Université de Montréal

Modèle informatique du coapprentissage des ganglions de la base et du cortex :

L'apprentissage par renforcement et le développement de représentations

par

François Rivest

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences

en vue de l'obtention du grade Ph.D.

en Informatique

Décembre 2009

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée :

Modèle informatique du coapprentissage des ganglions de la base et du cortex : L'apprentissage par renforcement et le développement de représentations

Présentée par :
François Rivest

a été évaluée par un jury composé des personnes suivantes :

Pascal Vincent, président-rapporteur
Yoshua Bengio, directeur de recherche
John Francis Kalaska, co-directeur de recherche
Paul Cisek, membre du jury
Richard Courtemanche, examinateur externe
Pierre Rainville, représentant du doyen de la FAS

Résumé

Tout au long de la vie, le cerveau développe des représentations de son environnement permettant à l'individu d'en tirer meilleur profit. Comment ces représentations se développent-elles pendant la quête de récompenses demeure un mystère. Il est raisonnable de penser que le cortex est le siège de ces représentations et que les ganglions de la base jouent un rôle important dans la maximisation des récompenses. En particulier, les neurones dopaminergiques semblent coder un signal d'erreur de prédiction de récompense. Cette thèse étudie le problème en construisant, à l'aide de l'apprentissage machine, un modèle informatique intégrant de nombreuses évidences neurologiques.

Après une introduction au cadre mathématique et à quelques algorithmes de l'apprentissage machine, un survol de l'apprentissage en psychologie et en neuroscience et une revue des modèles de l'apprentissage dans les ganglions de la base, la thèse comporte trois articles. Le premier montre qu'il est possible d'apprendre à maximiser ses récompenses tout en développant de meilleures représentations des entrées. Le second article porte sur l'important problème toujours non résolu de la représentation du temps. Il démontre qu'une représentation du temps peut être acquise automatiquement dans un réseau de neurones artificiels faisant office de mémoire de travail. La représentation développée par le modèle ressemble beaucoup à l'activité de neurones corticaux dans des tâches similaires. De plus, le modèle montre que l'utilisation du signal d'erreur de récompense peut accélérer la construction de ces représentations temporelles. Finalement, il montre qu'une telle représentation acquise automatiquement dans le cortex peut fournir l'information nécessaire aux ganglions de la base pour expliquer le signal dopaminergique. Enfin, le troisième article évalue le pouvoir explicatif et prédictif du modèle sur différentes situations comme la présence ou l'absence d'un stimulus (conditionnement classique ou de trace) pendant l'attente de la récompense. En plus de faire des prédictions très intéressantes en lien avec la littérature sur les intervalles de temps, l'article révèle certaines lacunes du modèle qui devront être améliorées.

Bref, cette thèse étend les modèles actuels de l'apprentissage des ganglions de la base et du système dopaminergique au développement concurrent de représentations temporelles dans le cortex et aux interactions de ces deux structures.

Mots-clés : Apprentissage par renforcement, apprentissage par différence temporelle, conditionnement classique, conditionnement de trace, cortex, dopamine, ganglions de la base, intervalle de temps, neuroscience informatique, représentation abstraite.

Abstract

Throughout lifetime, the brain develops abstract representations of its environment that allow the individual to maximize his benefits. How these representations are developed while trying to acquire rewards remains a mystery. It is reasonable to assume that these representations arise in the cortex and that the basal ganglia are playing an important role in reward maximization. In particular, dopaminergic neurons appear to code a reward prediction error signal. This thesis studies the problem by constructing, using machine learning tools, a computational model that incorporates a number of relevant neurophysiological findings.

After an introduction to the machine learning framework and to some of its algorithms, an overview of learning in psychology and neuroscience, and a review of models of learning in the basal ganglia, the thesis comprises three papers. The first article shows that it is possible to learn a better representation of the inputs while learning to maximize reward. The second paper addresses the important and still unresolved problem of the representation of time in the brain. The paper shows that a time representation can be acquired automatically in an artificial neural network acting like a working memory. The representation learned by the model closely resembles the activity of cortical neurons in similar tasks. Moreover, the model shows that the reward prediction error signal could accelerate the development of the temporal representation. Finally, it shows that if such a learned representation exists in the cortex, it could provide the necessary information to the basal ganglia to explain the dopaminergic signal. The third article evaluates the explanatory and predictive power of the model on the effects of differences in task conditions such as the presence or absence of a stimulus (classical versus trace conditioning) while waiting for the reward. Beyond making interesting predictions relevant to the timing literature, the paper reveals some shortcomings of the model that will need to be resolved.

In summary, this thesis extends current models of reinforcement learning of the basal ganglia and the dopaminergic system to the concurrent development of representation in the cortex and to the interactions between these two regions.

Keywords: Reinforcement learning, temporal-difference learning, classical conditioning, trace conditioning, cortex, dopamine, basal ganglia, interval timing, computational neuroscience, abstract representation.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

Abréviations informatiques et mathématiques:

ANN	Réseau de neurones artificiels	*Artificial neural networks*
ANOVA	Analyse de la variance	*Analysis of variance*
BP	Rétropropagation	*Backpropagation*
BPTT	Rétropropagation dans le temps	*Backpropagation through time*
ICA	Analyse en composantes indépendantes	*Independent component analysis*
LHS	Coté droit	*Left hand side*
LSTM	Longue mémoire a court terme	*Long short-term memory*
PCA	Analyse en composantes principales	*Principal component analysis*
PDP	Traitement parallèle distribué	*Parallel distributed processing*
RHS	Coté gauche	*Right hand side*
RL	Apprentissage par renforcement	*Reinforcement learning*
RTRL	Apprentissage récursif en temps réel	*Real-time recurrent learning*
SARSA	État-action-récompense-état-action	*State-action-reward-state-action*
TD	Différence temporelle	*Temporal-difference*
TDNN	Réseau de neurones à lignes de délai	*Time-delay neural network*

Abréviations neurologiques et psychologiques:

BG	Ganglions de la base	*Basal ganglia*
CS	Stimulus conditionné	*Conditioned stimulus*
CD	Noyau caudé	*Caudate nucleus*
DA	Dopaminergique	*Dopaminergic*
FI	Intervalle fixe	*Fixed-interval*
fMRI	Imagerie fonctionnelle par résonance magnétique	*Functional magnetic resonance imagery*
GP	Globus pallidus	*Globus pallidus*
GPe	Segment externe du globus pallidus	*Globus pallidus external segment*
GPi	Segment interne du globus pallidus	*Globus pallidus internal segment*
ISI	Intervalle interstimuli	*Interstimuli interval*

ITI	Intervalle interessais	*Intertrial interval*
LET	Théorie d'apprendre à chronométrer	*Learning to time theory*
LHA	Hypothalamus latéral	*Lateral hypothalamus*
LTD	Dépression à long terme	*Long term depression*
LTP	Potentialisation à long terme	*Long term potentiation*
MSN	Neurone moyen épineux	*Medium spiny neuron*
NAc	Noyau accumbens	*Nucleus accumbens*
PFC	Cortex prefrontal	*Prefrontal cortex*
PI	Intervalle de pointe	*Peak-interval*
PPTN/ PPTg	Noyau tegmental pédonculopontine	*Pedonculopontine tegmental nucleus*
PT	Putamen	*Putamen*
SBF	Fréquence de battement au striatum	*Striatal beat frequency*
SET	Théorie d'espérance scalaire	*Scalar expectancy theory*
SOM	Carte autoorganisée	*Self-organizing map*
SN	Substance noire	*Substantia nigra*
SNc	Substance noire compacte	*Substantia nigra pars compacta*
SNr	Substance noire *pars reticulata*	*Substantia nigra pars reticulata*
STN	Noyau sous-thalamique	*Subthalamic nucleus*
TAN	Neurone actif toniquement	*Tonically active neuron*
US	Stimulus non conditionné	*Unconditioned stimulus*
VTA	Aire ventrale tegmentale	*Ventral tegmental area*

Sigles d'organismes et de conférences:

FRSQ	Fonds de la recherche en santé Québec	
IRSC/ CIHR	Instituts de recherche en santé du Canada	*Canadian Institutes of Health Research*
NIPS	Systèmes neuraux de traitement d'information	*Neural Information Processing Systems*

Notation mathématique

D'une façon générale, la notation mathématique suivante sera utilisée :

v,w,θ,i,j,k,l,m,n	Les lettres minuscules sont des scalaires réels ou entiers.
$\boldsymbol{w}$	Les lettres minuscules grasses sont des vecteurs réels.
W, A	Les lettres majuscules sont des matrices réelles ou des ensembles.
$w_i, \boldsymbol{w_i}$	Un indice peut indiquer la position d'un scalaire dans un vecteur, un ensemble ou une séquence ou d'un vecteur dans une matrice, un ensemble ou une séquence.
$w_{i,t}$	S'il y a deux indices sur un scalaire, le premier indique la position dans le vecteur, et le deuxième la position du vecteur dans l'ensemble ou la séquence.
M,N,K,L	Les petites majuscules sont des scalaires entiers constants.
$f(),g(),h(), \pi()$	Les lettres minuscules avec parenthèses sont des fonctions scalaires.
$\boldsymbol{f()}$	Les lettres minuscules grasses avec parenthèses sont des fonctions vectorielles.
$E()$	Certaines lettres majuscules peuvent être des fonctions.
$f'()$	$'$ indique la dérivée de f.
$df()/du$	d/d indique la dérivée de f par rapport à u.
$\nabla_{\boldsymbol{u}} f$	∇ indique le gradient de f par rapport au vecteur $\boldsymbol{u}$.
$\partial \boldsymbol{f} / \partial x$	∂/∂ indique la dérivée partielle de $\boldsymbol{f}$ par rapport à x.
$E\{X\}$	$E\{\}$ signifie espérance de la variable aléatoire X.
$P(X)$	$P()$ signifie probabilité de l'évènement X.

À mes parents, ma femme et mes enfants

Pour leur soutien quotidien sans lequel cette entreprise aurait été impossible.

Remerciements

Je tiens tout d'abord à remercier mes directeurs de recherche Yoshua Bengio et John F. Kalaska. Ils m'ont permis d'entreprendre un projet multidisciplinaire ambitieux et m'ont supporté tout le temps nécessaire à l'atteinte de mes objectifs de recherche. Ils m'ont enseigné l'art de la recherche dans leur domaine respectif en plus de mettre à ma disposition les ressources nécessaires à mon projet. Ils m'ont donné toute la latitude d'étudier les problèmes qui m'intéressaient tout en me fournissant le support, la supervision et les outils nécessaires à mener à bien mes recherches. Je remercie particulièrement John qui m'a permis de devenir un chercheur en neuroscience alors que j'étais un étudiant en informatique. Il m'a enseigné par ses généreuses corrections, questions et commentaires, à rejoindre ce domaine de recherche.

J'aimerais aussi remercier mes précieux collaborateurs, Thomas R. Shultz, Doina Precup, Frédéric Dandurand, Jean-Philippe Thivierge, Marc G. Bellemare, Ouri Monchi et László Egri, sans qui plusieurs de mes publications (en Annexe IV) n'auraient jamais vue le jour. Leur contribution à mon travail dépasse largement les articles sur lesquels nous avons collaboré.

Mes recherches ont profité des discussions inspirantes avec mes collègues étudiants et professeurs du *Laboratoire d'informatique des systèmes adaptatifs* dont Douglas Eck, Balázs Kégl, Aaron Courville, Pascal Lamblin, Nicolas Chapados et James Bergstra, du groupe de discussions *Math-Neuro* dont Paul Cisek, Andrea Green, Sergiy Yakovenko, Pascal Fortier-Poisson, Valeriya Gritsenko, Christine Desmarais et Alexandre Pastor-Bernier, ainsi qu'avec Peter Shizgal de l'Université Concordia. Je remercie également Trevor Drew et le *Groupe de recherche sur le système nerveux central* qui par leurs séminaires, retraites, et autres activités m'ont permis de pleinement découvrir les neurosciences et d'y parfaire ma formation dans ce domaine. Ce groupe de recherche est une vraie mine d'or pour quiconque souhaite faire de la recherche en neuroscience. Les étudiants gagneraient beaucoup à mieux le connaître.

Je tiens à remercier Christine Desmarais, Pascal Lamblin, Elliot Ludvig, Maxime Lévesque, Constant Rivest, Nadia Gosselin-Kessiby, James Bergstra et Patrick Mercier ainsi que les évaluateurs anonymes, qui ont lu et commenté différents chapitres de cette thèse. Je suis particulièrement reconnaissant à Frédéric Dandurand et Pascal Fortier-Poisson qui, en plus de mes directeurs de recherche, ont généreusement accepté de servir de cobaye et qui ont lu et commenté la quasi-totalité du manuscrit. Écrire une thèse accessible aux étudiants et professeurs de deux domaines aussi différents que l'informatique et les neurosciences n'est pas une tâche simple. Je n'aurais pu réussir cet ouvrage sans leur aide. Je remercie également Marie-Claire Rivest et Nadine Michaud pour leur aide lors de la préparation de nombreuses figures et affiches. Merci aussi aux auteurs et éditeurs qui m'ont autorisé à reproduire certaines de leurs figures. Leurs contributions ont permis de rendre cette thèse plus compréhensible. Je suis cependant responsable pour toutes erreurs qui pourraient subsister.

J'ai aussi eu la chance d'être supporté financièrement pendant mes études doctorales par une bourse étudiante au Ph.D de *l'Équipe en voie de développement en neuroscience computationnelle* des IRSC ainsi que par une bourse pour projet collaboratif du *Groupe de recherche sur le système nerveux central* des FRSQ. Mes recherches ont aussi profité de la grappe de calcul du *Laboratoire d'informatique des systèmes adaptatifs*, du financement de la *Chaire de Recherche du Canada sur le Algorithmes d'Apprentissage Statistique* de Yoshua Bengio et des fonds d'opérations des IRSC de John Kalaska.

Finalement, cet ouvrage n'aurait jamais vu le jour sans les encouragements incessants de ma femme Marie-Claude, le support inconditionnel de mes parents, ainsi que la compréhension et les encouragements de mes amis et membres de ma famille qui ont toujours cru en mes aspirations.

Mille fois merci

Avant-propos

Cette thèse a pour thème l'étude du développement de représentations abstraites de l'environnement pendant l'apprentissage par renforcement. L'apprentissage par renforcement (par récompenses), bien que le nom d'un sous-domaine de l'intelligence artificielle, n'est pas d'origine informatique. Il trouve racine dans le conditionnement classique et instrumental en psychologie. Ce thème est le sujet d'étude de plusieurs disciplines, des neurosciences, jusqu'à l'économie. Afin de situer cet ouvrage dans son contexte et de bien définir son orientation, il est nécessaire d'expliquer le contexte multidisciplinaire de celle-ci ainsi que de bien identifier la différence entre tenter de créer un système intelligent (capable d'apprendre) et tenter de comprendre l'intelligence naturelle.

L'approche multidisciplinaire…

Alors que pendant longtemps les sciences se sont tranquillement divisées et spécialisées, l'arrivée de nouvelles technologies a aussi favorisé le mouvement inverse, le recoupement de disciplines aux approches différentes intéressées par des thèmes communs et pouvant s'entre-aider. La bio-informatique, qui consiste principalement à l'utilisation de l'informatique pour résoudre des problèmes dans l'étude de la biologie, et l'informatique linguistique, qui consiste au traitement automatique du langage humain, en sont des exemples. Les neurosciences sont un regroupement de biologistes, physiologistes, psychologues et autres, qui ont tous en commun l'étude du système nerveux à différents niveaux. La conférence nord-américaine de la *Society for Neuroscience* réunit chaque année plus de 20 000 scientifiques (selon ses statistiques). Similairement, les sciences cognitives, qui ont pour but l'étude de la pensée, regroupent des domaines aussi variés que la linguistique, la psychologie, la philosophie, l'informatique (dont l'intelligence artificielle) et les neurosciences. D'ailleurs, plusieurs disciplines, comme les neurosciences, profitent chaque jour des développements technologiques effectués dans d'autres disciplines. L'imagerie fonctionnelle par résonance magnétique, issue de la physique et en quête de meilleurs outils statistiques en est un exemple. Au fur et à mesure que les technologies se développent et que de nouvelles possibilités

s’ouvrent aux scientifiques, les besoins en expertise pour adapter ces technologies à leur champ d’études croissent.

C'est dans ce contexte d’enrichissement interdisciplinaire continu que les travaux de cette thèse ont évolué. Le thème de l'apprentissage par renforcement en est un pour lequel les chercheurs en psychologie, neuroscience, recherche opérationnelle, intelligence artificielle, robotiques et même finance ont réussi à communiquer et ainsi faire fructifier leurs idées en mettant leurs questions et leurs réponses en commun, comme font foi le *Multidisciplinary Symposium on Reinforcement Learning* qui a eu lieu à l’été 2009 à Montréal, ainsi que l’étendue de l’apprentissage par renforcement dans les différentes littératures.

Lorsque nous rencontrons les questions d'une autre discipline, il arrive parfois que notre formation nous suggère des solutions. Un tel partage d'information interdisciplinaire ne peut être que fructueux. Cependant, les objectifs, les questions, le contexte et les contraintes varient grandement d'une discipline à l'autre. Pire encore, le vocabulaire, à la limite le langage lui-même, est différent. La solution évidente des uns est parfois rejetée sans appel par les autres, à tort ou à raison, et c'est normal. Il est facile de douter du pouvoir des relations interdisciplinaires face à ces difficultés. Les chercheurs à cheval sur des domaines différents vont généralement dire que leurs présentations seront totalement différentes en fonction de leur auditoire, selon qu’'ils préparent une présentation devant les gens de la discipline A ou B; une démonstration évidente des difficultés de communication inhérentes au haut niveau d'avancement de chaque discipline. Il ne faut donc pas baisser les bras; que nous voulions apprendre d'une autre discipline, ou que nous souhaitions y contribuer, il faut être attentif aux différences de perspectives.

Construire, déconstruire ou reconstruire… quelle différence?

Bien que le sujet d’intérêt soit le même, la principale différence entre les chercheurs en intelligence artificielle et ceux en neuroscience et en science cognitive réside dans leurs objectifs. Ils ne poursuivent pas les mêmes buts, et par conséquent, ils ont des méthodologies et des critères d’évaluations bien différents. Cette section

fait un survol des différences entre ces deux approches au problème de l'apprentissage.

Robotique, intelligence artificielle et apprentissage machine

Dans ces disciplines, l'objectif est d'inventer ou de construire une machine capable de résoudre un problème. Le robot doit pouvoir se rendre du point A au point B, l'ordinateur doit pouvoir jouer aux échecs et gagner, ou le programme doit pouvoir reconnaitre des objets dans une image, etc.

La solution doit-elle ressembler à une solution qu'une espèce vivante a développée comme une aile, un œil ou un neurone? A priori, non! Par contre, il peut être très utile de s'en inspirer. Plusieurs chercheurs en intelligence artificielle le font et l'apprentissage par renforcement en est un exemple (voir section 4.4.1). Toutefois, l'objectif, c'est que la solution marche et soit la plus performante possible! On peut mesurer la performance des solutions et les comparer. Le robot réussit-il à aller du point A au point B, en combien de temps, en consommant combien d'énergie? L'ordinateur bat-il Kasparov? Combien de codes postaux le programme reconnait-il? Construirions-nous un avion qui bat des ailes comme un oiseau? Non! Considérant les moyens à notre disposition, ce n'est pas performant. Le contexte et les contraintes ne sont pas les mêmes que pour l'oiseau. Une bonne solution est une solution qui marche, et qui répond d'abord à des critères de performance : économie, durabilité, taux d'erreurs, etc.

