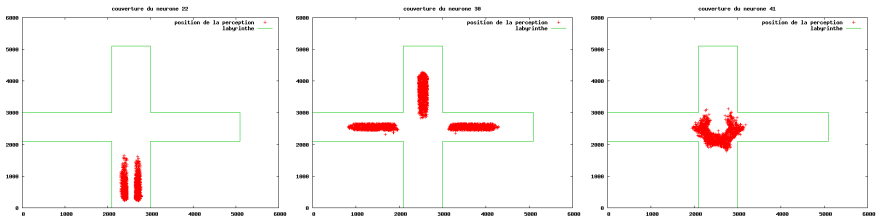


Fig. 3. Examples of clusterings found by the GWR self-organizing map. The pictures show, for three different AC experts, the positions of the robot for which the expert has the highest responsibility – thus, positions where the Actor-Critic expert is involved in the learning process.

In this work, the responsibility space of the Actor-Critic experts is determined by one of the following self-organizing maps (SOMs): the Kohonen Algorithm, the Growing Neural Gas, or the Growing When Required. We will describe here only essential aspects necessary for the comprehension of the method maps. Each map has a certain number of nodes, receives as an input the state space constituted of the same perception variables than the Actor-Critic model, and will autonomously try to categorize this state space. Training of the SOMs is processed as following:

```
Begin
  Initialize a fixed number of nodes (for the Kohonen
  Map) or 2 nodes for GNG and GWR algorithms;
  While (iteration < 50000)
    Move the robot randomly; //Actor-Critic disabled
    Try to categorize the current robot's perception;
    If (GNG or GWR) and (classification-error > threshold)
        Add a new node to the map;
    End if;
    Adapt the map;
  End;
  // After that, the SOM won't be adapted anymore
```

```
  While (trial < 600)
    Move the robot with the Actor-Critic (AC) model;
    Get the current robot's perception;
    Find the SOM closest node (k) to this perception;
    Set expert k responsibility to 1 and others to 0;
    Compute the learning rule and adapt synaptic weights of the AC;
  End;
End;
```

Parameters used for the three SOM algorithms are given in the appendix table. Figure 3 shows some examples of categorization of the state space obtained with a GWR algorithm. Each category corresponds to a small region in the plus-maze, where its associated Actor-Critic expert will have to learn. Notice that we set the parameters so that regions are small enough to train at least several experts, and large enough to require that some experts learn to select different actions successively inside the region.

5 Results

The results correspond to several experiments of 600 trials for each of the three different methods (11 with GWR, 11 with GNG, and 11 with Kohonen maps). Each experiment is run following the algorithmic procedure described in the previous section.

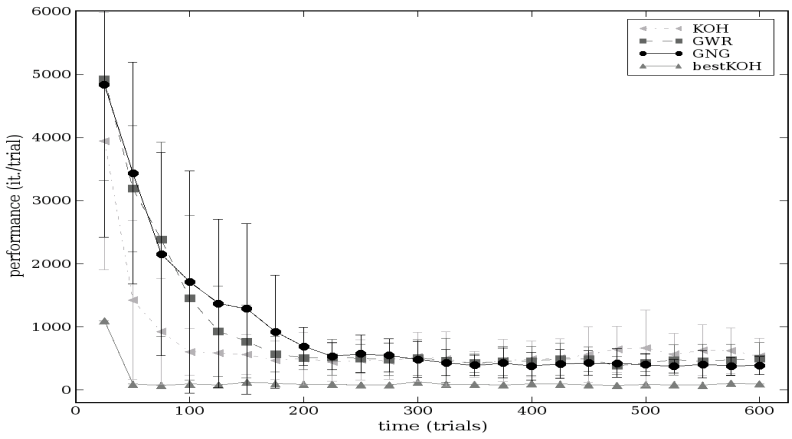
Table 1. Summarized performances of the methods applied to reinforcement learning.

Method	Average performance during second half of the experiment (nb iterations per trials)	Standard error	Best map's average performance
Hand-tuned map	93.71	N/A	N/A
KOH (n=11)	548.30	307.11	87.87
GWR (n=11)	459.72	189.07	301.76
GNG (n=11)	403.73	162.92	193.39

Figure 4 shows the evolution with time of the learning process of each method. In each case, the smallest number of iterations occurs around the 250th trial and remains stabilized. Table 1 summarizes the global performances averaged over the second half of the experiment – e.g. after trial #300. Performances of the three methods are comparable (Kruskall-Wallis test reveals no significant differences: $p > 0.10$). When looking at the maps' categorizations precisely and independently from the reinforcement learning process, measure of the maps' errors of categorization highlights that Kohonen maps provide a slightly worst result in general, even while using more neurons than the GWR and GNG algorithms. However, this doesn't seem to have consequences on the reinforcement learning process, since performances are similar. So, the Kohonen algorithm, whose number of experts is a priori set, is not better than the two others which recruit new experts autonomously.

Performances with GNG and GWR algorithms are not very different either. In their study, Marsland et al. [15] conclude that GWR is slightly better than the GNG algorithm in its original version. Here, we used a modified version of GNG [9]. In our simulations, the GNG recruited on average less experts than the GWR but had a classification error a little bigger. However, when applied to reinforcement learning, the categorizations provided by the two algorithms did not show major differences.

Fig. 4. Learning curves of the reinforcement learning experiments tested with different self-organizing maps.



Qualitatively, the three algorithms have provided the multi-module Actor-Critic with quite good experts' responsibility space clustering, and the animat managed to learn an appropriate sequence of actions to the reward location. However, performances are still not as good as a version of the model with hand-tuned synaptic weights. The latter has an average performance of 93.71 iterations per trial, which is characterized by a nearly “optimal” behavior where the robot goes systematically straight to the reward location, without loosing any time (except the regular trajectory deviation produced by the exploration function of the algorithm). Some of the best Kohonen maps and GNG maps reached similar nearly optimal behavior. As shown in table 1, the best Kohonen map got an average performance of 87.87 iterations per trial. Indeed, it seems that the categorization process can produce very variable reinforcement learning depending on the map built during the first part of the experiment.

6 Discussion

In this work, we have combined three different self-organizing maps with a mixture of Actor-Critic experts. The method was designed to provide an Actor-Critic model with autonomous abilities to recruit new expert modules for the learning of a reward-seeking task in continuous state space. Provided with such a control architecture, the simulated robot can learn to perform a sequence of actions in order to reach the reward. Moreover, gating Actor-Critic experts with our method strongly ressembles neural activity observed in the striatum – e.g. the input structure of the basal ganglia – in rat performing habit learning tasks in an experimental maze [10]. Indeed, the latter study shows striatal neurons' responses that are restricted to localized chunks of the trajectory performed by the rat in the maze. This is comparable with the clusters of experts' responsibilities shown in figure 3.
However, the performance of the model presented here remains weaker than a hand-tuned behavior. Indeed, the method produces very variables results, from maps with nearly optimal performance to maps providing unsatisfying robotics behavior.
Analysis of the maps created with our method shows that some of them are more appropriate to the task than others, particularly when the boundaries between two experts' receptive fields corresponds to a region of the maze where the robot should switch from one action to another in order to get the reward. As an example, we noticed that the majority of the maps obtained in this work had their expert closer to the reward location with a too large field of responsibility. As a consequence, the trunk of the global value function that this expert has to approximate is more complex, and the behavior to learn is more variable. This results in selecting inappropriate behavior in the field of this expert – for example, the robot selects the action “drinking” too far from reward location to get a reward. Notice however that this is not a problem with selecting several different actions in the same region of the maze, since some experts managed to learn to alternate between two actions in their responsibility zone, for example in the area close to the center of the plus-maze. A given expert

having limited computational capacities, its limitations occur when its region of responsibility is too large.

To improve the performance, one could suggest setting parameters of the SOM in order to increase the number of experts in the model. However, this would result in smaller experts' receptive field than those presented in figure 3. As a consequence, each expert would receive a nearly constant input signal inside its respective zone, and would need only to select one action. This would be computationally equivalent to the use of small fields place cells for the clustering of the state space of an Actor-Critic, which has been criticized by several authors [2], and would not be different than other algorithms where the winning node of a self-organizing map produces a discretization of the input space to a reinforcement learning process [18].
One could also propose to increase each expert-module's computational capacity. For instance, one could use a more complex neural network than the single linear neuron that we implemented for each expert. However, one cannot a priori know the task complexity, and no matter the number of neurons an expert possesses, there could still exist too complex situations.Moreover, “smart” experts having a small responsibility region could overlearn the data with poor generalization ability [7].

7 Perspective

In future work, we rather propose to enable the experts' gating to adapt slightly to the behavior of the robot. The management of experts should not be mainly dependent on the experts' performances in controlling behavior and estimating the reward value, as we have shown in previous work [11]. However, considering the categorization of the visual space as the main source of experts’ specialization, it could be useful to add information about the behavior in order for boundaries between two experts' responsibility regions to flexibly adapt to areas where the animat needs to switch its behavior. In [21], the robot's behavior is a priori set and stabilized, and constitutes one of the inputs to a mixture of experts having to categorize the sensory-motor flow perceived by a robot. In our case, at the beginning of the reinforcement learning process, when behavior is not yet stable, visual information could be the main source of experts’ specialization. Then, when the model starts to learn an appropriate sequence of actions, behavioral information could help adjusting the specialization. This would be similar to electrophysiological recordings of the striatum showing that, after extensive training of the rats, striatal neurons' responses tend to translate to particular “meaningful” portions of the behavioral sequences, such as the starting point and the goal location [10].

Acknowledgments

This research has been granted by the LIP6 and the European Project Integrating Cognition, Emotion and Autonomy (ICEA). The authors wish to thank Thomas Degris and Jean-Arcady Meyer for useful discussions.

References

Albertin, S. V., Mulder, A. B., Tabuchi, E., Zugaro, M. B., Wiener, S. I.: Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behavioral Brain Research*. 117(1-2) (2000) 173-83

Arleo, A., W. Gerstner,W.: Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics*, 83(3) (2000) 287-99

Baldassarre, G.: A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviors. *Journal of Cognitive Systems Research,* 3(1) (2002) 5-13

Doya, K., Samejima., K., Katagiri, K., Kawato, M.: Multiple model-based reinforcement learning. *Neural Computation,* 14(6) (2002) 1347-69

Filliat, D., Girard, B., Guillot, A., Khamassi, M., Lachèze, L., Meyer, J.-A. : State of the artificial rat Psikharpax. In: Schaal, S., Ijspeert, A., Billard, A., Vijayakumar, S., Hallam, J., Meyer, J.-A. (eds): *From Animals to Animats 8:*

Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior, Cambridge, MA. MIT Press (2004) 3-12
Fritzke, B.: A growing neural gas network learns topologies. In: Tesauro, G, Touretzkys, D.S., Leen, K.(eds): *Advances in Neural Information Processing Systems*, MIT Press, (1995) 625-32
Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. *Neural Computation* 4 (1992) 1-58
Gurney, K.,Prescott, T.J., Redgrave, P.: A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics,* 84 (2001) 401-10
Holmström, J.: Growing neural gas : Experiments with GNG, GNG with utility and supervised GNG. Master's thesis, Uppsala University (2002)
Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V., Graybiel, A.M.: Building neural representations of habits. *Science*, 286(5445) (1999) 1745-9
Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., Guillot, A: Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior, Special Issue Towards Artificial Rodents* 13(2) (2005) 131-48
Kohonen, T.: Self-organizing maps. Springer-Verlag, Berlin (1995)
Lee, J. K., Kim, I. H.: Reinforcement learning control using self-organizing map and multi-layer feed-forward neural network. In: *International Conference on Control Automation and Systems,* ICCAS 2003 (2003)
Meyer, J.-A., Guillot, A., Girard, B., Khamassi, M., Pirim, P., Berthoz, A.: The Psikharpax project: Towards building an artificial rat. *Robotics and Autonomous Systems* 50(4) (2005) 211-23
Marsland, S., Shapiro, J., Nehmzow, U.: A self-organising network that grows when required. *Neural Networks*, 15 (2002) 1041-58
Prescott, T.J., Redgrave, P., Gurney, K.: Layered control architectures in robots and vertebrates. *Adaptive Behavior, 7* (1999) 99-127
Schultz, W., Dayan, P., Montague, P. R. : A neural substrate of prediction and reward. *Science*, 275 (1997) 1593-9
Smith, A. J.: Applications of the self-organizing map to reinforcement learning. *Neural Networks* 15(8-9) (2002)1107-24
Sutton, R. S., Barto, A. G.: Reinforcement learning: An introduction. The MIT Press Cambridge, MA. (1998)
Tang, B., Heywood, M. I.: Shepherd, M.: Input Partitioning to Mixture of Experts. In: *IEEE/INNS International Joint Conference on Neural Networks*, Honolulu, Hawaii (2002) 227-32
Tani, J., Nolfi, S.: Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems.*Neural Networks*12(1999)1131-41

Appendix: Parameters table

Symbol	Value	Description
t	1 sec.	Time between two successive iterations of the model.
α	[50;100]	Time threshold to trigger the exploration function.
g	0.98	Discount factor of the Temporal Difference learning rule.
η	0.05 / 0.01	Learning rate of the Critic and the Actor respectively.
N	36	Number of nodes in Kohonen Maps.
η-*koh*	0.05	Learning rate in Kohonen Maps.
σ	3	Neighborhood radius in Kohonen Maps.
Ew, En	0.5, 0.005 / 0.1, 0.001	Learning rates in the GNG and GWR respectively.
a-max	100	Max. age in the GNG and GWR.
S		Threshold for nodes recruitment in the GNG.
α-*gng*, β-*gng*	0.5, 0.0005	Error reduction factors in the GNG.
λ	1	Window size for nodes incrementation in the GNG.
a-T	0.8	Activity threshold in the GWR
h-T	0.05	Habituation threshold in the GWR.

5. Conclusion on the role of the rat striatum in learning

In this chapter, we presented different results concerning the study of the role of the striatum in reward-based learning of the visual cue-guided strategy. Our electrophysiological results, recorded from the rat ventral striatum, show anticipatory responses consistent with the Critic part of an Actor-Critic model in the TD-learning theory. By designing a simple computational Actor-Critic model with varying input information (different levels of temporal, spatial and visual information) we reproduce several different electrophysiological responses recorded. This extends the classical Actor-Critic model of the striatum where the system was particularly designed to reproduce temporal properties of dopaminergic neurons (Khamassi et al., in revision).

Since TD-learning is a *model-free* reinforcement learning algorithm, our results are consistent with the hypothesis that part of the ventral striatum (including the shell) participates in the learning of the visual cue-guided model-free strategy – that is, procedural Stimulus-Response behaviors (Dayan, 2001).

Then we combined this Critic-like activity with a biological plausible Actor model of the basal ganglia, assumed to be anchored in the dorsolateral striatum.
This architecture provides a good performance in a simulated robotics version of the plus-maze task (Khamassi et al., 2005). We finally proposed an autonomous method for the coordination of Actor-Critic modules. This method combines self-organizing maps and a mixture of experts, and show some interesing generalization ability for robotics (Khamassi et al., 2006).

Assuming the architecture where different striatal territories learn different navigation strategies, whereas the prefrontal cortex is assumed to detect task changes that require a strategy shift, the next chapter presents our investigation of prefrontal neuronal activity during a changing rule Y-maze task. Our hypothesis is that a set of prefrontal neurons should detect changes in the task rule imposed by the experimenter without any external cue. Moreover, a set of prefrontal neurons should correlate with the current strategy, thus showing a possible involvement of the PFC in strategy selection following task rule changes.

CHAPTER 3 : BEHAVIORAL AND NEURONAL ENSEMBLE RECORDING OF THE MEDIAL PREFRONTAL CORTEX IN RATS LEARNING AND SHIFTING STRATEGIES

1. Introduction

In this chapter, we present a synopsis of results of initial analyses of data from our study of the prefrontal cortical neural ensemble activity in rats learning new task contingency rules in a Y-maze, during strategy shifting and sleep prior and after learning. This project captured most of the efforts done during this PhD period: from building, programming and configuring the Ymaze setup, training rats in a first version of the task, building electrodes and surgical implants; to doing the experiments, processing the recorded data (video tracking, spike sorting), and analysing the data.

Based on the literature presented in the first chapter, our hypothesis is that the rat medial prefrontal cortex (mPFC) participates in the detection of extradimensional rule changes on the basis of a low incidence of rewards. Moreover, mPFC could participate in strategy selection following such reward contingency changes, in order to facilitate learning of a new strategy when the previously engaged one is no longer optimal.
Thus, we recorded prefrontal neurons in rats learning a binary choice task in a Y-maze, where reward delivery is governed by task rules that may change without any external cue signal. The task contingencies imposed on the rats are, in sequence: go to the right arm, go to the lit arm, go to the left arm, go to the dark arm.
Since prefrontal cortical damage impairs extradimensional strategy shifts, we expected to find a set of prefrontal neurons whose change in activity would reflect detection of changes in the task rule, a possible mechanism in order to trigger extinction of the former rule and learning of a new one. Moreover, we anticipated that some neurons in mPFC would show a modulation in activity correlated with spontaneous learning of new strategies made by the rat in order to optimize rewards.
Some of this work has been presented at international meeting, and the manuscript of this work is still in preparation (Khamassi, et al.) for submission to the Journal of Neuroscience. So these results are presented in the form of a full chapter.

2. Electrophysiological recordings in PFC

Khamassi et al. (in preparation) PFC and strategy shifting

INTRODUCTION

Decision is a key phase in the generation of behavior (Berthoz, 2003a). By decision in studies of rats and robots, we mean the cognitive processing for action selection. The mechanism leading to choosing an action can be seen as the combination of two quite different processes: the first can be named *teleological* (Dickinson, 1980), or *goal-directed* because it selects actions based on the animal's (or animat's) current motivational state and goal (the animal can be hungry and hence look for food, or it may be thirsty and look for water), the animal's "knowledge" about the consequences of candidate actions and whether or not this activity may bring it closer to attain the goal. The second process is called *habit*: an automatic, learned, or species-specific stimulus-response association that occurs at least partially independently of the animal's established goal. Learning of a goal-directed behavior involves examining the data available, both from past experiences and from the current state of the environment, then selecting, or if necessary, devising a *strategy* well-suited to attain that goal.

To learn a new strategy, the subject has often to go through a trial-and-error process, where various possibilities are explored, while keeping track of the *reward value* of each of them. Any type of agent, natural or artificial, surviving and operating in a complex, changing environment, has an increased chance of survival if it is capable of *goal-directed behavior*: the Artificial Intelligence and robotics communities face this precise problem when trying to design devices capable of performing a task without supervision in a situation in which not all the circumstances, potential outcomes, and the required responses can be predicted in advance.

This project studies the neural basis of reinforcement-based navigation strategy learning and shifting. Our working hypothesis is that the brain solves these kinds of problems in a network of high-order structures, which are highly interconnected and hierarchical arranged, including the hippocampus, the prefrontal cortex and the striatum (see Granon and Poucet, 2000; Thierry et al., 2000; Uylings et al., 2003; Voorn et al., 2004; Vertes, 2006 for reviews). As reviewed in the first chapter, there is now a convergence in the literature suggesting that, while the striatum could participate in learning specific navigation strategies, the medial prefrontal cortex could be involved in the detection of task changes prompting for a shift in the current strategy. Moreover, the medial prefrontal cortex could participate in the selection of the new strategy to perform, based on the storage of information concerning the consecutive successes and errors the animal has made during the past trials.

Indeed, the rat mPFC, and particularly the prelimbic area (PL) show some functional homologies with the primate dorsolateral prefrontal cortex (see Uylings et al., 2003 for a review), which plays a role in flexible goal-directed behaviors (see Granon and Poucet, 2000; Cardinal et al., 2002 for reviews). The mPFC is involved in attentional processes (Muir et al., 1996; Birrell and Brown, 2000; Delatour and Gisquet-Verrier, 2000), in working-memory (Van Haaren et al., 1985; Brito and Brito, 1990; Kolb, 1990; de Brabander et al., 1991; Delatour and Gisquet-Verrier, 1996, 1999) and in the registering of the consequences of actions (Corbit and Balleine, 2003; Killcross and Coutureau, 2003; Dalley et al., 2004; Ostlund and Balleine, 2005; see Cardinal et al., 2002 for a review). Moreover, it seems that PL is not involved in simple tasks requiring only one of these three processes, but rather in complex tasks requiring the combination of several of these processes to promote flexible behavior (see Granon and Poucet, 2000; Gisquet-Verrier and Delatour, 2006 for reviews).

PL lesions impair behavioral flexibility in response to a change in the task rule (de Bruin et al., 1994; Birrell and Brown, 2000; Colacicco et al., 2002; McAlonan and Brown, 2003; Salazar et al., 2004; Lapiz and Morilak, 2006). Moreover, PL damage-induced impairment is significantly increased when the task requires shifting from one strategy to another, whether the initial strategy

has been learned (Granon and Poucet, 1995; Ragozzino et al., 1999a,b) or is spontaneously used by the animal (Granon et al., 1994).

Furthermore, a particular category of strategy shifts is impaired by mPFC lesions (referring to the different types of shifts defined in the first chapter). Whereas lesions of the orbitofrontal cortex are found to impair intradimensional shifts – for example when the stimuli associated with reward before and after the task rule change have the same modality (either visual, olfactive, spatial) (Kim and Ragozzino, 2005), lesions of mPFC impair extradimensional shifts while intradimensional shifts are spared (Joel et al., 1997; Birrell and Brown, 2000; Ragozzino et al., 2003).

Several electrophysiological studies have reported prefrontal neural activity reflecting some crucial parameters underlying flexible goal-directed behaviors, such as movement (Poucet, 1997; Jung et al., 1998), reward (Pratt and Mizumori, 2001; Miyazaki et al., 2004), working-memory (Baeg et al., 2003), spatial goals (Hok et al., 2005) and action-outcome contingencies (Mulder et al., 2003 ; Kargo et al., 2007). A recent study showed that functional connectivity between neurons within the mPFC was found to be highest at the early stage of a new learning phase following a task rule change (Baeg et al., 2007). However, no individual prefrontal neurons have yet been shown to reflect task rule changes or spontenous shifts in the strategy, hallmarks of prefrontal function.

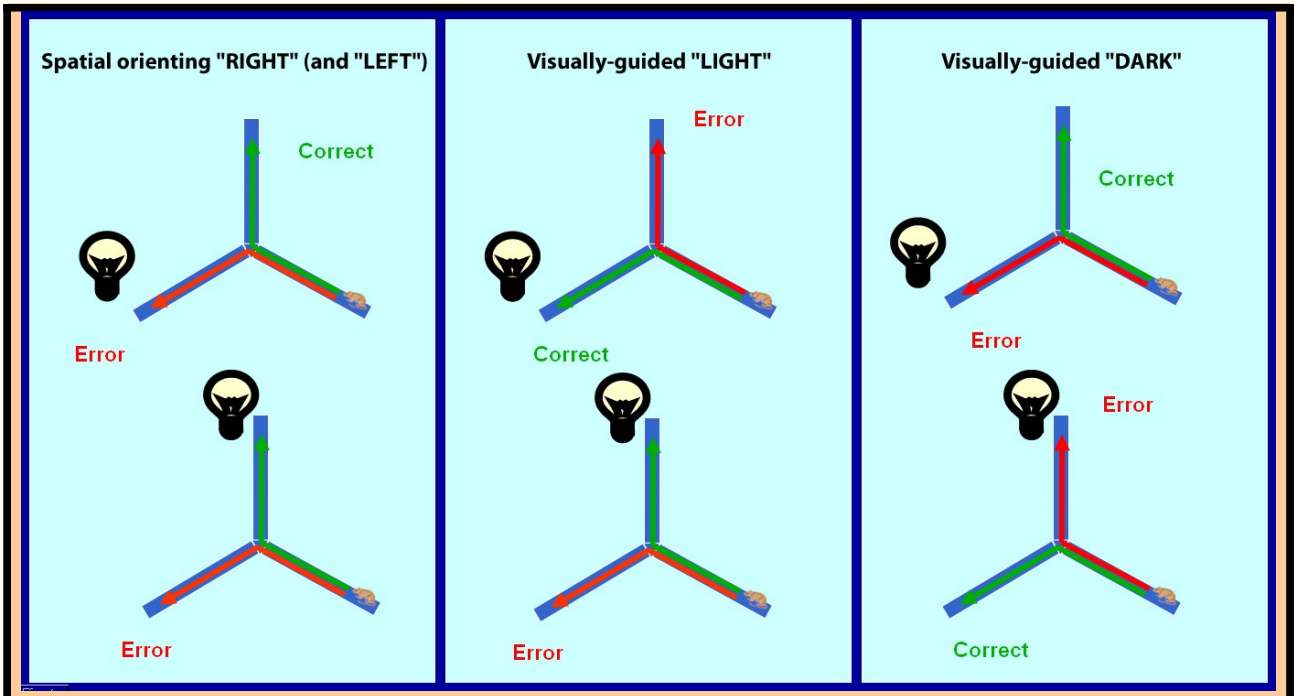


Figure 3.2.1: *Examples of different task strategies on the Y-maze. (Left) The rat has always to choose the arm to its right .(Center) The rat has to run to the illuminated arm. (Right) The rat has to go to the dark arm. The first case can be solved either by a praxic or a place strategy: the rewarded direction is always the same with respect to the rat's body, while the rewarded arm is always at the same spatial position. A light will also be present in this situation, acting as a confound. The two other cases are examples of visually guided navigation: there is a visual cue that is spatial in nature, since its position coincides with the rewarded location. D is an example of association between non-spatial stimuli (the sound and the light are presented in location between the choice arms) and a spatial response.*

Thus we have recorded ensembles of neurons from the medial prefrontal cortex, while the rat experiences extradimensional task rule changes unsignalled by external cues and must learn different task rules governing reward delivery in a Y-maze. The rat had a binary choice ("go left" or "go right"), and only one of these actions was rewarded. To infer the correct response, the rat had to take into account some external cues, and devise a mutable rule, or *strategy* (which could then change). The rule had to be discovered by a trial and error procedure. The external cues forming the

context of the choice appeared at various delays prior to choice, so that they had to be kept in the working memory buffer.

Our main hypotheses were that: First at a behavioral level, rats would successively try different strategies during learning, and then spontaneously shift their strategy when no longer consistently achieving rewards. Second, at the electrophysiological level, some neurons should detect changes in the task rule prompting learning of a new strategy. These neurons would show a transient change in activity in response to task rule changes, or would show transitions in activity synchronizes with these changes. Finally, some prefrontal neurons would show activity correlates with particular strategies, and thus show transitions in their activity when the animal shifts between strategies.

METHODS

The apparatus The Y-maze was formed by three arms separated by 120 degrees, with a cylindrical barrier at the center. The task to be performed by rats in this Y-maze is analogous to the Wisconsin Card Sorting Test, used for diagnosing neurological patients for prefrontal cortical damage (Berg, 1948 ; Grant et al., 1949 ; Heaton, 1993 ; Mansouri et al., 2006). This requires self-generated extradimensional strategy shifts. The present version of the test is a decision-making task where different rules govern the delivery of reward during consecutive blocks of trials. No clue was given to indicate the switch from one rule to another. In each block, rats had to learn by trial-and-error to adapt their behavior to the current rule in order to maximize the amount of reward received.

Strategies involved The rules were chosen to involve different memory systems relying on different categories of navigation strategies: on the one hand, a *visually-guided* strategy where a light cue indicates the presence or the absence of reward; on the other hand, a *spatial orienting* strategy where a certain position in space is associated with the presence of reward. The latter can either refer to a *place* or a *praxic* strategy, because we did not aim at disambiguating between the two: during the whole set of experiments, rats always started from the same departure arm. Thus, a left turn was always leading to the same position in space. However, neither praxic nor place strategies can engage the randomly positioned intramaze light cue used for the *visually-guided* strategy, thus permitting to test extradimensional shifts with this setup (e.g. from *visually-guided* to *spatial orienting*, and vice versa).

The task Rats started all trials from the *departure* arm (see figure 3.2.1), and after the central barrier was lowered, they had to select one of the two choice arms, then go to the end to receive a chocolate milk reward. As the barrier was lowered, a light went on in one of the two arms, randomly selected for each trial (figure 3.2.1). For each trial, the reward was available on only one arm. The baited arm was determined based on one of four possible contingency rules: 1) the right arm was always baited (*right* rule / *praxic-place* strategy), 2) the left arm was always baited (*left* rule / *praxic-place* strategy), 3) reward was available on the illuminated arm (*light* rule / *visually-guided* strategy), 4) reward was available on the non-illuminated arm (*dark* rule / *visually-guided* strategy); figure 1). After the rat faced the outcome of the trial (reward or nothing), he had to return back to the departure arm for next trial.

Once the rat has acquired the current rule (i.e., performance reached a criterion level of 10 consecutive rewarded trials, or 90% rewarded trials in the last 12 trials), the rule was then changed. As mentioned above, the change was not explicitly signalled to the rat in any way, so that it had to be inferred by the pattern of unrewarded trials.

Sessions Rats were trained in daily sessions. Each session consisted of 10 to 60 consecutive trials, stopping when the rat was no longer working. Since it happened that several daily sessions were required to learn certain task rules, there were sessions were no shift in the task rule was imposed. Thus, we will distinguish *shift-sessions* (e.g. sessions where a shift in the task rule occurred) from *non-shift-sessions*, and we will consider *pre-shift* and *post-shift* phases of a *shift-session*.