La définition même de « *simplicité* » (principe du *rasoir d'Ocam*) souvent utilisée en science nous encouragerait tout simplement à minimiser le nombre d'éléments utilisés. La règle la plus simple; le moins de systèmes que possible. En informatique, on regarde aussi le temps de calcul ou la mémoire requise. Moins de temps de calcul et moins de mémoire nécessaires sont des signes d'un meilleur algorithme. On parle de *complexité de calcul* pour parler du temps de calcul d'un algorithme ou de la complexité d'un problème à résoudre. Cependant, pour un algorithme, le mot complexité est ici synonyme de lenteur. Un algorithme basé sur une idée simple et facile à comprendre devra souvent effectuer beaucoup plus de calculs pour trouver la solution à un problème, il aura donc une grande complexité

calculatoire. À l'opposé, un algorithme qui résoudra le problème en moins d'opérations ou en prenant moins de mémoire devra souvent être plus ingénieux, et peut s'avérer très compliqué et difficile à analyser ou à comprendre.

Bref, la conception d'un système intelligent n'est contrainte par rien d'autre que notre capacité à les inventer et les limites technologiques. En principe, on choisit ce qui fonctionne le mieux, que le design résultant ressemble à un produit de la nature ou non. Cependant, la nature, et plus particulièrement les connaissances sur le fonctionnement du cerveau, sont une source importante d'inspiration pour plusieurs chercheurs de ce domaine.

Modélisation en science cognitive et neuroscience

Lorsque l'on fait de la modélisation en science cognitive, en neuroscience ou dans toute autre discipline scientifique étudiant des phénomènes naturels, l'objectif premier est de mieux comprendre le phénomène en question. Comment le lapin se rappelle-t-il du chemin entre le point A et B, pourquoi prend-il toujours le même chemin? Comment l'enfant apprend-il à parler? Comment reconnait-on les objets sur une photo? Pourquoi sommes-nous meilleurs dans certaines situations que dans d'autres?

Construire un modèle n'est pas un objectif en soi. Construire un modèle, c'est construire un outil pour nous aider à élucider le phénomène en question. Les critères d'évaluation d'un modèle de la reconnaissance d'images ne sont donc pas du tout les mêmes que les critères d'évaluation d'un algorithme de reconnaissance d'images, même si à première vue, ils font la même chose. Alors que l'algorithme doit reconnaître les images avec le moins d'erreurs possibles, le modèle doit plutôt reproduire les performances du système biologique qu'il tente d'expliquer. Il doit donc préférablement faire les mêmes erreurs, dans les mêmes situations. Mais la qualité de reproduction d'un phénomène donné n'est pas le seul critère d'évaluation d'un modèle. Un bon modèle est un modèle qui a aussi un bon *pouvoir explicatif*, mesure beaucoup plus subjective que le nombre d'erreurs sur un ensemble donné. Ici, « *simple* » veut souvent dire *facile à comprendre*. L'évaluation d'un modèle passe donc souvent par son pouvoir de prédiction. Dans quelle mesure le modèle me

permet-il d'imaginer ou de prédire le phénomène étudié sous différentes conditions. On effectue des prédictions à l'aide du modèle qui pourront par la suite être validées ou invalidées par des expériences scientifiques.

Par exemple, trois modèles peuvent représenter trois règles différentes ayant un sens pour le chercheur (trois hypothèses). On peut évaluer leur performance en regardant leur capacité à reproduire les données existantes. Peut-être seules deux règles sembleront suffisamment bonnes. Pour les évaluer encore mieux, on effectue des prédictions à l'aide des modèles pour trouver des situations où ils diffèrent. On peut finalement effectuer une expérience pour voir quel modèle colle mieux aux nouvelles données. L'un d'eux en sortira peut-être gagnant. Supposons maintenant que l'on veut mieux comprendre comment s'intègrent certains éléments biologiques constitutifs du phénomène. On peut aussi supposer un quatrième modèle, aussi performant quantitativement que le précédent, mais fait de deux ou trois règles au lieu d'une seule. Ces nouvelles règles devront représenter ces éléments constitutifs du système naturel étudié. Ce nouveau modèle permet d'analyser indépendamment ces éléments et leurs interactions et à ce niveau d'étude, il est meilleur que le précédent. De même que la mécanique newtonienne nous permet d'étudier les phénomènes macroscopiques, la mécanique quantique devient un outil préférable lorsque vient le temps d'étudier les particules plus petites et leurs effets les unes sur les autres.

La règle de la simplicité veut aussi dire inventer ou supposer le moins de processus que possible. Il est a priori préférable de partir des éléments dont on connait déjà l'existence, que d'en inventer de nouveau pour des raisons purement algorithmiques. Moins on doit inventer de nouveaux éléments, mieux c'est. C'est lorsque nous n'avons plus d'autres choix que de supposer l'existence de nouveaux éléments constitutifs que l'on fait les plus grandes découvertes, ou que l'on découvre simplement combien le modèle est erroné. Le modèle se veut une simplification de la réalité ayant pour objectif une meilleure compréhension de celle-ci.

Position et orientation de cette thèse

Les recherches présentées dans cette thèse ont débuté avec l'objectif de développer de nouveaux algorithmes d'apprentissage machine. Plus particulièrement,

l'objectif était de trouver une façon de permettre à un système d'apprentissage, que l'on guide à l'aide de récompenses et de punitions, de développer automatiquement une représentation abstraite, riche et stable de son environnement. L'idée générale était de s'inspirer de ce que l'on connait du cerveau et cela a conduit au premier article de cette thèse (Chapitre 6). Certaines bases de l'apprentissage par renforcement dans le cerveau sont connues des neurosciences. Cependant, les fondements mathématiques de sa grande capacité de développement d'abstractions demeurent mal compris. Pour cette raison, et bien d'autres, ce projet de recherche s'est transformé en un projet de modélisation de l'apprentissage dans le cerveau. L'objectif est donc devenu de mieux comprendre comment le cerveau développe des représentations abstraites dans un contexte d'apprentissage par récompense à l'aide du cadre mathématique et des outils fournit par l'apprentissage machine.

En cours de route, il est devenu évident que la représentation du passage du temps par le cerveau et sa position dans l'apprentissage était un problème non résolu et d'une grande importance pour pouvoir aller plus loin dans la compréhension de l'apprentissage du cerveau. Le sujet de recherche qui se trouvait donc au départ à cheval sur deux disciplines, soit l'apprentissage machine et les neurosciences de l'apprentissage, c'est tout à coup retrouvé au cœur d'un troisième domaine de recherche, le temps, sujet pour lequel une grande partie des connaissances proviennent encore de la psychologie du comportement. Comme cette thèse en est une d'informatique, utilisant cette dernière pour les neurosciences, et couvrant un large éventail de connaissances, deux chapitres intermédiaires complètent l'introduction et la revue de littérature traditionnelle. Le Chapitre 2 couvre principalement la partie informatique et l'apprentissage machine, alors que le Chapitre 3 couvre les bases générales de la psychologie et des neurosciences de l'apprentissage. Une attention particulière a été apportée pour tenter de rendre cette thèse accessible aux chercheurs de ces différentes disciplines.

Chapitre 1. Introduction

1.1 Description du problème

Depuis les débuts de l'informatique, il y a un peu plus d'une cinquantaine d'années, la possibilité d'une intelligence artificielle a fasciné et captivé les scientifiques de toutes disciplines (Russell & Norvig, 1995). Encore aujourd'hui, la question se pose toujours : construirons-nous un jour un être intelligent? Un tel projet implique la résolution de nombreux problèmes : perception de l'environnement, représentation des connaissances, raisonnement logique, utilisation d'un langage, etc. De tous ces problèmes, la capacité généralisée d'apprentissage est sûrement l'une des aptitudes les plus complexes et les plus difficiles à recréer dans un système qui se rapproche de l'humain en termes d'intelligence.

Mais comment l'être humain apprend-il? Dès son plus jeune âge, avant même qu'il soit né, l'enfant commence déjà à apprendre (Vasta, Haith, & Miller, 1999). Il commence à apprendre à reconnaître la voix de sa mère (DeCasper & Fifer, 1980) et sa langue maternelle (Mehler et al., 1988) par exemple. Après sa naissance, il doit apprendre à décomposer ce qu'il goûte, sent, touche, voit et entend. Ses yeux à eux seuls envoient à son cerveau un flux de données continu de plus de 20 mégaoctets par seconde (estimation très minimale[1]). Dès la première année, il devra apprendre à se servir de ses bras et de ses mains, puis à marcher. Chez l'humain, le positionnement de la main dans l'espace est un problème à sept degrés de liberté : trois pour l'épaule, deux pour le coude et deux pour le poignet (Rosenbaum, 1991; Zipser & Torres, 2007). Si l'on inclut les doigts, pour prendre un objet par exemple, la cinématique du bras pourrait compter plus de 22 degrés de liberté (Cerveri, Lopomo, Pedotti, & Ferrigno, 2005). L'enfant apprendra à se servir de ses membres à partir des capteurs internes (propriocepteurs) et cutanés et de ce qu'il voit, en présumant qu'il apprendra aussi à reconnaître ce qu'il voit (« Ah, ma main! »). Il devra aussi apprendre à utiliser

[1] Il y aurait entre 1.0 et 1.4 million de fibres nerveuses pour chaque œil. La fréquence de décharge d'un neurone peut aller jusqu'à 100 Hz, mais la capacité de discrimination temporelle (ou résolution temporelle) visuelle est d'environ 10 Hz. L'estimation ci-dessus est basée sur une résolution d'intensité de 1 bit à une fréquence d'échantillonnage du taux de décharge à 10 Hz, on aurait facilement pu en compter plus (Kandel, Schwartz, & Jessell, 2000).

son appareil vocal pour parler une langue, une abstraction complexe et infinie (Fromkin, Rodman, Hultin, & Logan, 1997) qu'il devra d'abord décortiquer et comprendre à partir de ce qu'il aura vu et entendu au cours de sa première année de vie.

Comment un enfant est-il capable d'apprendre tout cela? Notre compréhension actuelle de l'apprentissage chez l'humain n'est malheureusement pas beaucoup plus avancée que notre technologie pour fabriquer une telle capacité artificiellement. Il est généralement admis que les modifications des connexions synaptiques entre les neurones sont à la base de l'apprentissage (Teyler & DiScenna, 1984; Kandel et al., 2000). Toutefois, savoir ce qu'est un transistor ne nous révèle pas davantage toute la complexité d'un processeur multicœur, ni comment l'on passe de l'électricité au réalisme des jeux vidéos d'aujourd'hui. L'objectif de cette thèse est donc d'essayer de mieux comprendre l'apprentissage chez l'humain, ou du moins chez les mammifères, à l'aide des outils et des cadres théoriques développés en informatique et mathématiques. En particulier, comment développe-t-on des représentations abstraites dans un contexte d'apprentissage par récompense?

1.2 Justification de l'approche

Il peut sembler naïf de vouloir utiliser des théories développées pour l'apprentissage des machines pour mieux comprendre l'apprentissage dans le cerveau, alors que ces méthodes ne semblent pas arriver à la cheville des aptitudes humaines. Cependant, il faut comprendre que pour inventer l'apprentissage artificiel, les chercheurs ont dû définir mathématiquement le cadre du problème de l'apprentissage et ses limites. Plusieurs solutions à de petits problèmes ont aussi été trouvées. Les avantages, désavantages et limites de ces solutions sont maintenant connus. Le plus bel exemple démontrant bien l'utilité de cette approche est l'avènement de l'apprentissage par renforcement en apprentissage machine. À l'origine, Barto et Sutton (1982) voulaient créer un modèle de l'apprentissage chez les animaux à partir de l'étude de leur comportement. Ils établirent précisément le cadre théorique du problème d'apprentissage et développèrent un certain nombre de solutions au cœur desquelles un signal revenait presque toujours : le signal d'erreur de prédiction des

récompenses. Ces algorithmes ont été développés dans le contexte de l'apprentissage machine sur plusieurs années et devinrent un champ de recherche en soi : l'apprentissage machine par renforcement. Mais, ils ont aussi permis de modéliser un certain nombre de comportements animaux, entre autres en conditionnement classique ou pavlovien (Sutton & Barto, 1990). Près de quinze ans plus tard, des chercheurs en neuroscience qui enregistraient l'activité électrique de neurones chez le singe découvrirent des neurones qui semblaient encoder exactement ce signal d'erreur de prédiction de récompenses (Schultz, Dayan, & Montague, 1997). Depuis, cette région du cerveau, les neurones dopaminergiques des ganglions de la base, est devenue l'une des régions dont on comprend le mieux les capacités d'apprentissage grâce à des modèles mathématiques ancrés dans les concepts de l'apprentissage machine.

1.3 Clarifications

Il est intéressant, voire utile, de s'inspirer de l'apprentissage dans le cerveau pour faire avancer l'apprentissage machine. Mais l'objectif premier de cette thèse est plutôt, à l'inverse, d'étudier l'apprentissage dans le cerveau à l'aide du cadre et des outils mathématiques développés en apprentissage machine. Cette approche a pour but de jeter un regard différent sur l'apprentissage dans le cerveau et éventuellement d'aider les neuroscientifiques à en percer le mystère. Il est difficile, lorsque l'on enregistre un seul neurone, d'incorporer la richesse de toute l'information environnante pour interpréter son comportement. De plus, l'apprentissage ou l'adaptation peut se faire dans plusieurs neurones et sur de nombreuses synapses. L'objectif de cette thèse est d'essayer de tracer quelques grandes lignes directrices ou principes directeurs sur différentes populations de neurones de façon à mettre en contexte l'apprentissage respectif de chacune d'elles et leurs interactions, à l'aide de modèles neuroinformatiques.

Il est difficile de déterminer quel signal de quel algorithme un neurone donné encode. Certains signaux sont peut-être intracellulaires, tandis que d'autres sont peut-être distribués dans des populations entières de neurones. D'ailleurs, de par son évolution même, le cerveau a plus de chances de ressembler à une longue série de

correctifs, c'est-à-dire de petites améliorations utiles, plutôt qu'à un unique algorithme simple que l'on pourrait développer. Cependant, il semble que certaines populations de neurones puissent, d'un point de vue mathématique, résoudre des problèmes similaires ou effectuer des calculs semblables à certains algorithmes. C'est entre autres le cas des neurones dopaminergiques, qui semblent porter un signal d'erreur de prédiction de récompense, signal important dans les algorithmes d'apprentissage machine par renforcement. L'accent sera donc mis sur les concordances visibles et sur les discordances évidentes, plutôt que sur les menus détails.

L'apprentissage est aussi un domaine très large. Dans l'apprentissage machine, on considère souvent que la plus grande difficulté est de trouver la bonne représentation de l'information (Sutton & Barto, 1998; Bengio, 2009). Une fois ce problème résolu, apprendre à reconnaître un objet ou à sélectionner la meilleure action devient généralement un problème facile. Mais ces deux étapes, qui semblent ici distinctes et successives, ne sont probablement pas réalisées en totale isolation l'une de l'autre par le cerveau. Par notre interaction avec l'environnement, notre compréhension de celui-ci s'améliore au fur et à mesure que notre représentation de l'information fournie par nos sens s'enrichit. Nos décisions peuvent ainsi être basées sur des représentations de plus en plus abstraites, et nos actions devenir de plus en plus efficaces.

Aux moins deux régions du cerveau semblent jouer un rôle important dans ce coapprentissage de représentations abstraites et de quête de récompenses. Tout d'abord, le néocortex (que nous appellerons tout simplement le cortex) semble le siège de ces représentations abstraites. On y retrouve des populations de neurones sensibles aux différentes propriétés de ce que l'on voit, touche ou attend et d'autres liés à nos actions, au langage, etc. Ensuite, il y a les ganglions de la base et le système dopaminergique. Ce système semble jouer un rôle important dans l'apprentissage par renforcement (récompenses) chez l'animal en plus d'être celui modélisé par des algorithmes du même nom en apprentissage machine (voir Chapitre 4). Cette thèse se

limitera donc à l'étude de l'apprentissage dans ces deux systèmes et à leurs interactions.

1.4 Contribution

Un certain nombre de modèles de l'apprentissage pour différentes composantes du système nerveux central ont été développés au cours des cinquante dernières années dans le but de mieux les comprendre. Il y a, entre autres, des modèles du cervelet pour le contrôle moteur, des modèles de l'hippocampe comme mémoire associative et des modèles des ganglions de la base de l'apprentissage par renforcement. Il y a cependant peu de modèles adaptatifs du néocortex en comparaisons à la grande variété de représentations neurales que l'on y retrouve. Cette partie du cerveau est l'une des parties ayant le plus évolué chez les primates et les humains par rapport aux autres espèces. Tous ces modèles sont généralement étudiés en isolation par les spécialistes de chaque région concernée. Mais, il est aussi très important d'étudier leurs interactions. Toutes ces structures du système nerveux s'adaptent et communiquent entre elles et sont donc dans un milieu en constant changement. La contribution principale de cette thèse est de construire un modèle à partir duquel on peut commencer à étudier l'apprentissage de deux de ces structures (le cortex et les ganglions de la base) en interaction afin de mieux comprendre le développement de nouvelles représentations dans le contexte de l'apprentissage par récompense dans le cerveau. Les ganglions de la base sont parmi les mieux modélisés actuellement, et leurs modèles sont intrinsèquement reliés au cadre théorique de l'apprentissage machine par renforcement. Quant au néocortex, il semble la partie dont l'apprentissage demeure encore le plus énigmatique, bien qu'il s'agisse très probablement de l'un des systèmes les plus importants de l'apprentissage, principalement pour les aspects cognitifs et le développement de représentations de plus en plus abstraites dans le cerveau.

Une limite importante des modèles de l'apprentissage par renforcement est la représentation du passage du temps. Peu importe la tâche modélisée, soit le temps est découpé en évènements (les évènements ayant tous une durée unitaire), soit une représentation du temps faite sur mesures est insérée dans le modèle. Fait intéressant,

les bases de l'apprentissage d'intervalles de temps constants de l'ordre des secondes sont toujours méconnues. Il y a bien des modèles théoriques, mais les fondements neurobiologiques de cette aptitude demeurent inconnus (Buhusi & Meck, 2005). Les principaux modèles théoriques à saveur neurobiologique ne sont actuellement pas très convaincants ni très physiologiquement réalistes. Par contre, on retrouve des neurones sensibles à ces constantes dans le cortex (Leon & Shadlen, 2003; Reutimann, Yakovlev, Fusi, & Senn, 2004) et de telles représentations temporelles pourraient être apprises (Dragoi, Staddon, Palmer, & Buhusi, 2003; Hopson, 2003). Pour leur part, la majorité des modèles d'apprentissage par renforcement des ganglions de la base éludent complètement ce problème en fournissant une représentation temporelle faite sur mesure, même dans la situation simple de conditionnement classique ou pavlovien (voir Chapitre 4). En conditionnement classique appétitif, un stimulus est toujours suivi d'une récompense après un délai fixe. Aucune action n'est requise autre que de consommer la récompense. Certains résultats des modèles sont directement dépendants de la représentation temporelle choisie. L'étude de l'apprentissage de constantes temporelles est donc un sujet tout indiqué dans le contexte proposé ici. Une contribution importante de cette thèse est de montrer comment des représentations temporelles pourraient être apprises à l'aide d'une mémoire de travail dans le cortex dans différentes formes de conditionnement pavlovien. De plus, le modèle permet d'étudier l'interaction entre leur apprentissage et l'apprentissage dans les ganglions de la base. Le développement de représentations temporelles et l'apprentissage par renforcement dans ces structures sont les thèmes principaux de cette thèse.

1.5 Organisation

L'approche utilisée dans cette recherche étant fondamentalement multidisciplinaire, cette thèse comporte une longue introduction du matériel nécessaire à la compréhension des articles en plus de la revue de littérature habituelle. Le Chapitre 2 consiste en une introduction au cadre théorique de l'apprentissage machine ainsi qu'aux principales familles d'algorithmes du domaine pertinents à la modélisation discutées dans cette thèse. Le Chapitre 3 consiste en une introduction au

cadre habituel d'étude de l'apprentissage en psychologie et en neuroscience ainsi qu'une section spéciale sur les théories de la perception du temps. Le chapitre se termine sur une brève revue des différentes structures du système nerveux central jouant un rôle majeur dans l'apprentissage. Enfin, le Chapitre 4 couvre de façon plus exhaustive les ganglions de la base et le système dopaminergique. Cette structure du cerveau est au cœur des modèles d'apprentissage par renforcement, dont cette thèse est une extension directe. Le Chapitre 5 est un court chapitre méthodologique et justificatif précédant les articles.

Le premier article (Chapitre 6) (Rivest, Bengio, & Kalaska, 2005) est d'abord une contribution au domaine de l'apprentissage machine par renforcement inspirée du fonctionnement du cerveau. Mais, c'est aussi une *preuve de concept,* démontrant que le modèle en deux parties d'apprentissage non supervisé dans le cortex et d'apprentissage par renforcement dans les ganglions de la base développé dans cette thèse est un modèle valable de l'apprentissage dans le cerveau. Cet article démontre principalement qu'un tel modèle peut apprendre, bien que les entrées de chacun des systèmes n'aient pas une distribution stationnaire. C'est-à-dire, que les entrées changent avec le temps, puisqu'elles sont la sortie de systèmes adaptatifs qui servent d'entrées l'un à l'autre, comme dans le cerveau.

Le second article (Chapitre 7) (Rivest, Kalaska, & Bengio, 2010a) est la contribution principale de cette thèse. Il présente une extension des modèles d'apprentissage par renforcement des ganglions de la base en y ajoutant un modèle de l'apprentissage de la représentation de la tâche dans le cortex. Cet article utilise un modèle similaire au premier, cependant l'algorithme non supervisé représentant le cortex a été changé pour pouvoir aussi représenter la dynamique de l'environnement, dont le passage du temps. Il complète ainsi les modèles précédents des ganglions de la base en permettant non seulement d'étudier l'apprentissage de la récompense, mais en permettant d'y inclure du même coup l'apprentissage de l'environnement dans lequel cette récompense est obtenue. Centrés autour de conditionnement de trace, une variante où il y a un intervalle de temps sans stimulus entre le stimulus annonçant la récompense et l'arrivée de celle-ci, ces travaux remplacent la représentation

temporelle injectée artificiellement dans les modèles précédents par une représentation de la tâche acquise automatiquement. Le modèle, comme un animal dans une expérience, doit apprendre, non seulement à associer la récompense aux différentes étapes de la tâche, mais doit aussi apprendre à différencier ces étapes et les règles qui les régissent. Comme le temps est un élément important de la tâche utilisée dans l'article, le modèle apporte aussi une contribution importante au problème de l'acquisition de délais ou de constantes temporelles dans l'environnement.

Le troisième article (Chapitre 8) (Rivest, Kalaska, & Bengio, 2010b) étudie plus en profondeur le modèle, ses capacités de prédictions et son pouvoir d'explication sur une série de variantes du conditionnement classique. Le modèle suggère, entre autres, différentes représentations du temps en fonction des particularités de la tâche. En plus de faire des prédictions sur l'activité corticale et dopaminergique, il montre certaines similarités avec le comportement animal dans des situations qu'aucun autre modèle de ce type n'avait pu approcher auparavant. L'article met aussi en évidence certaines lacunes de la partie corticale du modèle issu de l'apprentissage machine qui pourront être améliorées dans le futur.

Finalement, une discussion générale et une brève conclusion termineront cette thèse au Chapitre 9. Les modèles développés dans cette thèse ouvrent de nouvelles perspectives sur l'étude de l'apprentissage de représentations riches et stables dans la quête de récompenses.

Chapitre 2. L'apprentissage machine

L'apprentissage machine fournira ici le cadre mathématique théorique qui sera utilisé pour l'étude de l'apprentissage dans le cerveau. Il y a trois grandes classes d'apprentissage machine (Duda, Hart, & Stork, 2000) : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. En apprentissage supervisé, on indique à l'algorithme la réponse exacte attendue, alors qu'en apprentissage non supervisé, il n'y a aucune réponse suggérée. L'apprentissage par renforcement se trouve entre les deux. De temps en temps, on donne une récompense ou une note à l'algorithme, sans plus.

Ce chapitre est divisé en trois sections, une pour chaque type d'apprentissage. Chaque section comprend une brève définition du cadre d'apprentissage, suivi d'une description de quelques algorithmes d'apprentissage proches de la biologie et ayant un lien important avec le propos de cette thèse. Certains de ces algorithmes serviront d'hypothèses pour construire un modèle de l'apprentissage dans le cerveau.

2.1 Apprentissage supervisé

2.1.1 Définition

L'apprentissage supervisé consiste à apprendre une fonction $\boldsymbol{f}: \mathfrak{R}^N \rightarrow \mathfrak{R}^M$ ou une fonction $f: \mathfrak{R}^N \rightarrow \{1, \ldots, M\}$. Le premier cas est un problème de régression et le second un problème de classification. Pour ce faire, l'algorithme reçoit des exemples formés d'une paire de vecteurs $(\boldsymbol{x_q}, \boldsymbol{d_q})$. Pour chaque vecteur d'entrées $\boldsymbol{x_q}$, il y a un vecteur de sorties désirées $\boldsymbol{d_q}$ correspondant, de sorte que l'algorithme doit apprendre une fonction $\boldsymbol{f()}$ telle que $\boldsymbol{f(x_q)} \approx \boldsymbol{d_q}$ pour toutes les paires présentées.

Si l'agent est un programme apprenant à jouer aux échecs de façon supervisée, auquel cas l'environnement est donc le jeu d'échecs, alors pour chaque configuration de jeu qu'on lui présentera, on lui fournira la valeur à accorder à la configuration (régression) ou le meilleur coup à jouer (classification).

2.1.2 Réseaux de neurones artificiels et rétropropagation

Les réseaux de neurones artificiels (ANN, de l'anglais *Artificial Neural Networks*) sont inspirés des neurones biologiques (voir Figure 2-1). À la base, un

neurone artificiel a un paramètre w_i pour chacune de ses entrées x_i. Ces paramètres, aussi appelés *poids,* représentent les connexions synaptiques afférentes avec d'autres neurones. Le neurone transforme ses entrées en utilisant ses paramètres et une fonction d'activation *h()* générant la sortie du neurone. On compare généralement la sortie des neurones artificiels à la fréquence de décharge de vrais neurones, « 0 » représentant l'absence de décharge et « 1 », la fréquence de décharge maximale. Les entrées sont donc sommées de façon pondérée par :

$net = w_0 b + \sum_{i=1}^{N} w_i x_i$, **Équation 2-1**

où N est le nombre d'entrées du neurone et $b = 1$. Ensuite, la fonction d'activation du neurone, généralement une sigmoïde telle que

$h(net) = \frac{1}{1+e^{-\alpha \cdot net}}$ ou $h(net) = \tanh(net)$, **Équation 2-2**

limite l'étendue des valeurs de sa sortie entre zéro et un. Lorsque l'on connecte plusieurs neurones ensemble, on obtient un réseau de neurones.

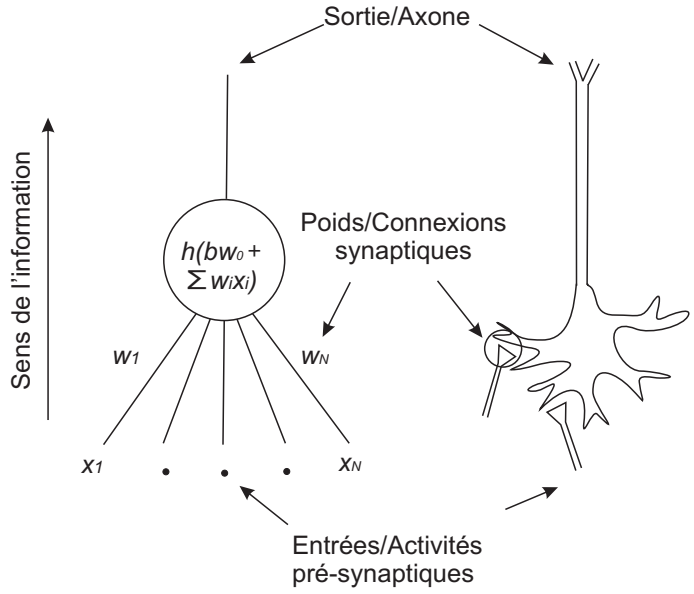


Figure 2-1 : Neurone artificiel à gauche et naturel à droite.

Les réseaux de neurones artificiels multicouches acycliques à rétropropagation (BP, de l'anglais *backpropagation*) ont été popularisés par LeCun (1985) et Rumelhart (1986) et le groupe de recherche PDP (de l'anglais *Parallel Distributed Processing*). Dans ces réseaux, les neurones sont organisés en couches, chacune recevant comme entrées les sorties de la couche précédente (Figure 2-2). Les couches

intermédiaires sont appelées couches cachées, et les entrées ne comptent généralement pas pour une couche. La dernière couche, la couche de sortie, peut contenir des fonctions d'activation linéaires dans le cas d'un problème de régression, ou sigmoïdes dans le cas d'un problème de catégorisation. Dans ce dernier cas, il y a généralement un neurone de sortie par catégorie. On peut soit choisir le neurone avec la plus grande activation comme étant la catégorie gagnante (technique appelée *hardmax* en anglais), ou interpréter les sorties comme des probabilités

$$y_j = P(\boldsymbol{x} \in j \mid W)$$ **Équation 2-3**

en utilisant une fonction d'activation exponentielle

$$h_j(net) = \frac{e^{net_j}}{\sum_{k=1}^{M} e^{net_k}}$$ **Équation 2-4**

pour le neurone de sortie *j* (technique appelée *softmax* en anglais) (Duda et al., 2000). Ces réseaux sont dits acycliques parce qu'il n'y a pas de connexions rétroactives d'une couche supérieure (plus près de la sortie) vers une couche inférieure (plus près de l'entrée). La rétropropagation est la façon dont les gradients utilisés pour corriger les poids sont calculés, le traitement statique de l'information produisant la sortie ne se fait que dans une seule direction.

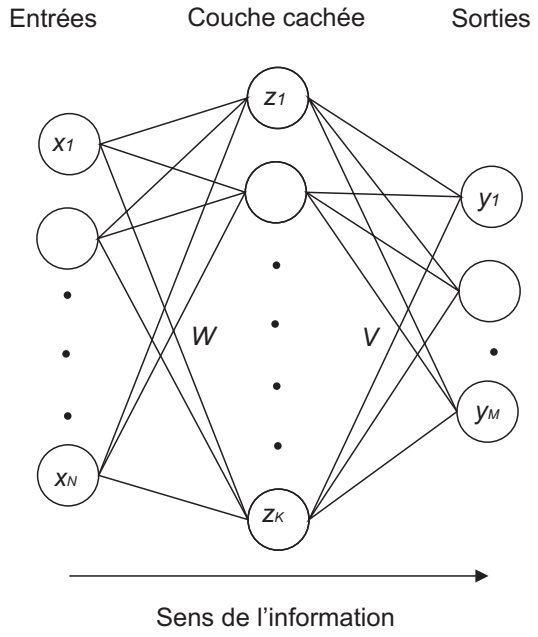


Figure 2-2 : Réseau de neurones acyclique à deux couches dont une couche cachée.

Dans cette architecture, avec deux couches dont une cachée, il est possible d'approximer n'importe quelle fonction continue et bornée avec une précision arbitraire (Cybenko, 1989; Hornick, Stinchcombe, & White, 1989), trois couches sont suffisantes pour approximer n'importe quelle fonction (Cybenko, 1988). Bien que ces réseaux puissent théoriquement approximer n'importe quelle fonction, trouver le nombre de neurones sur chaque couche et les valeurs des poids $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K\}$ et $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M\}$ nécessaires est un problème d'optimisation non linéaire très complexe. Il n'y a donc aucune garantie de trouver une solution satisfaisante, même si elle existe. La méthode d'entrainement par rétropropagation consiste à minimiser une fonction de coût, telle que l'erreur quadratique des neurones de sortie, en utilisant une descente de gradient (voir *Annexe I. Dérivée, gradient et optimisation*). Pour minimiser l'erreur quadratique des neurones de sortie, on calcule d'abord la somme des erreurs au carré :

$$E_q = \sum_{j=1}^{M} \left(d_{j,q} - y_{j,q}\right)^2 , \qquad \textbf{Équation 2-5}$$

où $y_{j,q}$ est la sortie du neurone j et $d_{j,q}$ la valeur cible pour l'exemple q. Des exemples d'entraînement de la forme $(\boldsymbol{x}_q, \boldsymbol{d}_q)$ sont nécessaires à cet apprentissage supervisé. Puis on détermine le gradient de cette fonction par rapport aux poids. On ajuste finalement les poids en soustrayant le gradient multiplié par un petit facteur $0 < \alpha < 1$ appelé le facteur d'apprentissage :

$$W \leftarrow W - \alpha \frac{\partial E_q}{\partial W} . \qquad \textbf{Équation 2-6}$$

Cette méthode cherche donc une fonction $\boldsymbol{f}()$, déterminée par le réseau de neurones, telle que $\boldsymbol{f}(\boldsymbol{x}_q) \approx \boldsymbol{d}_q$. Cependant, elle ne garantit pas de trouver la solution optimale globale, seulement un optimum local. Elle ne dicte pas non plus combien de neurone placer sur chaque couche. Il existe deux variantes importantes de cette méthode : la méthode *stochastique* et la méthode par *lots* (de l'anglais *batch*). Dans la méthode stochastique, les exemples sont présentés aléatoirement et les poids sont mis à jour après la présentation de chaque exemple $(\boldsymbol{x}_q, \boldsymbol{d}_q)$. La méthode par lots consiste à accumuler les gradients sur tout le lot d'exemples, puis à faire la mise à jour des poids en une seule opération :

$$W \leftarrow W - \frac{\alpha}{Q}\sum_{q=1}^{Q}\frac{\partial E_q}{\partial W}, \qquad \textbf{Équation 2-7}$$

où Q est le nombre d'exemples dans le lot.

Les réseaux de neurones artificiels sont souvent utilisés en psychologie connexionniste pour modéliser l'apprentissage chez l'humain. Un exemple classique, mais non psychologique, est *NetTalk* (Sejnowski & Rosenberg, 1987), un réseau ayant appris à prononcer les mots, c'est-à-dire à transformer les lettres des mots en phonèmes. Avant l'entrainement, le réseau produit des sons incompréhensibles, puis après un peu d'entrainement, il commence à babiller et à répéter certains sons, comme les bébés. Puis, après beaucoup d'entrainement, le réseau parle de façon compréhensible, bien qu'il mélange certaines paires de voyelles ou de consonnes semblables. Le réseau semble aussi robuste aux dommages, et peut réapprendre rapidement lorsqu'endommagé gravement par de gros changements de poids aléatoires. Pour une revue de la modélisation du développement cognitif et de l'apprentissage chez l'enfant et pour une comparaison des modèles connexionnistes et symboliques, voir (Shultz, 2003).

2.1.3 Séries temporelles et réseaux récurrents

Dans certains problèmes d'apprentissage, ce ne sont pas que les entrées actuelles $x_{i,t}$ (t = temps actuel) qui sont importantes, mais aussi les entrées aux temps précédents $x_{i,t-l}$ ($l > 0$). Par exemple, dans des problèmes de linguistique, le sens que l'on donne à un pronom peut dépendre des informations passées telles que le début de la phrase ou la phrase précédente. Dans la reconnaissance de phonèmes, le spectre des dernières 50 ms peut s'avérer suffisant pour reconnaître certains phonèmes, mais insuffisant pour des phonèmes qui dépendent de liaisons sonores ou de coarticulations. Dans ces cas, les réseaux acycliques ne sont pas très bien adaptés, à moins de maintenir les entrées accessibles un certain temps.

Une *époque* représente une série temporelle avec un début et une fin précise, par exemple : $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, …, $\boldsymbol{x_T}$. Un algorithme d'apprentissage peut fonctionner par époque, ou en temps continu. Le second cas est beaucoup plus difficile puisque le début et la fin des séries, s'il y a lieu, ne sont pas marqués. Par exemple lorsque quelqu'un parle, on entend un flux continu de sons. Le début et la fin des mots tout

comme le début et la fin des phrases sont des éléments que nous avons appris à reconnaître, mais qui ne sont pas marqués autrement que par le son; le temps ne s'arrête pas.

La solution la plus simple au problème des séries temporelles est généralement d'utiliser ce que l'on appelle des lignes de délai (TDNN, de l'anglais *Time-Delay Neural Networks*) (Lang, Waibel, & Hinton, 1990). Cette solution utilise une fenêtre de temps de quelques unités temporelles au lieu d'une seule sur les entrées et fait défiler les entrées devant cette fenêtre une unité de temps à la fois (voir Figure 2-3). C'est aussi équivalent à avoir une petite fenêtre temporelle d'une seule unité de temps couplée à une mémoire cache des dernières entrées dans l'ordre. Dans un réseau de neurones, on peut utiliser le même principe pour les unités des couches supérieures, celles-ci peuvent recevoir les entrées de la couche inférieure au temps actuel t, ainsi qu'une copie des activités des neurones aux temps précédents $t\text{-}l$, $0 \leq l \leq L$ pour un délai maximum donné L. Ces réseaux sont acycliques et peuvent être entraînés de façon supervisée en minimisation l'erreur quadratique par descente de gradient comme les réseaux précédents.