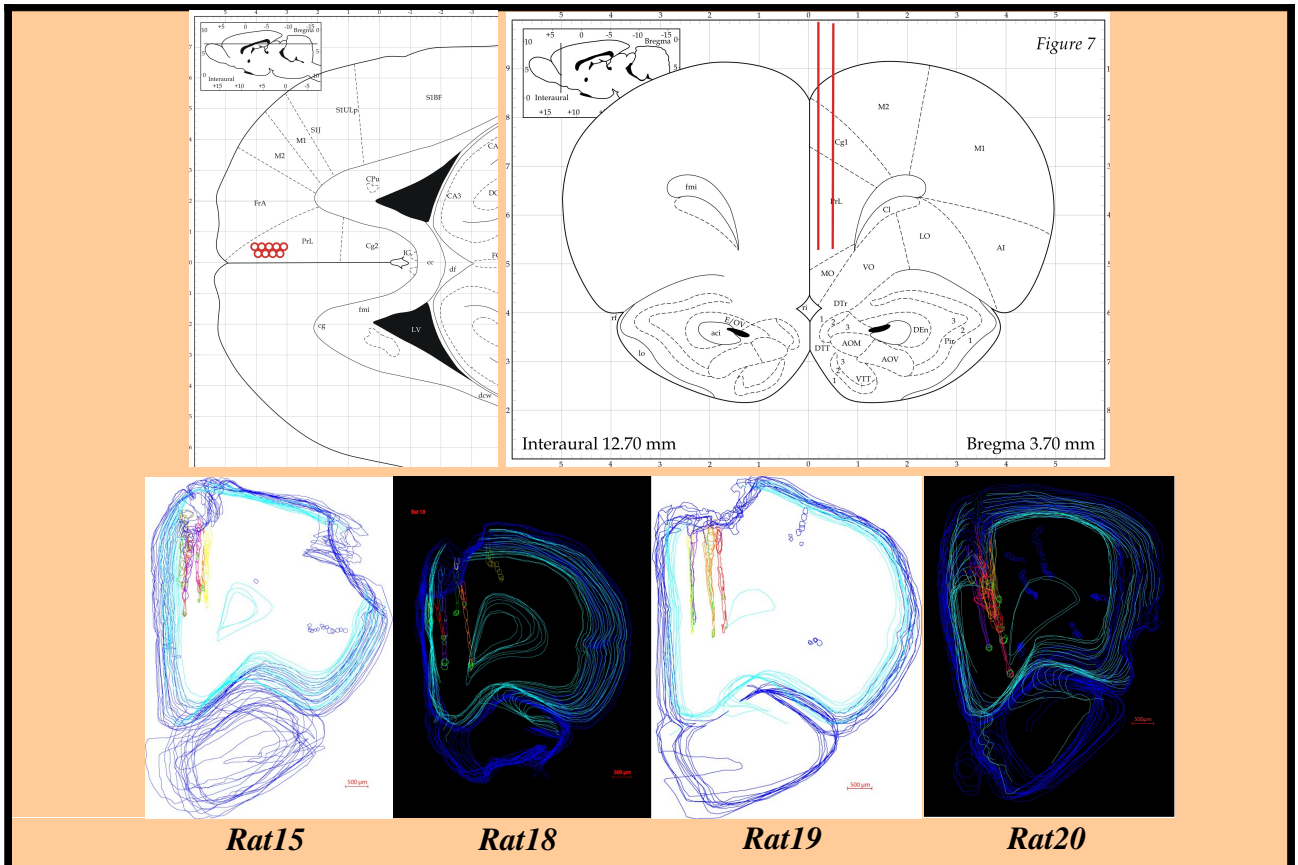


Figure 3.2.2: *Placement of prefrontal recording array and histological analyses of tetrode implantations with the Neurolucida system (thanks to N Quenech'du and Prof. D Hopkins). Top: Left: horizontal section showing the shape and placement of the array of guide tubes, with respect to the targeted cortical areas (PrL = Prelimbic). Right: Example of the expected trajectory of the electrodes through the prefrontal cortex. Bottom: Histological analyses of tetrode placements in four rats.*

Rats

Five Long-Evans male adult rats (225 to 275 g) were purchased (from the Centre d'Elevage René Janvier, Le Genest-St-Isle, France) and kept in clear plastic cages bedded with wood shavings. The rats were housed in pairs while habituating to the animal facility environment. They were weighed and handled each work day. Prior to training they were placed in separate cages and access to food was restricted to one daily feeding, between 2 to 4 hours after daily session, to maintain body weight at not less than 85% of normal values (as calculated for animals of the same age provided *ad libitum* food and water). The rats were examined daily for their state of health and were fed to satiation at the end of each work week. This level of food deprivation was necessary to motivate performance in the behavioral tasks, and the rats showed neither obvious signs of distress (excessive or insufficient grooming, hyper- or hypo-activity, aggressiveness) nor health problems. The rats were kept in an approved (City of Paris Veterinary Services) animal care facility in accordance with institutional (CNRS Comité Opérationnel pour l'Ethique dans les Sciences de la Vie), national (French Ministère de l'Agriculture, de la Pêche et de l'Alimentation No. 7186) and international (US National Institutes of Health) guidelines. A 12 hr/12 hr light/dark cycle was applied.

The data-acquisition technologies

Multiple single units were recorded simultaneously with an array of up to nine tetrodes – bundles of 4 insulated micro-wires (McNaughton et al., 1983; Recce and O'Keefe, 1989). Nine tetrodes were

placed in the right mPFC and three simple electrodes in the right ventral hippocampus (figure 3.2.2), using standard electrophysiological techniques. Spike waveforms were filtered between 600 and 6000 Hz, digitized using the Spike2 software and the Power 1401 device (Cambridge Electronic Device, UK) at 25 kHz and time-stamped. For spike recording, 32 samples at 32 kHz (1 ms total) were recorded whenever the signal exceeded a manually-set threshold (8 pre-trigger and 24 post-trigger samples). This being processed for the four wires of tetrodes, each recorded neural spike was stored as an ensemble of 128 (32 * 4) voltage values.
The signal recorded from the same tetrodes was also passed into a low-band filter (between 0 and 475 Hz) in order to extract Local Field Potentials (LFPs). The time-stamps of the behavioral events were integrated with the spike data on line. A video camera was synchronized with the data-acquisition software and monitored the consecutive positions of the animal during the experiment.

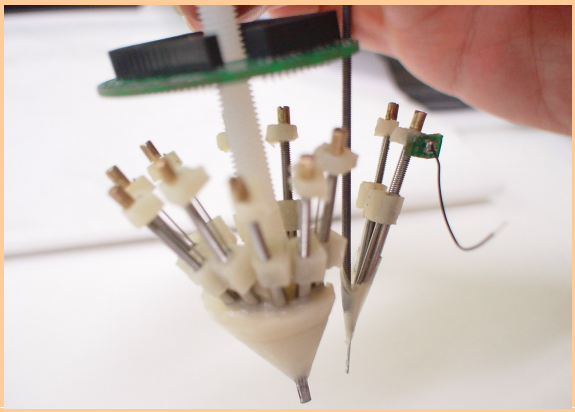


Figure 3.2.3: *The surgical implant for single-unit and EEG recording in the prefrontal cortex and the hippocampus. The system is composed of 2 separately implanted devices. The first (left) allows the positioning of 9 tetrodes in the prefrontal cortex, the second (right) drives two tetrodes in the ventral hippocampal region for EEG recording.*

At the end of experiments, a small electrolytic lesion was made through one lead of each tetrode (25 microA for 10 s.). Lesion sites were determined in Nissl-stained cryosections of formaldehyde (4%) perfused brains. Positions of recorded neurons are currently being determined with respect to the lesion site taking the distance travelled into account.

Signal processing

The digitized waveform signals are analysed after each session for action potential spike discrimination. This discrimination is made possible by the presence of four wires in each tetrode, which is similar to the use of several microphones in a room with several speakers for distinguishing off-line the relative positions of voices (see figure 3.2.4 for an illustration).

We first process the data with a custom python script for Principal Component Analysis in order to reduce the 128 dimensions space to 12 dimensions – three per tetrode wire: the first principal component for each wire corresponds roughly to the amplitude of electrical pulses captured by this wire. Then, we use the Klustakwik software using the Expectation-Maximization algorithm (Celeux and Govaert, 1992) as developed by Ken Harris (Harris et al., 2000) to do the spike-sorting – that is, to cut the 12-dimensional “cloud” of spikes into clusters of pulses emitted by single neurons (figure 3.2.5). Because of the limits of this method, it will always ultimately be necessary to revise the classification manually in a time-consuming procedure. Parameters of the EM algorithm were intentionally chosen to extract a high number of clusters (typically from 15 to 60) so that we then manually merge clusters that are likely to belong to the same cell. This method of spike-sorting provided an identification of 4 cells on average per tetrode per day.

In this thesis we did not distinguish pyramidal cells from interneuron. We rather systematically perform each statistical analysis on the ensemble of prefrontal neurons.

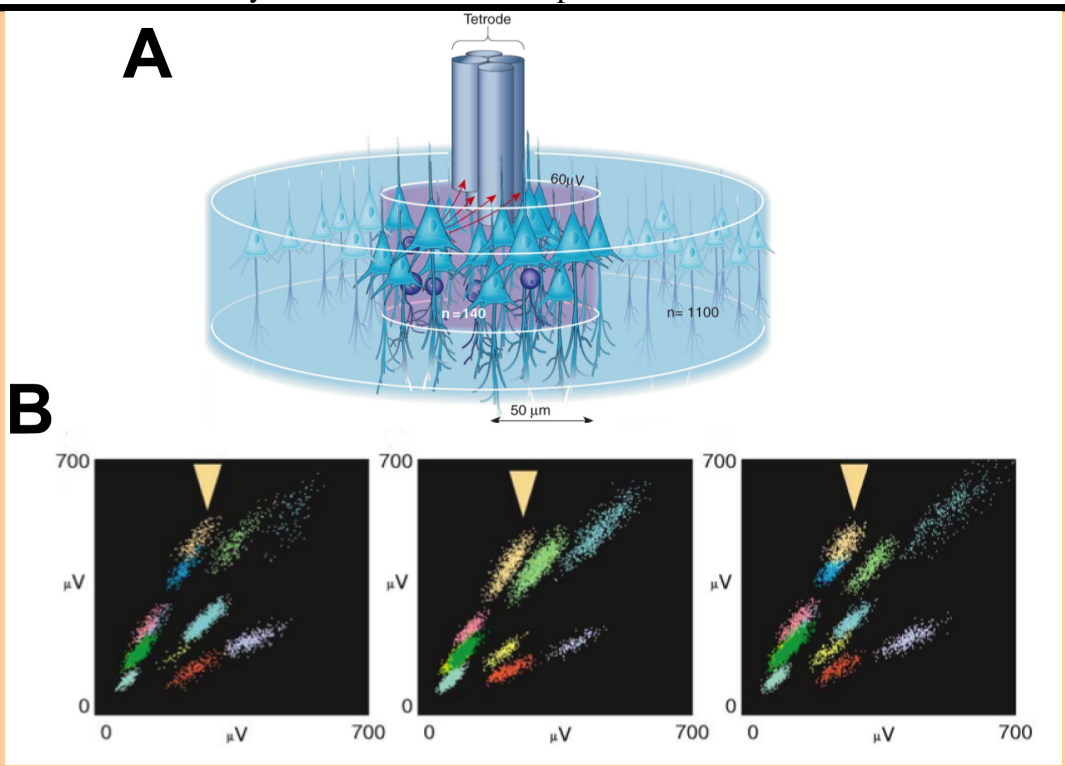


Figure 3.2.4: *Spike sorting problem for tetrode recordings (from Buzsaki, 2004). A. shows the volume of hippocampal tissue with cells that can be discriminated by the tetrode (purple). The light blue volume denotes the cells whose signal can be received by the tetrode, but with an insufficient signal-to-noise ratio for discrimination, thus contributing to background noise. B: shows the scattergram of waveform peak amplitudes with three projections showing each 2 tetrode channels. 10 cells could be discriminated, and are color-coded light blue.*

Automated behavior analysis

In order to correlate cell activity with various task events, the latter had to be extracted from the acquired data. The video tracking data of the instantaneous position of the animal were acquired using the MaxTRAQ software. Using this position information, we extracted the animal's instant of trial initiation, identified as '*START*', defined as the last acceleration in the departure arm. We also extracted the '*OUTCOME*' event which was defined as the instant when the animal reached the reservoir of the chosen arm. Once the animal entered an arm it was not permitted to turn back.

Other events such as the 'Light Onset' and 'Light Offset' of each trial were directly registered by the Spike2 software that generated them.

For the analyses presented here neuronal activity was examined from four trial periods based on these events: *preStart*, *earlyTrial*, *lateTrial*, *postOutcome*, corresponding respectively to the following time windows: [*START* – 2.5 sec; *START*], [*START*; *START* + 2.5 sec], [*OUTCOME* – 2.5 sec; *OUTCOME*], [*OUTCOME*; *OUTCOME* + 1.25 sec]. Note that, in the postOutcome period, the animal's behavior was not consistently the same for all trials: during correct trials, animals stayed for several seconds at the maze arm extremity in order to consume reward. In contrast, during error trials, animals quickly exited the reward zone, in order to go back to the departure arm and start a new trial. Thus, there is a risk of confound with motor correlates when analysing correlations between prefrontal neural activity and this task parameter.

The postOutcome period was hence restricted to a duration of 1.25 seconds, since in 96.39% of all trials recorded in the five animals (including error trials), rats remained at least for 1.25 sec at the reward location. Nonetheless any putative post-outcome variations of cell activity must be examined

closely for this potential confound.

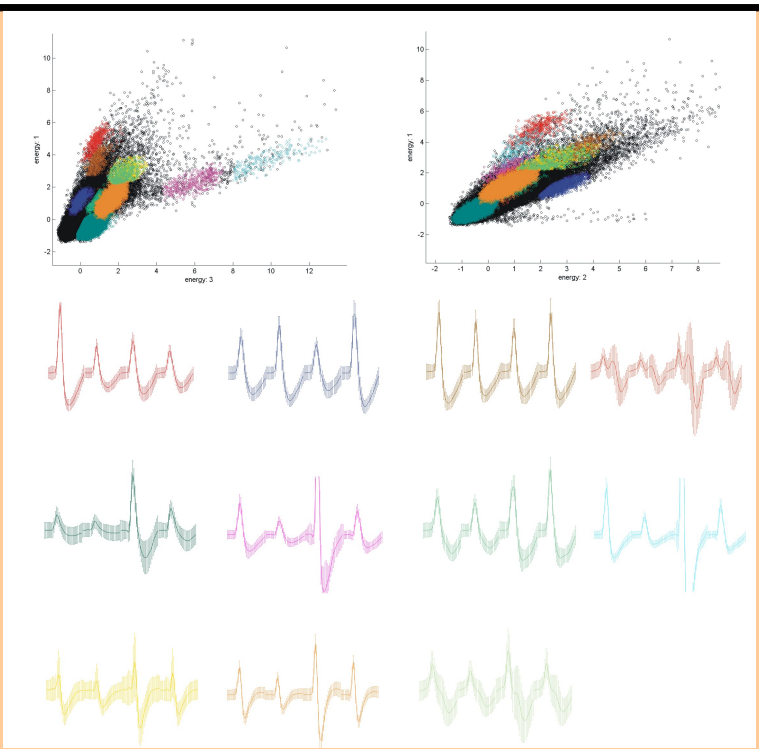


Figure 3.2.5: *Example of spike sorting from one of our recorded tetrodes. The amplitude and waveform recorded on the four channels of a tetrode may be used to discriminate spikes from several cells. In this example, we recorded units from the pyramidal layer of CA1. In the top part of the figure, the amplitude of all spikes recorded on three of channels are plotted pairwise. The different colors refer to the putative single units ("clusters") identified by the cluster discrimination program, and confirmed by the experimenter. This identification was based on PCA of these projection plots, as well as other projection plots involving the spike amplitude and other waveform parameters on the four channels. We identified 10 or 11 single cells from this recording (some uncertainty remains about two clusters (which ones?) with similar waveforms that may actually refer to the same cell). The bottom part of the figure shows the average waveforms on the four channels of the respective identified putative units.*

For the analyses each trial was characterized by variables such as the current task contingency (right, light, left or dark), the position of the light (right or left), the arm chosen by the animal (right or left), and the outcome of the trial (correct or error).
To systematically categorize the animal's behavior in term of "strategy" followed during blocks of N consecutive trials (for instance N=6), we used the following step-by-step method :

Let's consider the behavioral data of the first recording session Rat20/201219 :

#Trial	Task	LightPosition	Animal's choice	Correct/Error
1	R	R	R	C
2	R	R	L	E
3	R	L	R	C
4	R	L	L	E
5	R	L	R	C
6	R	R	R	C
7	R	L	L	E
8	R	L	R	C
9	R	R	R	C
10	R	L	R	C
11	R	L	L	E
12	R	R	R	C
13	R	L	R	C
14	R	L	R	C
15	R	R	R	C
16	R	R	R	C
17	R	R	R	C

where "R" means r*ight*, "L" means *left*, "C" means *correct*, and "E" means *error*.

STEP A : For each trial, possible strategies for the categorization of the rat current behavior are listed. In the case of session "Rat20/201219", this gives the following matrix :

#	Right	Left	Light	Dark	Altern
1	1	_	1	_	_
2	_	1	_	1	1
3	1	_	_	1	1
4	_	1	1	_	1
5	1	_	_	1	1
6	1	_	1	_	_
7	_	1	1	_	1
8	1	_	_	1	1
9	1	_	1	_	_
10	1	_	_	1	_
11	_	1	1	_	1
12	1	_	1	_	1
13	1	_	_	1	_
14	1	_	_	1	_
15	1	_	1	_	_
16	1	_	1	_	_
17	1	_	1	_	_

where "1" means that a strategy is possible at a given trial, and "_" means that a strategy is not possible at a given trial.

STEP B : Then we count from bottom to top the number of trials in each block. This produces the following matrix :

#	Right	Left	Light	Dark	Altern
1	1	_	1	_	_
2	_	1	_	2	4
3	1	_	_	1	3
4	_	1	1	_	2
5	2	_	_	1	1
6	1	_	2	_	_
7	_	1	1	_	2
8	3	_	_	2	1
9	2	_	1	_	_
10	1	_	_	1	_
11	_	1	2	_	2
12	6	_	1	_	1
13	5	_	_	2	_
14	4	_	_	1	_
15	3	_	3	_	_
16	2	_	2	_	_
17	1	_	1	_	_

STEP C : Then we look from top to bottom and keep only blocks whose size is bigger than N (here, N=6). If a "6" is found in the matrix, following trials are kept until a "_" is found. This gives the following matrix :

#	Right	Left	Light	Dark	Altern
1	_	_	_	_	_
2	_	_	_	_	_
3	_	_	_	_	_
4	_	_	_	_	_
5	_	_	_	_	_
6	_	_	_	_	_
7	_	_	_	_	_
8	_	_	_	_	_
9	_	_	_	_	_
10	_	_	_	_	_
11	_	_	_	_	_
12	6	_	_	_	_
13	5	_	_	_	_
14	4	_	_	_	_
15	3	_	_	_	_
16	2	_	_	_	_
17	1	_	_	_	_

Thus in this case there was no detectable strategy in trials 1-11, then the 'Right' strategy for trials 12-17.

STEP D (not shown) : Finally, if two different blocks are overlapping, we need to decide which of the two strategies best describes the animal's behavior. Two cases are possible: 1) if one block is "included" in the other (i.e. if the latter block starts first and ends last), then the former block is deleted; 2) if the two blocks have non overlapping trials (i.e. if one block starts first and the other ends last), then the two blocks are kept, which means that 2 different strategies are simultaneously possible during the overlapping trials.

Table 3.1 computes the probability of the rat executing a block of N consecutive trials at a single strategy by chance. We decided to consider that the rat is indeed following a certain strategy if it is so during a block of at least 6 consecutive trials, since arriving at consecutive rewarded trials by chance has a $p<0.05$. For the rest of the manuscript, only blocks of at least 6 consecutive trials following a given strategy will be considered.

blocks' min. size	*3 trials*	*4 trials*	*5 trials*	*6 trials*	*7 trials*	*8 trials*
prob. to get a block by chance	.25	.125	.0625	.0312	.0156	.0078

Table 3.1: *summary of the probabilities to find a block of N consecutive trials with the same strategy by chance. N = 3, 4, ... or 8 consecutive trials.*

Electrophysiological data analysis

One-way ANOVA was used to determine behavioral correlates. ANOVA results were considered significant at $p < 0.05$. The Student–Newman–Keuls test was employed for post hoc analyses.

The Wilcoxon-Mann Whitney test was used to determine the contributions of task parameters (the trial correctness, the light position, the animal's choice or the task contingency rule) on variations in the firing rate of the neurons. Using a Bonferonni correction, we considered a neuron as being significantly modulated by a given task parameter if the Wilcoxon-Mann Whitney test gave a $p < 0.01$ in one of the four task periods (*preStart*; *earlyTrial*; *lateTrial*; *postOutcome*).

Finally, the Wilcoxon-Mann Whitney test was used to determine the contributions of the possible behavioral strategies engaged by the animal on variations in the firing rate of the neurons. Using a Bonferroni correction we considered a neuron as being significantly modulated by behavioral strategies if the Wilcoxon-Mann Whitney test gave a $p < 0.002$ for one of the possible strategies (*Right*; *Left*; *Light*; *Dark*; Alternation) in one of the four task periods (*preStart*; *earlyTrial*; *lateTrial*; *postOutcome*).

BEHAVIORAL RESULTS

A total of 3322 trials were performed by 5 rats over 108 sessions. Sessions' length was comprised between 7 and 68 trials, with an average of 31 trials.

Rats were exposed at least to three different consecutive rules:

- 2 rats were exposed to two changes in task rules (from the right rule to the light rule, then from the light rule to the left rule);
- 2 rats were exposed to three changes in task rules;
- 1 rat was exposed to 17 changes in task rules.

On average rats took 3.8 experimental sessions, corresponding to 116 trials, to acquire a given task rule.

ELECTROPHYSIOLOGICAL RESULTS

Behavioral correlates

We recorded 2413 cells from the prefrontal cortex in these sessions. After excluding cells with an average rate of discharges during trial inferior to 0.3 Hz, we considered 1894 prefrontal neurons for statistical analyses. Within this ensemble of cells, an important subset showed a significant behavioral correlate: 70.41% showed activity that was significantly different in one of the trial periods (*preStart*, *earlyTrial*, *lateTrial*, *postOutcome)* using an ANOVA test $p < .05$ *(*Note that in this and subsequent tallies, the possibility that the same neuron was recorded from one day to the next is not taken into account, since this proved quite difficult to verify). This proportion of cells is comparable with previous results *(50 % : Pratt and Mizumori, 2001; 68 % : Mulder et al., 2003).* Figures 3.2.6 and 3.2.7 show examples of such neurons. The former has an activity which phasically increases around the time of the outcome. The latter shows a tonic activity with an inhibition around the time of the outcome.

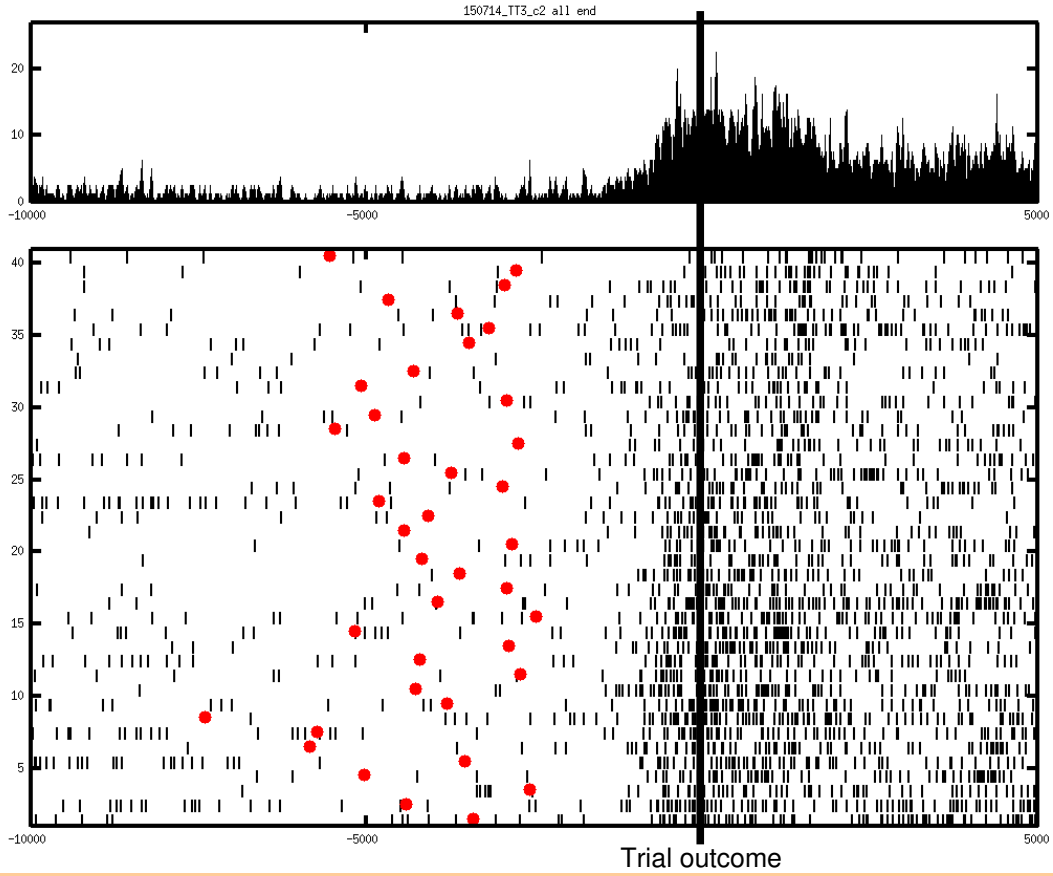


Figure 3.2.6: *Peri-event time histogram and raster display of a prefrontal neuron showing activity correlated with the trial outcome.* Activity of this neuron is synchronised with the time of the trial OUTCOME, that is, the time when the animal reaches the reward location. A window of 15 sec of activity is displayed (x-axis). The y-axis corresponds to consecutive trials through the session (from bottom to top). For each trial, a red dot represents the START time , and short black vertical traces represent action potentials emitted by the neuron. The upper part of the figure displays the histogram of activity of this neuron cumulated on all trials.

Interestingly, subsets of cells showed significant modulations of behaviorally correlated activity as functions of different task parameters: 14.41% were correlated with the choice of the animal (right or left), 10.82% were correlated with the reward, and 7.76% were correlated with the position of the light (Wilcoxon Mann-Whitney test, $p < .01$ with Bonferroni correction in one of the four trial periods: *preStart*, *earlyTrial*, *lateTrial*, *postOutcome*).

Figures 3.2.8 and 3.2.9 respectively show examples of a neuron with activity modulated by reward and another neuron correlated with the left-right choice of the animal. Both cases have a peri-outcome phasic response. In the former neuron, this response appears only during error trials, no matter which arm was chosen (the task rule was *Light*). In the latter neuron, the response appears only when the right arm is chosen, no matter if the choice was correct or not (task rule was *Light*).

Table 3.2 summarizes these results for each rat. Overall these responses demonstrate that the prefrontal cortex encodes the main parameters required to solve the task.

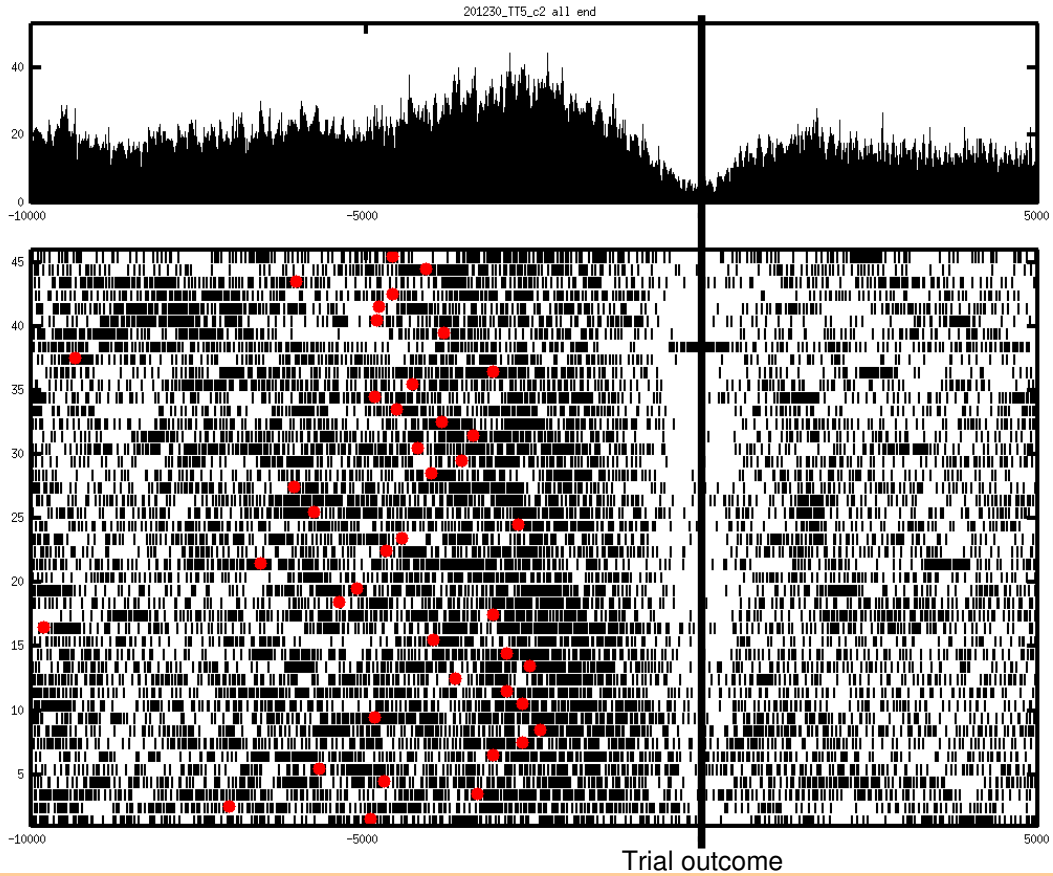


Figure 3.2.7: *Peri-event time histogram and raster display of a prefrontal neuron showing activity synchronised with the time of the trial outcome.* Same format as previous figure. The displayed neuron has an activity which transiently decreases around the time of the outcome.

number of cells	1894	83	379	556	205	671
	TOTAL	**Rat12**	**Rat15**	**Rat18**	**Rat19**	**Rat20**
behavioral correlates (1 test,p<0.05)	70.91%	50.60%	68.60%	69.78%	32.68%	87.33%
reward correlates (4 tests,p<.01)	10.82%	2.41%	7.92%	5.94%	5.37%	19.23%
choice correlates (4 tests,p<.01)	14.41%	1.20%	12.14%	11.69%	4.39%	22.65%
light correlates (4 tests,p<.01)	7.76%	6.02%	4.75%	7.01%	5.37%	11.03%

Table 3.2 : *Summary of the percentages of cells in each category.* p values were adjusted for Bonferroni corrections. Cells with *behavioral correlates* showed activity that was significantly different in one of the trial periods (*preStart*, *earlyTrial*, *lateTrial*, *postOutcome)* using an ANOVA test $p < .05$. Cells with *reward correlates* showed significant modulations of behaviorally correlated activity as functions of the trial correctness (rewarded or not) in at least of the four task periods (One Wilcoxon-Mann Whitney test per task period, $p < .01$. The same method was used to determine neurons with *choice correlates* (modulation by the arm chosen: right/left), and cells with *light correlates* (modulation by the position of light: right/left).

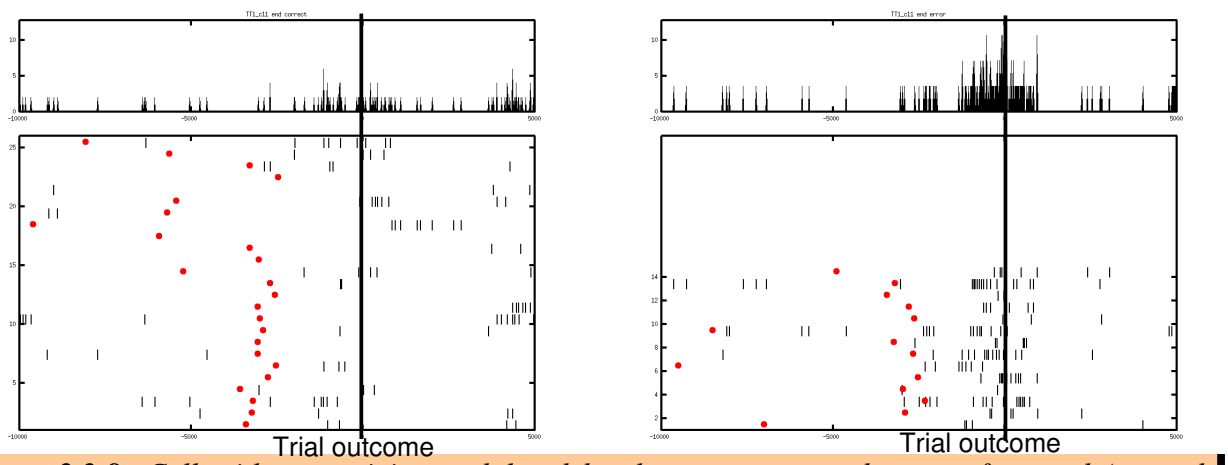


Figure 3.2.8: *Cell with an activity modulated by the presence or absence of reward (reward correlates).* Same format as figure 3.2.6. Left : correct trials; Right : errors. This cell showed an increase in activity starting before the trial outcome only during unrewarded trials. Since the cell was recorded during a session were the task rule was *Light*, unrewarded trials included both visits to left and right arms.

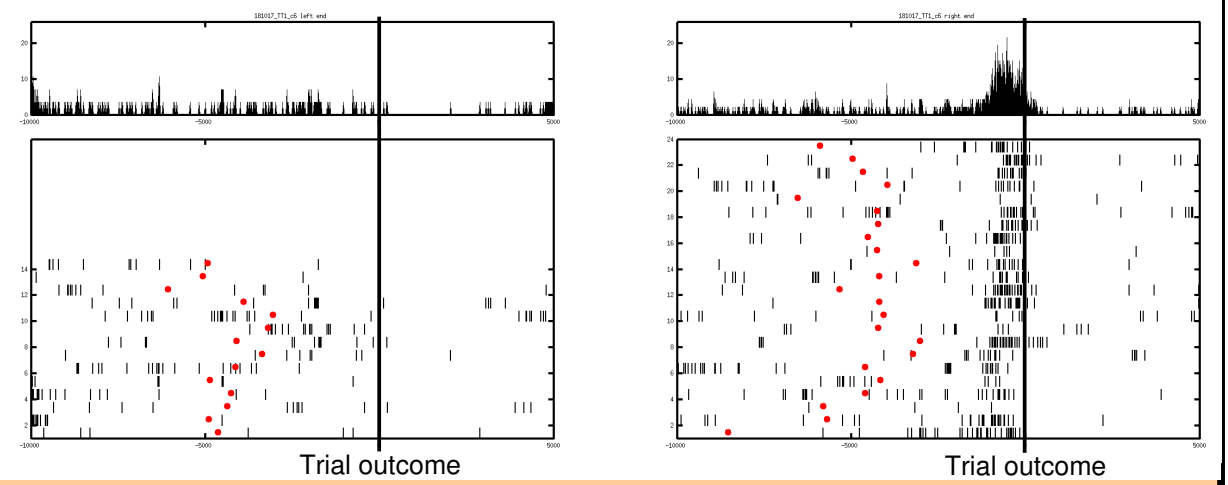


Figure 3.2.9: *Cell with an activity modulated by the chosen arm (choice correlate).* Same format as figure 3.2.6. Left : left trials; Right : right trials. This cell showed an increase in activity starting before the trial outcome only during visits to the right arm. Since the cell was recorded during a session were the task rule was *Light*, visits to the right arm included both rewarded and unrewarded trials.

Transitions in activity related to shifts in the task rule or in the behavioral strategy

A subset of prefrontal neurons showed abrupt changes in firing rate during the course of a recording session. This kind of transition was found in 12.25% of the prefrontal cells (i.e. 232 neurons). These changes included disappearance or appearance of a behavioral correlate, and change in magnitude in behaviorally correlated activity. In 133 cells these transitions in cell activity corresponded to strategy shifts spontaneously made by the animal. In 99 cells, these transitions corresponded to changes in the task rule. In addition, in 25 cells the activity was modulated by both changes in the task rule and by changes in the behavioral strategy. In most cases, there was only one apparent transition and the statistical effect was due to the proximity between the two events within the

session.
For the tests of cell activity modulations by a change in the task rule, only cells recorded during a *shift-session* could be evaluated. Among the 1894 initial active prefrontal cells, this leaves a total of 423 neurons recorded in 24 sessions. A cell was considered as significantly modulated by the task rule if its activity during one of the four trial periods (*preStart*, *earlyTrial*, *lateTrial*, *postOutcome)* was significantly different in trials before the switch compared to trials after the switch.
For tests of cell activity changes associated with a shift in the behavioral strategy, only cells recorded during sessions with blocks of at least 6 consecutive trials of at least one strategy were taken into account. Among the 1894 initial prefrontal cells, this leaves a total of 1365 neurons recorded in 76 sessions. A cell was considered as significantly modulated by the behavioral strategy if its activity was found as significantly modulated by one of the five possible strategies (*Right*; *Left*; *Light*; *Dark*; *Alternation*) during one of the four trial periods (*preStart*, *earlyTrial*, *lateTrial*, *postOutcome*), as compared with activity during the remainder of the session.
As a result, 26.48% of cells recorded during a *shift-session* were significantly modulated by a shift in the the task rule. In contrast, 11.58% of cells recorded during a session with blocks of at least 6 consecutive trials of the same strategy were significantly modulated by the behavioral strategy engaged by the animal.
Figures 3.2.10 to 3.2.12 show examples of raster displays and histograms of cells recorded in the same session that show such transitions in activity. These neurons display transitions correlated either with the change in the task rule (figures 3.2.12 and 3.2.13) or with the spontaneous change in the animal's strategy (figure 3.2.14). In this session, the rat started by performing a *light* strategy that he had started to learn during the previous session. At trial #11, the criterion being passed, the *left* task rule was imposed. The rat continued to perform the *light* strategy for about 12 trials. Then for 6 trials, the strategy was indeterminable. Finally, for the last 9 trials of the session, the animal performed a *left* strategy.
Strategy shifts could be divided in three different cases: 1) a shift from a strategy *A* to a strategy *B*; 2) a shift from a strategy *A* to an indeterminable strategy; or 3) a shift from an indeterminable strategy to a strategy *A*. Interestingly, among the 11.58% of neurons with a change in activity in relation to a shift in the animal's strategy, 70 cells were found during shifts where either strategy *A* or *B* was the *alternation* strategy. In contrast, respectively 46, 52, 39 and 9 neurons were found during shifts where either strategy *A* or *B* was the *right*, *light*, *left* and *dark* strategy. This suggests that the *alternation* strategy, which was not rewarded in our task, was encoded by more prefrontal neurons than other strategies. Moreover, the *dark* strategy was only poorly represented within the mPFC network that we recorded, which is consistent with the observation that the *dark* rule was the hardest to learn by our rats. However, we cannot exclude the possibility that the difference in these proportions is the consequence of a sampling effect, having only recorded a very small subpart of mPFC neurons.
Finally, the ensemble of transitions in cell activity were found to correspond to two different patterns. In one case, the activity was found to be higher after the shift than before, thus corresponding to an increase in the neuron's activity. In the other case, it was lower, thus corresponding to a decrease of activity.
Both cells with an increase and a decrease in activity were found. Figures 3.2.13 and 3.2.14 – which show two groups of 50 cells showing such kind of transitions in activity. The activity of the first 50 cells are synchronised with the change in the task rule (figure 3.2.13). The activity of the 50 other cells are synchronised with a shift in the animal's behavior (figure 3.2.14). In each figure, the upper part displays neurons with an decrease in activity following a shift, and the bottom part shows neurons with an increase in activity.
Strikingly, some cells which did not respond before the shift, started to fire during the trials following the shift. Symmetrically, some cells which used to respond before the shift almost stopped their activity after. The other cells showed a modulation of activity before compared to after the

shift.
These results suggest that the prefrontal cortex can detect changes in the strategy rule. Moreover, the prefrontal cortex encodes information concerning spontaneous strategy shifts performed by animals. As a consequence, the prefrontal cortex could possibly contribute to the selection of the strategy to perform at a given moment. We will see below that these elements are also crucial for neurocomputational models.

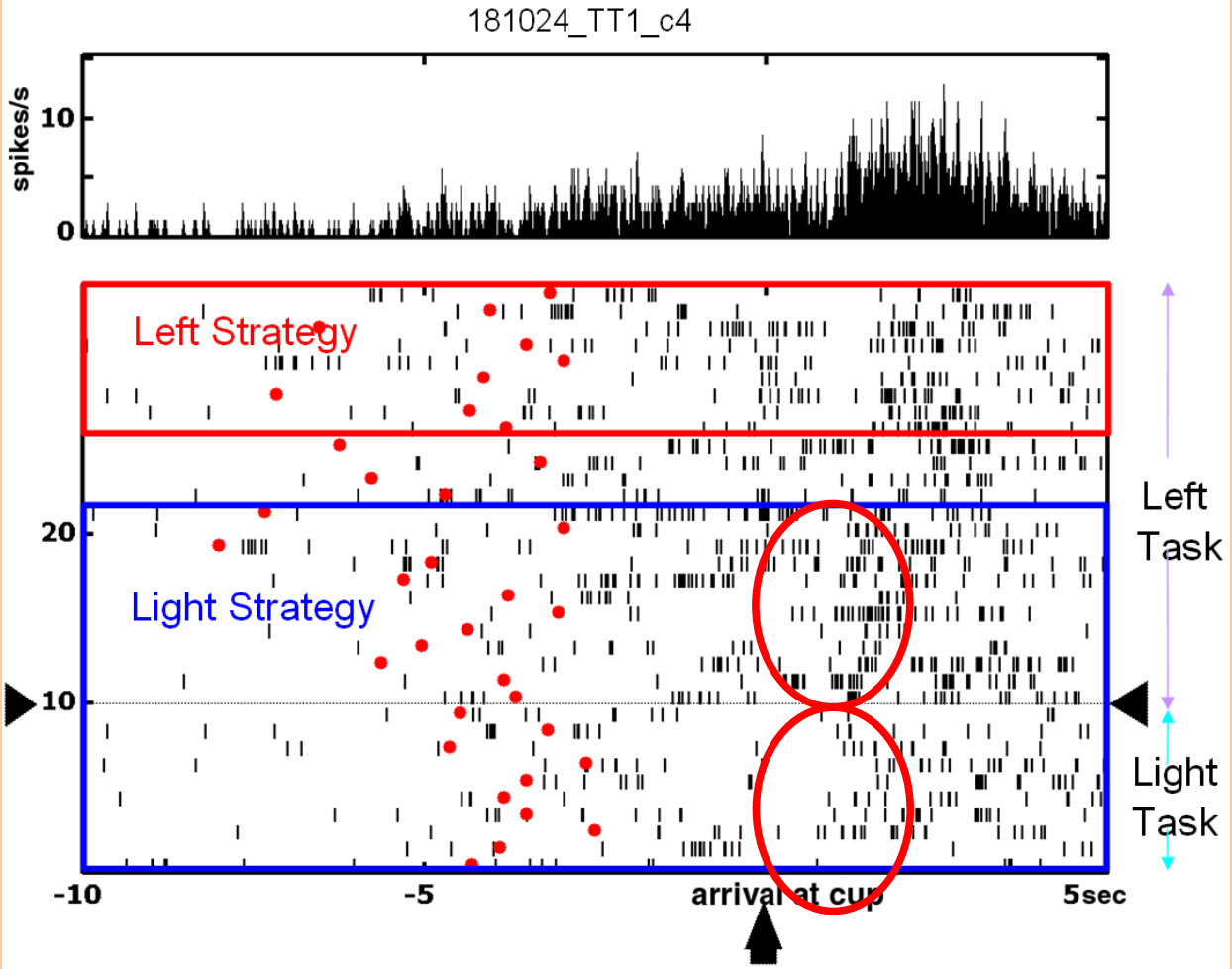


Figure 3.2.10: *Raster of a cell showing a transition in activity related to the change in task contingency.* Same format as figure 3.2.6. In the session presented here (session #181024), which is also used in figures 3.2.11 and 3.2.12, the rat started by performing a *light* strategy that he had started to learn during the previous session. At trial #11, the criterion being passed, the *left* task rule was imposed. The rat continued to perform the *light* strategy for about 12 trials. Then for 6 trials, the strategy was indeterminable. Finally, for the last 9 trials of the session, the animal performed a *left* strategy. The activity is synchronised with the time of the trial OUTCOME, that is, the time when the animal arrives at the reward location. This neuron showed a greater activity during the *postOutcome* period after the task rule change rather than before.

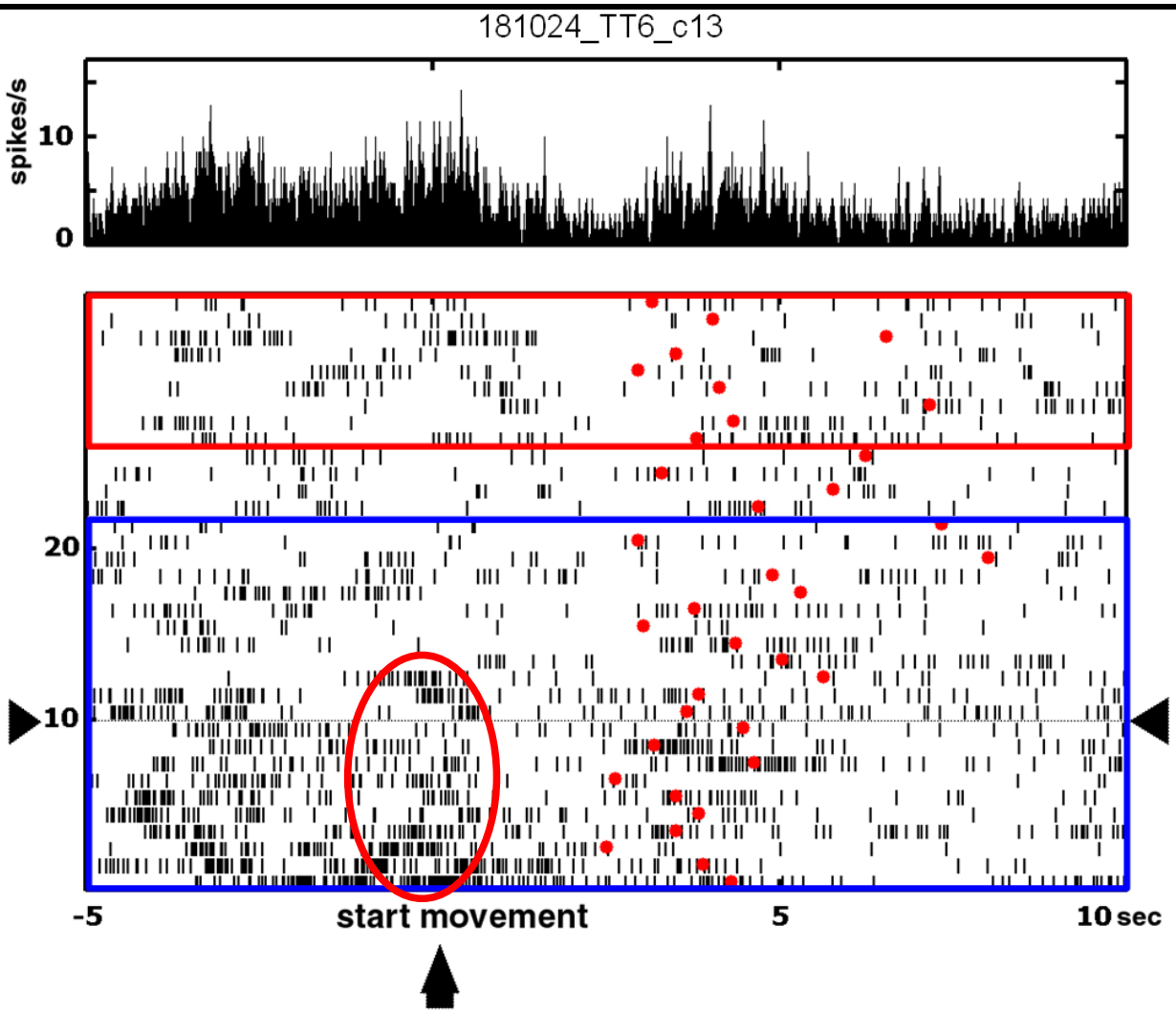


Figure 3.2.11: *Raster of a cell showing a transition in activity related to the change in task contingency.* Same format and same displayed session as figure 3.2.10. This neuron show a greater activity around the time of trial START before the task rule change rather than after. Interestingly, the neuron continued to burst around the START for three trials after the task rule change, as if late detecting the task rule change.

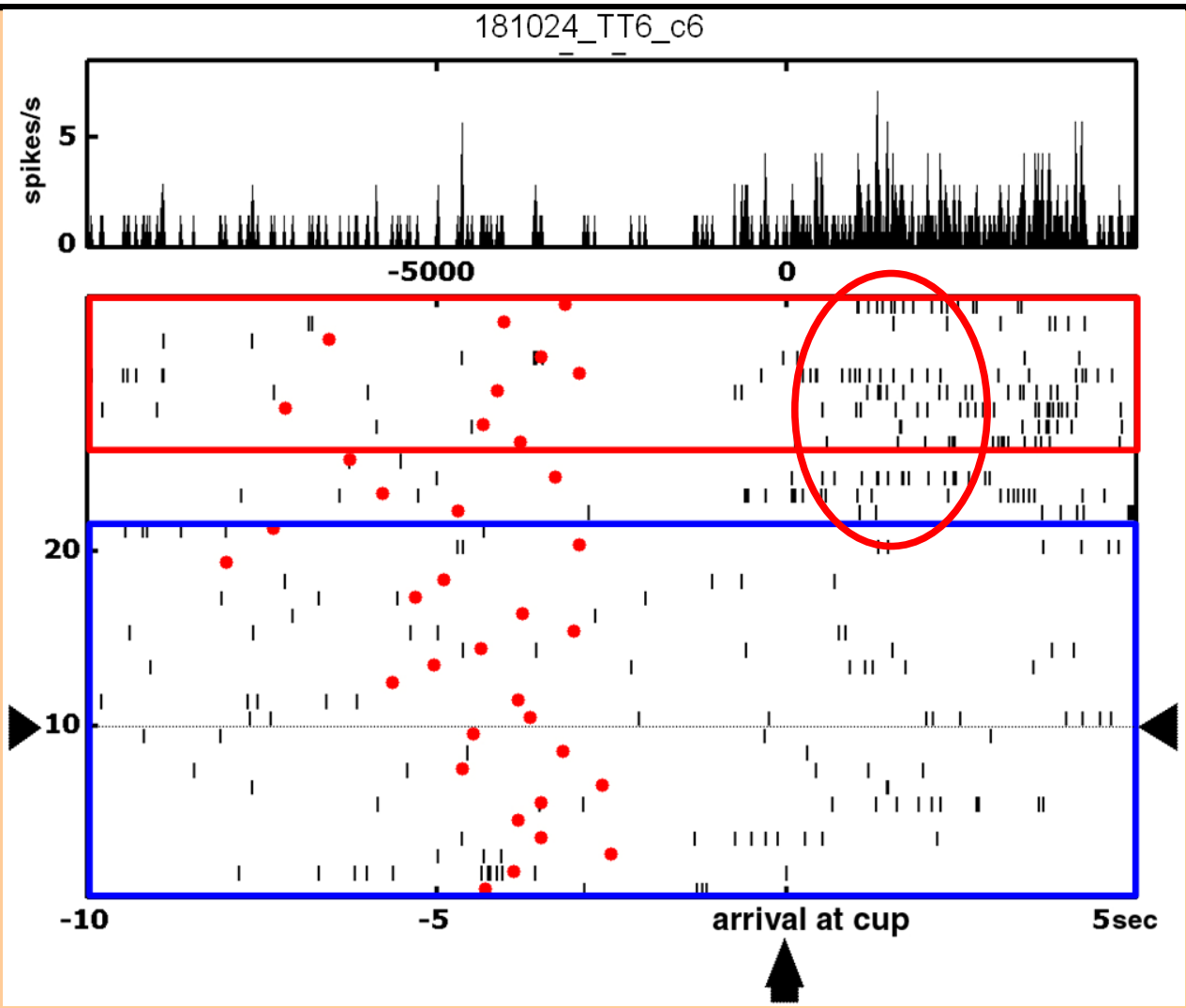
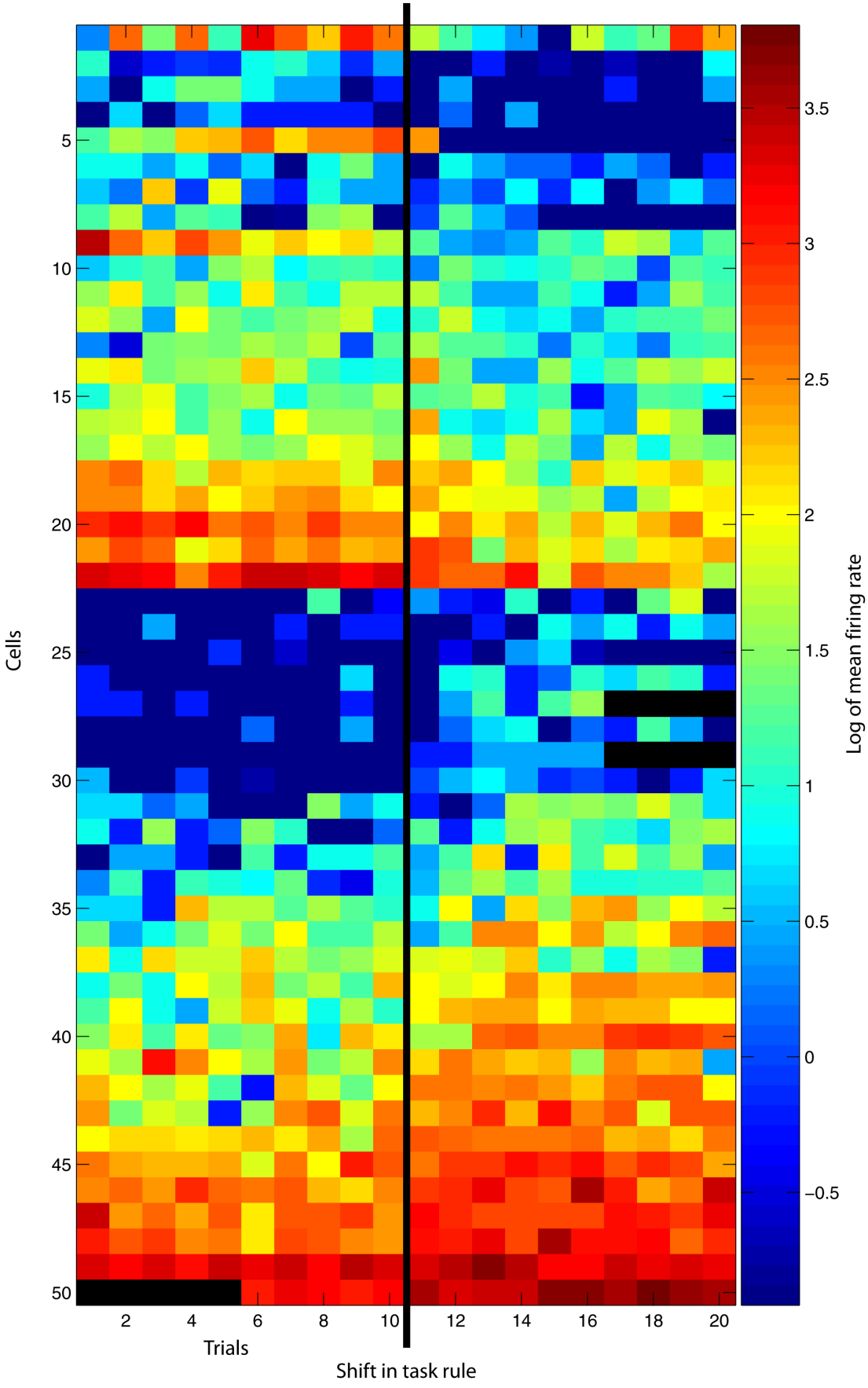


Figure 3.2.12: *Raster of cell showing a transition in activity related to the spontaneous change in the rat behavioral strategy.* Same format and same displayed session as figure 3.2.10. This neuron show a lower activity around the time of trial OUTCOME during trials when the *light* strategy was performed by the animal rather than after.

Figure 3.2.13 (next page): *50 cells recorded in 4 rats showing a transition in activity correlated with a shift in the task rule.* Each row corresponds to one neuron. As the cells were not recorded from the same sessions, they were all synchronised on the trial where the task rule shift occurred. Black areas are displayed when there was less than 10 trials before or after the shift.


Cells
5
10
15
20
25
30
35
40
45
50
2
4
6
8
10
12
14
16
18
20
Trials
Shift in task rule
3.5
3
2.5
2
1.5
1
0.5
0
−0.5
Log of mean firing rate

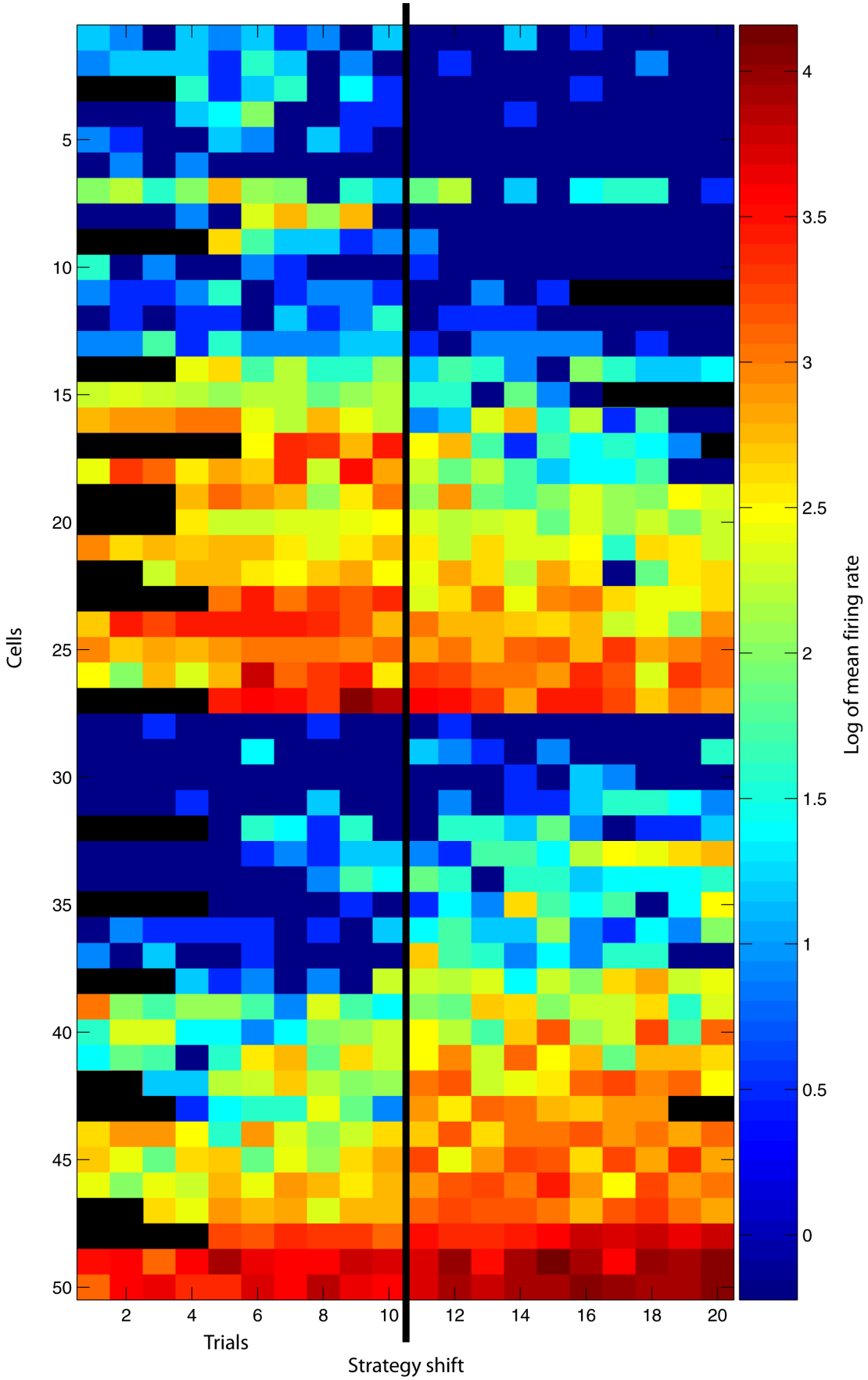

Cells
5
10
15
20
25
30
35
40
45
50
2
4
6
8
10
12
14
16
18
20
Trials
Strategy shift
Log of mean firing rate
4
3.5
3
2.5
2
1.5
1
0.5
0

Figure 3.2.14 (previous page): *50 cells recorded in 4 rats showing a transition in activity correlated with the animal's behavioral strategy shift.* Each row corresponds to one neuron. As the cells were not recorded from the same sessions, they were all synchronised on the trial where the strategy shift occurred Black areas are displayed when there was less than 10 trials before or after the shift.

Discussion

Our results show prefrontal cells with behavioral correlates, and correlates with the main task parameters (reward, choice, cue). These results are in line with previous electrophysiological recordings from the rat medial prefrontal cortex showing movements, reward and stimulus correlates (Poucet, 1997; Jung et al., 1998; Pratt and Mizumori, 2001; Mulder et al., 2003).
Moreover, we found cells with a transition in activity related to extradimensional changes in the environment's conditions for reward delivery and to changes in the animal's behavioral strategy. The former results confirm that mPFC can participate in the detection of extradimensional shifts (Birrell and Brown, 2000; Ragozzino et al., 2003). However, since we did not use intradimensional task rule shifts in our experimental design, we cannot conclude anything concerning the role of mPFC in intradimensional shifts, which the latter authors report as unimpaired by mPFC lesions. Moreover, further analyses should be done on our data to see whether more cells with a transition in activity in response to a shift in the animal's spontaneous strategy are found during extradimensional spontaneous shifts than during intradimensional ones.

Besides, the cell activity transitions we found in relation to task rule changes suggest that the prefrontal cortex could contribute to attentional processes such as the detection of such changes. This is consistent with the neuropsychological literature stating that lesions of medial prefrontal cortex result in attentional deficits in tasks where rats have to detect the change in task rules, and in impair strategy shifting in response to such changes (de Bruin et al., 1994; Granon et al., 1994; Granon and Poucet, 1995; Muir et al., 1996; Joel et al., 1997; Ragozzino et al., 1999a,b; Birrell and Brown, 2000; Delatour and Gisquet-Verrier, 2000; Colacicco et al., 2002; McAlonan and Brown, 2003; Ragozzino et al., 2003; Salazar et al., 2004; Lapiz and Morilak, 2006).
However, the latter lesion results leave at least two interpretations opened: 1) rats with mPFC lesion are impaired in the **detection** of rule changes requiring strategy shifts; 2) rats with mPFC lesions can detect rule changes, but are impaired in the acquisition or **selection** of the appropriate strategy in response to such changes.
Our results suggest that both could be possible. Indeed, a neural network model having to select and shift behavioral strategies could work in a similar manner than a model of action selection: 1) several groups of neurons would represent each of the possible strategies; 2) the neurons would receive input information concerning the task context and events; 3) an output network would receive a convergence of input from the first group, and would select the strategy with the highest associated neural activity – for example, see the *gating network* used for strategy selection in the model of Chavarriaga and colleagues (Chavarriaga et al., 2005a; Dollé, Khamassi, Guillot and Chavarriaga, 2006). The cells we recorded could be characterized by either the first and the third groups, whereas further investigation are necessary to distinguish between the two.
However, the proportion of cells with an activity correlated to task rule changes being higher than the cells correlated with the animal's strategy (respectively 26.48% and 11.58%), it is possible that the mPFC is more involved in the *detection* than in the *selection*. The basal ganglia is indeed a

possible candidate for subserving this function in interaction with mPFC, since it is considered as a central brain structure for *action selection* (Redgrave et al., 1999a).
Moreover, having assumed in previous chapters that the striatum could participate in learning of different navigation strategies, it could be possible that information about new strategy learning comes from the striatum to mPFC through prefronto-striatal loops.
However, more investigations will be required to determine if the prefrontal cortex can really subserve strategy selection, or if strategy-related activity is only an emerging property related to goal selection. Indeed, a previous study has reported spatial goal encoding in mPFC (Hok et al., 2005). In experimental designs where a particular strategy is appropriate to reach a particular goal in the environment, goal correlates and strategy correlates cannot be discriminated. So it would be interesting to design a task protocol where different strategies can lead to a same goal, and different goals can be reached using a same strategy. Recorded neural activity in mPFC during such a task would help understand the role of mPFC in this ensemble of goal-directed behaviors.
The different functions discussed here can coexist within mPFC. Inside the medial prefrontal cortex, there could be different segregated networks that subserve different functions related to goal-directed behavior. And previous activities reported in the mPFC could reflect functions that coexist with strategy selection. Notably, it has been shown that neighboring cells in the mPFC have very heterogeneous behavioral correlates and show a weak synchrony, suggesting that these neighboring cells process largely independent information (Jung et al., 2000).