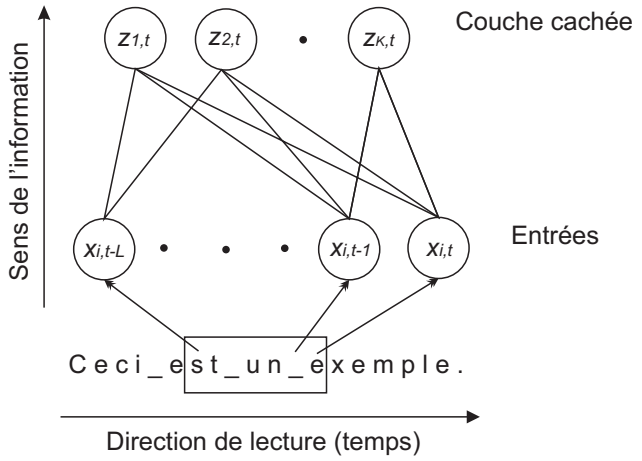


Figure 2-3 : Utilisation de L lignes de délai pour l'entrée i d'une première couche. Le rectangle représente la fenêtre temporelle d'entrées perceptibles par la couche cachée au temps t.

Cependant, cette solution a le défaut de sa qualité, étant donné que le réseau est acyclique, c'est-à-dire sans rétroaction, un neurone ne dépend toujours que des entrées courantes et précédentes, mais jamais de son état précédent ou de celui des autres neurones de la même couche. Que l'activité d'une unité cachée ou de sortie

puisse aussi dépendre de sa propre activité au temps précédent est une propriété importante (Elman, 1990). Une façon simple d'ajouter une telle rétroaction sans compliquer la règle d'apprentissage est d'ajouter ce que l'on appelle une *unité de contexte*. Cette unité n'est en fait qu'une copie de l'activité d'un neurone au temps précédent (*t-1*), placée dans la couche inférieure tel que montré sur la Figure 2-4. Alors que les TDNN utilisent L lignes de délai par neurones de la couche précédente pour avoir accès à leurs activités aux temps $t-1, \ldots, t-L$, les unités de contexte sont équivalentes à une seule ligne de délai pour chaque neurone de la même couche donnant ainsi accès à leur activité au temps précédent *t-1* uniquement. Cette forme de réseau s'appelle un réseau d'*Elman* (1990). Chaque neurone d'une couche a ainsi accès aux activations des neurones de cette couche au temps précédent via leurs entrées de la couche inférieure. Ce cycle dans le réseau augmente le niveau de complexité de ce que peut apprendre le réseau, sans complexifier la procédure d'apprentissage. Cependant, la propagation de l'erreur dans le temps ne recule que d'une unité de temps à la fois, puisque le gradient s'arrête à la copie du contexte et n'est pas calculé plus profondément le long de la chaîne du gradient.

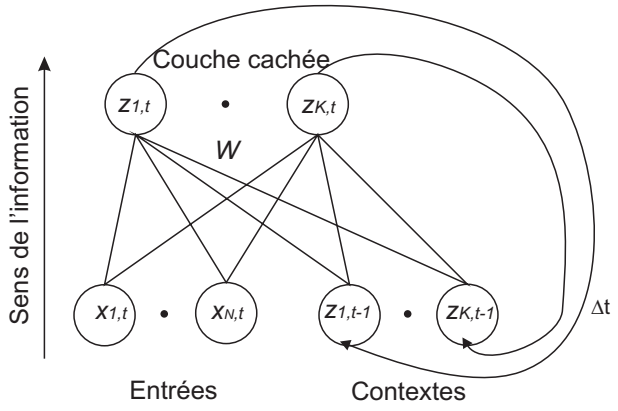


Figure 2-4 : Réseau d'Elman avec unités de contexte pour la couche cachée.

Il est toutefois possible de calculer le gradient complet, et donc de calculer la correction pour les poids, en considérant les activités de tous les temps précédents de l'époque. Pour un réseau ayant une couche contenant toutes les rétroactions, on *déroule* le réseau sur lui-même au lieu d'utiliser des unités de contextes. Ce réseau (Figure 2-5) calcule les sorties de la même façon que celui avec les unités de contexte

(Figure 2-4), et en ce sens ils sont équivalents. Par contre, pour pouvoir dérouler le gradient dans le temps, il faut garder en mémoire l'historique des activités des neurones à tous les temps précédents. Si un neurone caché est représenté par

$z_{k,t} = h\left(\sum_{l=1}^{N} w_{l,k} x_{l,t} + \sum_{l=1}^{K} v_{l,k} z_{l,t-1}\right)$, *h'(u) = dh(u)/du,* **Équation 2-8**

alors le gradient complet déroulé peut s'écrire sous une forme récursive, telle que

$$\frac{\partial z_{k,t}}{\partial w_{i,k}} = h'\left(\sum_{l=1}^{N} w_{l,k} x_{l,t} + \sum_{l=1}^{K} v_{l,k} z_{l,t-1}\right)\left[x_{i,t} + \sum_{l=1}^{K} v_{l,k} \frac{\partial z_{l,t-1}}{\partial w_{i,k}}\right]$$ **Équation 2-9**

pour les connexions en provenance des entrées et

$$\frac{\partial z_{k,t}}{\partial v_{j,k}} = h'\left(\sum_{l=1}^{N} w_{l,k} x_{l,t} + \sum_{l=1}^{K} v_{l,k} z_{l,t-1}\right)\left[z_{j,t-1} + \sum_{l=1}^{K} v_{l,k} \frac{\partial z_{l,t-1}}{\partial v_{j,k}}\right]$$ **Équation 2-10**

pour les connexions rétroactives. La différence avec les réseaux d'Elman réside dans le dernier terme des Équation 2-9 et Équation 2-10, qui rend ces équations récursives. Dans les réseaux d'Elman, la dernière sommation des gradients de $z_{l,t-1}$ est tout simplement tronquée à zéro, au lieu d'être déroulée récursivement. L'approche du gradient complet s'appelle la rétropropagation au travers du temps (BPTT, de l'anglais *backpropagation through time*). Elle est biologiquement peu probable et son coût en termes d'espace mémoire est proportionnel à la longueur des épisodes (ou séries) puisqu'il faut maintenir en mémoire l'historique des activités des neurones pour tous les temps précédents (Williams & Zipser, 1995).

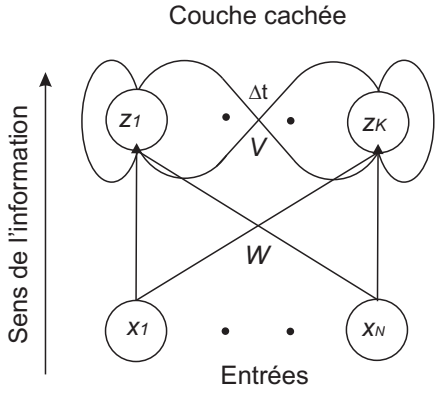


Figure 2-5 : Réseau avec une couche cachée incluant toutes les rétroactions possibles.

Une autre approche est d'apprendre les récurrences en temps réel (RTRL, de l'anglais *real-time recurrent learning*) en maintenant à jour les gradients cumulés au fil du temps $\partial z_{k,t}/\partial v_{i,j}$ de chaque neurone par rapport à chaque poids, incluant ceux des

autres neurones. Ceci évite de conserver un historique des activations précédentes, mais est beaucoup plus couteux en termes de mémoire et de calcul en fonction de la taille du réseau (Williams & Zipser, 1995).

Cependant, que ce soit par BPTT ou RTRL, il demeure très difficile d'entraîner des réseaux récurrents comme celui de la Figure 2-5. En effet, soit le gradient de l'erreur par rapport aux poids au travers du temps explose, soit il tend vers 0 exponentiellement rapidement en fonction du temps. En d'autres termes, l'apport réel du gradient pour corriger l'erreur due aux temps précédents t-τ, $\tau > 0$, tend rapidement vers zéro en fonction de τ (Bengio, Simard, & Frasconi, 1994; Hochreiter & Schmidhuber, 1997). Ces méthodes ne sont donc pas beaucoup plus puissantes que le gradient tronqué des unités de contexte.

2.1.4 Réseau de longue mémoire à court terme (LSTM)

Récemment, une nouvelle solution au problème de l'entrainement des réseaux de neurones sur des séries temporelles a été proposée : les réseaux de longue mémoire à court terme (LSTM, de l'anglais *long short-term memory*) (Hochreiter & Schmidhuber, 1997; Gers, Schmidhuber, & Cummins, 2000; Gers, Schraudolph, & Schmidhuber, 2002). Ces réseaux ont une architecture beaucoup plus sophistiquée que les précédents. Tout d'abord, plutôt que d'avoir plusieurs couches cachées, ces réseaux ont une seule couche cachée formée de plusieurs blocs mémoires parallèles (Figure 2-6). Ils peuvent être entraînés de façon supervisée par rétropropagation, en minimisant l'erreur quadratique des sorties par exemple, sur des séquences d'entraînement $\{(\boldsymbol{x}_1,\boldsymbol{d}_1), \ldots, (\boldsymbol{x}_T,\boldsymbol{d}_T)\}$ où $\boldsymbol{d}_t$ est le vecteur cible des sorties désirées pour le vecteur d'entrées $\boldsymbol{x}_t$ au temps t. Comme dans le cas des réseaux d'Elman, le gradient pour les rétroactions en dehors des blocs n'est pas déroulé au travers du passage du temps. Mais, au cœur des blocs, des neurones d'état ont une activation linéaire, ce qui leur permet d'avoir un gradient d'une forme simple, qui ne tend pas vers zéro au travers du temps. De plus, leur gradient peut être calculé sans devoir tenir en mémoire l'historique des activations aux temps t-l, $l > 1$ comme les autres réseaux récurrents décrits précédemment.

Dans les LSTM, les signaux d'entrée sont distribués vers chaque bloc mémoire en deux types d'entrées différents : l'entrée principale (à gauche des blocs mémoires sur la Figure 2-6) et l'entrée des portes (au bas des blocs mémoires sur la Figure 2-6). De la même façon, les sorties des blocs mémoires sont distribuées vers la couche de sortie, mais aussi par boucles rétroactives vers les entrées des blocs mémoires (entrées principales et entrées des portes). Seules quelques connexions de chaque type ont été dessinées sur la Figure 2-6 pour simplifier le schéma.

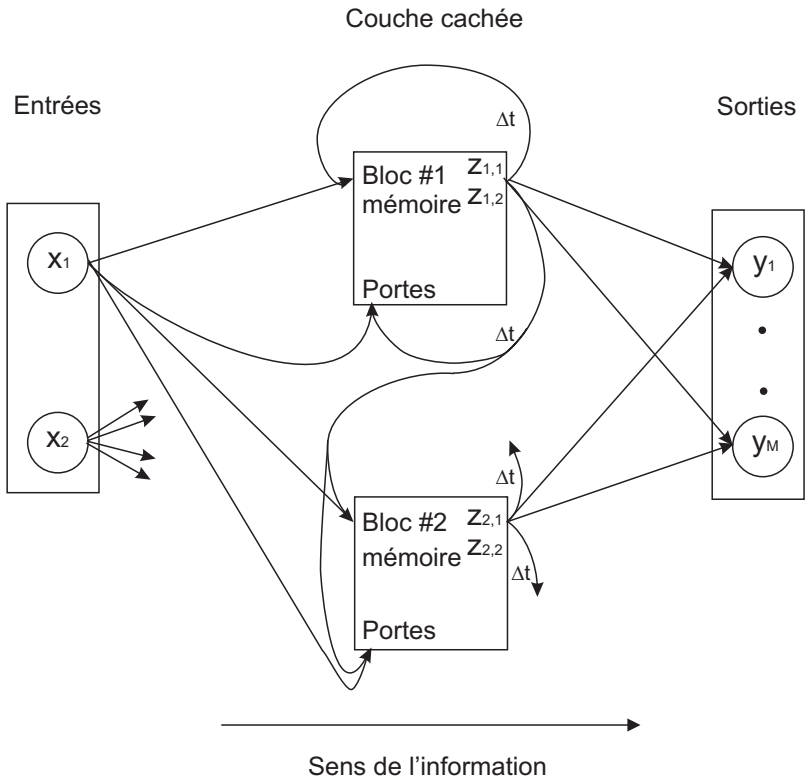


Figure 2-6 : Exemple de réseau de longue mémoire à court terme. Ce réseau a deux entrées (x_1, x_2) et deux blocs mémoires de deux cellules mémoires chacun.

Le cœur du réseau réside dans l'architecture des blocs mémoires (Figure 2-7). Les portes calculent chacune une sommation pondérée distincte des entrées qui est ensuite passée par une fonction d'activation dont la sortie varie entre zéro et un comme un neurone artificiel ordinaire. Cependant, les sorties de ces neurones servent de gains à différents endroits dans le bloc agissant un peu comme des interrupteurs. Pour chaque cellule mémoire du bloc, une sommation pondérée distincte des entrées est aussi calculée et passée par une fonction d'activation. Ce signal est ensuite

multiplié par la porte d'entrée. Si la porte d'entrée est ouverte, alors le signal entre dans la cellule mémoire. La valeur de la cellule mémoire dépend de deux termes : du signal d'entrée multiplié par la porte d'entrée et de la valeur de la cellule mémoire au

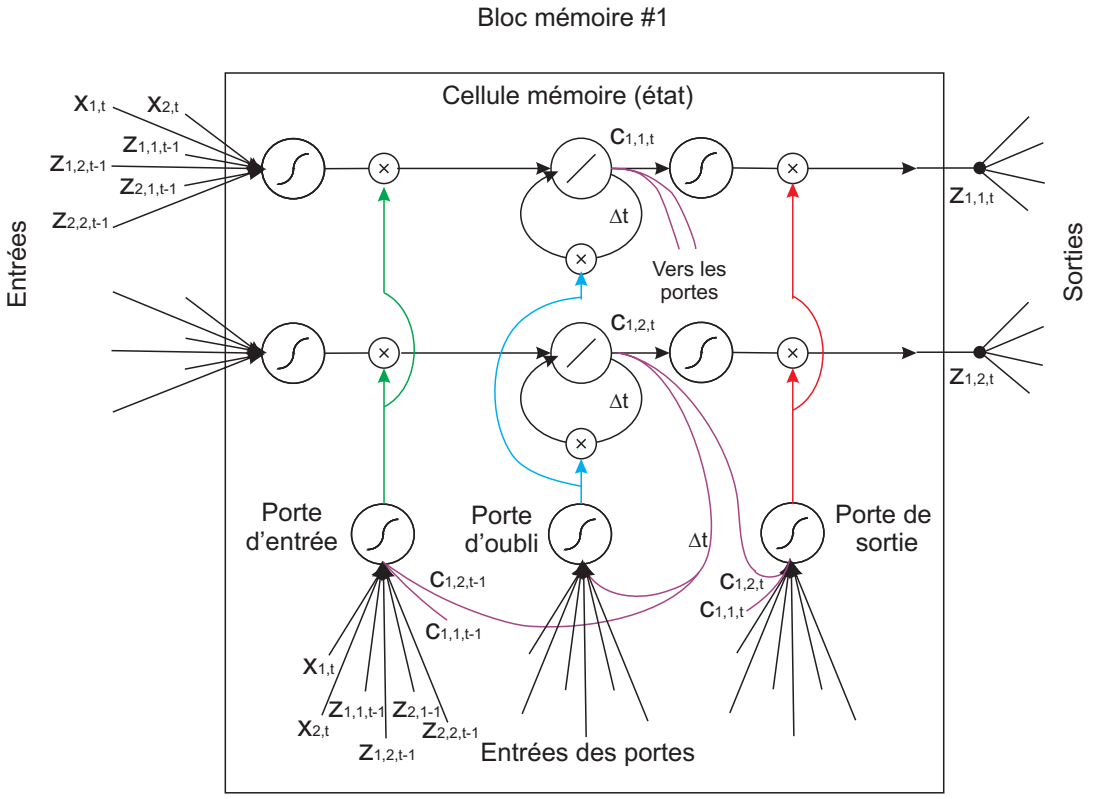


Figure 2-7 : Bloc de mémoire du réseau LSTM de la Figure 2-6. Ce bloc mémoire d'un réseau LSTM à deux entrées ($x_{1,t}$, $x_{2,t}$) contient deux cellules mémoires (ou cellules d'état) $c_{1,1,t}$ et $c_{1,2,t}$. Les sorties $z_{1,1,t}$ et $z_{1,2,t}$ servent d'entrées au temps suivant.

temps précédent multiplié par la porte d'oubli; lorsque cette dernière est fermée, la cellule perd la valeur qu'elle avait au temps précédent. Par exemple, si la porte d'oubli est fermée et la porte d'entrée ouverte, alors on a l'équivalent de la fonction « *sauvegarde dans la mémoire (STO)* » d'une calculatrice. Si les deux portes sont ouvertes, alors on a l'équivalent de la fonction « *ajouter à la mémoire (M+)* » d'une calculatrice. Si les deux portes sont fermées, nous avons la fonction « *effacer la mémoire (CLR)* ». Finalement, le signal de la cellule mémoire est encore une fois passé par une fonction d'activation avant d'être multiplié par la porte de sortie. Si cette dernière est fermée, le bloc produit une sortie de zéro, si elle est ouverte, alors la valeur des cellules mémoires peut être lue à la sortie du bloc, soit par les autres blocs au temps suivant, soit directement par la couche de sortie. Cette porte de sortie est

l'équivalent de la fonction « *lecture de mémoire (RCL)* » d'une calculatrice. Les portes sont contrôlées par leur somme pondérée des entrées du réseau ($x_{1,t}$, $x_{2,t}$, Figure 2-7) et des sorties au temps précédent des blocs mémoires ($z_{1,1,t-1}$, $z_{1,2,t-1}$, $z_{2,1,t-1}$, $z_{2,2,t-1}$, Figure 2-7). Chacune des sommations pondérées du réseau est faite de poids différents qui sont appris par rétropropagation. Une description plus détaillée de l'algorithme se trouve au Chapitre 7 et au Chapitre 8.

Bien que cette architecture semble extrêmement artificielle, elle reste plus plausible biologiquement que plusieurs des autres réseaux récurrents présentés ici. Tout d'abord, le gradient de chaque poids ne requiert que des signaux disponibles au temps présent ou précédent (*t* ou *t-1*), aucun gradient ni aucune activation de neurones à un temps *t-l, l>1* n'est requis. Ensuite, l'ajustement des poids à l'intérieur d'un bloc ne requiert aucun autre signal en provenance des autres blocs que celui qui y arrive déjà : ni gradient, ni poids, ni activité de neurones locaux à d'autres blocs. Seuls quelques signaux non locaux de la couche de sortie sont nécessaires à la mise à jour des poids. De plus, ces réseaux ressemblent bien à ce que l'on pourrait s'imaginer d'une *mémoire de travail* chez l'humain. C'est-à-dire la possibilité de mettre une information en mémoire, de l'y garder activement, ou de la remplacer par une autre nouvelle information. Les réseaux de longue mémoire à court terme ont d'ailleurs été récemment utilisés dans des travaux de modélisation de cette faculté (O'Reilly & Frank, 2006).