Finally, understanding how different contributions to goal-directed behavior can be subserved by different brain areas can constitute a major intuition towards the design of autonomous robots. Indeed, autonomy, flexibility and adaptation in existing robots is much weaker than rats' cognitive capacities (Meyer and Guillot, In press). Taking inspiration from the way the brain separates or merges goal selection, strategy selection, action planning and habit learning can help design efficient neural network architectures for the control of autonomous robots.

3. Towards a model for strategy shifting

To reproduce the way in which animals learned the consecutive shifting rules of the task, we used a Markovian-like model for strategy learning. Hidden-Markov models provide a Bayesian framework to learn to represent dynamic phenomena where the situation changes with time (Dayan and Yu, 2001). In such models, at any given state, the decision taken is strictly depending on the previous state.
In our case, the states will correspond to strategies that can be performed to solve the Y-maze task described in the previous section. The model will learn to select the appropriate strategy based on the history of previous trials. The model will be trained on real behavioral data performed by the rats in these experiments. That is, while replaying the sequence of trials performed by real rats and their outcomes, the model will make a proposition (or *prediction*) of the strategy to perform. This proposition will be stored in order to be compared with the choice of the animal. Then, the model will not be trained on its own errors, but rather on the errors made by the animal. As a consequence, at any given moment during the simulation, the model will have the same "experience" as the animal. This way, we can tune parameters from the model so that the model learns at a speed comparable to the animal's performance. If the model can reach a similar performance and can perform strategy shifts at similar task phases as the animal, then we can use the model in two different manners:

- First, the model would be considered as a good representation of the way the animal learned the task, and of the way the animal could decide to shift its own strategy;
- Second, parameters in the model could be compared to neuronal activity measured in the rat

prefrontal cortex, in order to see if there is any correlation.
This was the objective we intended by this modeling work. Yet, this work is underway, and we will present the preliminary results of this project.

METHODS

We implemented 5 states corresponding to the 5 identified possible strategies engaged by rats in our experiment. Let's note *S* as the ensemble of strategies. *S* = {*Right*; *Left*; *Light*; *Dark*; Alternation}. Each strategy was assigned with a confidence value $C_k(t)$ (with $k \in \{1;5\}$ and *t* corresponding to the trial number), initialised to zero: $\forall k, C_k(0)=0$.

The probability of transition from any state to the state with the current highest confidence is always equal to 1. In other words, when a strategy has the highest confidence value, it is systematically chosen to be performed by the model, without any stochasticity. When two states have the maximum confidence, they both have a probability of 0.5 to be chosen randomly, independently from the past history.

At each trial, the model makes a prediction simply based on the appropriate action corresponding to the current strategy with the maximal confidence. For example, if the strategy *Right* has the highest confidence, then model predicts a Right move.

Prediction : $P(t+1)=appropriateAction\left(argmax_k\left[C_k(t)\right]\right)$

Then the model is adapted based on the behavior of the rat during the current trial.

Learning : $\forall k \epsilon S', C_k(t+1)=C_k(t)+\eta.(1-E(t+1))$

where *S'* is the ensemble of possible strategies at trial *t*, computed in the following manner: if the rat went to the right while the light was also on the right, then *S'* = {*Right*;*Light*}.

where E(t+1) is equal to 1 if the animal made an error at trial t+1, and where η is the learning speed of the model .

Figure 3.2.6 shows a an example of progress of the procedure on 5 consecutive trials.

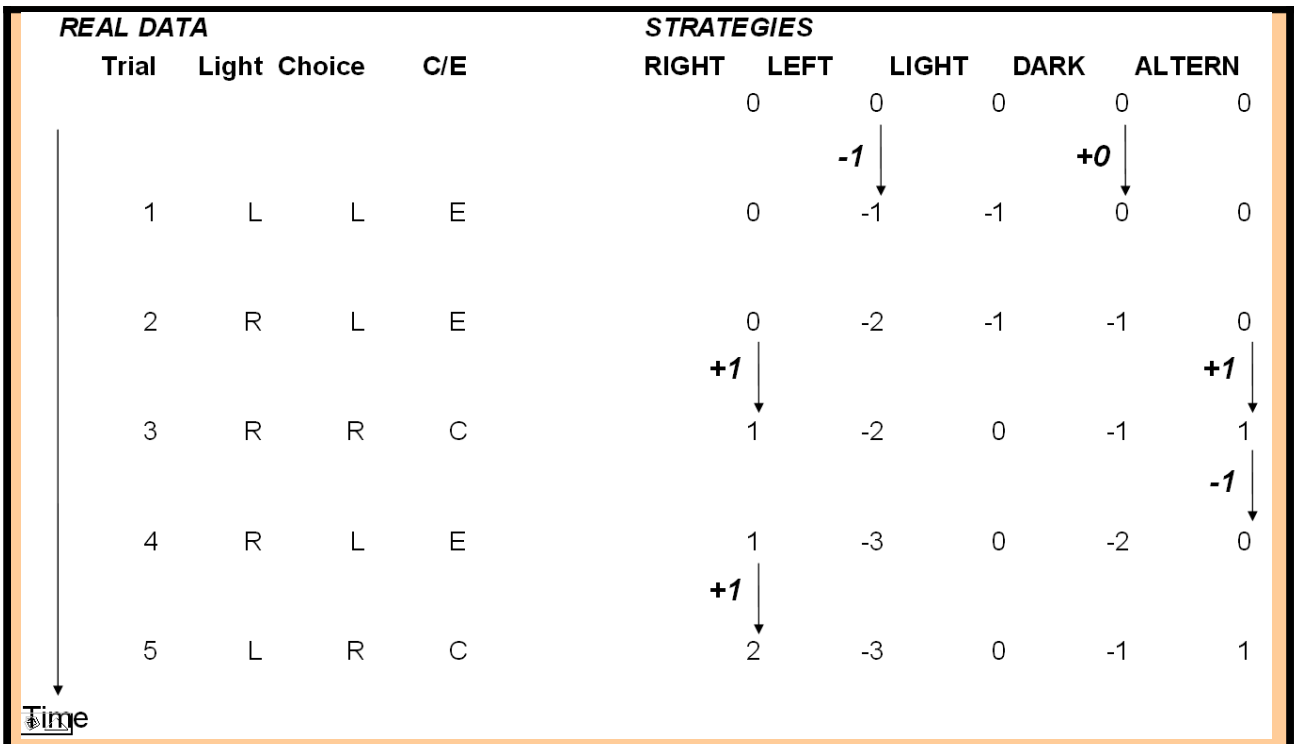


Figure 3.2.6: *Example of unfolding of the Confidence model on 5 consecutive trials.* The left part of the figure describes the task parameters of each trial during the real experiment. The right part describes the evolution of the strategies confidences in the model. Confidences are initiated to 0. Then each arrow describes a learning process targeting strategies that are consistent with the animal behavior at the current trial. The value marked near each arrow corresponds to the increment applied to the strategy confidence, depending on the correctness of the trial made by real animals.

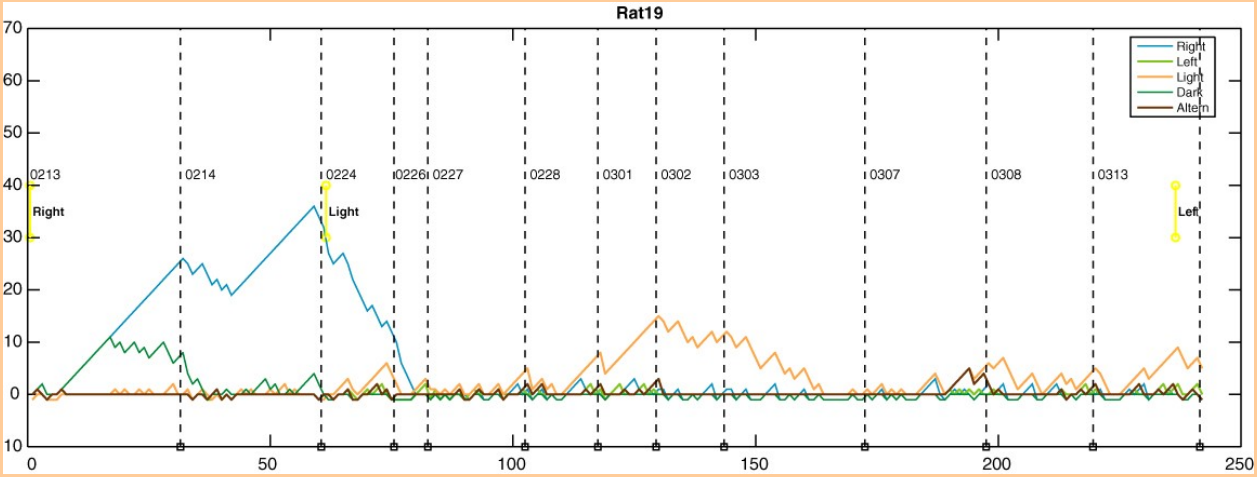


Table 3.2.7: *Sample of temporal evolution of the confidence value within the model.* The figure shows the simulation of the model on 12 consecutive sessions recorded in rat #19. The x-axis represents the consecutive trials. Vertical dashed lines separate data from different sessions. The y-axis represents the confidence values of the model (see inset). Each curve represents the confidence value of one strategy. Vertical yellow lines represent shifts in the task rule that occurred during the experiment with real rats.

Figure 3.2.7 displays a sample of results for simulation of the model with a certain set of parameters. The curves represent the temporal evolution of the confidence values associated to each strategy in the model. The latter was simulated on 12 consecutive sessions recorded in rat #19. In this version of the model, the appropriate strategy is learned more rapidly than the real rat. As shown on the figure, the curve representing the *right* strategy increases sharply before the task rule shift. Later, during sessions *0301*, *0302* and *0303*, the model has already learned the left strategy whereas the real rat, yet, did not: the orange curve has risen above other curves. Because there is a factor of forgetfulness in the model, even the confidence value of the appropriate strategy can decrease when the real animal has not performed it for long time. For example, this can been seen on figure 3.2.7 during session *0303* where the orange curve decreases.

In order to evaluate the performance of the model, we display the rat's behavior in a slightly different way than in the previous section: instead of representing the consecutive arm choices (Right/Left) made by animal at each trial, all possible strategies that the rat could be following at a given moment are displayed. If, on a given trial, the rat went to the right whereas the light was on the left, we assume that the rat could have been following two different strategies at this trial: the *Right* strategy or the *Dark* strategy. The process that translates the rat's behavior into successive possible strategies is the one described in STEP A in the behavioral analysis of the previous section.

Figure 3.2.17 shows such a representation for the behavioral data of rat #18. On the figure, the density of strategy blocks correspond to the consistency with which the animal has indeed performed each strategy. Interestingly, the figure shows clearly the density of the *left* strategy block remains high after the task rule had changed from *left* to *dark*. This means that the rat had persisted in performing the *left* strategy after the task rule change. Moreover, figure 3.2.17 displays moments when there is no clear dominant strategy in the rat behavior: for example between trials #90 and #120.

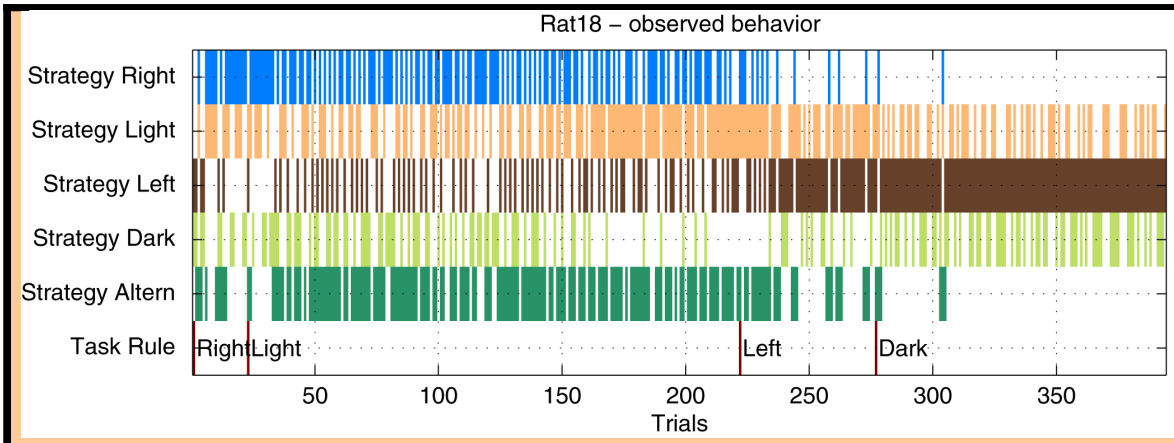


Figure 3.2.17: *Behavior of rat #18 observed during all sessions performed.* x-axis displays successive trials while y-axis labels the different « strategies » followed by the rat.

The matrix representation of the rat behavior can then be compared to the maximal confidences of the basic model at each trial. Figure 3.2.18 displays the latter simulated on the data recorded in rat #18. Although this plot and the rat behavior presented in the previous figure coincide only for 60% of trials, the two matrices look strikingly similar: Both matrices show denser strategy blocks around corresponding task rule changes compared to other periods. Furthermore, figure 3.2.18 show the same absence of clear dominant strategy between trials #90 and #120 as in figure 3.2.17.
Moreover, the basic model got a reward in 58% of the trials, which is very close to the average percentage of rewards got by real rats (58.3%). The similarities in the strategy-matrices and in the percentage of reward obtained demonstrate the quality of performance of the model in mimicking the rat behavior.

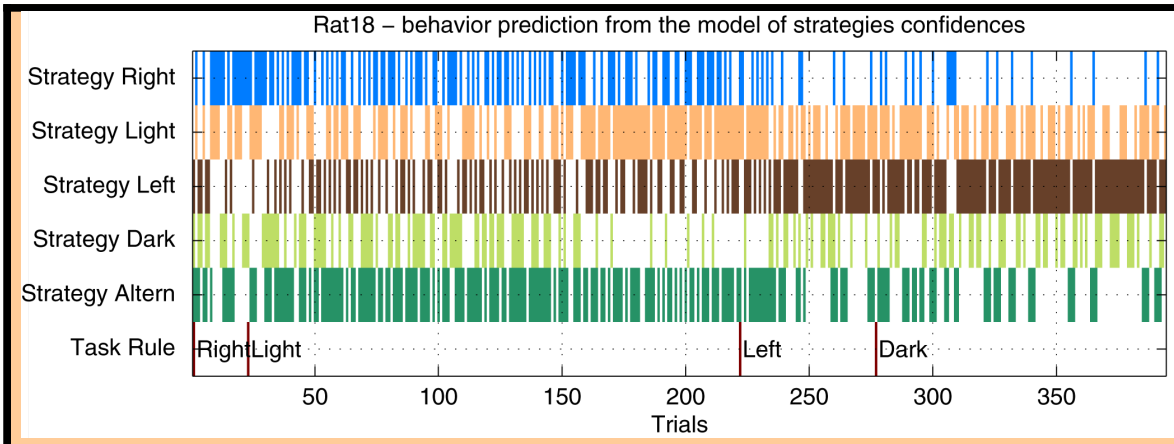


Figure 3.2.18: *behavior predicted by the basic model is 60% similar to the rat observed behavior.*

In order to compare these performances with a different case, figure 3.2.19 displays the maximal confidences in the “optimal” model. The latter matrix is about 50% similar to the real rats behavior. The figure shows well how the “optimal” model quickly learned the different consecutive tasks. The percentage of trials where the “optimal” model got a reward is about 98%, which is far different from the average of 58% of trials where real rats got a reward.

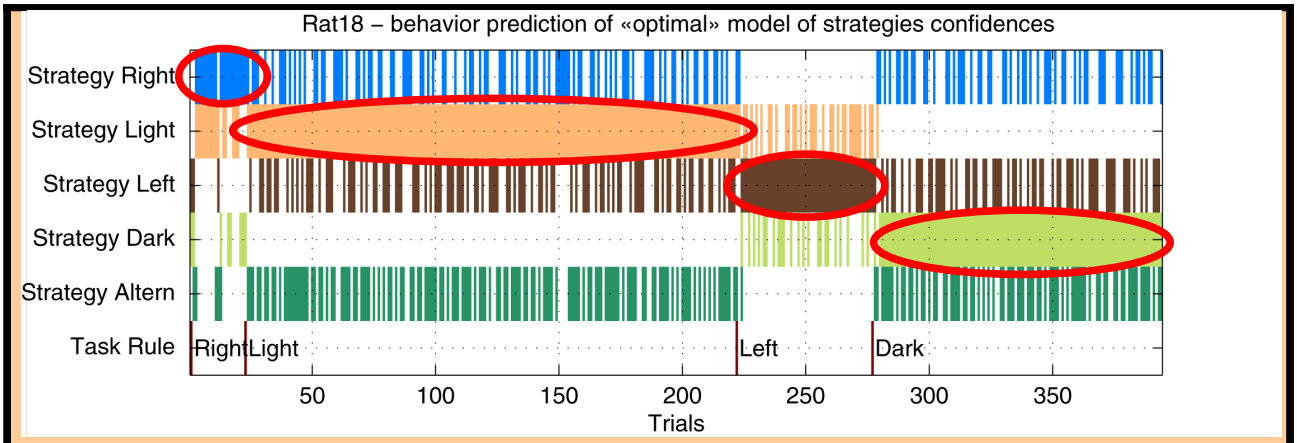


Figure 3.2.19: *By tuning parameters in the model, we obtain a model less similar to the rat actual behavior, but showing a more "optimal" behavior in the task resolution.*

In contrast, when tuning parameters in the model so that it learns to maximize reward, we obtain a model which displays a nearly optimal performance in that it quickly learns the strategy corresponding to the current task rule, and gets reward in a percentage of about 98% of the trials. Even if such « optimal » model is far from mimicking the animals' behavior, it will be interesting for us to see whether parameters of this « optimal » model are more correlated to prefrontal activity compared to the model which is closer to the rats behavior. So we will briefly describe the results concerning the « optimal » model.

Figure 3.2.19 displays the result of the simulation of this model. Clearly, only a few trials after each task rule shift, the « optimal » model can select the appropriate strategy. This results in strategy blocks with only little discontinuities.

Table 3.4 summarizes the percentage of resemblance and the percentage of reward obtained by each model with each rat's data. The "optimal" model got an average of 50.50% resemblance with the rats' choices while being very good at solving the task. In a certain manner, this "confirms" the observation that rats are not optimal at solving the previously described task.

MODEL		Real Rats	« Optimal » model		Basic model	
	nbTrials	**reward**	**similarity**	**reward**	**similarity**	**reward**
RAT20	1817	56,58	50,15	98,57	59,26	53,33
RAT15	570	53,51	49,05	98,42	55,16	57,89
RAT18	394	62,18	48,53	97,97	61,62	57,36
RAT19	252	66,27	53,65	97,62	60,63	63,1
RAT12	289	66,09	51,14	98,62	58,41	58,48
AVERAGE	**3322**	**58,29**	**50,50**	**98,24**	**59,02**	**58,03**

Table 3.4: *Behavioral results of the models. That is, for each version of the model, the table shows the percentage of total trials where the model predicted the same behavior than what the animal actually did (displayed as "SIMILARITY"), and the percentage of total trials where the model would have got a reward if it had been performing the task (i.e. the percentage of trials were the model's prediction was an appropriate behavior for the task resolution, displayed as "REWARD").*

In order to study if one of the two versions of the model could also describe part of the evolution of information coding in the mPFC during the task, we statistically tested prefrontal cell correlations with the model. The method employed is the same as the one described for analyses of cells showing a transition in activity correlated with a shift in behavioral strategies: a cell was considered

as significantly correlated with the model if its activity was correlated with the confidence of one the five strategies in the model during one of the four trial periods. A cell was considered as correlated with the confidence in a strategy if its activity was significantly higher or lower during trials were the strategy had a higher value than other strategies in the model, compared to trials were it has not. We used the Wilcoxon-Mann Whitney test with a Bonferroni correction, $p < .002$.
Table 3.5 summarizes the results of this statistical test. We found that a proportion of 2.01% cells showed an activity which was correlated with the strategy with optimal confidence in the basic model. Besides, 4.86% cells were correlated with the optimal model. These percentages are lower than the 11.58% cells reported in the previous section as correlated with strategies followed by the rats. Further analyses are required to see whether there were some cells correlated with the model which did not correlate with the strategies followed by real rats, nor with any other parameter of the task.

	nb cells	1894	83	379	556	205	671
		TOTAL	**Rat12**	**Rat15**	**Rat18**	**Rat19**	**Rat20**
Basic model (4*5 tests < 0.002)		**2,01**	**2,41**	**1,32**	**1,08**	**0,49**	**3,58**
Optimal model (4*5 tests < 0.002)		**4,86**	**0**	**2,64**	**3,06**	**0,98**	**9,39**

Table 3.5: *Electrophysiological results of the models. That is, the % of cells correlated to the state of each version of the model. More precisely, a cell is considered to be correlated with the state of a model if it's activity, considered in a given trial period, is significantly different at trials where the model is in the state "Strategy A" compared with trials where the model is in a different state (e.g. state "Strategy B" or state "No strategy").*

DISCUSSION

The results of our model are yet preliminary. They show that rats' behavior observed in our Y-maze experiment can be approximated in simulation. This seems to argue in favor of the hypothesis that rats behavior could be, at least partially, modeled using bayesian inferences. In line with this idea, evidence is accumulating suggesting that the primate cortical network can implement Bayesian inference (Deneve et al., 2001; Deneve and Pouget, 2004; Doya et al., 2007; Samejima and Doya; 2007).
Moreover, it could be worth pursuing in improving the model and testing its correlation with prefrontal activity. Principally because several theories of the prefrontal cortex-basal ganglia system suggest that the prefrontal cortex could learn to perform a given task and propose appropriate actions to perform, whereas the basal ganglia would bias these decisions through a more stochastic action selection process (Gurney et al., 2001b). In contrast, other theories suggest that the basal ganglia adapt its activity in relation to a given task faster than the prefrontal cortex, and then would gate the latter in order to provide it with appropriate decisions (Frank et al., 2001; Pasupathy and Miller, 2005). So it remains an open question whether the prefrontal cortex and the basal ganglia is a quicker encoder of « optimal » task parameters. So it will be interesting to continue to investigate this issue by desiging « optimal » models in opposition to models similar to the animal behavior, and to study whether a difference between prefrontal and striatal activity could be in the model they are the most correlated to.

4. Other collaborative work in the frame of this project

Other analyses of our data are now been processed by other members of the research team. Some of the main issues addressed by their analyses concerns the interaction between the prefrontal cortex and the hippocampus during the task, and the processes of memory consolidation during sleep following the task.

Several abstracts presenting some results of these analyses are given in the appendix of this document. These abstracts will be presented at the SfN meeting this year.

CHAPTER 4 : GENERAL DISCUSSION

This thesis presented our contribution to the understanding of the roles of the rat striatum and medial prefrontal cortex (mPFC) in action selection, decision making and other cognitive processes for navigation strategies learning and shifting. For this purpose, experiments were designed in which:

- rats had to *learn different reward-seeking tasks* and to *encode various sensorimotor associations* to achieve them – *i.e.* to *perform different strategies* for navigating towards goals.
  * In the *plus-maze* task, rats had to learn to localize different amounts of rewards in different arms of the maze, and to recall that only lit arms provided reward. Reward distributions were changed within each session and were explicitly trained in learning trials (which were recorded along with recall trials).
  * In the *Y-maze* task, always starting from the same departure arm, rats had to learn different rules employing cues of different stimulus dimensions along consecutive blocks: reward was located on the right (or left) arm; reward was located at the lit (or dark) arm.
- rats had to *detect changes* in the task rule imposed without any explicit signal. This required recall of the previously learned strategy best for the new situation, or, if none is appropriate, to proceed with a new *learning process*.
  * In the *Y-maze* task, each time rats passed a performance criterion, the task rule was changed. Rats had to detect such changes based on a lower incidence of reward obtained, and to learn the new task rule.

Based on these experimental designs, our objectives were :

- to better understand the complementarity of the mPFC and the striatal activity in relation to these behavioral processes;
- to evaluate this activity in terms of a better understanding of the prefronto-striatal system's involvement in navigation;
- to apply this toward developping biomimetic models for action selection in navigation systems for robotics;
- to further the dialog between experimental and computational neurosciences.

The following sections recall our results and discuss each of these points while 4.0 summarizes our contribution.

1. Principal novel observations and interpretations

1.1 Within the framework of the *plus-maze* task:

- by analysing electrophysiological data recorded in the Ventral Striatum (VS), we found neurons with **activity anticipating rewards** – discharging phasically prior to each element in a successive series of rewards, both when the rat approached and was immobile at the goal. These neurons were found in a network including the medial shell and core of nucleus accumbens as well as, the ventromedial caudate, zones receiving PFC and/or hippocampal inputs;
- by reproducing the different patterns of this reward anticipation activity observed in different VS neurons with an Actor-Critic model, we showed that this activity was consistent with the hypothesis that **part of VS can play the role of a Critic** – i.e. a driver of reward-based learning of Stimulus-Response (S-R) associations within the striatum.
- we interpreted the VS reward anticipatory neurons recorded as **not precisely encoding temporal information concerning stimuli** – which would permit a precise timing of

dopaminergic neurons' responses to the absence of a predicted reward. We rather propose that **different groups of striatal neurons constitute different Actor-Critic modules with access to different cue, contextual and temporal information**. We predict that dopaminergic neurons recorded in tasks such as the *plus-maze* task should reflect this modularity, some of them by marking a pause in activity in relation to the animal's erroneous expectancy of an extra reward after the last one.

- in a robotics simulation, after restricting the plus-maze task to a "purely" S-R task – suppressing the requirement for localizing different amounts of reward – we designed an Actor-Critic model where VS (together with dorsal striatal striosomes) is the Critic which drives learning, whereas matrisomes in the Dorsolateral Striatum (DLS) constitute the Actor which memorizes S-R associations as habits. With this model, the kind of reward anticipation activity mentioned above enables the animat to **learn a behavioral sequence that lead to the reward**, yet using different Actor-Critic submodules which are specialized in different parts of the task.
- by using a machine learning method (namely *growing self-organizing maps*) to automatically categorize the animat's visual perceptions in the plus-maze task, and by combining this method with a multi-module Actor-Critic system, we have shown that this striatum-inspired architecture for S-R learning can have interesting **generalization abilities potentially applicable for the field of navigation in autonomous robots**.

These results strengthen the hypothesis that part of VS could contribute to the acquisition of procedural navigation strategies (here a cue-guided strategy) and could be segregated into learning modules which apply to different parts of the environment.

Since VS reward anticipatory activity could be reproduced with a Temporal Difference learning model – which is a *model-free* reinforcement learning algorithm (Sutton and Barto, 1998) found to be suitable for describing S-R strategies (Daw et al., 2005) –, the involvement of VS could be dedicated to learning *model-free* cue-guided strategies (cue-guided and possibly others) – i.e., without building a world model, as defined in chapter 1.

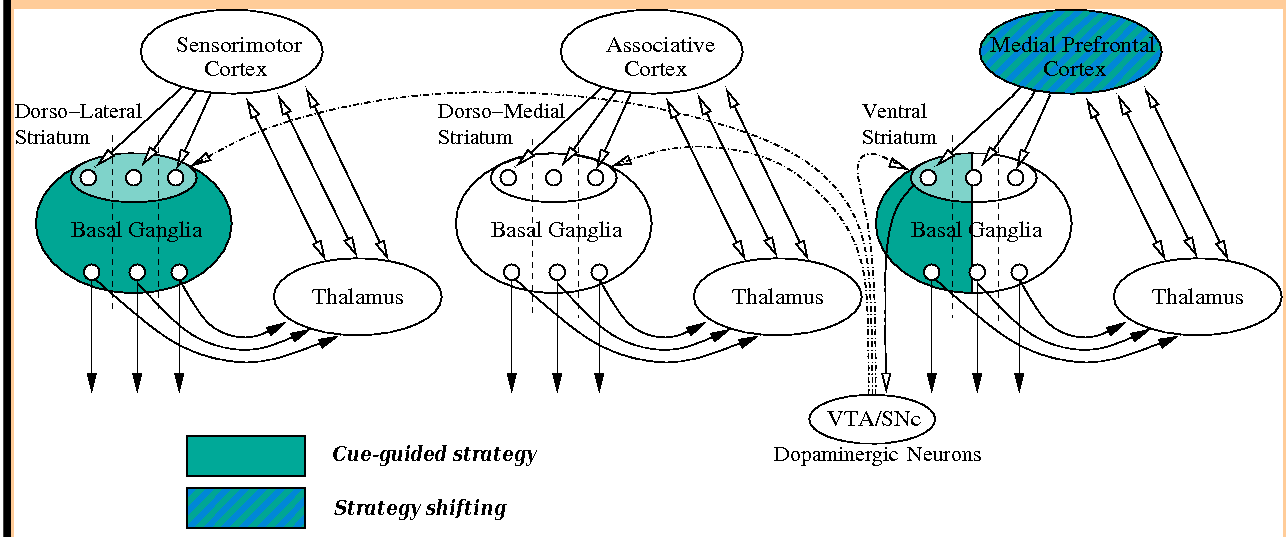


Figure 4.0 : *Summarized contributions.* Our main contributions consisted in 1) electrophysiologically confirming the role of the ventral striatum in cue-guided strategy learning and 2) the role of medial prefrontal cortex in strategy shifting; 3) modelling the roles of the dorsolateral and ventral striatum in cue-guided strategy learning.

1.2 Within the framework of the *Y-maze* task:

- from neuronal ensemble recordings of the medial prefrontal cortex (mPFC) in rats learning and shifting navigation strategies, we found:
  * a set of neurons with **transitions in activity in response to extradimensional changes in the task rule**, thus likely to participate in the **detection of task rule changes**;
  * another set of neurons with activity transitions associated with the **current strategy spontaneously performed by the animal**, which could participate in **strategy selection**.
  These activities were recorded during periods where behavioral automatization was avoided by changing the task rule as soon as rats had learned a given rule. Hence, the neural activities reported could be associated with ***goal-directed behavior***, rats having to continuously relearn action-outcome contingencies.
- by modelling, at the behavioral level, the way rats learned to perform appropriate strategies and to shift them, we found that part of the ***animal's behavior could be described by a Markovian model,*** and we proposed a preliminary contribution to the understanding of which formal rules govern strategy shifting processes in the rat;
- Moreover, we found a small subset of ***prefrontal neurons whose activity was correlated with the state of the latter model***, hence suggesting that prefrontal neural activity could participate in such strategy selection.

These results strengthen the hypothesis that the mPFC could contribute to attentional processes required to detect extradimensional task rule changes (here, between cue-guided and spatial orientation strategies) and could also contribute to flexible strategy selections following such task changes.

In our Y-maze experiments, since we deliberately tried to prevent habituation of newly learned strategies and rapidly imposed new rules (in 27 cases), after these rule changes, the number of trials required for the animals to abandon the previously learned strategy was most often very low. Except in the case of one task rule change after which the animal persisted in performing the previous strategy for 268 trials (and thus could be interpreted as *habitual*), in the other cases, it always took less than 30 trials, with an average of 10 trials necessary to abandon the previous strategy. As a consequence, it appears reasonable to think that the acquired rule-following behavior of the animals recorded in our Y-maze task was globally *goal-directed* (i.e. *model-based*). Thus prefrontal cellular activity reflecting behavioral strategy shifting could indeed reflect extradimensional shifts between ***model-based*** cue-guided and spatial orienting strategies rather than *model-free* ones. This is consistent with previous proposals that the mPFC is involved in goal-directed decisions and flexible behavior planning (see Granon and Poucet, 2000; Cardinal et al., 2002; Killcross and Coutureau, 2003; Uylings et al., 2003; Balleine et al., 2007 for reviews).

2. Implications for the prefronto-striatal system

These results are consistent with the functional architecture presented in the first chapter (figure 4.1). More precisely, our results demonstrate and validate neural mechanisms within this architecture.

2.1 Ventral Striatum

These results do not exclude the possibility that VS could also participate in learning of strategies other than procedural navigation, as suggested by the report that VS lesions impair both egocentric and allocentric strategies (DeLeonibus et al., 2005). However, the correspondence between the VS neurons that we recorded and a cue-guided Critic model (with neither spatial nor temporal correlates) only permits us to support the contention that VS could participate in learning of cue-guided strategies.

These results do not exclude the possibility that VS could also participate in model-based strategies, or more generally in goal-directed behaviors. Indeed, other neurons previously recorded in the plus-maze task show *goal-approach* correlates which could participate in such a function (Mulder et al., 2004). Similarly to reward anticipatory neurons reported in this thesis, goal-approach cells were found to be distributed over the medial shell, the ventromedial caudate and the core. As a consequence, this does not permit us to find a sharp and clear functional distinction between core and shell neurons, which is in accord with the suggestion of a finer subdivision of VS suggested by some authors (Heimer et al., 1997; Ikemoto, 2002).
Theses results do not exclude the possibility that other parts of the striatum (or more generally, other brain areas) also participate in the expression of the Critic function. Indeed, as described in the first chapter, some authors have previously postulated that striosomes within the dorsal striatum could participate in this (Houk et al., 1995; see Joel et al., 2002 for a review). Moreover, the closely associated orbitofrontal cortex is another candidate for participation in the Critic function, since reward anticipatory activity has also been found in the rat orbitofrontal cortex (Schoenbaum et al., 2003; Roesch et al., 2006; van Duuren et al., 2007).
In order to better understand the precise complementary roles of different striatal territories in navigation, it would be of interest to record simultaneously in DLS, DMS, shell and core during learning and shifting different navigation strategies, as is programmed by our team in the immediate future.

2.2 Medial Prefrontal Cortex

Concerning the medial prefrontal cortex, while our results support the notion that mPFC could participate in extradimensional (ED) shifts, this does not exclude the possibility that mPFC also subserve intradimensional (ID) shifts. However, previous studies have reported that mPFC lesions do not impair ID shifts, but only alter ED shifts (Birrell and Brown, 2000; Ragozzino et al., 2003). Complementarily, orbitofrontal cortex (OFC) lesions are found to impair ID shifts (McAlonan and Brown, 2003; Kim and Ragozzino, 2005). Interestingly, the firing of OFC neurons is also thought to represent the current behavioral strategy, particularly when the outcome predicted by a stimulus is altered or reversed (Schoenbaum et al., 2000). Taken together with our finding of both task change and behavioral strategy responsive neurons in mPFC, this suggests that, both in OFC and mPFC, task shifting and behavioral strategy elaboration are two functions that seem to be tightly related.
Moreover, our results do not exclude that parts of the ventral striatum receiving PFC inputs could also participate in strategy shifting, which is indeed supported by a previous report of task selective neurons from our team (Shibata et al., 2001).
Finally, our results do not exclude that the mPFC could also participate in learning of goal-directed behaviors (e.g. model-based strategies). Indeed, correlates of spatial goals (Hok et al., 2005), and of action-outcome contingencies (Mulder et al., 2003; Kargo et al., 2007) in mPFC neurons support this role. However, the post-training expression of goal-directed behaviors relying on such components does not appear to depend on the mPFC (Ostlund and Balleine, 2005). Thus, there remains further work to be done to help understand the way these functions integrate within the mPFC. Further analyses of our data on the mPFC are underway: we are studying the evolution of neural activity with learning, and comparing the coherence between local field potentials in the mPFC and in the hippocampus at different task phases. Moreover, we are analysing how the communication between the prefrontal cortex and the hippocampus during sleep recorded after daily sessions can reflect performance-dependent memory consolidation.

Conclusion: implications for neuromimetic models of navigation to be used in the EC ICEA integrated project

In this thesis, we pursued a pluridisciplinary work which required both an analytic approach (behavioral and neurophysiological studies) and a synthetic one (computational modeling and simulated robotics). Our contribution to the European Comunnity funded ICEA (Integrating Cognition, Emotion and Autonomy) project came within the context of this integrative framework.
One of the goals of this project is to design a bioinspired integrated architecture combining several navigation strategies for improving the decision autonomy of artificial systems evolving in unknown environments. Our work have some implications for such an architecture inspired by the prefronto-striatal system.

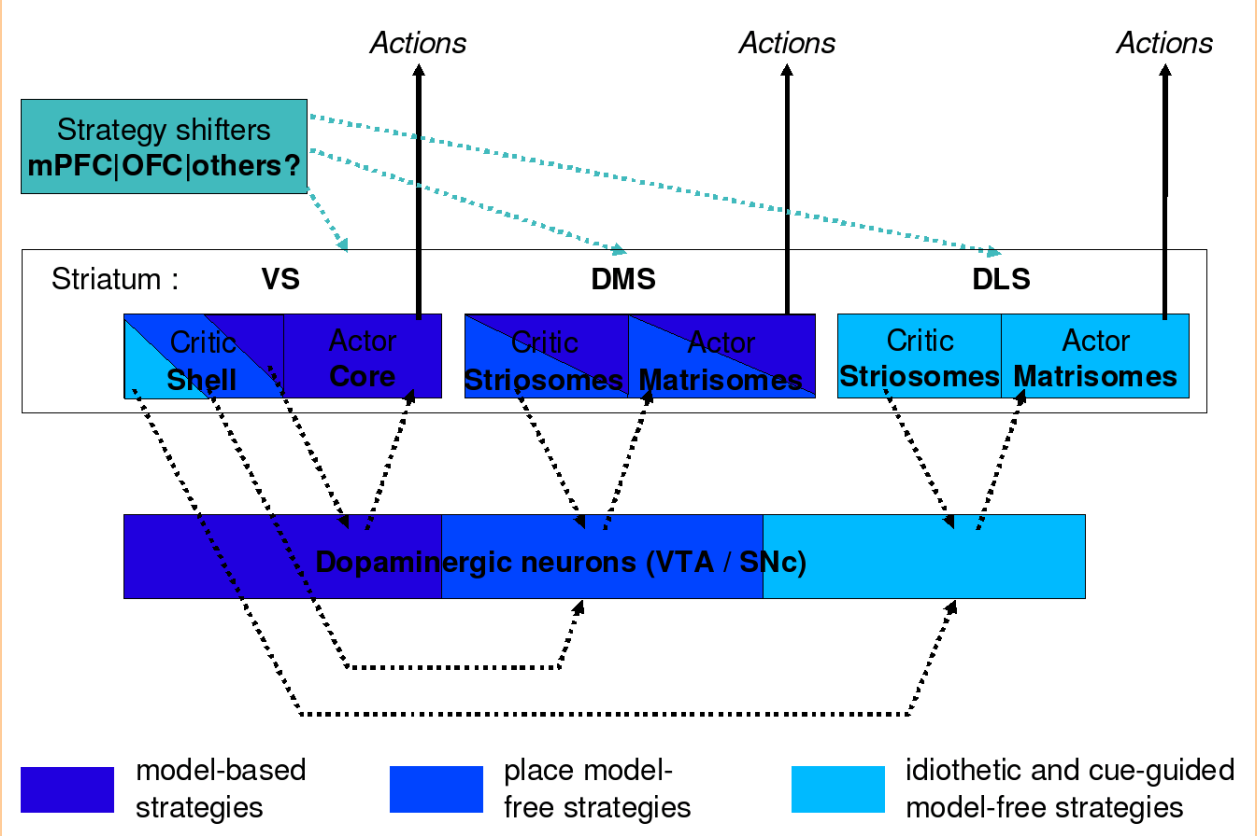


Figure 4.1 : *Schematic hypothesized functional architecture of part of the prefronto-striatal system.* Different territories of the striatum (receiving their respective cortical inputs) would subserve different types of navigation strategies. Striosomes within each striatal territory are assumed to play the role of a *Critic* driving reinforcement learning, whereas matrisomes play the *Actor* controlling action selection. Within this architecture dopaminergic neurons are assumed to encode reinforcement learning signals based on reward prediction errors. The dorsomedial striatum, not studied here, could either be involved in model-free strategies, in model-based ones, or in both. Our results are consistent with the identification of the shell as a Critic for model-free strategies. Moreover, we showed that the mPFC could participate in the detection of task changes and could subserve strategy shifting following such changes. VS, Ventral Striatum; DMS, Dorsomedial Striatum; DLS, Dorsolateral Striatum; mPFC, medial prefrontal cortex; VTA, Ventral Tegmental Area; SNc, Substantia Nigra pars compacta.

First, parallel learning processes could be simultaneously engaged in motor, associative and limbic

cortico-striatal loops for the model-based and model-free performance of praxic, cue-guided and place navigation strategies, depending on available allothetic or idiothetic data.
We have proposed a computational model of cue-guided model-free strategy based on a reinforcement learning Actor-Critic framework, involving both motor and limbic loops.
Whereas some Actor-Critic models involve VS as a Critic, others do not (see Joel et al., 2002 for a review). Our results showing activity anticipating reward suggest that VS could indeed contribute to such a function. For the control architecture employed in ICEA, it would be worthwhile to integrate such a reinforcement learning function of the ventral striatum within the existing model of the basal ganglia (Humphries et al., 2006). However, our robotics model of reinforcement learning inspired by the striatum still has some performance limitations and would require improvement before being able to deal with complex robotics tasks. In this purpose, it would be worthwhile to take inspiration from reinforcement learning theoretical work coordinating different learning modules: such as hierarchical reinforcement learning (Morimoto and Doya, 1998, 2001; Barto and Mahadevan, 2003; Elfwing et al., 2003; Barto et al., 2004; Haruno and Kawato, 2006), and macro-actions (McGovern et al., 1998; Precup and Sutton, 1998; DePisapia and Goddard, 2003). Moreover, it would be valuable to take into account work done to apply reinforcement learning to the case where an agent has to deal with several motivations (Kaplan and Oudeyer, 2003; Oudeyer et al., 2005; Konidaris and Barto, 2006).
Then, the ICEA architecture will have also to be provided with learning mechanisms for model-based strategies. This is indeed an ongoing collaboration between our team and other members of the ICEA consortium, on the one hand, bymodeling of more biologically plausible models building hippocampal and prefrontal representations and, on the other hand, by precisely modeling limbic loops and their interaction with the others – taking inspiration from our hypothesized architecture of figure 4.1 (http://www.iceaproject.eu/).
Whereas our results on the prefrontal cortex provide a first clue to how to implement biologically plausible extradimensional (ED) strategy shifting mechanisms for the ICEA project – by means of a markovian decision process –, there remains to study how the rodent brain performs intradimensional (ID) strategy shifts, extending the current approach to studies of the orbitofrontal cortex.
Finally, in all existing bioinspired computational models, spontaneous alternations between different strategies that we observed in our experiments are never taken into account. However there could be an adaptive mechanism playing an important role in such a decision process. Investigating which dynamics in the navigation architecture could produce such behavioral variabilities may be an interesting perspective for enhancing the adaptability of artificial systems facing unknown environments.

APPENDIX

1. Other articles

Zugaro et al. (2004) Head-Direction cells

Zugaro, Arleo, Déjean, Burguière, Khamassi and Wiener (2004). Rat anterodorsal thalamic head direction neurons depend upon dynamic visual signals to select anchoring landmark cues. *European Journal of Neuroscience*, 20:530-6.

http://animatlab.lip6.fr/papers/zugaro04_parallax.pdf

Filliat et al. (2004) The Psikharpax Project

Filliat, Girard, Guillot, Khamassi, Lachèze and Meyer (2004). State of the artificial rat Psikharpax. In Schaal, S., Ijspeert, A., Billard, A., Vijayakumar, S., Hallam, J., and Meyer, J.-A., editors, *From Animals to Animats 8: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, pages 3-12, Cambridge, MA. MIT Press.

http://uei.ensta.fr/filliat/papers/SAB2004_psikharpax.pdf

Khamassi et al. (2004) TD-learning models

Khamassi, Girard, Berthoz and Guillot (2004). Comparing three critic models of reinforcement learning in the basal ganglia connected to a detailed actor in a S-R task. In Groen, F., Amato, N., Bonarini, A., Yoshida, E., and Kröse, B., editors, *Proceedings of the Eighth International Conference on Intelligent Autonomous Systems*, pages 430-437, Amsterdam, The Netherlands. IOS Press.

http://animatlab.lip6.fr/papers/IAS8_MK.pdf

Meyer et al. (2005) The Psikharpax Project

Meyer, Guillot, Girard, Khamassi, Pirim and Berthoz (2005). The Psikharpax project: Towards building an artificial rat. *Robotics and Autonomous Systems, 50(4):211-223*.

http://animatlab.lip6.fr/papers/Meyer_RAS.pdf

Battaglia et al. (In press) The Hippocampo-prefonto-cortico-striatal system

Battaglia and Peyrache and Khamassi and Wiener (In press) Spatial decisions and neuronal activity in hippocampal projection zones in prefrontal cortex and striatum. In Mizumori S (Ed.) *Hippocampal Place Fields: Relevance to Learning and Memory.*

2. Other abstracts

Arleo et al. (2004) Head-Direction cells

Arleo, Déjean, Boucheny, Khamassi, Zugaro and Wiener (2004). Optic field flow signals update the activity of head direction cells in the rat anterodorsal thalamus. *Society for Neuroscience Abstracts, San Diego, USA*.

Dollé et al. (2006) Model of strategy shifting

Dollé, Khamassi, Guillot and Chavarriaga (2006). Coordination of learning modules for competing navigation strategies into different mazes. Poster presented at the workshop « Multisensory Integration and Concurrent Parallel Memory Systems for Spatial Cognition, Nineth International Conference on the Simularion of Adaptive Behavior (SAB06).

http://www.idiap.ch/~rchava/sab06wk/posters/DolleKhGuCh06.pdf

SfN Abstract 2007
Control/Tracking Number: 2007-A-43388-SfN
Activity: Scientific Abstract
Current Date/Time: 5/15/2007 12:04:21 PM

Increased firing rate and theta modulation in medial prefrontal neurons during episodes of high coherence in the theta band of hippocampal/prefrontal local field potentials (LFP) in behaving rats.

***K. BENCHENANE**[1], A. PEYRACHE[1], M. KHAMASSI[1,2], P. TIERNEY[3], V. DOUCHAMPS[1], F. P. BATTAGLIA[4], S. I. WIENER[1];
[1]LPPA, Col. de France - CNRS, Paris, France; [2]Inst. des Systemes Intelligeants et Robotiques, Univ. Pierre et Marie Curie, Paris VI, Paris, France; [3]Inserm u667, Col. de France, Paris, France; [4]SILS-APCN, Univ. van Amsterdam, Amsterdam, The Netherlands

Abstract: The functional linkage of hippocampus and prefrontal cortex (PFC) has been shown as, for instance, phase locking of PFC neurons to the hippocampal theta rhythm. Moreover, coherence in the theta band between hippocampal and prefrontal local field potentials (LFPs) was found to be increased in correct trials in a spatial working memory task. In these studies, rats were over-trained and performed a simple spatial task. As PFC is strongly implicated in learning and behavioral flexibility, we designed a task to elucidate the interaction between hippocampus and PFC in learning and cross-modal strategy shifts.
Hippocampal LFP and medial prefrontal neurons and LFP were recorded and analyzed in three rats learning four successive reward contingencies on a Y maze: go to the arm on the right, go to the lit arm, go left, then go to the dark arm (one maze arm was randomly selected to be lit on each trial). Rats returned to the start arm after each trial.
A robust theta rhythm at 6-8 Hz was observed in the PFC LFP. During learning, high hippocampal-PFC coherence (values>0.7) in the theta band (5-10 Hz) was observed. This occurred principally at the decision point in the maze, suggesting heightened communication between hippocampus and PFC at the moment of behavioural choice.
Over all sessions, 776 PFC neurons were recorded. Statistical significance of theta modulation was analysed by the Rayleigh test. The strength of modulation of PFC neurons by theta (modRatio) was defined as the ratio between the magnitude of the sine-wave fitting the spikes' phase histograms and its baseline. According to this method, 273 PFC neurons (35%) were found to be significantly modulated by hippocampal theta (Rayleigh test, $p<0.05$), with a mean modRatio of 7.7+/-0.079 percent.
Interestingly, the PFC neuronal firing rates were increased by 71% on average during periods of high hippocampal-PFC theta coherence (t-test, $p<0.0001$). Moreover, the magnitude of the modRatio increased by 32% (9.8+/-0.099 percent, t-test $p<0.001$) during these periods. These data show that coherence in the theta band between hippocampus and PFC is related to the rat's behavior. Furthermore the hippocampal-prefrontal coherence selectively activates a subpopulation of cells in the PFC. This provides evidence of selective control of hippocampal-PFC functional binding at the level of LFP rhythms and at the level of single cell activity. This could be important in the timely transmission of signals from hippocampus to PFC for learning and consolidation.

SfN Abstract 2007
Control/Tracking Number: 2007-A-42327-SfN
Activity: Scientific Abstract
Current Date/Time: 5/15/2007 10:28:00 AM

Rat medial prefrontal cortex neurons are modulated by both hippocampal theta rhythm and sharp wave-ripple events.

A. PEYRACHE[1], K. BENCHENANE[1], M. KHAMASSI[1,2], V. DOUCHAMPS[1], P. L. TIERNEY[3], F. P. BATTAGLIA[4], *S. I. WIENER[5];
[1]LPPA, CNRS-College de France, Paris, France; [2]Inst. des Systemes Intelligents et Robotiques, Univ. Pierre et Marie Curie, Paris VI, Paris, France; [3]Inserm u667, Col. de France, Paris, France; [4]SILS-APCN, Univ. van Amsterdam, Amsterdam, The Netherlands; [5]LPPA, CNRS-College De France LPPA, 75005 Paris, France

Abstract: Mnemonic functions are attributed to hippocampus and to one of its principal projection zones, the prefrontal cortex. In order to elucidate the functional neural processing in this pathway medial prefrontal cortex neurons and local field potentials (LFP) and hippocampal LFP were recorded simultaneously in five freely moving rats during 98 recording sessions in a Y maze and in previous and subsequent sessions of quiet repose. In 35% of the 2230 cells analysed, action potentials were significantly phase modulated by hippocampal theta during task performance (Rayleigh test, $p<0.05$) as shown previously. (Modulation by theta was defined as the ratio between the magnitude of the sine-wave fitting the phase histograms and its baseline.) Furthermore, in 21%, firing rates increased (11%) or decreased (10%) during hippocampal ripples occuring during previous and subsequent resting periods (t-test, $p<0.05$). (Modulation by sharp waves is taken as the logarithm of the ratio between mean firing rate of a cell in a window surrounding (+/- 25 ms) ripples' peak and the mean firing rate of the cell in a window lasting from 1 s to 50 ms before ripples' peak.) In 10% of the cells there was significant modulation by both theta and ripples, and the amplitude of these respective modulations was significantly correlated (Pearson's correlation test, $p<0.05$). This correlation may correspond to the strength of hippocampal afferences to the respective neurons and their local circuits, suggesting that the hippocampal/prefrontal interaction is mediated by the same population of prefrontal cells both during sleep and active behavior.

SfN Abstract 2007
Control/Tracking Number: 2007-A-110087-SfN
Activity: Scientific Abstract
Current Date/Time: 5/15/2007 10:03:51 AM

Time course of reactivation of memory-related cell ensembles in the rat medial prefrontal cortex during sleep.

***F. P. BATTAGLIA**[1], A. PEYRACHE[2], K. BENCHENANE[2], M. KHAMASSI[2,3], V. DOUCHAMPS[2], P. L. TIERNEY[4], S. I. WIENER[2];
[1]SILS-APCN, Univ. van Amsterdam, Amsterdam, The Netherlands; [2]LPPA, CNRS Col. de France, Paris, France; [3]Inst. des Systemes Intelligents et Robotiques, Univ. Pierre et Marie Curie, Paris VI, Paris, France; [4]Inserm u667, Col. de France, Paris, France

Abstract: The prefrontal cortex is implicated in the flexible learning of stimulus-outcome associations, which are consolidated in memory during offline periods. Reactivation of memory traces, in the form of the reinstatement of experience-related activity in prefrontal cell assemblies during sleep could be the basis for such a consolidation process. To study this, we developed a novel analysis which allows to follow the time course of task-related reactivation in simultaneously recorded cell ensembles. The correlation matrix of binned spike trains from multiple cells is decomposed in its principal components, the largest of which represents groups of cells whose activity was highly correlated during the reference recording period. The instantaneous cell pair co-activation matrix during sleep, weighted by the coefficients in a given principal component, and averaged over all cell pairs, can then be taken as a measure of the reactivation of the cell assembly corresponding to that principal component at a given time.
We analyzed medial prefrontal ensembles from five rats while learning a set-switching task on a Y-maze, and in rest sessions preceding and following the task. In 62 out of 86 sessions, cell assembly reactivation was significantly greater ($p<0.05$) during slow wave sleep (SWS) after the session than in SWS before. There was a significant correlation (Pearson's correlation test, $p<0.05$) between the eigenvalues associated with the principal components during task performance (indicating the strength of the encoding) and the increased re-activation in post-task SWS (compared to pre-task SWS). Moreover, in 67 out of 86 sessions, co-activation was correlated ($p<0.05$) with the power of both delta and spindle cortical oscillations, and it was much weaker during rest periods that were classified as non-sleep. The increased co-activation in the post-experience sleep was attributable to discrete bouts of activation, typically 2-5 seconds in duration. This new technique permits to precisely follow the time course of neural ensemble re-activation. These data demonstrate that theta, ripple-sharp waves and spindles are important for prefrontal post-task SWS reactivation, a possible neural ensemble basis for memory consolidation.

3. Supplemental material of the VS-reward article

(Annex I)

The Temporal-Difference (TD) learning algorithm was developed in the field of optimal control theory and provides an efficient method for an embedded agent (animat, robot or other artifact) to learn to assemble a sequence of actions enabling it to optimize reinforcement (e,g., reward) in a given environment (Sutton and Barto. 1998). This approach addressed the problem that rewards may arrive considerably later than the initial neural activity, too late to modify the appropriate synapses (the 'credit assignment problem'). TD learning has since been successfully used to describe reinforcement learning mechanisms in basal ganglia networks, but mainly for single rewards. It has been implemented in simulations where dopaminergic neurons compute reinforcement signals (Schultz et al. 1997), while striatal neurons compute reward anticipation (Suri and Schultz 2001). A given task is represented as a discretized series of timesteps. At each timestep, the agent occupies a particular position (or *state*) in the environment, perceives a set of signals (e.g., internal signals about motivation, or visual information about the environment), and selects an action. When the agent reaches a reward location and selects an appropriate action, it receives a reward and strengthens the neural connections leading to this state.

Instead of requiring memorization of a lengthy sequence of actions to eventually be reinforced when a reward is achieved – which is costly in terms of numbers of computations and memory requirements – the TD algorithm proposes an efficient and elegant method to reinforce appropriate state and signal prompted actions towards a reward. The reinforcement signal is computed on the basis of the difference between the value of the states at two consecutive timesteps (hence the name 'temporal-difference learning'). The value of a given state *S* is considered to be the value of reward which is expected (or predicted) to be received in the future, starting from this state, and is noted V*(S)*. If the action A_{t-1} is performed in state S_{t-1}, and then at time *t,* the expected reward value V in state S_t is higher than that of S_{t-1} – (that is $V_t(S_t) > V_{t-1}(S_{t-1})$) –, then action A_{t-1} is reinforced, and the value of state S_{t-1} is increased. The effective reinforcement signal that drives this learning process is given by the following equation:

$$\hat{r}_t = r_t + \gamma . V_t(S_t) - V_{t-1}(S_{t-1}) \qquad (1)$$

where r_t is the reward achieved at time *t*, and γ is a *discount factor* ($0 < \gamma < 1$) which limits the capacity to take into account rewards in the far future. At each time step *t*, this reinforcement signal is used to update the probability of choosing action *A* in state *S*, and to update the amount of reward that state *S* "predicts" according to the following equations :

$$P(A|S) \mathrel{+}= \hat{r}_t \text{ and } V(S) \mathrel{+}= \hat{r}_t \qquad (2) \text{ and } (3)$$

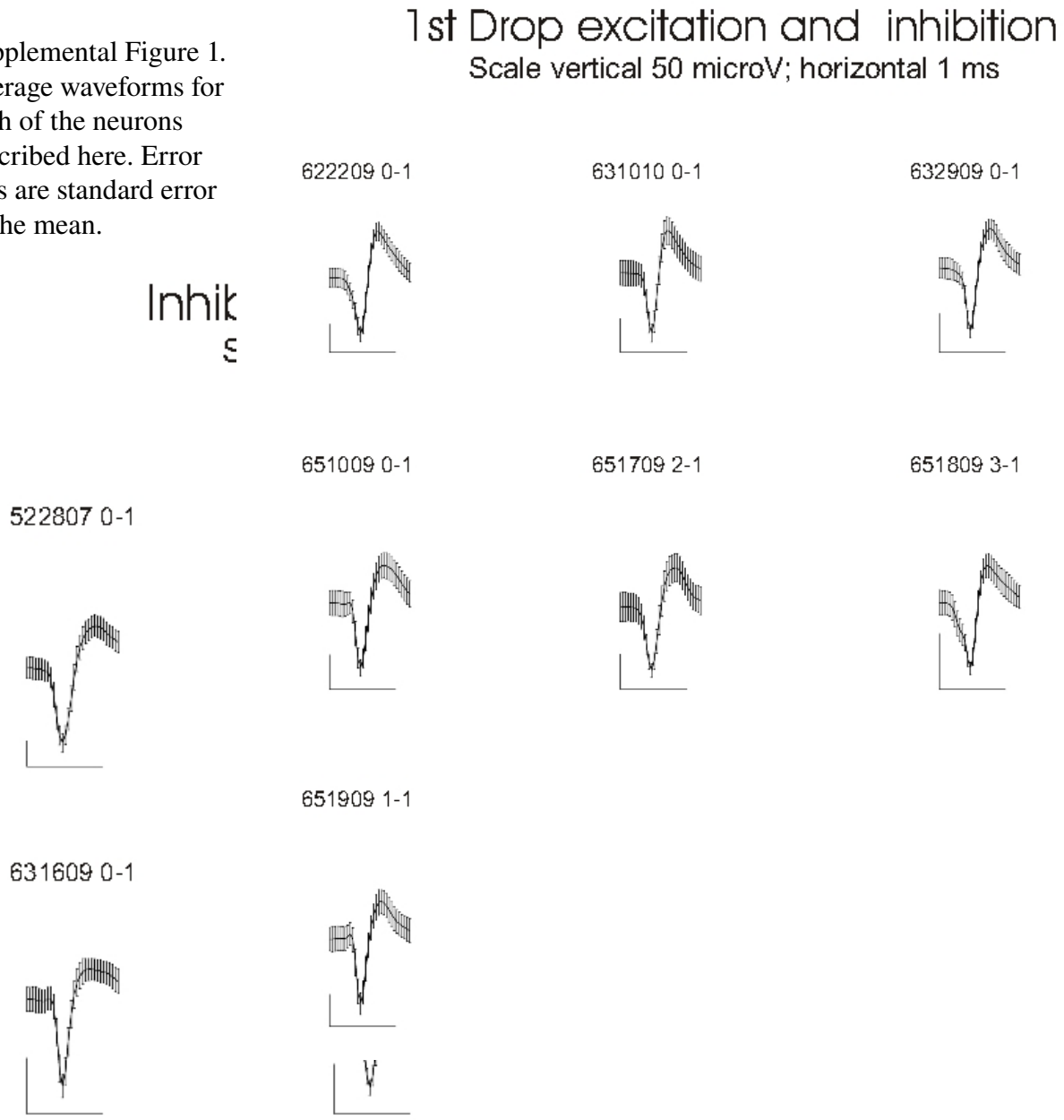
where += means "is incremented by".

It remains to be verified whether an algorithm of this type is actually implemented in the vertebrate brain. Nevertheless it provides an initial intelligible framework to understand a possible way to learn a sequence of actions towards a reward. Its simplicity and efficiency support its compatibility with the constraints of natural selection.

(END OF ANNEX I)

Supplemental figures

Supplemental Figure 1. Average waveforms for each of the neurons described here. Error bars are standard error of the mean.