2.2 Apprentissage non supervisé

2.2.1 Définition

L'apprentissage non supervisé peut aussi passer par l'apprentissage d'une fonction de la forme $\boldsymbol{f}: \Re^N \rightarrow \Re^M$ ou $\boldsymbol{f}: \Re^N \rightarrow \{1, \ldots, M\}$. Cependant, seuls des vecteurs d'entrées $\boldsymbol{x_q}$ sont fournis à l'algorithme, il ne reçoit aucun exemple de sorties $\boldsymbol{d_q}$. À la place, il cherche généralement à trouver quelque chose de représentatif dans l'information fournie, de façon à trouver une représentation plus naturelle, plus utile ou plus efficace des données. Il peut servir à découvrir des caractéristiques importantes des données, à prédire les caractéristiques des situations à venir (séries temporelles), à modéliser la distribution des données, ou même à générer de

nouveaux exemples (dans ce cas, on parle de modèle génératif). Il peut aussi être combiné à un apprentissage supervisé (Hinton, Osindero, & Teh, 2006; Bengio, Lamblin, Popovici, & Larochelle, 2007; Erhan, Manzagol, Bengio, Bengio, & Vincent, 2009; Larochelle, Bengio, Louradour, & Lamblin, 2009) ou par renforcement (Foster & Dayan, 2002; Bakker, Linaker, & Schmidhuber, 2002; Khamassi, Martinet, & Guillot, 2006) (voir aussi section 2.3.6).

Si l'on reprend l'exemple de l'agent qui apprend à jouer aux échecs, il pourrait développer une représentation de la position des pièces. Celle-ci, au lieu d'être une description brute, serait une représentation décrivant la présence ou l'absence de sous-configurations typiques telle que la présence de trois pièces qui se protègent mutuellement. Il pourrait aussi simplement développer des catégories de configurations, ou des valeurs pour celles-ci. Plus généralement, la fonction ***f()*** peut représenter la distribution de probabilité des $\boldsymbol{x_q}$ ou des caractéristiques importantes de cette distribution.

2.2.2 Distribution des données et maximisation de vraisemblance

Une première approche statistique en entrainement non supervisé est de tenter de reconstruire la distribution qui génère les données. Supposons un ensemble de vecteurs observés, des échantillons indépendants, $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_Q\}$, alors la vraisemblance des échantillons est donnée par :

$$P(D|\boldsymbol{\theta}) = \prod_{q=1}^{Q} p(\boldsymbol{x_q}|\boldsymbol{\theta}) \qquad \textbf{Équation 2-11}$$

où $\boldsymbol{\theta}$ est le vecteur de paramètres de la fonction de distribution *p()*. Il suffit de choisir une fonction *p(),* un mélange de gaussiennes par exemple, et de trouver les paramètres $\boldsymbol{\theta}$ qui maximisent $P(D|\boldsymbol{\theta})$. Ceci génère la distribution de la forme choisie la plus vraisemblable, c'est-à-dire la distribution qui a le plus de chance d'être à l'origine des données observées. La Figure 2-8 donne un exemple incluant des données et le mélange de gaussiennes la plus vraisemblable pour celles-ci. Il n'est pas nécessaire d'avoir un modèle paramétrique de la distribution *p()*. Il est possible de trouver une distribution non paramétrique, par exemple, à l'aide d'un réseau de neurones artificiels probabiliste (voir Duda et al., 2000).

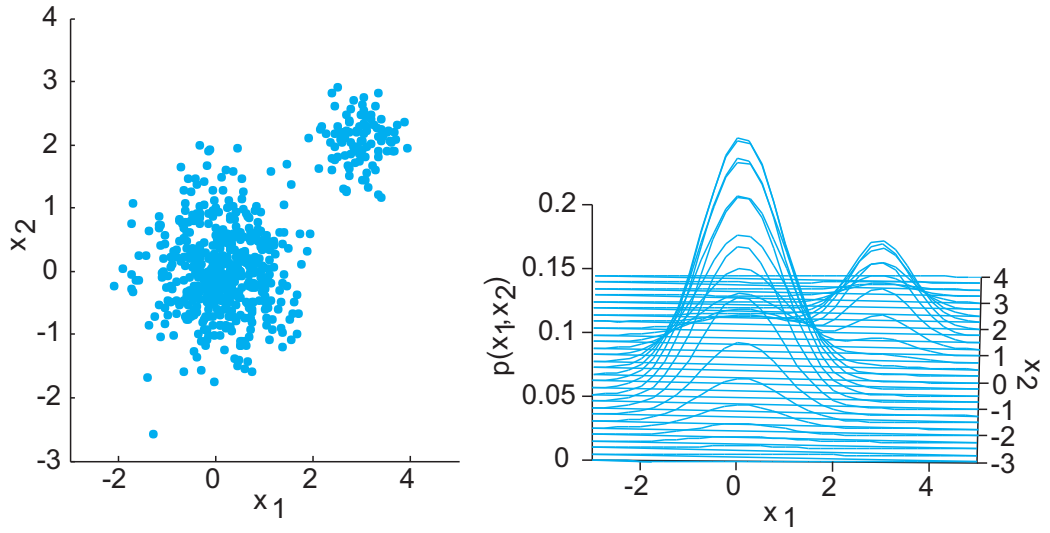


Figure 2-8 : Mélange de gaussiennes. Exemple de données à gauche provenant d'une distribution ressemblant à un mélange de gaussiennes accompagnées à droite de la distribution trouvée : le mélange de gaussiennes le plus vraisemblablement à l'origine des données.

2.2.3 Analyse en composantes principales (PCA)

Une autre approche consiste à trouver un sous-espace de la distribution des données dans lequel se retrouve la majorité des données. Par exemple, notre système solaire, bien qu'il soit dans un espace tridimensionnel, est toujours représenté dans un plan. On peut expliquer une grande partie des données, des positions des planètes et de la rotation autour du Soleil dans ce plan sans s'encombrer de la troisième dimension.

L'analyse en composantes principales (PCA, de l'anglais *Principal Component Analysis*) consiste donc à trouver ce sous-espace et une transformation linéaire W pour réécrire les coordonnées des points dans cet espace, tels que $\boldsymbol{y_q} = W\boldsymbol{x_q}$ où $\boldsymbol{x_q}$ est un échantillon dans l'espace original des données et $\boldsymbol{y_q}$ sa projection dans ce nouvel espace de plus petite dimension. La Figure 2-9 donne un exemple de données tridimensionnelles pouvant être représenté en deux dimensions. La méthode consiste à trouver les vecteurs propres de la matrice de covariance des échantillons $\boldsymbol{x_q} \in D$ et de les ordonner dans l'ordre décroissant de leur valeur propre. Ensuite, on choisit une borne inférieure telle qu'on ne considère que les vecteurs propres dont la valeur propre est supérieure à cette borne. Le nombre de vecteurs propres conservés

détermine le nombre de dimensions de ce nouvel espace. C'est ainsi que l'on réduit le nombre de dimensions des données tout en conservant le maximum d'information.

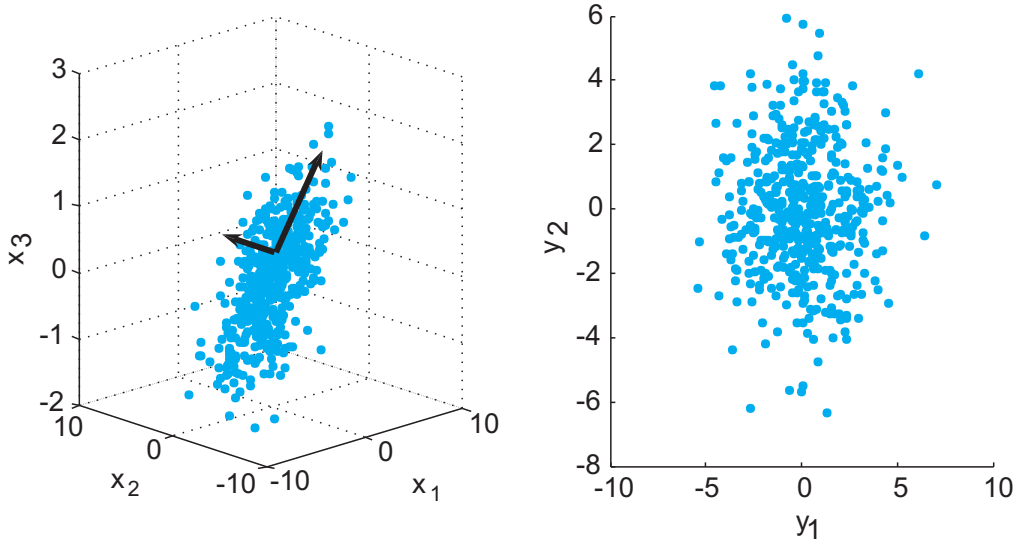


Figure 2-9 : Réduction de dimension par PCA. Exemple de données tridimensionnelles, à gauche, pouvant être représentées en deux dimensions, à droite, dans le plan formé à partir des deux vecteurs du graphique de gauche.

Cette analyse peut se faire d'une façon biologiquement plausible à l'aide de règles semblables à celle d'Oja (1982) pour un seul neurone telle que

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha\left(y_q \boldsymbol{x}_q - y_q^2 \boldsymbol{w}\right) \qquad \textbf{Équation 2-12}$$

où $\boldsymbol{w}$ est le vecteur de poids du neurone y, et q l'indice du vecteur d'entrées $\boldsymbol{x}_q \in D$. Oja a démontré que l'on pouvait avoir un neurone qui représente la position de $\boldsymbol{x}_q$ dans la dimension principale à l'aide d'une simple règle d'apprentissage hebbienne. Hebb a postulé que la connexion entre deux neurones doit se renforcer si ceux-ci sont actifs en même temps (Hebb, 1949), ce qui revient à baser l'adaptation d'une connexion sur la corrélation de ses neurones.

Ce principe peut être généralisé pour M composantes principales en utilisant M neurones. Après un certain temps d'entrainement, ces neurones coderont les entrées $\boldsymbol{x}_q$ dans un sous-espace principal. Pour un traitement complet sur l'implémentation en réseaux de neurones artificiels de la PCA ou sur les réseaux basés sur des règles hebbiennes similaires ou d'autres statistiques de deuxième ordre, voir (Diamantaras & Kung, 1996), en particulier le chapitre 4.

2.2.4 Analyse en composantes indépendantes (ICA)

La PCA se limite cependant à des statistiques de second ordre, c'est-à-dire à analyser les corrélations. Elle ne permet pas d'extraire de statistiques plus approfondies des données. C'est ce que fait l'analyse en composantes indépendantes (ICA, de l'anglais *Independent Component Analysis*) (Comon, 1994; Bell & Sejnowski, 1995). Cette méthode statistique consiste à rechercher les variables indépendantes à l'origine des données observées. Ce problème est aussi appelé la séparation aveugle de sources (BSS, de l'anglais *Blind Source Separation*). Par exemple, nos oreilles nous fournissent deux signaux sonores x_1 et x_2 qui sont le mix des sons provenant de différentes sources telles que notre collègue qui nous parle, une voiture qui passe, les oiseaux qui chantent, les passants qui discutent, etc. C'est ce que l'on appelle *le problème de la soirée cocktail* où tout le monde parle en même temps. L'objectif est de trouver la fonction $\boldsymbol{g^{-1}()}$ nous permettant de reconstruire les signaux sonores de chacune de ces sources indépendantes : s_1 pour le collègue, s_2 pour la voiture, s_3 pour l'oiseau, et ainsi de suite.

On suppose donc qu'un certain nombre d'échantillons sources $\boldsymbol{s_q} \in \Re^M$ sont à l'origine des données $\boldsymbol{x}_q \in \Re^N$ observées suite à une transformation $\boldsymbol{g()}$ tel que $\boldsymbol{x_q} = \boldsymbol{g}(\boldsymbol{s_q})$. Le but de cette approche est de construire une approximation $\boldsymbol{f}: \Re^N \rightarrow \Re^M$ de la fonction inverse $\boldsymbol{g^{-1}()}$ de telle sorte que l'on puisse reconstruire les échantillons $\boldsymbol{s}_q$, sources des données observées $\boldsymbol{x}_q$ tel que $\boldsymbol{f}(\boldsymbol{x}_q) \approx \boldsymbol{s}_q$. Cette méthode requiert d'optimiser une fonction, par exemple la maximisation de vraisemblance, ou la minimisation d'information mutuelle, qui permettra de maximiser l'indépendance des sorties $y_{1,q} = f_1(\boldsymbol{x}_q), \ldots, y_{M,q} = f_M(\boldsymbol{x}_q)$. Par conséquent, les dimensions des données qui en seront extraites seront des variables indépendantes ou presque. Si les sources sont indépendantes et si $\boldsymbol{s_q}$ et $\boldsymbol{y_q}$ sont de même dimension, alors la transformation $\boldsymbol{f()}$ devrait permettre de recouvrer les sources originales.

Pour reprendre l'exemple précédent, on peut indexer les données par le temps t. Dans ce cas, $x_{1,t}$ et $x_{2,t}$ deviennent les échantillons du son par les deux oreilles et $s_{1,t}$, $s_{2,t}$ et $s_{3,t}$ deviennent les échantillons du son produit par le collègue, la voiture et l'oiseau au temps t respectivement. Les oreilles entendent chacune un mix de ses trois

signaux $x_{i,t} = g_i(\boldsymbol{s_t})$. Après entrainement, la fonction $(y_{1,t}, y_{2,t}, y_{3,t}) = \boldsymbol{f(x_t)}$ permet de séparer ces signaux puisqu'ils sont sensiblement indépendants.

Cette méthode ne fonctionne pas avec des données ayant une distribution gaussienne, qui est entièrement définie par des statistiques d'ordre un et deux, soit la moyenne et la variance. Il est généralement recommandé de commencer par normaliser en fonction de ces statistiques, c'est-à-dire centrer-réduire les données, avant de faire une ICA.

Alors qu'en analyse en composantes principales on cherche toujours à réduire le nombre de dimensions des données (base sous-complète, de l'anglais *under-complete*), en analyse de composantes indépendantes on peut aussi parler de bases surcomplètes (de l'anglais *over-complete*), où le nombre de dimensions de $\boldsymbol{y_q}$ est plus grand que celui de $\boldsymbol{x_q}$ ($M > N$). Dans l'exemple précédent, il n'y avait que deux oreilles tandis qu'il y avait trois sources sonores dans l'environnement.

Il est difficile d'évaluer la plausibilité biologique de ces algorithmes, mais ils semblent faire de bons modèles du cortex visuel primaire. En effet, il a été démontré que les neurones issus d'un entrainement par ICA, dans un contexte où les entrées ressemblent à ce que les neurones du cortex visuel reçoivent, répondent d'une façon similaire à ces derniers aux stimuli visuels (Doi, Inui, Lee, Wachtler, & Sejnowski, 2003; Hyvarinen, Gutmann, & Hoyer, 2005). Il est cependant intéressant de constater que l'ICA peut être faite à partir du principe de maximisation de l'information des sorties. Ce principe, appelé *Infomax* (Bell & Sejnowski, 1995), a aussi été utilisé bien avant par Linsker (1988) pour créer un réseau de neurones simple de style hebbien. Comme pour l'ICA, les réseaux de Linsker ont aussi servi de modèle du cortex visuel. Pour un traitement exhaustif des différentes approches pour trouver les signaux indépendants, voir (Hyvarinen, Karhunen, & Oja, 2001).

2.2.5 L'autosupervision

Il y a des problèmes non supervisés qui peuvent être formulés comme des problèmes supervisés. Par exemple, un réseau peut tenter d'apprendre à reproduire les entrées en sorties. Dans ce cas, la paire d'entrainement $(\boldsymbol{x_q}, \boldsymbol{d_q})$ est donnée par $(\boldsymbol{x_q}, \boldsymbol{x_q})$. Pour un réseau acyclique à une couche cachée entrainé par la rétropropagation, appelé

autoencodeur dans ce cas, la couche cachée apprend l'équivalent de la PCA, si le nombre de neurones cachés ne dépasse pas N et si les neurones sont linéaires. Des versions multicouches plus performantes d'autoencodeurs ont aussi été développées (Hinton & Salakhutdinov, 2006; Vincent, Larochelle, Bengio, & Manzagol, 2008).

Un second exemple est la *prédiction de séries temporelles,* où le problème consiste à apprendre à reproduire les entrées d'une série temporelle (voir section 2.1.3). Dans ce cas particulier, la paire d'entrainement $(\boldsymbol{x}_q, \boldsymbol{d}_q)$ est donnée par $(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$. Plus généralement, on peut tenter de prédire, $\boldsymbol{x}_{t+1}$ à partir de $\{\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots\}$. Cette technique permet d'apprendre une pièce musicale, telle qu'une chanson par exemple, comme une séquence. Une fois apprise, il ne suffit que d'entendre les premières notes pour jouer toute la chanson, les notes jouées en sortie servant d'entrées pour produire les notes suivantes. Cette façon de faire représente en partie la vision autoassociative et prédictive qu'ont Hawkins et d'autres de la mémoire et du néocortex (Rao & Sejnowski, 2001; Hawkins & Blakeslee, 2004).

2.3 Apprentissage par renforcement

2.3.1 Définition

À la base, l'apprentissage par renforcement passe aussi par l'apprentissage d'une fonction $\boldsymbol{f}: \mathbb{R}^N \rightarrow \mathbb{R}^M$ ou $\boldsymbol{f}: \mathbb{R}^N \rightarrow \{1, \ldots, M\}$. Cependant, à l'opposé des deux autres formes d'apprentissage, qui reçoivent soit la bonne réponse $\boldsymbol{d}_q$ en supervisé, soit aucune réponse en non supervisé, l'agent en apprentissage par renforcement reçoit un signal d'erreur ou de *récompense* à l'occasion. De plus, l'apprentissage par renforcement diffère des deux précédents par le fait qu'il ne peut fonctionner qu'en ligne, c'est-à-dire que les données d'entrainement ne peuvent pas réellement être préparées à l'avance, à moins d'avoir un modèle de l'environnement au préalable. En effet, l'agent (défini par la fonction $\boldsymbol{f()}$) ne donne pas simplement une réponse, il effectue une action. Celle-ci influence l'environnement et les entrées suivantes qu'il observera. L'agent doit donc interagir avec son environnement pour apprendre les conséquences de ses actes à long terme et maximiser ses récompenses.

L'algorithme doit donc apprendre une politique d'action $f()$, notée ici par $\pi: S \rightarrow A$ où $A = \{a_1, \ldots, a_A\}$, qui retourne généralement une action y, notée ici par

$a_t = \pi(\boldsymbol{s_t})$ pour action. Aucune action cible n'est fournie avec l'état $\boldsymbol{x_t}$, noté ici par $\boldsymbol{s_t} \in S$ pour état, lors de l'entrainement[2]. De plus, suite à son action, l'agent reçoit un signal de récompense $r_{t+1} \in \mathfrak{R}$ avec le nouvel état résultant $\boldsymbol{x_{t+1}}$. Par conséquent, en apprentissage par renforcement, il y a généralement une séquence de la forme $\ldots \rightarrow \boldsymbol{s_t} \rightarrow a_t \rightarrow r_{t+1} \rightarrow \boldsymbol{s_{t+1}} \rightarrow a_{t+1} \rightarrow r_{t+2} \rightarrow \ldots$ où les états $\boldsymbol{s_t}$ et les récompenses r_t sont fournis par l'environnement à l'agent et les actions a_t sont fournies à l'environnement par l'agent (Figure 2-10). Le signal de renforcement r_t peut toutefois provenir de l'agent lui-même, auquel cas le signal de récompense est dit *intrinsèque*.