General increased activity during drinking

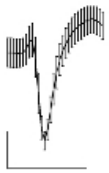
Scale vertical 50 microV; horizontal 1 ms

522307 0-2

522807 2-1

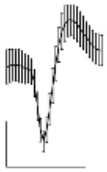
540108 0-1

550508 03



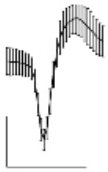
560708 1-1

620110a 0-2



632609 4-1

632909 3-2



633009 2-3

First Drop excitatory

Scale vertical 50 microV; horizontal 1 ms

522107 2-1

562307 2-1

612009 1-1

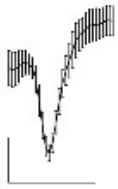
620809 1-1

620909 1-2

620909 2-1

621109 1-1

622609 0-1



632909 4-2

651509 0-1

651609 1-1

652109 0-1

Peridrop excitation all drops

Scale vertical 50 microV; horizontal 1 ms

522107 0-2

523107 0-1

542307 0-1

621609 2-2

631509 1-1

631609 1-1

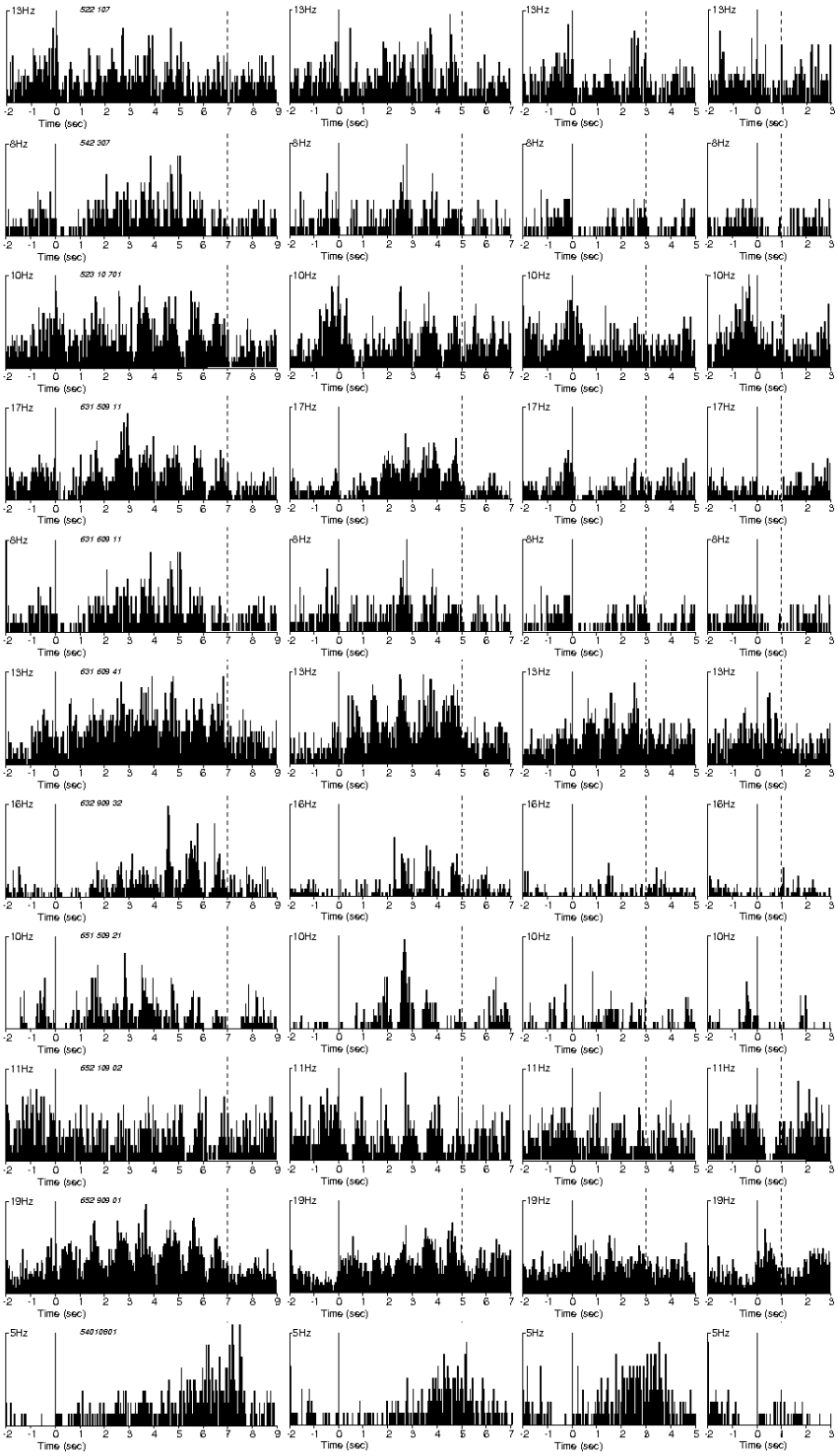
651509 2-1

651709 0-1

652109 0-2

652409 0-1

652909 0-1



Supplemental Figure 2. Further examples of neurons with reward anticipatory activity. Each row shows the activity histogram for a single neuron. Time zero corresponds to delivery of the first drop of water (also indicated by a *continuous vertical bar*). The *dashed vertical bar* corresponds to 1 s after delivery of the final drop of water, thus the respective columns correspond to 7, 5, 3 and 1 drops of water. Cell identifiers are inset in the first column.

BIBLIOGRAPHY

Aberman JE and Salamone JD (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience*, 92(2):545-52.

Adams S, Kesner RP and Ragozzino ME (2001). Role of the medial and lateral caudate-putamen in mediating an auditory conditional response association. *Neurobiol Learn Mem*, 76(1):106-16.

Aizman O, Brismar H, Uhlen P, Zettergren E, Levey AI, Forssberg H, Greengard P and Aperia A (2000). Anatomical and Physiological Evidence for D1 and D2 Dopamine Receptors Colocalization in Neostriatal Neurons. *Nature Neuroscience*, 3(3):226-30.

Albertin SV, Mulder AB, Tabuchi E, Zugaro MB and Wiener SI (2000). Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behav Brain Res,* 117: 173-83.

Albin RL, Young AB and Penney JB (1989). The functional anatomy of basal ganglia disorders. *Trends in Neuroscience*, 12:366-75.

Aldridge JW and Berridge KC (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *J Neurosci*, 18(7):2777-87.

Alexander GE and Crutcher MD (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci*, 13(7):266-71.

Alexander GE, Crutcher MD and DeLong MR (1990). Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Prog Brain Res*, 85:119-46.

Amemori K and Sawaguchi T (2006). Rule-dependent shifting of sensorimotor representation in the primate prefrontal cortex. *Eur J Neurosci*, 23(7):1895-909.

Amiez C, Joseph JP and Procyk E (2005). Anterior cingulate error-related activity is modulated by predicted reward. *Eur J Neurosci*, 21(12):3447-52.

Aosaki T, Graybiel AM and Kimura M (1994). Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science*, 265(5170):412-5.

Aosaki T, Tsubokawa H, Ishida A, Watanabe K, Graybiel AM and Kimura M (1994). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *J Neurosci* 14: 3969-84.

Aosaki T, Kimura M and Graybiel AM (1995). Temporal and spatial characteristics of tonically active neurons of the primate's striatum. *J Neurophys* 73: 1234-52.

Apicella P, Ljungberg T, Scarnati E and Schultz W (1991a). Responses to reward in monkey dorsal and ventral striatum. *Exp Brain Res*, 85(3):491-500.

Apicella P, Scarnati E and Schultz W (1991b). Tonically discharging neurons of monkey striatum respond to preparatory and rewarding stimuli. *Exp Brain Res* 84: 672-5.

Apicella P, Legallet E and Trouche E (1996). Responses of tonically discharging neurons in monkey striatum to visual stimuli presented under passive conditions and during task performance. *Neurosci Lett* 203: 147-50.

Apicella P, Ravel S, Sardo P and Legallet E (1998). Influence of predictive information on responses of tonically active neurons in the monkey striatum. *J Neurophysiol*, 80(6):3341-4.

Apicella P, Scarnati E, Ljungberg T and Schultz W (1992). Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *J Neurophysiol*, 68(3):945-60.

Apicella P, Scarnati E and Schultz W (1991). Tonically discharging neurons of monkey striatum respond to preparatory and rewarding stimuli. *Exp Brain Res*, 84(3):672-5.

Arbib MA and Dominey PF (1995). Modeling the roles of basal ganglia in timing and sequencing saccadic eye movements. In: *Models of information processing in the basal ganglia,* edited by Houk JC, Davis JL, Beiser D, pp 149-62, Cambridge, MA: MIT Press.

Arleo A (2000). Spatial Learning and Navigation in Neuro-Mimetic Systems, Modeling the Rat Hippocampus. Department of Computer Science, Swiss Federal Institute of Technology Lausanne, EPFL, Switzerland.

Arleo A (2005). Neural bases of spatial cognition and information processing in the brain. Habilitation à Diriger des Recherches. Université Pierre et Marie Curie, Paris 6, France.

Arleo A, Déjean C, Boucheny C, Khamassi M, Zugaro MB and Wiener SI (2004). Optic field flow signals update the activity of head direction cells in the rat anterodorsal thalamus. *Society for Neuroscience Abstracts, San Diego, USA*.

Arleo A and Gerstner W (2000). Spatial Cognition and Neuro-Mimetic Navigation: A Model of the Rat Hippocampal Place Cell Activity. *Biological Cybernetics*, Special Issue on Navigation in Biological and Artificial Systems, 83:287-99.

Arleo A and Rondi-Reig L (In press). Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. In F. Dolins and R. Mitchell (eds.), Spatial Perception and Spatial Cognition, chapter 11, Cambridge University Press.

Arleo A and Smeraldi F and Gerstner W (2004). Cognitive navigation based on nonuniform Gabor space sampling, unsupervised growing networks, and reinforcement learning. *IEEE Trans Neural Netw*, 15(3):639-52.

Atkeson CG and Santamaria JC (1997). A comparison of direct and model-based reinforcement learning. In International Conference on Robotics and Automation, pages 3557-64.

Averbeck BB and Lee D (2007). Prefrontal neural correlates of memory for sequences. *J Neurosci*, 27(9):2204-11.

Baddeley A (1996). The fractionation of working memory. *Proc Natl Acad Sci U S A*, 93:13468-72.

Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT and Jung MW (2003). Dynamics of population code for working memory in the prefrontal cortex. *Neuron*, 40(1):177-88.

Baeg EH, Kim YB, Kim J, Ghim JW, Kim JJ and Jung MW (2007). Learning-induced enduring changes in functional connectivity among prefrontal cortical neurons. *J Neurosci*, 27(4):909-18.

Baldassarre G (2002). A Modular Neural-Network Model of the Basal Ganglia's Role in Learning and Selecting Motor Behaviors. *Journal of Cognitive Systems Research*, 3(1):5-13.

Baldassarre G (2002b). A biologically plausible model of human planning based on neural networks and Dyna-PI models. In M. Butz, O. Sigaud and P. Gérard, editors, Proceedings of the Workshop on Adaptive Behaviour in Anticipatory Learning Systems, pages 40-60.

Baldassarre G and Parisi D (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In Meyer *et al.* (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior,* Supplement Volume (pp.131-39). The MIT Press, Cambridge, MA.

Baldassarre G (2003). Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In: *Adaptive behavior in anticipatory learning system*s, edited by Butz M, Sigaud O, Gerard P, pp. 179-200. Berlin: Springer Verlag.

Balkenius C (1994). Biological learning and artificial intelligence. Technical report. Lund University Cognitive Studies – LUCS 30. ISSN 1101–8453.

Balleine BW (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol Behav*, 86(5):717-30.

Balleine BW and Dickinson A (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37:407-19.

Balleine BW, Doya K, O'Doherty J and Sakagami M (Eds.) (2007). Reward and decision making in corticobasal ganglia networks. *Annals of the New York Academy of Sciences*, Vol. 1104.

Balleine B and Killcross S (1994). Effects of ibotenic acid lesions of the nucleus accumbens on instrumental action. *Behav Brain Res*, 65(2):181-93.

Banquet JP, Gaussier P, Quoy M, Revel A and Burnod Y (2005). A hierarchy of associations in hippocampo-cortical systems: cognitive maps and navigation strategies. *Neural Comput*, 17(6):1339-84.

Bar-Gad I, Havazelet-Heimer G, Goldberg JA, Ruppin E and Bergman H (2000). Reinforcement-driven dimensionality reduction--a model for information processing in the basal ganglia. *J Basic Clin Physiol Pharmacol*, 11(4):305-20.

Barnes TD, Kubota Y, Hu D, Jin DZ and Graybiel AM (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* 437: 1158-61.

Barto AG (1995). Adaptive critics and the basal ganglia. In: *Models of Information Processing in the Basal Ganglia*, edited by Houk JC, Davis JL, Beiser DG, pp 215-32. Cambridge, MA: MIT.

Barto AG and Mahadevan S (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, 13:341-79.

Barto AG, Singh S and Chentanez N (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the Third International Conference on Development and Learning (ICDL 2004)*.

Battaglia FP, Peyrache A, Benchenane K, Khamassi M, Douchamps V, Tierney PL and Wiener SI (2007). Time course of reactivation of memory-related cell ensembles in the rat medial prefrontal cortex during sleep. *Soc Neurosci Abstr.* 2007-A-110087-SfN.

Battaglia FP, Peyrache A, Khamassi M and Wiener SI (In press). Spatial decisions and neuronal activity in hippocampal projection zones in prefrontal cortex and striatum.. In S. Mizumori, editor, Hippocampal Place Fields: Relevance to Learning and Memory.

Battaglia FP, Sutherland GR and McNaughton BL (2004a). Hippocampal sharp wave bursts coincide with neocortical "up-state" transitions. *Learn Mem*, 11(6):697-704.

Battaglia FP, Sutherland GR and McNaughton BL (2004b). Local sensory cues and place cell directionality: additional evidence of prospective coding in the hippocampus. *J Neurosci*, 24(19):4541-50.

Bear MF, Cooper LN and Ebner FF (1987). A physiological basis for a theory of synapse modification. *Science*, 237(4810):42-8.

Bechara A, Damasio H and Damasio AR (2000). Emotion, decision making and the orbitofrontal cortex. *Cereb Cortex*, 10(3):295-307.

Bellman RE (1957). A Markov decision process. *Journal of Mathematical Mech.*, 6:679-84.

Benchenane K, Peyrache A, Khamassi M, Tierney P, Douchamps V, Battaglia FP and Wiener SI (2007). Increased firing rate and theta modulation in medial prefrontal neurons during episodes of high coherence in the theta band of hippocampal/prefrontal local field potentials (LFP) in behaving rats. *Soc Neurosci Abstr.* 2007-A-43388-SfN.

Berg EA (1948). A simple objective test for measuring flexibility in thinking. *J Gen Psychol*, 39:15-22.

Berns GS and Sejnowski TJ (1996). The Neurobiology of Decision Making, « How The Basal Ganglia Make Decision », pages 101-13. Springer-Verlag.

Berns GS and Sejnowski TJ (1998). A computational model of how the basal ganglia produce sequences. *J Cogn Neurosci*, 10(1):108-21.

Berridge KC and Robinson TE (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res Brain Res Rev*, 28(3):309-69.

Berthoz A (2003a). La Décision. Edition Odile Jacob, Paris, France.

Berthoz A (2003b). Stratégies cognitives et Mémoire spatiale. In: Philosophies de la perception, Phénoménologie, grammaire et sciences cognitives. Eds: J. Bouveresse et J-J. Rosat. Collège de France, Paris, Odile Jacob, pp101-9.

Berthoz A, Viaud Delmon I and Lambrey S (2003b). Spatial memory during navigation: What is being stored, maps or movements? In: Galaburda AM, Kosslyn SM, Christen Y (eds) The languages of the brain. Harvard University Press, Cambridge Mass; pp 288-306.

Bidaud Ph (2000). Micro-robotique - La science au présent, Encyclopaedia Universalis.

Biegler R, Morris RGM (1993). Blocking in the spatial domain with arrays of discrete landmarks. *J Exp Psychol Anim Behav Process*, 25:334–51.

Birrell JM and Brown VJ (2000). Medial Frontal Cortex Mediates Perceptual Attentional Set Shifting in the Rat. *The Journal of Neuroscience*, 20(11):4320–4.

Blair HT and Sharp PE (1995). Anticipatory head direction signals in anterior thalamus: Evidence for a thalamocortical circuit that integrates angular head motion to compute head direction. *Journal of Neuroscience*, 15(9):6260-70.

Block AE, Dhanji H, Thompson-Tardif SF and Floresco SB (2007). Thalamic-Prefrontal Cortical-Ventral Striatal Circuitry Mediates Dissociable Components of Strategy Set Shifting. *Cerebral Cortex*, 17:1625-36.

Blodgett HC (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology, 4*(8), 113-34.

Botreau F, El Massioui N, Cheruel F and Gisquet-Verrier P (2004). Effects of medial prefrontal cortex and dorsal striatum lesions on retrieval processes in rats. *Neuroscience*, 129(3):539-53.

Bouret S and Sara SJ (2004). Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *Eur J Neurosci*, 20(3):791-802.

Bowman EM, Aigner TG and Richmond BJ (1996). Neural signals in the monkey ventral striatum related to motivation for juice and cocaine rewards. J Neurophysiol, 75(3):1061-73.

Brasted PJ, Humby T, Dunnett SB and Robbins TW (1997).Unilateral lesions of the dorsal striatum in rats disrupt responding in egocentric space. *J Neurosci*, 17(22):8919-26.

Brito GNO and Brito LSO (1990). Septohippocampal system and the prelimbic sector of frontal cortex: A neuropsychological battery analysis in the rat. *Behav Brain Res*, 36:127-46.

Brodmann K (1895). Ergebnisse über die vergleichende histologi. *K Anat Anz Suppl*, 41:157-216.

Brooks, R. A (1991a). New approaches to robotics, *Science*, 253:1227–32.

Brooks (1999). Cambrian Intelligence: The Early History of the New AI MIT Press (A Bradford Book).

Brown RM, Crane AM and Goldman PS (1979). Regional distribution of monoamines in the cerebral cortex and subcortical structures of the rhesus monkey: concentrations and in vivo synthesis rates. *Brain Res*, 168:133-50.

Brown L and Sharp F (1995). Metabolic Mapping of Rat Striatum: Somatotopic Organization of Sensorimotor Activity. *Brain Research*, 686:207-22.

Brown J, Bullock D and Grossberg S (1999). How the Basal Ganglia Use Parallel Excitatory and Inhibitory Learning, or Incentive Salience ? *Brain Research Reviews*, 28:309-69.

Brown JW, Bullock D and Grossberg S (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, **17**, 471-510.

Bunney BS, Chiodo LA and Grace AA (1991). Midbrain Dopamine System Electrophysiological Functioning: A Review and New Hypothesis. *Synapse*, 9:79-84.

Burgess N, Recce M and O'Keefe J (1994). A Model of Hippocampal Function. *Neural Networks*, 7(6/7):1065-81.

Burgess N, Jeffery KJ and O'Keefe J (1999). Integrating Hippocampal and Parietal Functions: a Spatial Point of View. In Burgess, N. *et al.* (Eds), *The Hippocampal and Parietal Foundations of Spatial Cognition*, pp. 3-29, Oxford University Press, UK.

Burguière E (2006). Role du cervelet dans la navigation: Etude du mécanisme cellulaire de dépression synaptique à long terme des fibres parallèles. Thèse de l'Université Pierre et Marie Curie, Paris 6, France.

Burns LH, Annett L, Kelley AE, Everitt BJ and Robbins TW (1996). Effects of lesions to amygdala, ventral subiculum, medial prefrontal cortex, and nucleus accumbens on the reaction to novelty: implication for limbic-striatal interactions. *Behav Neurosci*, 110(1):60-73.

Butz MV, Sigaud O and Gérard P (Eds.) (2003). Anticipatory behavior in adaptive learning systems: Foundations, theories, and systems. Springer.

Buzsaki G, Geisler C, Henze DA and Wang XJ (2004). Interneuron Diversity series: Circuit complexity and axon wiring economy of cortical interneurons. *Trends Neurosci*, 27(4):186-93.

Callaway CW and Henriksen SJ (1992). Neuronal firing in the nucleus accumbens is associated with the level of cortical arousal. *Neuroscience*, 51(3):547-53.

Calton JL, Stackman RW, Goodridge JP, Archey WB, Dudchenko PA and Taube JS (2003). Hippocampal place cell instability after lesions of the head direction cell network. *J Neurosci*, 23(30):9719-31.

Canal CE, Stutz SJ and Gold PE (2005). Glucose injections into the dorsal hippocampus or dorsolateral striatum of rats prior to T-maze training: Modulation of learning rates and strategy selection. *Learning & Memory*, 12:367-74.

Cardinal RN, Pennicott DR, Sugathapala CL, Robbins TW and Everitt BJ (2001). Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science*, 292(5526):2499-501.

Cardinal RN, Parkinson JA, Hall J and Everitt BJ (2002). Emotion and Motivation: The Role of the Amygdala, Ventral Striatum and Prefrontal Cortex. *Neuroscience Biobehavioral Reviews*, 26(3):321-52.

Carelli RM and Deadwyler SA (1997). Cellular mechanisms underlying reinforcement-related processing in the nucleus accumbens: electrophysiological studies in behaving animals. *Pharmacol Biochem Behav*, 57(3):495-504.

Carelli RM (2002). Nucleus accumbens cell firing during goal-directed behaviors for cocaine vs. 'natural' reinforcement. *Physiol Behav*, 76(3):379-87.

Carpenter AF, Georgopoulos AP and Pellizzer G (1999). Motor cortical encoding of serial order in a context-recall task. *Science*, 283(5408):1752-7.

Celeux G and Govaert G (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, Special issue on optimization techniques in statistics,* 14 (3):315-32.

Centonze D, Picconi B, Gubellini P, Bernardi G and Calabresi P (2001). Dopaminergic control of synaptic plasticity in the dorsal striatum. *Eur J Neurosci*, 13(6):1071-7.

Chang JY, Zhang L, Janak PH and Woodward DJ (1997). Neuronal responses in prefrontal cortex and nucleus accumbens during heroin self-administration in freely moving rats. *Brain Res*, 754(1-2):12-20.

Chang JY, Chen L, Luo F, Shi LH and Woodward DJ (2002). Neuronal responses in the frontal cortico-basal ganglia system during delayed matching-to-sample task: ensemble recording in freely moving rats. *Exp Brain Res*, 142(1):67-80.

Chang Q and Gold PE (2003). Switching memory systems during learning: changes in patterns of brain acetylcholine release in the hippocampus and striatum in rats. *J Neurosci*, 23(7):3001-5.

Chavarriaga R (2005). Spatial learning and navigation in the rat : a biomimetic model. PhD thesis, EPFL, Lausanne, Switzerland.

Chavarriaga R, Strösslin T, Sheynikhovich D and Gerstner W (2005a). A computational model of parallel navigation systems in rodents. *Neuroinformatics* 3: 223-42.

Chavarriaga R, Strösslin T, Sheynikhovich D and Gerstner W (2005b). Competition between cue response and place response: A model of rat navigation behaviour. *Connection Science*, 17(1-2):167-83.

Cheng J and Feenstra MG (2006). Individual differences in dopamine efflux in nucleus accumbens shell and core during instrumental learning. *Learn Mem*, 13(2):168-77.

Chevalier G and Deniau JM (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci*, 13(7):277-80.

Churchland P and Sejnowski TJ (1995). The Computational Brain.

Coizet V, Comoli E, Westby GWM and Redgrave P (2003). Phasic activation of substantia nigra and the ventral tegmental area by chemical stimulation of the superior colliculus: an electrophysiological investigation in the rat. *Eur J Neurosci*, 17:28-40.

Colacicco G, Welzl H, Lipp H and Würbel H (2002). Attentional set-shifting in mice: modification of a rat paradigm, and evidence for strain-dependent variation. *Behav. Brain Res.* 132, pp. 95–102.

Collett TS, Cartwright BA and Smith BA (1986). Landmark learning and visuo-spatial memories in gerbils. *Journal of Comparative Physiology A, 158,* 835–51.

Colombo PJ, Davis HP and Volpe BT (1989). Allocentric spatial and tactile memory impairments in rats with dorsal caudate lesions are affected by preoperative behavioral training. *Behav Neurosci*, 103(6):1242-50.

Colwill RM and Rescorla RA (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *J Exp Psychol Anim Behav Process*, 11:120 –32.

Comoli E, Coizet V, Boyes J, Bolam JP, Canteras NS, Quirk, RH, Overton PG and Redgrave P (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events. *Nat Neurosci*, 6(9):974-80.

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

Cook D and Kesner RP (1988). Caudate nucleus and memory for egocentric localization. *Behav Neural Biol*, 49(3):332-43.

Corbit LH, Muir JL and Balleine BW (2001). The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *J Neurosci*, 21(9):3251-60.

Corbit LH and Balleine BW (2003). The role of prelimbic cortex in instrumental conditioning. *Behav Brain Res*, 146(1-2):145-57.

Cornuéjols A, Miclet L and Kodratoff Y (2002). ApprentissageArtificiel, France, « Apprentissage de réflexes par renforcement », pages 483-510.

Coulom R (2002). Reinforcement learning using neural networks, with applications to motor control. PhD thesis, Institut National Polytechnique de Grenoble.

Coutureau E and Killcross S (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behav Brain Res*, 146(1-2):167-74.

Cressant A, Muller and Poucet B (1997). Failure of centrally placed objects to control the firing fields of hippocampal place cells. *J Neurosci*, 17(7):2531-42.

Cromwell HC and Schultz W (2003). Effects of expectations for different reward magnitudes on neuronal activity in

primate striatum. *J Neurophys* 89: 2823-38.

Dalley JW, Theobald DE, Bouger P, Chudasama Y, Cardinal RN and Robbins TW (2004). Cortical cholinergic function and deficits in visual attentional performance in rats following 192 IgG-saporin-induced lesions of the medial prefrontal cortex. *Cereb Cortex*, 14(8):922-32.

Daw ND (2003). *Reinforcement Learning Models of the Dopamine System and Their Behavioral Implications*. PhD Thesis, Carnegie Mellon University, Pittsburgh, PA.

Daw ND, Touretzky DS and Skaggs WE (2002). Representation of reward type and action choice in ventral and dorsal striatum in the rat. *Soc Neurosci Abstr* 28: 765.11.

Daw ND, Niv Y and Dayan P (2005a). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8: 1704-11.

Daw ND, Niv Y and Dayan P (2006). Recent breakthroughs in basal ganglia research, New-York, USA, « Actions, policies, values, and the basal ganglia ». Nova Science Publishers Inc., New-York, USA.

Daw ND and Doya K (2006). The computational neurobiology of learning and reward. *Curr Opin Neurobiol*, 16(2):199-204.

Dayan P (1999). Unsupervised learning. In Wilson, RA and Keil, F, editors. *The MIT Encyclopedia of the Cognitive Sciences.*

Dayan P (2001). Motivated reinforcement learning. In: *Advances in Neural Information Processing Systems,* edited by Dietterich TG, Becker S, Ghahramani Z, pp 11-8. Cambridge, MA: MIT Press.

Dayan P and Abbott LF (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems. MIT Press.

Dayan P and Balleine BW (2002). Reward, motivation, and reinforcement learning. *Neuron*, 36(2):285-98.

Dayan P and Sejnowski TJ (1994). TD(lambda) Converges with Probability 1. *Machine Learning*, 14(1):295-301.

Dayan P and Yu AJ (2001). ACh, Uncertainty, and Cortical Inference. In T.G. Dietterich, S. Becker and Z. Ghahramani, editors, Proceedings of the 14th International Conference on Advances in Neural Information Processing Systems. Cambridge, MA. MIT Press.

Dayawansa S, Kobayashi T, Hori E, Umemo K, Tazumi T, Ono T and Nishijo H (2006). Conjunctive effects of reward and behavioral episodes on hippocampal place-differential neurons of rats on a mobile treadmill. *Hippocampus*, 16(7):586-95.

de Brabander JM, de Bruin JPC and Van Eden CG (1991). Comparison of the effects of neonatal and adult medial prefrontal cortex lesions on food hoarding and spatial delayed alternation. *Behav Brain Res*, 42:67-75.

de Bruin, JPC, Sanchez-Santed F, Heinsbroek RPW, Donker A and Postmen P (1994). A behavioural analysis of rats with damage to the medial prefrontal cortex using Morris water-maze: Evidence for behavioural flexibility, but not for impaired spatial navigation. *Brain Research*, 652:323-33.

DeCoteau WE and Kesner RP (2000). A double dissociation between the rat hippocampus and medial caudoputamen in processing two forms of knowledge. *Behav Neurosci*, 114(6):1096-108.

Degris T (2007). Apprentissage par renforcement dans les processus de décision Markoviens factorisés. PhD thesis, Université Pierre et Marie Curie, Paris 6, France.

Degris T, Sigaud O, Wiener SI and Arleo A (2004). Rapid response of head direction cells to reorienting visual cues: A computational model. *Neurocomputing*, 58-60(C):675-82.

Degris T, Sigaud O and Wuillemin P-H (2006). Learning the Structure of Factored Markov Decision Processes in Reinforcement Learning Problems. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 257-64, Pittsburgh, Pennsylvania.

Dehaene S and Changeux JP (1997). A hierarchical neuronal network for planning behavior. *Proc Natl Acad Sci U S A*, 94(24):13293-8.

Dehaene S and Changeux JP (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Prog Brain Res*, 126:217-29.

Delatour B and Gisquet-Verrier P (1996). Prelimbic cortex specific lesions disrupt delayed-variable response tasks in the rat. *Behav Neurosci*, 110(6):1282-98.

Delatour B and Gisquet-Verrier P (1999). Lesions of the prelimbic-infralimbic cortices in rats do not disrupt response selection processes but induce delay-dependent deficits: evidence for a role in working memory? *Behav Neurosci*, 113(5):941-55.

Delatour B and Gisquet-Verrier P (2000). Functional role of rat prelimbic-infralimbic cortices in spatial memory: evidence for their involvement in attention and behavioural flexibility. Behav Brain Res, 109(1):113-28.

De Leonibus E, Oliverio A and Mele A (2005). A study on the role of the dorsal striatum and the nucleus accumbens in allocentric and egocentric spatial memory consolidation. *Learn Mem*, 12(5):491-503.

Deneve S, Latham PE and Pouget A (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826-31.

Deneve S and Pouget A (2004). Bayesian multisensory integration and cross-modal spatial links, *Journal of Neurophysiology (Paris)*, 98 (1-3), 249-58.

Deniau JM, Mailly P, Maurice N and Charpier S (2007). The pars reticulata of the substantia nigra: a window to basal ganglia output. *Prog Brain Res*, 160:151-72.

Deniau JM, Menetrey A and Thierry AM (1994). Indirect nucleus accumbens input to the prefrontal cortex via the substantia nigra pars reticulata: a combined anatomical and electrophysiological study in the rat. *Neuroscience*, 61(3):533-45.

De Pisapia N and Goddard NH (2003). A neural model of frontostriatal interactions for behavioral planning and action chunking. *Neurocomputing*, 52-54:489-95.

Desban M, Kemel ML, Glowinski J and Gauchy C (1993). Spatial organization of patch and matrix compartments in the rat striatum. *Neuroscience*, 57(3):661-71.

Desimone R and Duncan J (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, 18:193-222.

D'Esposito M, Detre JA, Alsop DC, Shin RK, Atlas S and Grossman M (1995). The neural basis of the central executive system of working memory. *Nature*, 378(6554):279-81.

Devan BD and White NM (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. *J Neurosci*, 19(7):2789-98.

Dias R and Aggleton JP (2000). Effects of selective excitotoxic prefrontal lesions on acquisition of nonmatching- and matching-to-place in the T-maze in the rat: differential involvement of the prelimbic-infralimbic and anterior cingulate cortices in providing behavioural flexibility. *Eur J Neurosci*, 12(12):4457-66.

Di Chiara G (2002). Nucleus accumbens shell and core dopamine: differential role in behavior and addiction. *Behav Brain Res*, 137(1-2):75-114.

Dickinson A (1980). Contemporary animal learning theory. Cambridge: Cambridge University Press.

Dickinson A and Balleine B (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1), 1–18.

Dickinson A and Balleine B (2002). The role of learning in motivation. In Stevens(*Handbook of Experimental Psychology, Vol. 3: Learning, Motivation and Emotion*. 3rd edn. (ed. Gallistel, C.R.) pp. 497-533, Wiley, New York.

d' Hooge R and de Deyn Peter P (2001). Applications of the Morris water maze in the study of learning and memory.

Brain Research reviews, 36, 60-90.

Dollé L, Khamassi M, Guillot A and Chavarriaga R (2006). Coordination of learning modules for competing navigation strategies into different mazes. Poster presented at the *Workshop Multisensory Integration and Concurrent Parallel Memory Systems for Spatial Cognition, Nineth International Conference on the Simulation of Adaptive Behavior.*

Dominey PF and Arbib MA (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb Cortex*, 2:153-75.

Doya K (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw*, 12(7-8):961-74.

Doya K (2000a). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr Opin Neurobiol*, 10(6):732-9.

Doya K (2000). Reinforcement Learning in Continuous Time and Space. *Neural Computation*, 12:219-45.

Doya K (2001). Robotic neuroscience: A synthetic approach to the brain. *Neuroscience Research*, Supplement, 24, S16.

Doya K, Ishii S, Pouget A and Rao RPN (Eds.) (2007). Bayesian brain: Probabilistic approaches to neural coding. MIT Press.

Doya K, Samejima K, Katagiri K and Kawato M (2002). Multiple model-based reinforcement learning. *Neural Computation* 14(6): 1347-69.