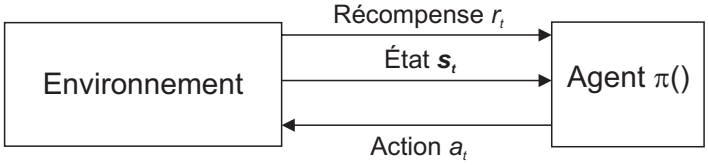


Figure 2-10 : Schéma de l'interaction entre l'agent et l'environnement.

Par exemple, l'agent d'échecs, lors d'une partie, recevrait la configuration actuelle, retournerait une action, et suite à cette action, l'environnement lui retournerait une récompense. Après que l'opposant ait joué, l'environnement fournirait à l'agent la nouvelle configuration pour qu'il choisisse à nouveau une action et ainsi de suite. Cette description du problème suppose que l'information disponible à n'importe quel moment *t* est complète. C'est-à-dire que les probabilités de l'état suivant ne dépendent d'aucunes autres variables non observables et sont indépendantes des états antérieurs. C'est la condition markovienne.

Cependant, la réalité est généralement tout autre. Par exemple, un joueur de poker ne voit pas le jeu des autres joueurs. Par conséquent, il n'observe pas $\boldsymbol{s_t}$, mais uniquement une partie de l'information dénotée $\boldsymbol{o_t} \in O$. On parle alors de problème partiellement observable (Figure 2-11). Pour le moment, supposons que $\boldsymbol{o_t} \equiv \boldsymbol{s_t}$.

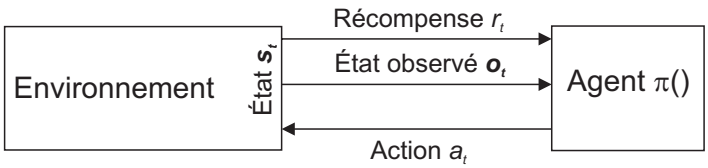


Figure 2-11 : Schéma de l'interaction entre l'agent et l'environnement partiellement observable.

[2] Le changement de notation est nécessaire ici pour suivre la notation standard dans la littérature sur l'apprentissage par renforcement.

La difficulté en apprentissage par renforcement est que la récompense peut n'arriver qu'à la fin. Par exemple, l'agent d'échecs peut recevoir une récompense de « 0 » après chaque coup, à l'exception d'un « -1 » ou d'un « 1 » après sa dernière action selon qu'il perde ou qu'il gagne la partie, respectivement. Dans ce cas, l'agent doit essayer de comprendre quelles actions ont mené à ce résultat. C'est le problème de l'assignation du crédit : à quelles actions associer la récompense. Lorsqu'un problème a un début et une fin précis, comme une partie d'échecs, on parle de problèmes épisodiques, où une partie correspond à un *épisode*. Un épisode est comparable à une époque dans les séries temporelles.

2.3.2 Apprentissage d'une fonction de valeur

Le but de l'agent est donc d'apprendre pour chaque état la ou les actions maximisant la somme des récompenses à venir. Mathématiquement, l'agent veut apprendre la politique $\pi()$ tel que $E_\pi\{\sum_{t=0}^{T}\gamma^t r_{t+1}\}$ est maximale et où $0 < \gamma \leq 1$ ($\gamma < 1$ si $T = \infty$) est le facteur de remise, c'est-à-dire que plus une récompense est éloignée dans le futur, moins elle a de valeur en ce moment. Le truc consiste à apprendre une fonction de valeur (de l'anglais *value function*) $V{:}S \rightarrow \Re$ ou $Q{:}(S, A) \rightarrow \Re$ qui estime ou prédit la somme des récompenses à venir pour un état $\boldsymbol{s_t}$, ou une paire d'état-action $(\boldsymbol{s_t}, a_t)$ plutôt que d'apprendre directement la politique $\pi()$. Si l'on a une bonne estimation pour la politique optimale, alors on peut agir de façon optimale. Par exemple, une politique $\pi()$ *greedy* qui utiliserait la fonction de valeur $V^\pi()$ serait une politique qui choisirait toujours, parmi les actions possibles, l'action a_t menant à l'état suivant $\boldsymbol{s_{t+1}}$ ayant la plus haute estimation $V^\pi(\boldsymbol{s_{t+1}})$. Évidemment, pour utiliser cette politique, il faut aussi connaître quelle sera l'état résultant pour chaque action. Au moins trois approches peuvent-être utilisées pour apprendre une fonction de valeur.

2.3.3 Apprentissage par différence temporelle (TD)

Supposons donc que nous avons une estimation $V^\pi(\boldsymbol{s_t})$ plus ou moins bonne de la somme des récompenses à venir et que nous aimerions corriger cette fonction pour qu'elle soit plus exacte pour l'état $\boldsymbol{s_t}$. Nous souhaitons que

$$V^\pi(\boldsymbol{s_t}) \approx E_\pi\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\}, \gamma < 1. \qquad \textbf{Équation 2-13}$$

On peut décomposer $V^\pi()$ tel que

$V^{\pi}(\boldsymbol{s_t}) \approx r_{t+1} + E_{\pi}\{\sum_{k=1}^{\infty} \gamma^k r_{t+k+1}\}$, **Équation 2-14**

$V^{\pi}(\boldsymbol{s_t}) \approx r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}})$. **Équation 2-15**

On peut supposer que $V^{\pi}(\boldsymbol{s_{t+1}})$ est meilleure que $V^{\pi}(\boldsymbol{s_t})$ et s'en servir pour corriger cette dernière:

$V^{\pi}(\boldsymbol{s_t}) \leftarrow V^{\pi}(\boldsymbol{s_t}) + \alpha\delta_{t+1}$ où $0 < \alpha < 1$ et où **Équation 2-16**

$\delta_{t+1} = r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}}) - V^{\pi}(\boldsymbol{s_t})$, **Équation 2-17**

soit la différence entre l'estimation originale $V^{\pi}(\boldsymbol{s_t})$ et la nouvelle estimation $r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}})$, termes de gauche et de droite de l'Équation 2-15 respectivement. Une fois que nous avons observé la récompense r_{t+1} et l'état $\boldsymbol{s_{t+1}}$, nous pouvons donc corriger l'estimation $V^{\pi}(\boldsymbol{s_t})$ pour l'état $\boldsymbol{s_t}$.

Cette technique s'appelle l'apprentissage par différence temporelle (TD, de l'anglais *temporal-difference*) (Sutton, 1988) puisqu'elle est basée sur la différence d'estimations de $V^{\pi}(\boldsymbol{s_t})$ aux temps t et $t+1$. Le signal d'erreur δ (Équation 2-17) est aussi appelé signal de renforcement effectif, puisque c'est lui, et non la récompense elle-même, qui détermine s'il y aura un renforcement de l'action en augmentant l'estimation de récompense. On peut aussi voir cette règle d'apprentissage comme la minimisation de la différence temporelle au carré,

$E_{\pi}\left\{\left(V^{\pi}(\boldsymbol{s_t}) - \left(r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}})\right)\right)^2\right\}$, **Équation 2-18**

par descente de gradient, dans ce cas, δ est le gradient.

La même règle peut être utilisée pour apprendre une fonction de valeur $Q^{\pi}()$ sur les paires d'états-action $(\boldsymbol{s_t}, a_t)$, auquel cas l'algorithme s'appelle SARSA pour $\boldsymbol{s_t} \rightarrow a_t \rightarrow r_{t+1} \rightarrow \boldsymbol{s_{t+1}} \rightarrow a_{t+1}$. Dans ce cas, l'Équation 2-17 devient :

$\delta_{t+1} = r_{t+1} + \gamma Q^{\pi}(\boldsymbol{s_{t+1}}, a_{t+1}) - Q^{\pi}(\boldsymbol{s_t}, a_t)$. **Équation 2-19**

2.3.4 Apprentissage hors politique

Les deux algorithmes précédents, TD et SARSA, sont dits dépendants de la politique, car la fonction de valeur $V^{\pi}()$ ou $Q^{\pi}()$ ainsi apprise est la fonction de valeur pour la politique $\pi()$ utilisée, et non pour la politique optimale. Une approche légèrement différente, la plus courante pour apprendre la politique optimale, s'appelle *Q-Learning* (Watkins & Dayan, 1992) et utilise la règle de mise à jour suivante :

$$Q^*(\boldsymbol{s_t}, a_t) \leftarrow Q^*(\boldsymbol{s_t}, a_t) + \alpha\delta_{t+1} \quad \textbf{Équation 2-20}$$

$$\delta_{t+1} = r_{t+1} + \gamma \max_{a'}\{Q^*(\boldsymbol{s_{t+1}}, a')\} - Q^*(\boldsymbol{s_t}, a_t) \quad \textbf{Équation 2-21}$$

L'Équation 2-21 est semblable aux Équation 2-17 et Équation 2-19 de TD et SARSA respectivement. À l'opposé de ces dernières, la règle de mise à jour de *Q-Learning* (Équation 2-21) ne dépend pas de l'estimation pour l'action a_{t+1} sélectionnée, mais de celle de la meilleure action estimée $a' \in A$. Donc, même en choisissant une action qui n'est pas nécessairement la meilleure selon la politique optimale, la méthode apprend quand même à prédire la somme des récompenses pour cette dernière, et non pour la politique $\pi()$ utilisée, sous certaines conditions mathématiques.

2.3.5 Acteur-critique

Dans les deux approches précédentes, la politique est souvent directement dépendante de la fonction de valeur. Par exemple, une politique *ε-Greedy* est une politique qui choisit la meilleure action selon la fonction de valeur $V^\pi()$ ou $Q^\pi()$ la plupart du temps, et qui choisit une action au hasard avec probabilité ε, pour un petit ε. Une approche différente consiste à avoir deux fonctions indépendantes pour la politique $\pi()$ et sa fonction de valeur $V^\pi()$ et d'utiliser le même signal d'erreur δ pour corriger les deux fonctions simultanément (Figure 2-12). Dans cette approche, dite *acteur-critique* (Sutton & Barto, 1998), la politique $\pi()$ est appelé l'*acteur*, et la fonction de valeur $V^\pi()$ avec son signal d'erreur δ forme le *critique*.

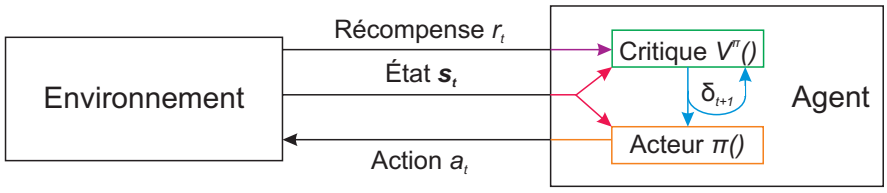


Figure 2-12 : Schéma de l'interaction acteur-critique.

Par exemple, supposons que l'état $\boldsymbol{s_t}$ puisse être représenté par un vecteur $\boldsymbol{s_t} \in \Re^N$ et que l'approximation de la fonction $V^\pi()$ soit donnée par une fonction linéaire de la forme

$$V^\pi(\boldsymbol{s_t}) = \boldsymbol{w_c}^T\boldsymbol{s_t}, \quad \textbf{Équation 2-22}$$

où $\boldsymbol{w_c}$ est un vecteur de poids indicé *c* pour critique. Supposons aussi que la politique $\pi()$ soit définie par un *softmax* (voir section 2.1.2, Équation 2-4) sur les fonctions

$$\pi_a(\boldsymbol{s_t}) = \boldsymbol{w_a}^T \boldsymbol{s_t}, \qquad \textbf{Équation 2-23}$$

où les $\boldsymbol{w_a}$ sont les vecteurs de poids correspondant à chaque action $a \in A$, l'action avec la plus haute valeur ayant le plus de chance de l'emporter. Dans ce cas, on peut définir les règles de mises à jour suivantes pour le critique et l'acteur respectivement:

$$\boldsymbol{w_c} \leftarrow \boldsymbol{w_c} + \alpha\delta_{t+1}\boldsymbol{s_t}, \qquad \textbf{Équation 2-24}$$

$$\boldsymbol{w_{a_t}} \leftarrow \boldsymbol{w_{a_t}} + \alpha\delta_{t+1}\boldsymbol{s_t}, \qquad \textbf{Équation 2-25}$$

où a_t est l'action gagnante au temps t.

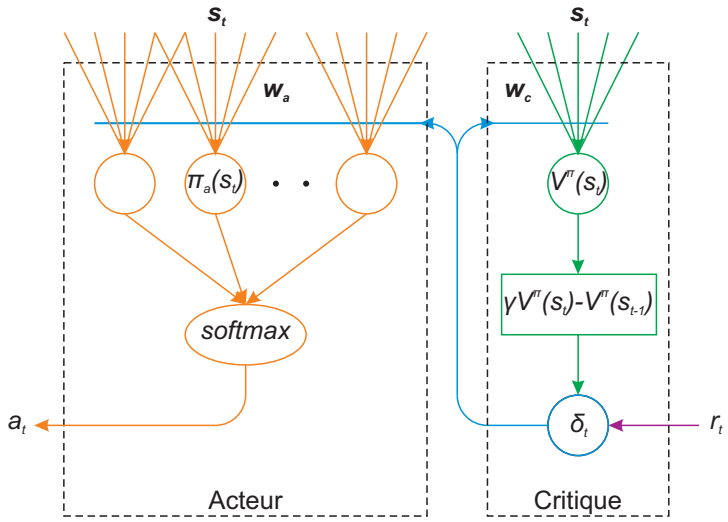


Figure 2-13 : Implémentation neurale simple du modèle acteur-critique.

La règle du critique (Équation 2-24) est en fait une descente de gradient sur le signal d'erreur au carré (Équation 2-18) par rapport aux paramètres $\boldsymbol{w_c}$, et revient donc à minimiser l'erreur de l'approximation de $V^\pi()$. Quant à la règle de l'acteur (Équation 2-25), elle revient à renforcer les poids entre le vecteur d'entrée $\boldsymbol{s_t}$ et l'action gagnante a_t en fonction de l'erreur de prédiction qui en résulte δ_{t+1}. Intuitivement, si $\delta_{t+1} > 0$, alors globalement l'action a_t a mené à plus de récompenses que prévu par $V^\pi(\boldsymbol{s_t})$. On veut donc augmenter les chances de choisir cette même action a_t la prochaine fois que l'on se retrouve dans la même situation $\boldsymbol{s_t}$. Pour ce faire, l'Équation 2-25 ajuste les poids de l'action gagnante a_t en fonction de la grandeur de δ_{t+1} et du vecteur d'état $\boldsymbol{s_t}$. Similairement, si $\delta_{t+1} < 0$, alors l'action a mené à moins de récompenses que prévu. Les poids liant l'état $\boldsymbol{s_t}$ à l'action a_t sont alors diminués, de façon à défavoriser cette action au profit des autres actions la

prochaine fois que la situation se présente. On peut facilement imaginer $V^{\pi}()$ et $\pi_a()$ être de simples neurones comme à la Figure 2-13 (Barto, 1995).

2.3.6 Représentations et abstractions structurales

Les algorithmes précédents sont garantis de converger vers la meilleure solution pour $V^{\pi}()$, $Q^{\pi}()$ ou $\pi()$ (selon l'algorithme) sous un certain nombre de conditions mathématiques (Sutton, 1988; Dayan, 1992; Watkins & Dayan, 1992; Jaakkola, Jordan, & Singh, 1994; Tsitsiklis, 1994). Parmi celles-ci, l'environnement doit toujours être markovien ($P(\boldsymbol{s'}|\boldsymbol{s_t},a_t,\boldsymbol{s_{t-1}},a_{t-1},\ldots)=P(\boldsymbol{s'}|\ \boldsymbol{s_t},a_t)$), stationnaire ($P(\boldsymbol{s_k}|\boldsymbol{s_i},a_j,t\,)=P(\boldsymbol{s_k}|\boldsymbol{s_i},a_j,t\,')$), et complètement observable ($\boldsymbol{o_t} \equiv \boldsymbol{s_t}$). Les fonctions $V^{\pi}()$ et $Q^{\pi}()$ doivent généralement être tabulaires ou linéaires par rapport à $\boldsymbol{s_t}$ et a_t. Les différents états et actions doivent être visités infiniment souvent et le facteur d'apprentissage α doit décroître de façon appropriée. Dans certains cas, il est aussi possible de passer par l'apprentissage des probabilités de transitions $P(\boldsymbol{s_k}|\boldsymbol{s_i},a_j)$. C'est ce que l'on appelle apprendre un modèle de l'environnement (Sutton & Barto, 1998).

Un des problèmes majeurs en apprentissage machine, particulièrement en apprentissage par renforcement, est celui de la représentation des données (Sutton & Barto, 1998). Les algorithmes en apprentissage par renforcement supposent une représentation suffisamment riche et stable de l'environnement et des actions. Souvent, trouver la bonne représentation représente 99 % de la solution à un problème d'apprentissage. Les algorithmes présentés dans la section 2.3 n'ont pas accès aux états précédents, à moins de l'inclure dans la représentation à l'aide de lignes de délai ou d'une autre forme de mémoire. Ils n'ont pas non plus d'information sur le temps écoulé depuis un évènement. L'état actuel, c'est ce qu'ils observent. Ils sont donc en situation partiellement observable ($\boldsymbol{o_t} \neq \boldsymbol{s_t}$, Figure 2-11). Si certaines informations sur l'état courant peuvent être inférées à partir d'évènements passés, alors il faut leur fournir un modèle de l'environnement capable de faire ces inférences. C'est la condition markovienne, c'est-à-dire que les décisions que peuvent apprendre ces algorithmes ne peuvent dépendre que de leur état courant. Si l'on veut leur permettre d'utiliser de l'information non directement observable sur l'état courant, alors il faut leur fournir une représentation de l'état qui est enrichie de cette

information. Si l'agent a une mémoire des observations précédentes, alors l'état de l'agent (parfois appelé sa croyance, ou *belief state*, $\boldsymbol{b_t} \in B$) peut être plus riche que l'état observable ($\boldsymbol{o_t}$), tout en étant possiblement différent de l'état réel de l'environnement ($\boldsymbol{s_t}$) (Figure 2-14). Quoi qu'il en soit, la qualité de la solution trouvée par ces algorithmes dépend de la représentation ou du modèle qu'a l'agent de l'environnement. Créer ces représentations automatiquement reste cependant un problème très difficile et toujours d'actualité.

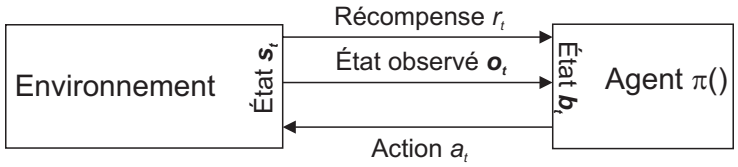


Figure 2-14 : Schéma de l'interaction entre l'agent ayant sa propre croyance sur l'état d'un environnement partiellement observable.

Il y a deux champs d'études importants en apprentissage par renforcement pour développer de meilleures représentations automatiquement (Foster & Dayan, 2002). Le premier, discuté ci-dessus, est l'*abstraction structurale*, où l'on cherche une représentation plus riche de l'état. Pour ce faire, on peut tenter de construire des entrées représentant des combinaisons complexes d'états ayant un point commun à un niveau plus élevé que les données brutes, à l'instar des méthodes d'apprentissage non supervisé. Par exemple, on pourrait vouloir avoir une variable pour indiquer la présence d'un objet *x*, peu importe l'orientation visuelle ou l'éclairage ambiant. Ce genre de représentation requiert plusieurs niveaux de traitement et de représentations intermédiaires (Bengio, 2009). Une autre façon d'enrichir les représentations consiste à construire une représentation contenant plus d'information que ce qui est directement observable à chaque instant *t* (c'est-à-dire que $\boldsymbol{b_t} \neq \boldsymbol{o_t}$), soit en maintenant un historique du passé (semblable aux lignes de délai, TDNN, section 2.1.3), soit en utilisant l'information de l'état précédent ($\boldsymbol{b_{t-1}}$) (à l'instar des réseaux d'Elman, section 2.1.3) ou en essayant de prédire le futur (par exemple, autosupervision et séries temporelles, section 2.2.5) (Littman, Sutton, & Singh, 2002; Sutton & Tanner, 2005). Une telle représentation pourrait idéalement nous indiquer la présence de l'objet *x* même lorsqu'il est caché par un autre objet.

Le second type d'abstraction est l'*abstraction temporelle* (rien à voir avec le *temps*), où l'on cherche à développer des macros, c'est-à-dire des méta-actions qui encapsulent une séquence d'actions simples permettant d'atteindre un état situé à plusieurs actions de distance. Par exemple, une voiture intelligente pourrait apprendre une commande *« stationnement en parallèle »* qui comprendrait une séquence d'actions simples pour effectuer la manœuvre. Le propos de cette thèse se limitera toutefois à la construction d'abstractions structurales (dans le cerveau) et non celles de méta-actions. Pour un traitement plus complet de l'apprentissage par renforcement, voir (Sutton & Barto, 1998).

2.3.7 La place de l'apprentissage machine dans cette thèse

Le principal objectif de cette thèse est de tenter de mieux comprendre comment le cerveau crée de telles abstractions structurales (section 2.3.6) dans un contexte d'apprentissage par renforcement. L'apprentissage machine est d'une importance capitale dans ce projet puisqu'il fournit le cadre mathématique théorique nécessaire à l'étude de ce phénomène. De plus, différents algorithmes serviront à modéliser des hypothèses sur le développement de représentations dans le cerveau pendant l'apprentissage par renforcement. Ces modèles ainsi créés seront simulés sur des tâches artificielles afin d'évaluer leur efficacité comme méthode d'apprentissage. Ils seront aussi simulés sur des tâches représentant des conditions semblables à celles d'animaux en laboratoire afin d'être comparés aux données comportementales et neurophysiologiques existantes. Dans cette optique, les algorithmes ayant des similitudes avec les neurones seront favorisés. De cette façon, il sera possible de comparer l'activité de ces neurones artificiels à celle de vrais neurones. Finalement, d'autres simulations seront effectuées afin de faire des prédictions pouvant éventuellement être vérifiées en laboratoire.

Chapitre 3. L'apprentissage animal

La psychologie définit aussi différentes formes d'apprentissage et il y a différentes structures dans le système nerveux. Cependant, il n'y a pas toujours de relations claires entre une forme d'apprentissage telle que définie en psychologie et une structure du cerveau en particulier. Par exemple, l'apprentissage non associatif tel que la désensibilisation, pourrait a priori se produire au niveau de presque n'importe quel neurone. Par contre, l'apprentissage explicite (section 3.1.1) semble dépendre de structures précises dont l'hippocampe. De plus, on commence à peine à isoler les bases neurologiques de l'acquisition de certaines connaissances, comme celle des délais fixes de l'ordre des secondes (Buhusi & Meck, 2005). Il est donc important de bien comprendre ces différents éléments afin de pouvoir intégrer les données comportementales et neurophysiologiques dans un modèle mathématique commun.

Ce chapitre est donc divisé en trois parties : la première décrit brièvement les différentes formes d'apprentissage telles qu'étudiées en psychologie, la deuxième, les modèles théoriques les plus importants, et la troisième, les différentes structures du cerveau importantes dans l'apprentissage ainsi que leurs liens avec les différentes formes d'apprentissage. Ce chapitre n'est pas une revue exhaustive des théories et données sur l'apprentissage, mais un survol des différents concepts avec un peu plus de détails sur ceux liés à cette thèse.

3.1 Formes d'apprentissage en psychologie

Les différentes formes d'apprentissage et de mémoire étudiées en psychologie (Figure 3-1) ressemblent a priori peu aux types d'apprentissage machine. En psychologie, l'apprentissage est le processus par lequel on acquiert des connaissances alors que la mémoire est le processus par lequel les connaissances sont encodées, enregistrées et rappelées. On distingue tout d'abord la mémoire déclarative de la mémoire implicite. La première comprend généralement des choses que l'on peut verbaliser ou se rappeler de façon consciente et ne sera pas l'objet de ce projet de recherche. La seconde comprend l'apprentissage procédural et principalement deux sous formes d'apprentissage : non associatif et associatif. Cette section fait un survol

de ces formes d'apprentissage, avec plus de détails sur l'apprentissage associatif, c'est-à-dire les conditionnements classique et opérant.

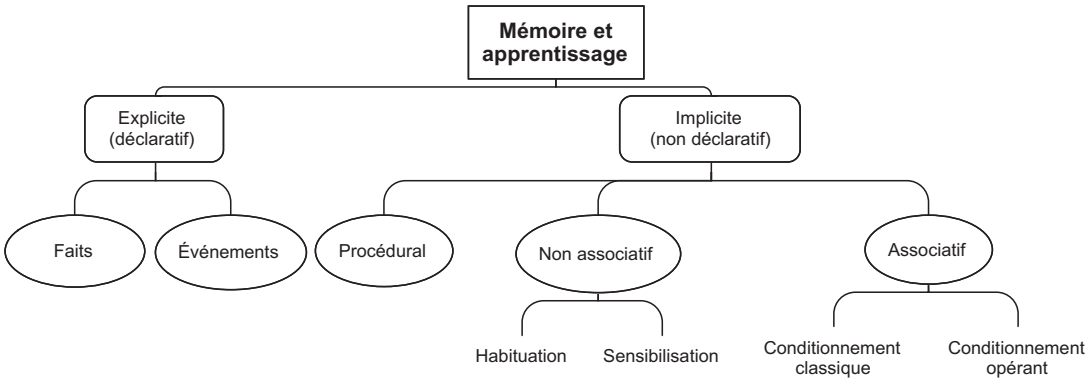


Figure 3-1 : Formes d'apprentissage en psychologie (selon Kandel et al., 2000).

3.1.1 Apprentissage (et mémoire) déclaratif(s)

L'apprentissage déclaratif, ou explicite, est la mémorisation symbolique ou la génération de règles verbales. « En quelle année êtes-vous nés? » « Quel est votre nom? » Les réponses à ces questions sont toutes considérées comme des souvenirs explicites. C'est d'ailleurs la base d'une grande partie du système d'éducation où les connaissances sont transmises de façon explicite. Plus particulièrement, on parle de *mémoire épisodique* pour parler de la mémoire des évènements, comme dans l'exemple : « Qu'avez-vous fait hier soir? »

3.1.2 Apprentissage procédural

L'apprentissage procédural peut être décrit comme le développement d'habiletés, et d'une certaine façon, il peut contenir de l'apprentissage associatif et non associatif. Un chien qui a appris à faire le beau, un rat qui a appris à trouver la sortie d'un labyrinthe, un enfant qui apprend à manger seul, un adulte qui apprend à conduire, quelqu'un qui s'entraine au tennis, sont tous des exemples d'apprentissage procédural.

D'autre part, l'apprentissage procédural inclut le phénomène d'*automatisation* de certaines connaissances ou compétences, dont apprendre à conduire est un exemple. Lorsque quelqu'un apprend à conduire, l'instructeur lui dit de regarder régulièrement dans ses rétroviseurs, de penser à ralentir, etc. Au début, cette personne doit continuellement penser à toutes ces actions. Mais après un certain temps, la

conduite automobile devient une seconde nature, et toutes ces actions se font presque automatiquement.

3.1.3 Apprentissage non associatif : l'habituation

L'apprentissage non associatif est l'effet sur un organisme de la présentation à répétition d'un stimulus quelconque. On reconnaît principalement deux formes d'apprentissage non associatif : l'habituation et la sensibilisation.

L'habituation (ou désensibilisation) est le fait de s'habituer à un stimulus. Par exemple, un bébé auquel on présente un jouet peut, après un certain temps, se désintéresser de ce jouet. Un animal qui devient stressé à la vue d'un étranger peut, à force de le côtoyer, ne plus en être affecté outre mesure. C'est un phénomène important qui illustre la capacité d'adaptation à l'environnement. L'effet de *surprise*, qui survient lorsque l'on observe quelque chose d'imprévu, est utilisé pour étudier les connaissances des enfants. Si l'enfant accroît son intérêt pour un stimulus anormal ou qu'il en est surpris, on considère qu'il a la connaissance lui permettant de trouver le stimulus qu'il observe comme étant anormal. Il possède donc une représentation pour laquelle ce stimulus ne semble pas convenir. En psychologie, l'habituation et la surprise servent souvent de mesures indirectes de la représentation interne de certaines connaissances, spécialement chez les bébés (par exemple, voir Baillargeon, 1986; Aguiar & Baillargeon, 1998).

La sensibilisation est le phénomène inverse. C'est le fait de devenir plus sensible à un stimulus répété. Par exemple, si l'on entend de façon répétée un mot que l'on ne reconnaît pas au début, après quelques répétitions on le reconnaîtra facilement. Ce type d'adaptation est souvent à court terme. Si par exemple quelqu'un vous lit une liste de mots, puis vous pose des questions auxquelles certains de ces mots peuvent être une réponse, ils seront souvent employés plutôt que d'autres que vous auriez utilisés si l'on ne vous avait pas lu cette liste. Dans ce cas, on parle aussi parfois d'*amorçage* (*priming* en anglais).

3.1.4 Apprentissage associatif : le conditionnement classique

À l'opposé de l'idée d'associer des paires de stimuli symboliques, comme des mots, il est ici question de l'apprentissage associatif non explicite; c'est-à-dire de

l'association, généralement due à la répétition, de stimuli et de leurs effets. Ce type d'apprentissage est aussi appelé le conditionnement. Il y en a deux sous-formes : le conditionnement classique, et le conditionnement instrumental ou conditionnement opérant.

En conditionnement classique, on associe à un nouveau stimulus la réponse physiologique normale à un second stimulus. Le nouveau stimulus s'appellera le stimulus conditionné (CS, de l'anglais c*onditioned stimulus*). Le second stimulus, comme de la nourriture, s'appelle le stimulus non conditionné (US, de l'anglais *unconditioned stimulus*), puisque la réponse physiologique, comme la salivation, est présente sans entrainement. L'exemple le plus connu est celui des chiens de Pavlov (1960). Lorsqu'on leur présente leur repas (le stimulus non conditionné), les chiens se mettent à saliver (la réponse non conditionnée). Si l'on sonne une cloche avant leur repas, après quelques jours, ils saliveront au son de la cloche. De la même façon, un enfant qui se bouche les oreilles au bruit d'un aspirateur, en viendra à se boucher les oreilles à la vue de celui-ci, avant qu'on ne l'actionne, à moins bien sûr, qu'il ne s'y habitue (ce qui arrivera fort probablement).

En général, le début du CS doit arriver avant le début de l'US. S'il arrive après, il peut y avoir un effet inhibiteur à la place, car il peut être considéré comme annonçant la fin de l'US au lieu de son arrivée. De plus, le CS et l'US sont généralement aussi présents au même moment pendant un petit laps de temps (Figure 3-2, en haut à gauche). Si le CS disparaît avant l'arrivée de l'US, on parle alors d'un conditionnement *de trace* (Figure 3-2, en haut à droite). Cette forme de conditionnement est généralement moins efficace (Gallistel & Gibbon, 2000) que le conditionnement classique. Lorsque l'on fait une série de conditionnement à répétition (Figure 3-2, en bas), le rapport (τ_{ITI}/τ_{ISI}) entre le délai interessais τ_{ITI} (ITI, de l'anglais *intertrial interval*) et le délai interstimuli τ_{ISI} (ISI, de l'anglais, *interstimuli interval*) devient important. Plus ce rapport est petit, plus le conditionnement est difficile. Selon certaines théories (Gallistel & Gibbon, 2000), lorsque le rapport est petit (disons $\tau_{ITI} < \tau_{ISI}$), le CS ne semble pas prédire un taux de récompense beaucoup

plus grand que le bruit de fond. Par contre, si $\tau_{ITI} \gg \tau_{ISI}$, alors le taux de récompense associé au CS devient beaucoup plus grand que le taux de récompense de base.

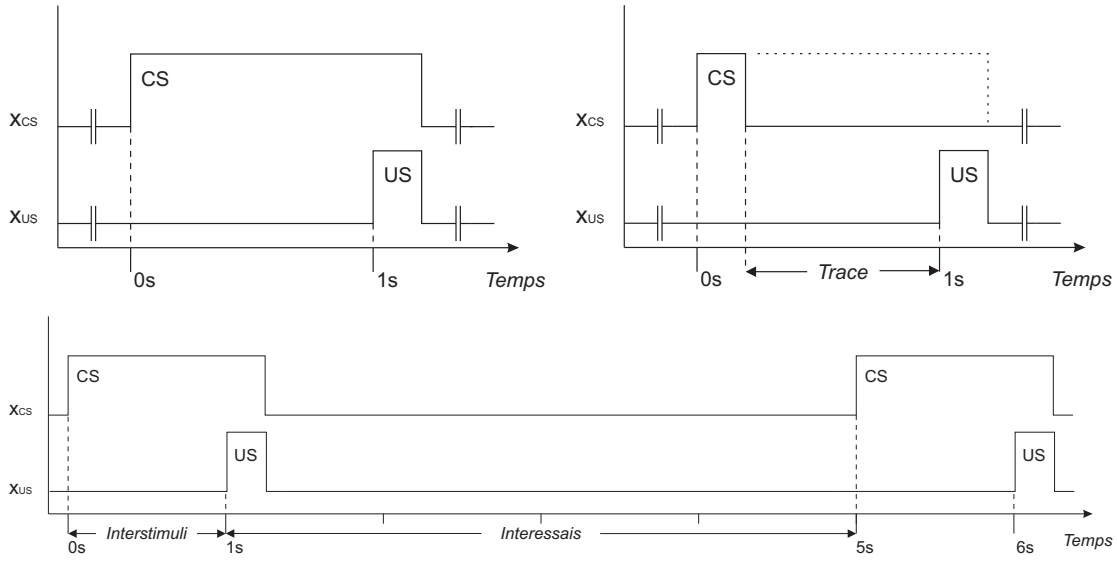


Figure 3-2 : Conditionnement classique. En haut à gauche, conditionnement classique à délai fixe de 1 sec. En haut à droite, conditionnement de trace à délai fixe de 1 sec. En bas, série d'essais à délai interstimuli fixe de τ_{ISI} = 1 sec avec une période d'interessais de τ_{ITI} = 4 sec.

Il existe aussi d'autres effets importants. Par exemple, soit A et B, deux stimuli. Un animal qui est d'abord conditionné sur la séquence A→US, puis sur la séquence B→A ou B→A→US, développera une réponse conditionnée au stimulus B (*second conditionnement*). Par contre, un animal qui est d'abord conditionné sur la séquence A→US, puis entrainé sur la séquence AB→US, où les stimuli A et B sont présentés simultanément, ne développera pas de conditionnement au stimulus B. Celui-ci est dit *masqué* ou *bloqué* par le stimulus A, qui pour l'animal, explique déjà l'arrivée de l'US. Un animal peut aussi apprendre des règles non linéaires. Par exemple, il peut être conditionné à A→US, B→US et AB→∅, auquel cas la réponse à AB disparaîtra.

Ceci n'est qu'un survol des principaux effets d'intérêts pour cette thèse. Il y a toute une panoplie d'études sur le préconditionnement, reconditionnement et l'extinction, et les effets qui sont généralement observés. Pour un traitement plus complet, voir (Dickinson & Mackintosh, 1978; Gallistel & Gibbon, 2000).

3.1.5 Apprentissage associatif : le conditionnement opérant

Le conditionnement instrumental (ou opérant) est une forme un peu différente, où l'on utilise le conditionnement comme instrument pour apprendre une tâche à un animal. Au lieu d'associer un nouveau stimulus à une association stimulus-réponse existante, on associe une récompense à une action de l'animal (Skinner, 1938). Par exemple, supposons une simple cage munie d'un levier contrôlant un distributeur de nourriture. Lorsqu'un rat appuie sur le levier (l'action), un peu de nourriture devient accessible à celui-ci (la récompense ou le renforcement positif). Dans ce cas, l'animal viendra à associer l'action qu'il vient de poser avec la récompense qu'il reçoit tout de suite après. Dans une forme plus complexe, on peut rendre le renforcement dépendant d'une action de l'animal et du contexte (ou stimulus). Par exemple, si le rat appuie sur le levier lorsqu'une lumière est allumée, il peut recevoir de la nourriture, et lorsque la lumière est éteinte, recevoir une décharge électrique. Ici, ce n'est plus seulement l'association d'un stimulus à une réponse physiologique, mais plutôt l'association d'une action, ou d'une séquence d'actions, de l'animal à une récompense. Le cadre mathématique de l'apprentissage par renforcement (section 2.3) convient relativement bien au conditionnement instrumental (incluant le conditionnement classique).

On peut aussi étudier la capacité d'un animal à mesurer le temps en le récompensant pour appuyer sur le levier seulement après un certain délai. La forme la plus simple de ce paradigme est l'horaire à intervalle fixe (FI, de l'anglais *fixed-interval*) (Skinner, 1938). Dans ce protocole, l'animal peut appuyer sur le levier autant qu'il le souhaite, mais ne reçoit une récompense que pour la première pression après un certain délai fixé par l'expérimentateur. À ce moment, une récompense est donnée et le chronomètre est redémarré pour un nouvel intervalle. On marque généralement le début d'un essai par un stimulus et on laisse une période d'interessais sans stimulus entre la récompense et le début de l'essai suivant comme pour le conditionnement (Figure 3-2, en bas). Après entrainement, on observe généralement une réponse moyenne de plus en plus soutenue au fur et à mesure que l'on se rapproche du délai fixé (Figure 3-3, coté gauche).