Doya K and Uchibe E (2005). The Cyber Rodent Project: Exploration of Adaptive Mechanisms for Self-Preservation and Self-Reproduction. *Adaptive Behavior*, vol. 13, no. 2, pages 149-60.

Dreher JC and Burnod Y (2002). An integrative theory of the phasic and tonic modes of dopamine modulation in the prefrontal cortex. *Neural Netw*, 15(4-6):583-602.

Drewe EA (1974). The effect of type and area of brain lesion on Wisconsin card sorting test performance. *Cortex*, 10:159-70.

Edito (2005). The practice of theoretical neuroscience. *Nature Neuroscience*, 8(12):1627.

Ekstrom AD, Kahana MJ, Caplan JB, Fields TA, Isham EA, Newman EL and Fried I (2003). Cellular networks underlying human spatial navigation. *Nature*, 425:184–8.

Elfwing S, Uchibe E and Doya K (2003). An evolutionary approach to automatic construction of the structure in hierarchical reinforcement learning. In *Proc. of the Genetic and Evolutionary Computation Conference*, pp. 507-9.

Eschenko O and Mizumori SJ (2007). Memory influences on hippocampal and striatal neural codes: effects of a shift between task rules. *Neurobiol Learn Mem*, 87(4):495-509.

Etienne AS and Jeffery KJ (2004). Path integration in mammals. *Hippocampus*, 14:180–92.

Fenu S, Bassareo V and Di Chiara G (2001). A role for dopamine D1 receptors of the nucleus accumbens shell in conditioned taste aversion learning. *J Neurosci*, 21(17):6897-904.

Ferbinteanu J and McDonald RJ (2001) Dorsal/ventral hippocampus, fornix, conditioned place preference. *Hippocampus* 11:187–200.

Filliat D, Girard B, Guillot A, Khamassi M, Lachèze L and Meyer J-A (2004). State of the artificial rat Psikharpax. In Schaal *et al.* (Eds), *From Animals to Animats 8: Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior,* pp. 2-12. The MIT Press, Cambridge, MA.

Fiorillo CD, Tobler PN and Schultz W (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898-902.

Floresco SB, Ghods-Sharifi S, Vexelman C and Magyar O (2006). Dissociable roles for the nucleus accumbens core and shell in regulating set shifting. *J Neurosci*, 26(9):2449-57.

Foster D, Morris R and Dayan P (2000). Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus,* 10: 1-16.

Frank MJ, Loughry B and O'Reilly RC (2001). Interactions Between Frontal Cortex and Basal Ganglia in Working Memory: A Computational Model. *Cognitive, affective and behavioral neuroscience*, 1(2):137-60.

Fritzke B (1995). A growing neural gas network learns topologies. In: Tesauro G, Touretzkys DS, Leen K (Eds): *Advances in Neural Information Processing Systems*, MIT Press, pp. 625-32.

Franz MO and Mallot HA (2000). Biomimetic robot navigation. *Robotics and Autonomous Systems*, 30:133-53.

Fritts ME, Asbury ET, Horton JE and Isaac WL (1998). Medial prefrontal lesion deficits involving or sparing the prelimbic area in the rat. *Physiology & Behavior*, 64:373-80.

Fukuda M, Tamura R, Kobayashi T, Nishijo H and Ono T (1997). Relationship between change of movement direction and activity of hippocampal place cells. *Neuroreport*, 8(15):3207-11.

Fuster JM (1997). *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe, (3rd ed.),* Philadelphia: Lippincott-Raven.

Fyhn M, Molden S, Witter MP, Moser EI and Moser MB (2004). Spatial representation in the entorhinal cortex. *Science*, 305(5688):1258-64.

Gallistel CR (1990). *The organisation of learning*. Cambridge, MA: MIT Press.

Gallup HF and Diamond L, (1960). Transfer of Double Alternation Behavior of Rats in a Temporal Maze. *The American Journal of Psychology*, 73(2):256-61.

Gardiner TW and Kitai ST (1992). Single unit activity in the globus pallidus and neostriatum of the rat during performance of a trained head movement. *Exp Brain Res* 88: 517-30.

Gaussier P, Leprêtre S, Joulain C, Revel A, Quoy M and Banquet JP (1998). Animal and robots learning: experiments and models about visual navigation. In *Seventh European Workshop on Learning Robots-EWLR'98*, Edinburgh, UK.

Gaussier P, Leprêtre S, Quoy M, Revel A, Joulain C and Banquet JP (2000). Experiments and models about cognitive map learning for motivated navigation. Interdisciplinary approaches to robot learning. *Robotics and Intelligent Systems Series*, World Scientific.

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

Geman S, Bienenstock E and Doursat R (1992). Neural networks and the bias/variance dilemma. *Neural Computation* 4, 1-58.

Genovesio A, Brasted PJ and Wise SP (2006). Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. *J Neurosci*, 26(27):7305-16.

Gerfen CR (1984). The neostriatal mosaic: Compartmentalization of corticostriatal input and striatonigral output systems. *Nature*, 311:461-4.

Gerfen CR (1992). The neostriatal mosaic: Multiple levels of compartmental organization in the basal ganglia. *Annu. Rev. Neurosci.* 15:285-320.

Gerfen CR, Herkenham M and Thibault J (1987). The Neostriatal Mosaic. II. Patch- and Matrix- Directed Mesostriatal Dopaminergic and Non-Dopaminergic Systems. *Journal of Neuroscience*, 7:3915-34.

Gerfen CR and Wilson CJ (1996). The basal ganglia. In L.W. Swanson, A. Björklund and T. Hökfelt, editors, *Handbook of chemical neuroanatomy*. Elsevier Science B.V.

Gerstner W and Kistler WM (2002). Spiking Neuron Models. Single Neurons, Populations, Plasticity. Cambridge University Pres.

Goldman-Rakic PS (1987). Circuitry of primate prefrontal cortex and the regulation of behavior by representational memory. In: F. Plum and V. Moutcastle, Editors, *Handbook of Physiology, The Nervous System* Vol. 5, American Physiological Society, pp. 373–417.

Girard B (2003). Intégration de la Navigation et de la Sélection de l'Action dans une Architecture de Contrôle Inspirée des Ganglions de la Base. AnimatLab-LIP6, Université Pierre et Marie Curie, Paris, France.

Girard B, Cuzin V, Guillot A, Gurney KN and Prescott TJ (2002). Comparing a bio-inspired robot action selection mechanism with winner-takes-all. In Hallam B, Floreano D, Hallam J, Hayes G and Meyer J-A, editors, *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, pages 75-84. Cambridge, MA. MIT Press.

Girard B, Cuzin V, Guillot A, Gurney KN and Prescott TJ (2003). A Basal Ganglia inspired Model of Action Selection Evaluated in a Robotic Survival Task. *Journal of Integrative Neuroscience,* 2(22), 179-200.

Girard B, Filliat D, Meyer J-A, Berthoz A and Guillot A (2004). An integration of two control architectures of action selection and navigation inspired by neural circuits in the vertebrates: The Basal ganglia. *Connectionist Models of Cognition and Perception II, Proceedings of the Eighth Neural Computation and Psychology Workshop*, pages 72-81.

Girard B, Filliat D, Meyer J-A, Berthoz A and Guillot A (2005). Integration of navigation and action selection functionalities in a computational model of cortico-basal ganglia-thalamo-cortical loops. *Adaptive Behavior*, 13(2):115-30.

Gisquet-Verrier P and Delatour B (2006). The role of the rat prelimbic/infralimbic cortex in working memory: not involved in the short-term maintenance but in monitoring and processing functions. *Neuroscience*, 141(2):585-96.

Goldstein RZ, Leskovjan AC, Hoff AL, Hitzemann R, Bashan F, Khalsa SS, Wang GJ, Fowler JS and Volkow ND (2004). Severity of neuropsychological impairment in cocaine and alcohol addiction: association with metabolism in the prefrontal cortex. *Neuropsychologia*, 2004;42(11):1447-58.

Grace AA (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience*, 41(1):1-24.

Grafman J (2002), The structured event complex and the human prefrontal cortex. In: D.T. Stuss and R.T. Knight, Editors, *Principles of Frontal Lobe Function*, Oxford University Press, pp. 292–310.

Granon S, Hardouin J, Courtière A and Poucet B (1998). Evidence for the involvement of the rat prefrontal cortex in sustained attention. *Quaterly J of Experimental Psychology*, 51B:219-33.

Granon S and Poucet B (1995). Medial prefrontal lesion in the rat and spatial navigation. *Behavioral Brain Research*, 109:474-484.

Granon S and Poucet B (2000). Involvement of the rat prefrontal cortex in cognitive functions: A central role for the prelimbic area. *Psychobiology*, 28(2):229-37.

Granon S, Save E, Buhoht MC and Poucet B (1996). Effortful information processing in a spontaneous spatial situation by rats with medial prefrontal lesions. *Behav Brain Res*, 78:147-54.

Granon S, Vidal C, Thinus-Blanc C, Changeux JP and Poucet B (1994). Working memory, response selection, and effortful processing in rats with medial prefrontal lesions. *Behav Neurosci*, 108:883-91.

Grant DA, Jones OR and Tallantis B (1949). The relative difficulty of the number, and color concepts of a Weigle-type problem. *J Exp Psychol*, 39:552-7.

Graybiel AM (1995). The basal ganglia. *Trends Neurosci*, 18(2):60-2.

Graybiel AM (1998). The basal ganglia and chunking of action repertoires. *Neurobiol Learn Mem* 70: 119-136.

Graybiel AM and Kimura M (1995). Adaptive neural networks in the basal ganglia. In: *Models of information processing in the basal ganglia,* edited by Houk JC, Davis JL, Beiser D, pp 103-16, Cambridge, MA: MIT Press.

Greenberg, N (2001). The Neuroethology of Paul MacLean : frontiers and convergences, « The Past and Future of the Basal Ganglia ».

Grobéty MC (1990). Importance de la dimension verticale de l'espace locomoteur dans l'orientation spatiale du rat de laboratoire. Thèse de Doctorat Faculté des Sciences de l'Université de Lausanne.

Groenewegen HJ, Vermeulen-Van der Zee E, te Kortschot A and Witter MP (1987). Organization of the projections from the subiculum to the ventral striatum in the rat. A study using anterograde transport of Phaseolus vulgaris leucoagglutinin. *Neuroscience*, 23(1):103-20.

Groenewegen HJ, Wright CI, Beijer AV (1996). The nucleus accumbens: gateway for limbic structure to reach the motor system? *Prog Brain Res* 107: 485-511.

Guazzelli A, Corbacho FJ, Bota M and Arbib MA (1998). Affodances, motivations and the world graph theory.. Adaptive Behavior : Special issue on biologically inspired models of spatial navigation, 6(3-4):435-71.

Guigon E, Dreher J-C, Guigon E and Burnod Y (2002b). A model of prefrontal cortex dopaminergic modulation during the delayed alternation task. *Journal of Cognitive Neuroscience* 14(6):853-65.

Guigon E, Koechlin E and Burnod Y (2002a). Short-term memory. In: *The Handbook of Brain Theory and Neural Networks*, 2nd ed (Arbib MA, ed), pp 1030-4. Cambridge: MIT Press.

Guillot A and Meyer J-A (1994). Computer simulation of adaptive behavior in animats. In *Computer Animation' 94*, Thalmann and Thalmann, editors. Pages 121-31.

Guillot A and Meyer J-A (2001). The Animat Contribution to Cognitive Systems Research. *Journal of Cognitive Systems Research*, 2(2):157-65, 2001.

Guillot A and Meyer J-A (2003). La contribution de l'approche animat aux sciences cognitives. *In Cognito*, A(1):1-26.

Gurney KN, Prescott TJ and Redgrave P (2001a). A Computational Model of Action Selection in the Basal Ganglia. I. A new functional anatomy. *Biological Cybernetics*. 84, 401-10.

Gurney KN, Prescott TJ and Redgrave P (2001b). A Computational Model of Action Selection in the Basal Ganglia. II. Analysis and simulation of behavior. *Biological Cybernetics*. 84, 411-23.

Gurney KN, Prescott TJ, Wickens JR and Redgrave P (2004). Computational models of the basal ganglia: from robots to membranes. *Trends Neurosci*, 27(8):453-9.

Haber SN, Fudge JL and McFarland NR (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci* 20: 2369-82.

Hadj-Bouziane F, Frankowska H, Meunier M, Coquelin PA and Boussaoud D (2006). Conditional visuo-motor learning and dimension reduction. *Cogn Process*, 7(2):95-104.

Hafting T, Fyhn M, Molden S, Moser MB and Moser EI (2005).Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801-6.

Hall J, Parkinson JA, Connor TM, Dickinson A and Everitt BJ (2001). Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. *Eur J Neurosci*, 13(10):1984-92.

Hamilton DA, Rosenfelt CS and Wishaw IQ (2004). Sequential control of navigation by locale and taxon cues in the Morris water task. *Behavioural Brain Research, 154*, 385-97.

Hannesson DK, Vacca G, Howland JG and Phillips AG (2004). Medial prefrontal cortex is involved in spatial temporal order memory but not spatial recognition memory in tests relying on spontaneous exploration in rats. *Behav Brain Res*, 153(1):273-85.

Harris KD, Henze DA, Csicsvari J, Hirase H and Buzsaki G (2000). Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J Neurophysiol*, 84(1):401-14.

Hartley T and Burgess N (2005). Complementary memory systems: competition, cooperation and compensation. *Trends in Neurosciences*, 28(4):169-70.

Haruno M and Kawato M (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Netw*, 19(8):1242-54.

Hassani OK, Cromwell HC and Schultz W (2001). Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *J Neurophysiol*, 85(6):2477-89.

Hasselmo ME (2005). A model of prefrontal cortical mechanisms for goal-directed behavior. *J Cogn Neurosci*, 17(7):1115-29.

Haykin and Simon (1999). "9. Self-organizing maps", *Neural networks - A comprehensive foundation*, 2nd edition, Prentice-Hall.

Heaton RK (1993). Wisconsin Card Sorting Test Manual. Odessa, FL:PAR.

Hebb DO (1949). The organization of behavior: a neurophysiological theory. Wiley, New York.

Heimer L, Alheid GF, de Olmos JS, Groenewegen H, Haber S, Harlan RE and Zahm, D (1997). The accumbens: beyond the core-shell dichotomy. *Journal of Neuropsychiatry*, 9:354-81.

Hikosaka O, Sakamoto M and Usui S (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J Neurophys* 61: 814-32.

Hikosaka O, Miyashita K, Miyachi S, Sakai K and Lu X (1998). Differential roles of the frontal cortex, basal ganglia, and cerebellum in visuomotor sequence learning. *Neurobiol Learn Mem*, 70(1-2):137-49.

Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, Miyachi S and Doya K (1999). Parallel neural networks for learning sequential procedures. *Trends Neurosci*, 22(10):464-71.

Hikosaka O, Takikawa Y and Kawagoe R. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol Rev*, 80(3):953-78.

Hok V, Save E, Lenck-Santini PP and Poucet B (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proc Natl Acad Sci U S A*, 102(12):4602-7.

Hok V, Lenck-Santini PP, Roux S, Save R, Muller RU and Poucet B (2007). Goal-related activity in hippocampal place cells. *J Neurosci*, 27(3):472-82, 2007.

Hollerman JR, Tremblay L and Schultz W (1998a). Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J Neurophysiol*, 80(2):947-63.

Hollerman JR and Schultz W (1998b). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci*, 1(4):304-9.

Hollerman JR, Tremblay L and Schultz W (2000). Involvement of basal ganglia and orbitofrontal cortex in goal-directed behavior. *Prog Brain Res*, 126:193-215.

Holmström J. (2002). Growing neural gas: Experiments with GNG, GNG with utility and supervised GNG. Master's thesis, Uppsala University.

Honig W (1978). Studies of working memory in the pigeon. In Hulse S. et al. (Eds.) *Cognitive processes in animal behavior*, pp. 211-48, Hillsdale NJ: Lawrence Erlbaum.

Honzik CH (1936). The sensory basis of maze learning in rats. *Comp. Psychol. Monog*, 13:113.

Horvitz JC (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*,

96(4):651-6.

Horvitz JC, Stewart T and Jacobs BL (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res*, 759(2):251-8.

Houk JC, Adams JL and Barto AG (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: *Models of information processing in the basal ganglia,* edited by Houk JC, Davis JL, Beiser D, pp 249-70, Cambridge, MA: MIT Press.

Humphries MD (2003). High level modelling of dopamine mechanisms in striatal neurons *Technical Report ABRG 3 2*.

Humphries MD and Gurney KN (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network* **13**:131-56.

Humphries MD, Gurney K and Prescott TJ (2007). Is there a brainstem substrate for action selection? *Philos Trans R Soc Lond B Biol Sci*.

Humphries MD, Stewart RD and Gurney KN (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J Neurosci*, 26(50):12921-42.

Ikemoto S (2002). Ventral striatal anatomy of locomotor activity induced by cocaine, (D)-amphetamine, dopamine and D1/D2 agonists. *Neurosc*i 113: 939-55.

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

Ito M and Kano M (1982). Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex. *Neurosci Lett*, 33:253-8.

Itoh H, Nakahara H, Hikosaka O, Kawagoe R, Takikawa Y and Aihara K (2003). Correlation of primate caudate neural activity and saccade parameters in reward-oriented behavior. *J Neurophysiol* 89: 1774-83.

Jacobs RA, Jordan MI, Nowlan SJ and Hinton GE (1991). Adaptive Mixture of Local Experts. *Neural Computation*, 3:79-87.

Janak PH, Chen MT and Caulder T (2004). Dynamics of neural coding in the accumbens during extinction and reinstatement of rewarded behavior. *Behav Brain Res* 154: 125-35.

Joel D, Niv Y and Ruppin E (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw* 15(4-6): 535-47.

Joel D and Weiner I (2000). The Connections of the Dopaminergic System with Striatum in Rats and Primates: An Analysis with respect to the Functional and Compartmental Organization of the Striatum. *Neuroscience*, 96:451-74.

Joel D, Weiner I and Feldon J (1997). Electrolytic lesions of the medial prefrontal cortex in rats disrupt performance on an analog of the Wisconsin Card Sorting Test, but do not disrupt latent inhibition: Implications for animal models of schizophrenia. *Behav Brain Res*, 85:187-201.

Jog MS, Kubota Y, Connolly CI, Hillegaart V and Graybiel AM (1999). Building neural representations of habits. *Science*, 286(5445):1745-9.

Jung MW, Qin Y, Lee D and Mook-Jung I (2000). Relationship among discharges of neighboring neurons in the rat prefrontal cortex during spatial working memory tasks. *J Neurosci*, 20(16):6166-72.

Jung MW, Qin Y, Lee D and McNaughton BL (1998). Firing characteristics of deep layer neurons in prefrontal cortex in rats performing spatial working memory tasks. *Cereb Cortex*, 8:437-50.

Jung MW, Wiener SI and McNaughton BL (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. *J Neurosci*, 14(12):7347-56.

Kakade S and Dayan P (2002). Dopamine: generalization and bonuses. *Neural Netw*, 15(4-6):549-59.

Kaelbling LP, Littman ML and Moore AW (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237-85.

Kaplan F and Oudeyer PY (2003). Motivational principles for visual know-how development. In CG Prince, L Berthouze, H Kozima, D Bullock, G Stojanov and C Balkenius (Eds.) *Proceedings of the Third International Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotics Systems*, pp. 73-80, Edinburgh, Scotland, Lund University Cognitive Studies.

Kargo WJ, Szatmary B and Nitz DA (2007). Adaptation of prefrontal cortical firing patterns and their fidelity to changes in action-reward contingencies. *J Neurosci*, 27(13):3548-59.

Kawagoe R, Takikawa Y and Hikosaka O (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nat Neurosci,* 1: 411-16.

Kawagoe R, Takikawa Y and Hikosaka O (2003). Reward-predicting activity of dopamine and caudate neurons - a possible mechanism of motivational control of saccadic eye movement. *J Neurophysiol*, 91:1013-24.

Kawato M (1999). Internal models for motor control and trajectory planning.*Current Opinion in Neurobiology*,9:718-27.

Kelley AE (1999). Neural integrative activities of nucleus accumbens subregions in relation to learning and motivation. *Psychobiology*, 27(2):198-213.

Kelley AE, Smith-Roe SL and Holahan MR (1997). Response-reinforcement learning is dependent on N-methyl-D-aspartate receptor activation in the nucleus accumbens core. *Proc Natl Acad Sci U S A*, 94(22):12174-9.

Kermadi I and Joseph JP (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *J Neurophysiol*, 74(3):911-33.

Kermadi I, Jurquet Y, Arzi M and Joseph JP (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Exp Brain Res*, 94(2):352-6.

Kesner RP, Bolland BL and Dakis M (1993). Memory for spatial locations, motor responses, and objects: triple dissociation among the hippocampus, caudate nucleus, and extrastriate visual cortex. *Exp Brain Res*, 93(3):462-70.

Kesner RP and Rogers J (2004). An analysis of independence and interactions of brain substrates that subserve multiple attributes, memory systems, and underlying processes. *Neurobiol Learn Mem*, 82(3):199-215.

Khamassi M (2003). Un modèle d'apprentissage par renforcement dans une architecure de contrôle de la sélection chez le rat artificiel Psikharpax. Université Pierre et Marie Curie, Paris, France.

Khamassi M, Girard B, Berthoz A and Guillot A (2004). Comparing three critic models of reinforcement learning in the basal ganglia connected to a detailed actor in a S-R task. In F. Groen, A. Amato, A. Bonarini, E. Yoshida and B. Kröse, editors, Proceedings of the Eighth International Conference on Intelligent Autonomous Systems, pages 430-7. Amsterdam, The Netherlands. IOS Press.

Khamassi M, Lachèze L, Girard B, Berthoz A and Guillot A (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adapt Behav*, *Spec Issue Towards Artificial Rodents* 13(2):131-48.

Khamassi M, Martinet L-E and Guillot A (2006). Combining self-organizing maps with mixture of experts: Application to an Actor-Critic model of reinforcement learning in the basal ganglia. In: *From Animals to Animats 9. Proceedings of the Ninth International Conference on Simulation of Adaptive Behavior* edited by Nolfi S, Baldassare G, Calabretta R, Hallam J, Marocco D, Meyer J-A, Miglino O, Parisi D, pp 394-405. Springer - Lecture Notes in Artificial Intelligence 4095.

Khamassi M, Mulder AB, Tabuchi E, Douchamps V and Wiener SI (submitted to J Neurophysiol, in revision). Actor-Critic models of reward prediction signals in the rat ventral striatum require multiple input modules.

Killcross AS and Coutureau E (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex*, 13(4):400-8.

Kim J and Ragozzino ME (2005). The involvement of the orbitofrontal cortex in learning under changing task

contingencies. *Neurobiol Learn Mem*, 83(2):125-33.

Kimura M (1986). The role of primate putamen neurons in the association of sensory stimuli with movement. *Neurosci Res*, 3(5):436-43.

Kimura M (1990). Behaviorally contingent property of movement-related activity of the primate putamen. *J Neurophysiol*, 63(6):1277-96.

Kimura M (1995). Role of basal ganglia in behavioral learning. *Neurosci Res*, 22(4):353-8.

Kimura M (1986). The role of primate putamen neurons in the association of sensory stimuli with movement. *Neurosci* Res 3: 436-43.

Kimura M, Rajkowski J and Evarts E (1984). Tonically discharging putamen neurons exhibit set-dependent responses. USA 81: *Proc Nat Acad Sci*: 4998-5001.

Knierim JJ, Kudrimoti HS and McNaughton BL (1995). Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, 15(3):1648-59.

Koechlin E, Ody C and Kouneiher F (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181-5.

Koechlin E and Summerfield C (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn Sci*, 11(6):229-35.

Kohonen T (1995). Self-organizing maps. Springer-Verlag, Berlin.

Kolb B (1990). Animal models for human PFC-related disorders. In Uylings HBM, Van Eden CG, de Bruin JPC, Corner MA, Feenstra MPG (Eds.) *The prefrontal cortex: its structure, function and pathology*. *Progress in brain research*, (85):501-19.

Konidaris GD and Barto AG (2006). An adaptive robot motivational system. *From Animals to Animats 9: Proceedings of the Nineth International Conference on the Simulation of Adaptive Behavior*, pages 346-56.

Konishi S, Nakajima K, Uchida I, Kameyama M, Nakahara K, Sekihara K and Miyashita Y (1998). Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nat Neurosci*, 1(1):80-4.

Krech D (1932). The genesis of "hypotheses" in rats. University of California. *Publications in Psychology*, 6, 45-64.

Kuipers T (1982). Approaching Descriptive and Theoretical Truth. *Erkenntnis* 18(3):343-78.

Lapiz MD and Morilaz DA (2006). Noradrenergic modulation of cognitive function in rat medial prefrontal cortex as measured by attentional set shifting capability. *Neuroscience*, 137(3):1039-49.

Lauwereyns J, Watanabe K, Coe B and Hikosaka O (2002a). A neural correlate of response bias in monkey caudate nucleus. *Nature*, 418(6896):413-7.

Lauwereyns J, Takikawa Y, Kawagoe R, Kobayashi S, Koizumi M, Coe B, Sakagami M and Hikosaka O (2002b). Feature-based anticipation of cues that predict reward in monkey caudate nucleus. *Neuron*, 33(3):463-73.

Lavoie AM and Mizumori SJ (1994). Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Res* 638: 157-168.

Lee JK and Kim IH (2003). Reinforcement Learning Control Using Self-Organizing Map and Multi-Layer Feed-Forward Neural Network. In *International Conference on Control Automation and Systems, ICCAS 2003*.

Leonard B and McNaughton BL (1990). Rat: Conceptual, behavioral, and neurophysiological perspectives. In RP Kesneret, DS Olton (Eds.), *Neurobiology of comparative cognition*, Chapt.13, Lawrence Erlbaum Associates.

Leutgeb S, Ragozzino KE and Mizumori SJ (2000). Convergence of head direction and place information in the CA1 region of hippocampus. *Neuroscience*, 100(1):11-9.

Lindman HR (1974). *Analysis of Variance in Complex Experimental Designs*. San Francisco: W. H. Freeman and Co.

Marsland S, Shapiro J, Nehmzow U (2002). A self-organising network that grows when required. *Neural Networks*, 15:1041-58.

Ljungberg T, Apicella P and Schultz W (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol*, 67(1):145-63.

Mansouri FA, Matsumoto K and Tanaka K (2006). Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog. *J Neurosci*, 26(10):2745-56.

Martin PD and 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. *Behav Brain Res* 116: 23-38.

Matsumoto N, Minamimoto T, Graybiel AM and Kimura M (2001). Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *J Neurophysiol,* 85: 960-76.

Matsumoto K, Suzuki W and Tanaka K (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301(5630):229-32.

Matthews DB, Ilgen M, White AM and Best PJ (1999). Acute ethanol administration impairs spatial performance while facilitating nonspatial performance in rats. *Neurobiol Learn Mem*, 72(3):169-79.

Maurice N, Deniau JM, Menetrey A, Glowinski J and Thierry AM (1997). Position of the ventral pallidum in the rat prefrontal cortex-basal ganglia circuit. *Neuroscience*, 80(2):523-34.

Maurice N, Deniau JM, Glowinski J and Thierry AM (1999). Relationships between the prefrontal cortex and the basal ganglia in the rat: physiology of the cortico-nigral circuits. *Journal of Neuroscience*, 19(11):4674-81.

McAlonan K and Brown VJ (2003). Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. *Behav Brain Res*, 146(1-2):97-103.

McClure SM, Daw ND and Montague PR (2003). A computational substrate for incentive salience. *Trends Neurosci*, 26(8):423-8.

McDonald RJ and White NM (1994). Parallel information processing in the water maze: evidence for independent memory systems involving dorsal striatum and hippocampus. *Behavioral Neural Biology*, 61(3):260–70.

McDonald RJ and White NM (1995). Hippocampal and nonhippocampal contributions to place learning. *Behavioral Neuroscience, 109,* 579–93.

McGeorge AJ and Faull RM (1989). The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience*, 29:503-37.

McGovern A, Precup D, Ravindran B, Singh S and Sutton RS (1998). Hierarchical optimal control of MDPs. Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems, pp. 186-91.

McIntyre CK, Marriott LK and Gold PE (2003). Patterns of brain acetylcholine release predict individual differences in preferred learning strategies in rats. *Neurobiol Learn Mem,* 79: 177–83.

McNaughton BL (1989). Neural Mechanisms for Spatial Computation and Information Storage. In Nadel *et al.* (Eds), *Neural Connections, Mental Computations*, chapter 9, pp.285-350, MIT Press, Cambridge, MA.

McNaughton BL, Battaglia FP, Jensen O, Moser EI and Moser MN (2006). Path integration and the neural basis of the 'cognitive map'. *Nat Rev Neurosci*, 7(8):663-78.

McNaughton BL, O'Keefe J and Barnes CA (1983). The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. *J Neurosci Methods*, 8(4):391-7.

Meyer J-A (1996). *Artificial Life and the Animat Approach to Artificial Intelligence* in Artificial Intelligence, Boden M,

editors. Pages 325-54.

Meyer J-A and Guillot A (1991). Simulation of adaptive behavior in animats: review and prospect. Proceedings of the first international conference on simulation of adaptive behavior (From animals to animats), Pages: 2-14, Paris, France, MIT Press.

Meyer J-A and Guillot A (In press). Biologically-inspired robots. In *Handbook of Robotics*. Springer-Verlag.

Meyer J-A, Guillot A, Girard B, Khamassi M, Pirim P and Berthoz A (2005). The Psikharpax project: Towards building an artificial rat. *Robotics and Autonomous Systems,* 50(4):211-23.

Miller EK and Cohen JD (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, 24:167-202.

Milner B (1963). Effects of different brain lesions on card sorting. *Arch Neurol*, 9:90.

Mink JW (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50(4):381-425.

Mirenowicz J and Schultz W (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J Neurophysiol*, 72(2):1024-7.

Miyachi S, Hikosaka O, Miyashita K, Karadi Z and Rand MK (1997). Differential roles of monkey striatum in learning of sequential hand movement. *Exp Brain Res*, 115(1):1-5.

Miyazaki K, Mogi E, Araki N and Matsumoto G (1998). Reward-quality dependent anticipation in rat nucleus accumbens. *Neuroreport,* 9: 3943-8.

Miyazaki K, Miyazaki KW and Matsumoto G (2004). Different representation of forthcoming reward in nucleus accumbens and medial prefrontal cortex. *Neuroreport*, 15(4):721-6.

Mizumori SJ, Yeshenko O, Gill KM and Davis DM (2004). Parallel processing across neural systems: implications for a multiple memory system hypothesis. *Neurobiol Learn Mem*, 82(3):278-98.

Mogenson GJ, Jones DL and Yim CY (1980). From motivation to action: Functional interface between the limbic system and the motor system. *Prog Neurobiol,* 14: 69-97.

Montague PR, Dayan P and Sejnowski TJ (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci,* 6(5): 1936-47.

Montes-Gonzalez F, Prescott TJ, Gurney KN, Humphries M and Redgrave P (2000). An Embodied Model of Action Selection Mechanisms in the Vertebrate Brain. In Meyer *et al*. (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* (pp.157-66). The MIT Press, Cambridge, MA.

Morgan MA, Schulkin J and LeDoux JE (2003). Ventral medial prefrontal cortex and emotional perseveration: the memory for prior extinction training. *Behav Brain Res*, 146(1-2):121-30.

Morimoto J and Doya K (1998). Hierarchical reinforcement learning of low-dimensional sub-goals and high-dimensional trajectories. In Proceedings of the Fifth International Conference on Neural Information Processing, Burke, VA, Vol. 2, IOS Press, Amsterdam, 1998, pp. 850–3.

Morimoto J and Doya K (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, 36(1):37-51.

Morris RGM (1981). Spatial localization does not require the presence of local cues. *Learn Motiv*, 12:239–60.

Moschovakis AK, Scudder CA and Highstein SM (1996). The microscopic anatomy and physiology of the mammalian saccadic system. *Prog Neurobiol*, 50(2-3):133-254.

Muir JL, Everitt BJ and Robbins TW (1996). The cerebral cortex of the rat and visual attentional function: Dissociable effects of mediofrontal, cingulate, anterior dorsolateral and parietal cortex lesions on a five-choice serial reaction

time task. *Cerebral Cortex*, 6:470-81.

Mulder AB, Gijsberti Hodenpijl M and Lopes da Silva FH (1998). Electrophysiology of the hippocampal and amygdaloid projections to the nucleus accumbens of the rat: convergence, segregation and interaction of inputs. *J Neurosci* 18: 5095-102.

Mulder AB, Nordquist RE, Orgut O and Pennartz CM (2003). Learning-related changes in response patterns of prefrontal neurons during instrumental conditioning. *Behav Brain Res*, 146(1-2):77-88.

Mulder AB, Shibata R, Trullier O and Wiener SI (2005). Spatially selective reward site responses in tonically active neurons of the nucleus accumbens in behaving rats. *Exp Brain Res,* 163: 32-43.

Mulder AB, Tabuchi E and Wiener SI (2004). Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur J Neurosci,* 19 : 1923-32.

Muller RU, Kubie JL, Bostock EM, Taube JS, Quirk GJ (1991). Spatial firing correlates of neurons in the hippocampal formation of freely moving rats. In: Paillard, J (Ed.), Brain and Space. Oxford University Press, pp. 296–333.

Muller RU, Ranck Jr JB and Taube JS (1996). Head direction cells: properties and functional significance. *Curr Opin Neurobiol.* 6, 196–206.

Muller RU, Poucet B, Fenton AA and Cressant A (1999). Is the hippocampus of the rat part of a spacialized navigational system? *Hippocampus*, 9(4):413-22.

Mushiake H, Saito N, Sakamoto K, Itoyama Y and Tanji J (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, 50(4):631-41.

Nakahara H, Doya K and Hikosaka O (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J Cogn Neurosci*, 13(5):626-47.

Nakahara H, Itoh H, Kawagoe R, Takikawa Y and Hikosaka O (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron*, 41(2):269-80.

Nakahara K, Hayashi T, Konishi S and Miyashita Y (2002). Functional MRI of macaque monkeys performing a cognitive set-shifting task. *Science*, 295(5559):1532-6.

Nicola SM, Yun IA, Wakabayashi KT and Fields HL (2004). Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task. J *Neurophysiol* 91: 1840-65.

Niv Y, Daw ND, Joel D and Dayan P (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, 191(3):507-20.

Niv Y, Duff MO and Dayan P (2005). Dopamine, uncertainty and TD learning. *Behav Brain Funct*, 1:6.

Northcutt RG and Kaas JH (1995). The emergence and evolution of mammalian neocortex. *Trends Neurosci*, 18:373-9.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K and Dolan RJ (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science,* 304(5669): 452-4.

O'Keefe J (1990). A computational theory of the hippocampal cognitive map. *Prog Brain Res,* 83:301–12.

O'Keefe J and Dostrovsky J (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res*, 34(1):171-5.

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

Olton DS, Becker JT and Handelmann GE (1979). Hippocampus, space, and memory. *Behav Brain Res*, 2:313-66.

O'Reilly RC and Frank MJ (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput*, 18(2):283-328.

Ostlund SB and Balleine BW (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *J Neurosci*, 25(34):7763-70.

Otani S (2004). *Prefrontal Cortex: From Synaptic Plasticity to Cognition.* Boston: Kluwer Acadamic.

Oudeyer PY, Kaplan F, Hafner VV and Whyte A (2005). The playground experiment: Task-independent development of a curious robot. Proceedings of the 2005 AAAI Spring Symposium on Developmental Robotics, pages 42-7. Stanford, USA.

Overton PG, Coizet V, Dommett EJ and Redgrave P (2005). The parabrachial nucleus is a source of short latency nociceptive input to midbrain dopaminergic neurons in the rat. *Soc Neurosci Abstracts*, Program No 301.5.

Packard MG (1999). Glutamate infused posttraining into the hippocampus or caudate–putamen differentially strengthens place and response learning. *PNAS*, 96(22):12881–6.

Packard MG, Hirsh R and White NM (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *J Neurosci*, 9(5):1465-72.

Packard MG and Knowlton BJ (2002). Learning and memory functions of the Basal Ganglia. *Annu Rev Neurosci*, 2002;25:563-93.

Packard MG and McGaugh JL (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience,* 106, 439–46.

Packard MG and McGaugh JL (1996). Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning. Neurobiology of Learning and Memory, 65:65–72.

Parent A and Hazrati LN (1995a). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res Brain Res Rev*, 20(1):91-127.

Parent A and Hazrati LN (1995b). Functional anatomy of the basal ganglia. II. The place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Res Brain Res Rev*, 20(1):128-54.

Parkinson JA, Olmstead MC, Burns LH, Robbins TW and Everitt BJ (1999). Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity by D-amphetamine. *J Neurosci*, 19(6):2401-11.

Parron C, Poucet B and Save E (2004). Entorhinal cortex lesions impair the use of distal but not proximal landmarks during place navigation in the rat. *Behav Brain Res*, 154(2):345-52.

Pasupathy A and Miller RK (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433(7028):873-6.

Paxinos G and Watson C (1998). *The Rat Brain in Stereotaxic Coordinates*. NY: Acad Press.

Pearce JM, Roberts AD and Good M (1998). Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature*, 396(6706):75-7.

Pennartz CMA (1996). The Ascending Neuromodulatory Systems in Learning by Reinforcement: Comparing Computational Conjectures with Experimental Findings. *Brain Research Reviews*, 21:219-45.

Pennartz CMA (1997). Reinforcement learning by Hebbian synapses with adaptive thresholds. *Neuroscience*, 81(2):303-19.

Pennartz CM, Groenewegen HJ and Lopes da Silva FH (1994). The nucleus accumbens as a complex of functionally distinct neuronal ensembles: an integration of behavioural, electrophysiological and anatomical data. *Prog Neurobiol*, 42(6):719-61.

Pennartz CM, Lee E, Verheul J, Lipa P, Barnes CA and McNaughton BL (2004). The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *J Neurosci*, 24(29):6446-56.

Pennartz CMA, McNaughton BL and Mulder AB (2000). The glutamate hypothesis of reinforcement learning. *Prog*

Brain Res, 126:231-53.

Perez-Uribe A (2001). Using a Time-Delay Actor-Critic Neural Architecture with Dopamine-like Reinforcement Signal for Learning in Autonomous Robots. In Wermter *et al.* (Eds), *Emergent Neural Computational Architectures based on Neuroscience: A State-of-the-Art Survey* (pp. 522-33). Springer-Verlag, Berlin.

Petrides M (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex, *Philos Trans R Soc Lond B Biol Sci,* 351, pp. 1455–61.

Petrides M, Alivisatos B, Evan AC and Meyer E (1993). Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing, *Proc. Natl. Acad. Sci. U. S. A.* 90, pp. 873–7.

Peyrache A, Benchenane K, Khamassi M, Douchamps V, Tierney PL, Battaglia FP and Wiener SI (2007). Rat medial prefrontal cortex neurons are modulated by both hippocampal theta rhythm and sharp wave-ripple events. *Soc Neurosci Abstr.* 2007-A-42327-SfN.

Phillips GD, Setzu E and Hitchcott PK (2003). Facilitation of appetitive Pavlovian conditioning by d-amphetamine in the shell, but not the core, of the nucleus accumbens. *Behav Neurosci*, 117(4):675-84.

Ploeger GE, Spruijt BM and Cools AR (1994). Spatial localization in the Morris water maze in rats: acquisition is affected by intra-accumbens injections of the dopaminergic antagonist haloperidol. *Behav Neurosci*, 108(5):927-34.

Poldrack RA and Packard MG (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia* 41: 245–51.

Posner MI and Snyder CRR (1975). Attention and cognitive control. In *Information Processing and Cognition*, ed. Solso RL and Hillsdale NJ: Erlbaum.

Potegal M (1969). Role of the caudate nucleus in spatial orientation of rats. *J Comp Physiol Psychol*, 69(4):756-64.

Poucet B (1984). Evaluation of connectedness by cats in path-selection problems. *Perceptual and Motor Skills*, 58:51-4.

Poucet B (1989). Object exploration, habituation, and response to a spatial change in rats following septal or medial frontal cortical damage. *Behav Neurosci*, 103(5):1009-16.

Poucet B (1997). Searching for spatial unit firing in the prelimbic area of the rat medial prefrontal cortex. *Behav Brain Res*, 84:151-9.

Poucet B (1993). Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psychol. Rev.* 100: 163–82.

Poucet B and Hermann T (1990). Septum and medial frontal cortex contribution to spatial problem-solving. *Behav Brain Res*, 37(3):269-80.

Poucet B and Hermann T (2001). Exploratory patterns of rats on a complex maze provide evidence for topological coding. *Behav Processes*, 53(3):155-62.

Poucet B, Lenck-Santini PP, Hok V, Save E, Banquet JP, Gaussier P and Muller RU (2004). Spatial navigation and hippocampal place cell firing: the problem of goal encoding. *Rev Neurosci*, 15(2):89-107.

Poucet B, Lenck-Santini PP, Paz-Villagran V and Save E (2003). Place cells, neocortex and spatial navigation: a short review. *J Physiol (Paris)*, 97(4-6):537-46.

Pratt WE and Mizumori SJ (2001). Neurons in rat medial prefrontal cortex show anticipatory rate changes to predictable differential rewards in a spatial memory task. *Behav Brain Res*, 123(2):165-83.

Precup D and Sutton RS (1998). Multi-time models for temporally abstract planning. *Advances in Neural Information Processing Systems*, 10, pp. 1050-6, MIT Press, Cambride, MA.

Prescott TJ (1996). Spatial representation for navigation in animats. *Adaptive Behavior*, Vol. 4, No. 2, 85-123.

Prescott TJ and Humphries MD (2007). Who dominates who in the dark basements of the brain? *Behav Brain Sci*, 30(1):104-5.

Prescott TJ, Redgrave P and Gurney K (1999). Layered control architectures in robots and vertebrates. *Adaptive Behavior,* 7:99-127.

Preuss TM (1995). Do rats have a prefrontal cortex? The Rose-Woolsey-Akert program reconsidered. *J Cogn Neurosci*, 7:1-24.

Procyk E, Tanaka YL and Joseph JP (2000). Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nat Neurosci*, 3(5):502-8.

Pych JC, Chang Q, Colon-Rivera C and Gold PE (2005). Acetylcholine release in hippocampus and striatum during testing on a rewarded spontaneous alternation task. *Neurobiology of Learning and Memory*, 84: 93–101.

Quirk GJ, Russo GK, Barron JL and Lebron K (2000). The role of ventromedial prefrontal cortex in the recovery of extinguished fear. *J Neurosci*, 20(16):6225-31.

Ragozzino ME and Choi D (2003). Dynamic changes in acetylcholine output in the medial striatum during place reversal learning. *Learn Mem*, 11(1):70-7.

Ragozzino ME, Detrick S and Kesner RP (1999). Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. *J Neurosci*, 19(11):4585-94.

Ragozzino ME and Kesner RP (2001). The role of rat dorsomedial prefrontal cortex in working memory for egocentric responses. Neurosci Lett, 308(3):145-8.

Ragozzino ME, Kim J, Hassert D, Minniti N and Kiang C (2003). The contribution of the rat prelimbic-infralimbic areas to different forms of task switching. *Behav Neurosci*, 117(5):1054-65.

Ragozzino ME, Ragozzino KE, Mizumori SJ and Kesner RP (2002). Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning. *Behav Neurosci*, 116(1):105-15.

Ragozzino ME, Wilcox C, Raso M and Kesner RP (1999). Involvement of rodent prefrontal cortex subregions in strategy switching. *Behav Neurosci*, 113(1):32-41.

Ranck JBJ (1984). Head-direction cells in the deep cell layers of dorsal presubiculum in freely moving rats. Society Neuroscience Abstracts.

Ravel S, Legallet E and Apicella P (1999). Tonically active neurons in the monkey striatum do not preferentially respond to appetitive stimuli. *Exp Brain Res*, 128(4):531-4.

Ravel S, Legallet E and Apicella P (2003). Responses of tonically active neurons in the monkey striatum discriminate between motivationally opposing stimui. *J Neurosci* 23: 8489-97.

Ravel S, Sardo P, Legallet E and Apicella P (2006). Influence of spatial information on responses of tonically active neurons in the monkey striatum. *J Neurophysiol*, 95(5):2975-86.

Recce ML and O'Keefe JO (1989). The tetrode: a new technique for multiunit extracellular recording. Soc Neurosci Abstr.

Redgrave P and Gurney K (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat Rev Neurosc*i, 7(12):967-75.

Redgrave P, Prescott TJ and Gurney K (1999a). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009-23.

Redgrave P, Prescott TJ and Gurney K (1999b). Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci*, 22(4):146-51.

Redish AD (1999). Beyond the Cognitive Map: From Place Cells to Episodic Memory (Cambridge, MA: MIT Press).

Redish AD and Touretzky DS (1998). The role of the hippocampus in solving the morris water maze. *Neural Computation,* 10(1), 73–111.

Restle F (1957). Discrimination of cues in mazes: A resolution of the "place-vs.-response" question. *Psychological Review, 64,* 217–28.

Reynolds JN, Hyland BI and Wickens JR (2001). A cellular mechanism of reward-related learning. *Nature*, 413(6851):67-70.

Ritter H, Martinetz T and Schulten K (1992). Neural Computation and Self-Organizing Maps; An Introduction. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Roberts WA, Cruz C and Tremblay J (2007). Rats take correct novel routes and shortcuts in an enclosed maze. *J Exp Psychol Anim Behav Process*, 33(2):79-91.

Robbins TW and Everitt BJ (1992). Functions of dopamine in the dorsal and ventral striatum. *Seminars in The Neurosciences*, 4:119-27.

Roesch MR, Taylor AR and Schoen G (2006). Value representation. *Neuron*, 51(4):509-20.

Rolls ET, Thorpe SJ and Maddison SP (1983). Responses of striatal neurons in the behaving monkey. 1. Head of the caudate nucleus. *Behav Brain Res*, 7(2):179-210.

Sakagami M and Niki H (1994). Spatial selectivity of go/no-go neurons in monkey prefrontal cortex. *Exp Brain Res*, 100(1):165-9.

Salazar RF, White W, Lacroix L, Feldon J and White IM (2004). NMDA lesions in the medial prefrontal cortex impair the ability to inhibit responses during reversal of a simple spatial discrimination. *Behav Brain Res*, 152(2):413-24.

Samejima K and Doya K (2007). Multiple Representations of Belief States and Action Values in Corticobasal Ganglia Loops. *Ann. N.Y. Acad. Sci.* 1104: 213–28.

Samejima K, Ueda Y, Doya K and Kimura M (2005). Representation of action-specific reward values in the striatum. *Science* 310: 1337-40.

Sargolini F, Fyhn M, Hafting T, McNaughton BL, Witter MP, Moser MB and Moser EI (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758-62.

Satoh T, Matsumoto N, Nakai S, Satoh T, Minamimoto T and Kimura M (2003). Dopamine neurons encode teaching signals for learning reward-based decision strategy. *International Congress Series*, 1250:311-8.

Schmitzer-Torbert N and Redish AD (2004). Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J Neurophysiol* 91: 2259-72.

Schoenbaum G, Chiba AA and Gallagher M (2000). Changes in functional connectivity in orbitofrontal cortex and basolateral amygdala during learning and reversal training. *J Neurosci*, 20(13):5179-89.

Schoenbaum G, Setlow B, Saddoris MP and Gallagher M (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron,* 39: 855-67.

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

Schultz W (2001). Reward signaling by dopamine neurons. *Neuroscientist*, 7(4):293-302.

Schultz W, Apicella P, Scarnati E and Ljungberg T (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J Neurosci* 12: 4595-610.

Schultz W, Apicella P and Ljungberg T (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task. *Journal of Neuroscience*, 13(3):900-13.

Schultz W, Dayan P and Montague PR (1997). A neural substrate of prediction and reward. *Science* 275: 1593-9.

Schultz W, Romo R, Ljungberg T, Mirenowicz J, Hollerman JR and Dickinson A (1995). Reward-related signals carried by dopamine neurons. In: *Models of Information Processing in the Basal Ganglia*, edited by Houk JC, Davis JL, Beiser DG, pp 233-48.Cambridge, MA: MIT.

Setlow B and McGaugh JL (1998). Sulpiride infused into the nucleus accumbens posttraining impairs memory of spatial water maze training. *Behav Neurosci*, 112(3):603-10.

Setlow B, Schoenbaum G and Gallagher M (2003). Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron*, 38(4):625-36.

Shallice T (1988). From Neuropsychology to Mental Structure, Cambridge University Press.

Shallice T (1996). The neuropsychology of prospective memory. In Brandimonte M, Einstein GO and McDaniel MA (Eds.) *Prospective memory: theory and applications*. Mahwah, NJ: Lawrence Erlbaum, p. 319-25.

Sherry DF and Schacter DL (1987). The Evolution of Multiple Memory Systems. *Psychological Review*, 94(4):439-54.

Sheynikhovich D, Chavarriaga R, Strosslin T and Gerstner W (2006). Adaptive sensory processing for efficient place coding. *Neurocomputing*, 69(10-12):1211-4.

Shibata R, Mulder AB, Trullier O and Wiener SI (2001). Position sensitivity in phasically discharging nucleus accumbens neurons of rats alternating between tasks requiring complementary types of spatial cues. *Neurosci* 108: 391-411.

Shidara M, Aigner TG and Richmond BJ (1998). Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J Neurosci*, 18(7):2613-25.

Shidara M and Richmond BJ (2004). Differential encoding of information about progress through multi-trial reward schedules by three groups of ventral striatal neurons. *Neurosci Res*, 49(3):307-14.

Shiffrin RM and Schneider W (1984). Automatic and controlled processing revisited. *Psychol Rev*, 91(2):269-76.

Shima K and Tanji J (1998a). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science*, 282(5392):1335-8.

Shima K and Tanji J (1998b). Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. *J Neurophysiol*, 80(6):3247-60.

Shirakawa O and Ichitani Y (2004). Prolonged initiation latency in Morris water maze learning in rats with ibotenic acid lesions to medial striatum: effects of systemic and intranigral muscimol administration. *Brain Res*, 1030(2):193-200.

Sigaud O (2004). Comportements adaptatifs pour des agents dans des environnements informatiques complexes. Habilitation à diriger des recherches, Universite Pierre et Marie Curie, Paris, France.

Smith AJ (2002). Applications of the Self-Organizing Map to Reinforcement Learning. *Neural Networks*, 15(8-9):1107-24.

Sporns O and Alexander WH (2002). Neuromodulation and Plasticity in an Autonomous Robot. *Neural Networks*, 15:761-74.

Steck SD and Mallot HA (2000). The role of global and local landmarks in virtual environment navigation. Presence Teleoperators 9, 69-83.

Strösslin T (2004). *A Connectionnist Model of Spatial Learning in the Rat.* PhD thesis, EPFL, Swiss Federal Institute of Technology, Swiss.

Strösslin T and Gerstner W (2003). Reinforcement learning in continuous state and action space. In Artificial Neural Networks – ICANN 2003.

Strösslin T, Sheynikhovich D, Chavarriaga R and Gerstner W (2005). Robust self-localisation and navigation based on hippocampal place cells. *Neural Netw*, 18(9):1125-40.

Suri RE (2002). TD models of reward predictive responses in dopamine neurons. *Neural Netw*, 15(4-6):523-33.

Suri RE, Bargas J and Arbib MA (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103(1):65-85.

Suri RE and Schultz W (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp Brain Res*, 121(3):350-4.

Suri RE and Schultz W (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871-90.

Suri RE and Schultz W (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comput* 13: 841-62.

Sutherland RJ and Hamilton DA (2004). Rodent spatial navigation: at the crossroads of cognition and movement. *Neurosci Biobehav Rev*, 28(7):687-97.

Sutherland RJ and Rodriguez AJ (1989). The role of the fornix/fimbria and some related subcortical structures in place learning and memory. *Behav Brain Res*, 32(3):265-77.

Sutton RS (1988). Learning to Predict by the Methods of Temporal Differences. Machine Learning:9-44.

Sutton RS (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Seventh International Machine Learning Workshop*, pages 216-24. Morgan Kaufmann, San Mateo, CA.

Sutton RS (1997). http://www.cs.ualberta.ca/~sutton/book/6/node7.html.

Sutton RS and Barto AG (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Sutton RS, Barto AG and Williams RJ (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazine*, 12:19-22.

Tabuchi ET, Mulder AB and Wiener SI (2000). Position and behavioral modulation of synchronization of hippocampal and accumbens neuronal discharges in freely moving rats. *Hippocampus* 10: 717-28.

Tabuchi E, Mulder AB and Wiener SI (2003). Reward value invariant place responses and reward site associated activity in hippocampal neurons of behaving rats. *Hippocampus* 13: 117-132.

Taha SA and Fields HL (2005). Encoding of palatability and appetitive behaviors by distinct neuronal populations in the nucleus accumbens. *J Neurosci* 25: 1193-202.

Takikawa Y, Kawagoe R and Hikosaka O (2002). Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. *J Neurophysiol* 87: 508-15.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y and Yamawaki S (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci* 7: 887-93.

Tang B, Heywood MI and Shepherd M (2002). Input Partitioning to Mixture of Experts. In *International Joint Conference on Neural Networks*, pp. 227-32, Honolulu, Hawaii.

Tani, Arie H, Ogata T, Tani J, and Sugano S (2007). Reinforcement learning of a continuous motor sequence with hidden states. *Advanced Robotics*, Special Issue on Robotic Platforms for Research in Neuroscience, VSP and Robotics Society of Japan, 21(10):1215-29.

Tani J and Nolfi S (1999). Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-motor Systems. *Neural Networks*, 12(7-8):1131-41.

Tanji J and Hoshi E (2001). Behavioral planning in the prefrontal cortex. *Curr Opin Neurobiol*, 11:164-70.

Tanji J and Shima K (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature*, 371(6496):413-6.

Taube JS (1995). Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *Journal of Neuroscience*, 15(1):70-86.

Taube JS, Muller RU and Ranck Jr JB (1990a). Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *J Neurosci*, 10(2):436-47.

Taube JS, Muller RU and Ranck Jr JB (1990b). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J Neurosci*, 10(2):420-35.

Thierry AM, Gioanni Y, Degenetais E and Glowinski J (2000). Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus* 10: 411-9.

Thinus-Blanc C (1996). *Animal spatial cognition*. World Scientific Press, Singapore.

Thorndike EL (1911). Animal intelligence: Experimental studies. New York: Macmillan.

Tierney PL, Degenetais E, Thierry AM, Glowinski J and Gioanni Y (2004). Influence of the hippocampus on interneurons of the rat prefrontal cortex. *Eur J Neurosci*, 20(2):514-24.

Tolman EC (1948). Cognitive maps in rats and men. *Psych Rev,* 55:189–208.

Tremblay L, Hollerman JR and Schultz W (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J Neurophys* 80: 964-77.

Trobalon JB, Miguelez D, McLaren IPL and Mackintosh NJ (2003). Intradimensional and Extradimensional Shifts in Spatial Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 29(2):143-52.

Trullier O (1998). Elaboration et Traitement des Représentations Spatiales Servant à la Navigation chez le Rat. AnimatLab & LPPA-Collège de France, Université Pierre et Marie Curie, Paris, France.

Trullier O and Meyer J-A (1997). Place sequence learning for navigation. In W. Gerstner, A. Germond, M. Haasler and J.D. Nicoud, editors, Proceedings of the Seventh International Conference on Artificial Neural Networks, pages 757-62. Lausanne, Switzerland.

Trullier O, Wiener SI, Berthoz A and Meyer J-A (1997). Biologically based artificial navigation systems: review and prospects. *Prog Neurobiol*, 51(5):483-544.

Uchibe E and Doya K (2004). Competitive-cooperative-concurrent reinforcement learning with importance sampling. In Proceedings of International Conference on Simulation of Adaptive Behavior: From animals and animats (SAB2004) (pp. 287–96).

Uylings HB, Groenewegen HJ and Kolb B (2003). Do rats have a prefrontal cortex? *Behav Brain Res*, 146(1-2):3-17.

van Duuren E, Escamez FA, Joosten RN, Visser R, Mulder AB and Pennartz CM (2007). Neural coding of reward magnitude in the orbitofrontal cortex of the rat during a five-odor olfactory discrimination task. *Learn Mem*, 14(6):446-56.

Van Haaren F, de Bruin JPC, Heinsbroek RPW and Van de Poll NE (1985). Delayed spatial reponse alternation: Effects of delay-interval duration and lesions of the medial prefrontal cortex on response accuracy of male and female Wistar rats. *Behav Brain Res*, 18:481-8.

Vertes RP (2006). Interactions among the medial prefrontal cortex, hippocampus and midline thalamus in emotional and cognitive processing in the rat. *Neuroscience*, 142(1):1-20.

Voorn P, Vanderschuren LJ, Groenewegen HJ, Robbins TW and Pennartz CM (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci*, 27(8):468-74.

Watanabe M (1996). Reward expectancy in primate prefrontal neurons. *Nature*, 382(6592):629-32.

Watanabe M, Hikosaka K, Sakagami M and Shirakawa S (2007). Reward expectancy-related prefrontal neuronal activities: are they neural substrates of "affective" working memory? *Cortex*, 43(1):53-64.

Watkins CJCH (1989). Learning from delayed rewards. PhD thesis, Cambridge University.

Watkins CJCH and Dayan P (1992). Technical Note: Q-Learning. *Machine Learning*, 8(3-4):279-92.

Watson J (1913). Psychology as the Behaviorist Views it. *Psychological Review*, 20, 158-77.

Webb B and Consi TR (Eds) (2001). Biorobotics. Using biorobotics to explore animal behavior and brain function.

Whishaw IQ and Mittleman G (1986). Visits to starts, routes, and places by rats (Rattus norvegicus) in swimming pool navigation tasks. *J Comp Psychol*, 100(4):422-31.

White NM (1997). Mnemonic functions of the basal ganglia. *Current Opinion in Neurobiology*, 7:164–9.

Wickens J (1997). Basal ganglia: Structure and computations. *Network: Computation in Neural Systems*, 8:77-109.

Wickens JR, Budd CS, Hyland BI and Arbuthnott GW (2007a). Striatal contributions to reward and decision making: making sense of regional variations in a reiterated processing matrix. *Ann N Y Acad Sci*, 1104:192-212.

Wickens JR, Horvitz JC, Costa RM and Killcross S (2007b). Dopaminergic mechanisms in actions and habits. *J Neurosci*, 27(31):8181-3.

Wickens J and Kötter R (1995). Cellular models of reinforcement learning. In: *Models of information processing in the basal ganglia,* edited by Houk JC, Davis JL, Beiser D, pp 187-214, Cambridge, MA: MIT Press.

Wiener SI (1993). Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J Neurosci* 13: 3802-17.

Wiener SI (1996). Spatial, behavioral and sensory correlates of hippocampal CA1 complex spike cell activity: implications for information processing functions. *Prog Neurobiol*, 49(4):335-61.

Wiener SI, Paul CA and Eichenbaum H (1989). Spatial and behavioral correlates of hippocampal neuronal activity. *J Neurosci*, 9(8):2737-63.

Wiener S and Schenk F (2005). Behavioral studies of directional orientation in developing and adult animals. In: Taube J, Wiener S. ed. *Head direction cells and the neural mechanisms underlying directional orientation*. pp. 247-74, MIT Press.

Wiener SI, Shibata R, Tabuchi E, Trullier O, Albertin SV and Mulder AB (2003) Spatial and behavioral correlates in nucleus accumbens neurons in zones receiving hippocampal or prefrontal cortical inputs. In T. Ono et al, (eds), *Cognition and Emotion in the Brain* , Elsevier, NY.

Williams GV, Rolls ET, Leonard CM and Stern C (1993). Neuronal responses in the ventral striatum of the behaving macaque. *Behav Brain Res*, 55(2):243-52.

Willingham DB, (1998). What differentiates declarative and procedural memories: Reply to Cohen, Poldrack, and Eichenbaum (1997). *Memory*, 6 (6): 689-99.

Wilson SW (1991). The Animat Path to AI. Proceedings of the first international conference on simulation of adaptive behavior (From animals to animats), Pages: 2-14, Paris, France, MIT Press.

Wilson DI and Bowman EM (2004). Nucleus accumbens neurons in the rat exhibit differential activity to conditioned reinforcers and primary reinforcers within a second-order schedule of saccharin reinforcement. *Eur J Neurosci,* 20: 2777-88.

Wilson DI and Bowman EM (2005). Rat nucleus accumbens neurons predominantly respond to the outcome-related properties of conditioned stimuli rather than their behavioral-switching properties. *J Neurophysiol,* 94: 49-61.

Winer BJ (1971). *Statistical Principles in Experimental Design. 2nd ed*. New York: McGraw-Hill.

Yang CR and Mogenson GJ (1984). Electrophysiological responses of neurones in the accumbens nucleus to hippocampal stimulation and the attenuation of the excitatory responses by the mesolimbic dopaminergic system. *Brain Res,* 324: 69-84.

Yeshenko O, Guazzelli A and Mizumori SJ (2004). Context-dependent reorganization of spatial and movement representations by simultaneously recorded hippocampal and striatal neurons during performance of allocentric and egocentric tasks. *Behav Neurosci*, 118(4):751-69.

Yin HH and Knowlton BJ (2004). Contributions of striatal subregions to place and response learning. *Learn Mem*, 11(4):459-63.

Yin HH and Knowlton BJ (2006). The role of the basal ganglia in habit formation. *Nat Rev Neurosci*, 7(6):464-76.

Yin HH, Knowlton BJ and Balleine BW (2004). Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning. *European Journal of Neuroscience*, 19(1):181-9.

Yin HH, Knowlton BJ and Balleine BW (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur J Neurosci*, 22(2):505-12.

Yin HH, Ostlund SB, Knowlton BJ and Balleine BW (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci*, 22(2):513-23.

Zahm DS and Brog JS (1992). On the significance of subterritories in the "accumbens" part of the rat ventral striatum. *Neuroscience*, 50(4):751-67.

Ziemke T (2007). What's life got to do with it? In: Chella and Manzotti (eds.). Artificial Consciousness (pp. 48-66). Exeter: Imprint Academic.

Ziemke T (2005). Cybernetics and Embodied Cognition: On the Construction of Realities in Organisms and Robots. *Kybernetes*, 34(1/2), 118-28.

Zugaro MB, Arleo A, Dejean C, Burguiere E, Khamassi M and Wiener SI (2004). Rat anterodorsal thalamic head direction neurons depend upon dynamic visual signals to select anchoring landmark cues. *Eur J Neurosci*, 20(2):530-6.

Zugaro MB, Berthoz A and Wiener SI (2001). Background, but not foreground, spatial cues are taken as references for head direction responses by rat anterodorsal thalamus neurons. *J Neurosci*, 21(14):RC154.