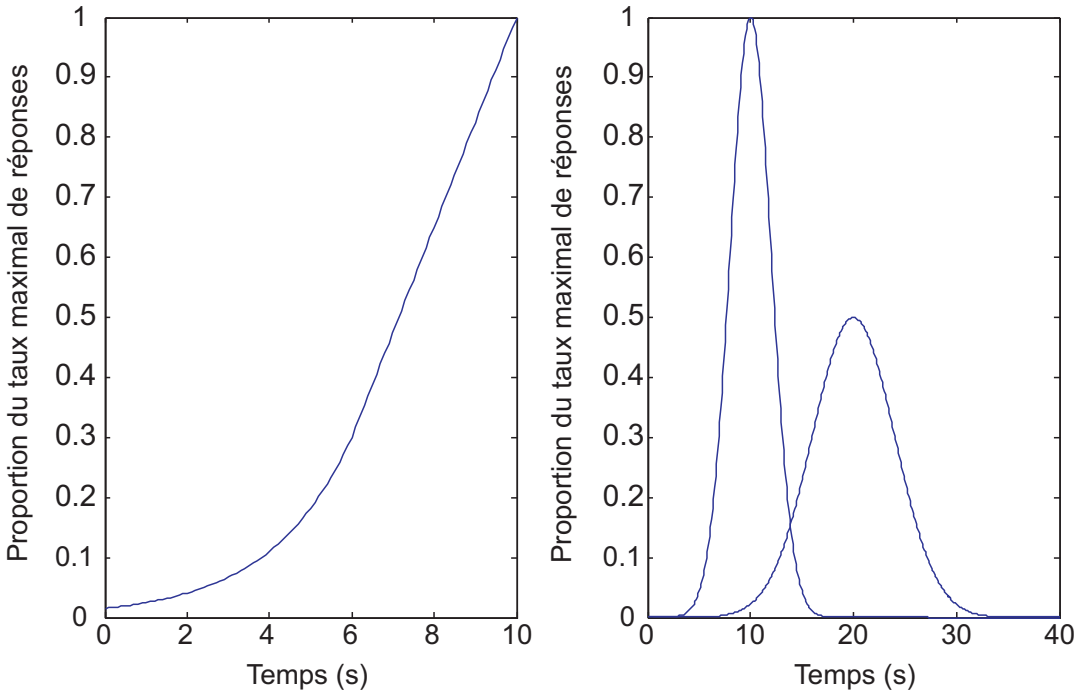


Figure 3-3 : Courbe de réponse en fonction du temps écoulé depuis le début de l'essai. À gauche : courbe de réponse en fonction du temps écoulé dans un horaire FI. À droite : courbes de proportion de réponses en fonction du temps écoulé pour des délais de 10 sec et 20 sec sur les essais tests d'un horaire PI, les deux expériences ayant le même délai interessais. (Données simulées.)

Une version plus élaborée, appelée intervalle de pointe (PI, de l'anglais *peak-interval*) (Roberts, 1981), consiste à remplacer certains essais par des essais de test. Pour ces essais, le stimulus reste allumé plus ou moins trois fois la durée du délai habituel et aucune récompense n'est donnée, peu importe le comportement de l'animal. Ainsi, l'animal peut apprendre que non seulement il aura une récompense après un certain délai, mais qu'en plus, si la récompense n'arrive pas après ce même délai, elle n'arrivera pas du tout. On peut ainsi voir quand l'animal décide d'arrêter d'appuyer en l'absence de récompense. Après une très longue période d'entrainement, on observe généralement une courbe de réponse moyenne en fonction du temps écoulé symétriquement distribuée autour du délai fixé (Figure 3-3, à droite). L'incertitude quant à la durée de l'intervalle est généralement proportionnelle à la durée du stimulus (loi de Weber, Figure 3-3, à droite), et ce, pour des durées allant de plusieurs secondes à plusieurs heures (Meck, 2003).

Pour une brève introduction aux procédures FI et PI, voir (Hopson, 2003). Pour une analyse plus récente sur le développement des courbes de réponse dans ces procédures, voir (Balci et al., 2009).

3.2 Modèles théoriques de l'apprentissage animal

Il y a quelques modèles théoriques importants qui tentent d'expliquer les comportements en situation d'habituation et de conditionnement. Ces modèles ne sont pas basés sur des fondements neurobiologiques, mais il reste nécessaire de les survoler rapidement. Il y a tout d'abord les modèles d'habituation et de conditionnement de Rescorla-Wagner et de Pearce, le modèle de conditionnement classique et opérant par différence temporelle (TD) et finalement le modèle de la théorie d'espérance scalaire sur les intervalles de temps. Des idées de modèles neuronaux de représentation du temps seront aussi présentées.

3.2.1 Rescorla-Wagner, TD, et Pearce

Le modèle Rescorla-Wagner (Rescorla & Wagner, 1972) est un des modèles de base du conditionnement classique et de certains effets d'habituation. Ce modèle associe une valeur de récompense *V()* au stimulus lors du conditionnement et fournit une règle pour son acquisition. Soit un vecteur de stimuli $\boldsymbol{s} = [s_1, \ldots, s_N]$ (CSs) et une récompense (US) r :

$$V(\boldsymbol{s}) = \sum_{i=1}^{N} w_i s_i, \qquad \textbf{Équation 3-1}$$

$$w_i \leftarrow w_i + \alpha\delta s_i, \text{ où} \qquad \textbf{Équation 3-2}$$

$$\delta = r - V(\boldsymbol{s}) \qquad \textbf{Équation 3-3}$$

et α un facteur d'apprentissage.

Ce modèle est équivalent à faire une descente de gradient stochastique sur les paramètres w_i pour minimiser le carré de l'erreur de prédiction de la récompense r $(r-V(\boldsymbol{s}))^2$. On peut le voir comme une forme d'apprentissage supervisé (section 2.1). Pour simuler un essai, il suffit d'appliquer la règle avec le vecteur $\boldsymbol{s}$ approprié. Par exemple, si seul le stimulus s_1 est visible, $\boldsymbol{s} = [1, 0, \ldots, 0]$. Plusieurs des effets de conditionnement classique peuvent être reproduits à l'aide de ce modèle, dont l'effet de masque (voir section 3.1.4). D'autres, comme le second conditionnement, ne le sont pas.

Le modèle TD, le même algorithme qu'au Chapitre 2 (section 2.3.3), permet de tenir compte de plus d'effets que le modèle Rescorla-Wagner. Dans l'exemple du second conditionnement, supposons un conditionnement à A→US, suivis du conditionnement à B→A ou B→A→US. Le second conditionnement ne peut être appris par le modèle Rescorla-Wagner, puisque A n'est pas une récompense (r). Par contre, la règle de TD, reproduite ici en Équation 3-4, le peut. Après le premier conditionnement, la valeur de récompense associée au stimulus A est incluse dans la règle d'apprentissage via le terme $V(\boldsymbol{s_{t+1}})$. C'est ce terme qui différentie le modèle Rescorla-Wagner (Équation 3-3) du modèle TD. Parmi les nombreux articles sur le sujet, mentionnons (Sutton & Barto, 1990).

$$\delta_{t+1} = r_{t+1} + \gamma V(\boldsymbol{s_{t+1}}) - V(\boldsymbol{s_t}) \qquad \textbf{Équation 3-4}$$

D'autres situations ne peuvent tout simplement pas être modélisées par ces modèles sans ajout, dont la combinaison de règles A→US, B→US, AB→∅ (voir section 3.1.4). Cette règle requiert d'avoir une autre entrée représentant la conjonction AB. Il y a un certain nombre de solutions au problème de représentation, pour différentes approches, voir (Pearce, 1994). Pour une revue récente des modèles, voir (Pearce & Bouton, 2001).

3.2.2 Théorie d'espérance scalaire

Le temps est une dimension critique dont on doit tenir compte pour apprendre les régularités dans notre monde en perpétuel mouvement. Cependant, les fondements de cette capacité pour des délais de l'ordre des secondes ou minutes restent encore méconnus (Buhusi & Meck, 2005). La théorie d'espérance scalaire (traduction libre de SET, de l'anglais *scalar expectancy theory*), vieille de plus de 30 ans (Gibbon, 1977), reste encore aujourd'hui celle expliquant le plus grand nombre de phénomènes, cependant, elle ne donne quasi aucune indication sur leurs fondements neurobiologiques.

Le modèle SET consiste principalement en une horloge qui *tic* à chaque petit délai Δt, un accumulateur, une mémoire, et une décision (Figure 3-4). Un signal spécifique indiquant le début de l'intervalle permet à l'accumulateur (a_t) d'accumuler des *tics*. À chaque fois, il est comparé à la même valeur (m^*) tirée de la mémoire de

référence au début de l'accumulation pour décider s'il est temps de prendre action ou non. Un bruit stochastique est ajouté au niveau de la décision en choisissant une nouvelle borne (b_t) ayant une distribution normale (moyenne $\mu = 0$, écart type $\sigma = 1$) à chaque instant t. La réaction (pour prendre la récompense) débute lorsque

$$(a_t - m^*)/m^* \geq b_t. \qquad \textbf{Équation 3-5}$$

Lorsqu'il y a récompense, la valeur de l'accumulateur (a_t) est ajoutée à la mémoire. C'est cette structure qui permet au modèle d'être invariant en fonction de l'échelle de temps et donc de suivre la loi de Weber (section 3.1.5, Figure 3-3, à droite). En effet, si l'on réécrit l'Équation 3-5,

$$(a_t - m^*) \geq b_t\, m^*, \qquad \textbf{Équation 3-6}$$

on observe que la borne du test (maintenant donnée par $b_t\ m^*$) a une distribution normale dont l'écart type est proportionnel à m^*. Donc que l'erreur de temps de la décision sera proportionnelle à la durée de l'intervalle m^*. Pour une revue exhaustive de ce modèle et des données qu'il couvre voir (Gallistel & Gibbon, 2000).

3.2.3 Modèles de représentations neuronales du temps

On reconnait généralement que le cerveau peut maintenir le temps à trois différentes échelles sans aide extérieure, chacune étant probablement soutenue par des mécanismes différents (Buhusi & Meck, 2005). Il y a l'échelle des millisecondes, pour le contrôle moteur, la parole et la musique par exemple et dans laquelle le cervelet joue possiblement un rôle important. Ensuite, il y a l'échelle de la seconde aux minutes, discutée ici. Puis finalement, l'échelle de la journée, incluant la faim, le rythme éveil-sommeil, etc., possiblement basée sur des mécanismes moléculaires et contrôlée par l'expression de différents gènes. On en connaît cependant relativement peu sur les bases neurobiologiques de la mesure du temps à l'échelle des secondes et le sujet suscite encore de nombreux débats (Dragoi et al., 2003; Hopson, 2003; Gallistel, 2003; Ivry & Spencer, 2004; Buhusi & Meck, 2005; Ivry & Schlerf, 2008).

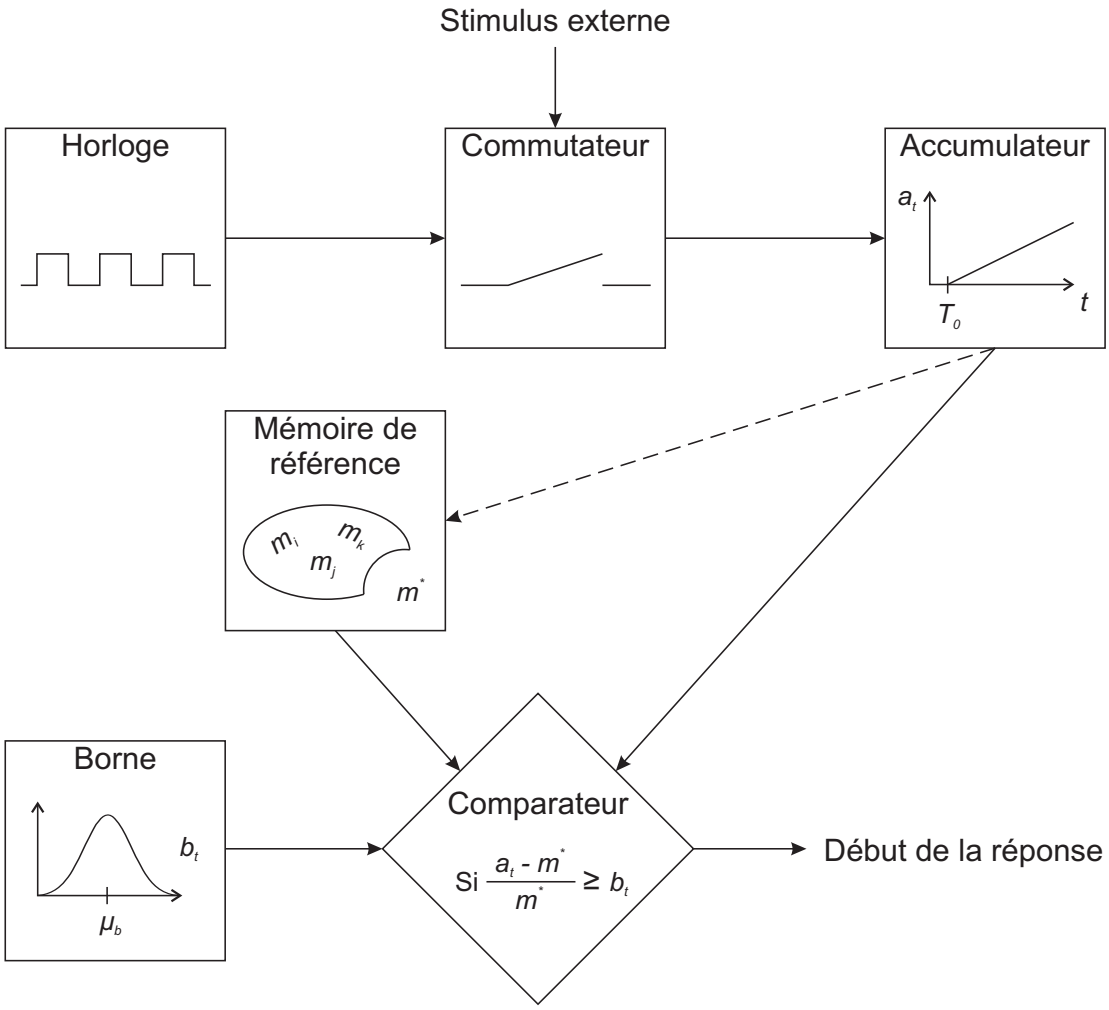


Figure 3-4 : Schéma du modèle SET.

Au niveau de la seconde et des minutes, on peut séparer les représentations neuronales imaginables en trois grandes classes (Durstewitz, 2004). Premièrement, on peut imaginer un code généré par une population de neurones bistables où chaque neurone est soit très actif, soit peu actif. Dans cette population, plus le temps s'écoule, plus il y a de neurones actifs (Figure 3-5). On peut aussi, imaginer que certains neurones puissent jouer un rôle d'intégrateur temporel, en augmentant leurs activités au fur et à mesure que le temps s'écoule à partir d'un stimulus quelconque (Figure 3-6). Ce second type d'activités a été observé à de nombreuses reprises dans le système nerveux dans différentes expériences où il y a un délai fixe de l'ordre des secondes (Niki & Watanabe, 1979; Funahashi, Bruce, & Goldman-Rakic, 1989; Komura et al., 2001; Romo, Brody, Hernandez, & Lemus, 1999; Lucchetti & Bon,

2001; Brody, Hernandez, Zainos, & Romo, 2003; Leon & Shadlen, 2003; Reutimann et al., 2004; Lucchetti, Ulrici, & Bon, 2005; Lebedev, O'Doherty, & Nicolelis, 2008).

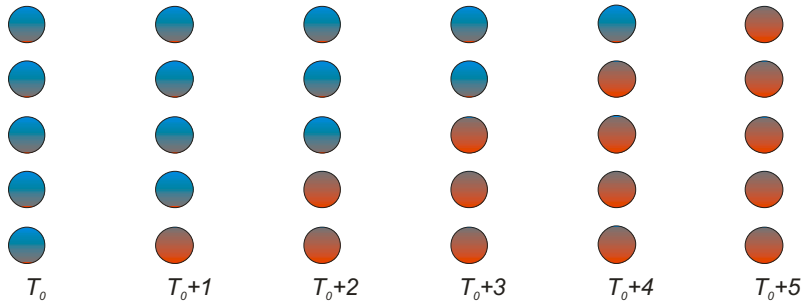


Figure 3-5 : Représentation neuronale du temps par le nombre de neurones bistables actifs.

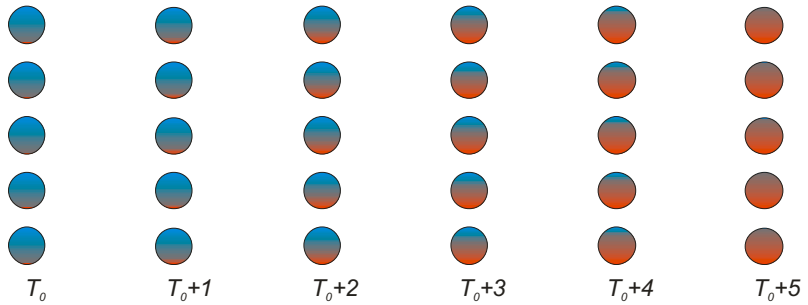


Figure 3-6 : Représentation neuronale du temps par une population d'intégrateurs temporels.

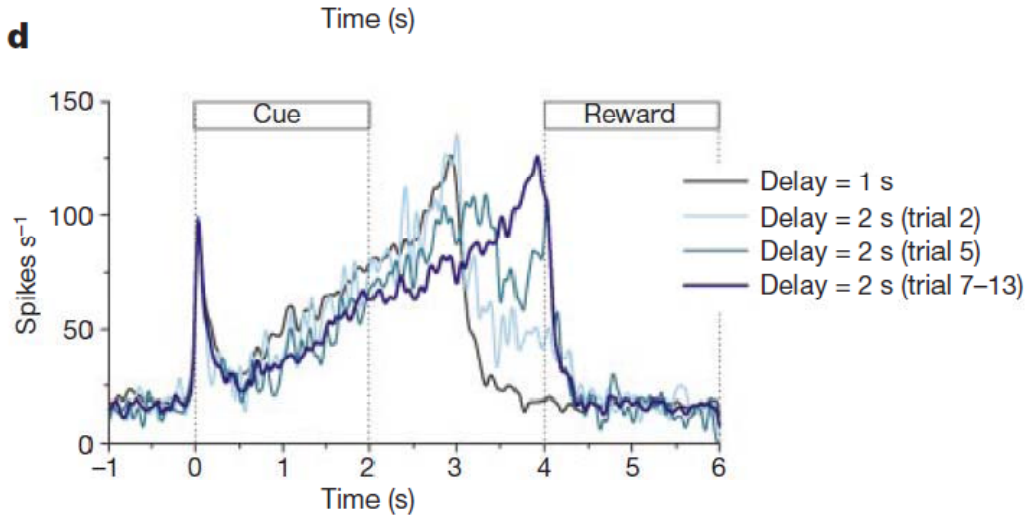


Figure 3-7 : Exemple d'activité de neurones qui augmente en fonction du délai. Dans cet exemple, l'intervalle entre le CS et la récompense (US) a soudainement passé de 1 sec à 2 sec. On peut y observer l'évolution de l'activité du neurone pendant l'adaptation au nouvel intervalle. Reproduit avec permission de Macmillan Publishers Ltd: *Nature* (Komura et al., 2001), copyright (2001).

Finalement, on peut aussi penser à une représentation quelconque construite sur une séquence d'états neuronaux sans accumulation claire telle une série de neurones actifs les uns après les autres dans un ordre précis (Figure 3-8). Cette dernière semble cependant plus réaliste pour de très courtes durées de moins de 500 ms (Buonomano, 2005; Karmarkar & Buonomano, 2007). Cette idée ressemble aux lignes de délai (section 2.1.3) qui semblent exister à l'échelle des millisecondes (Carr & Konishi, 1990). Des représentations du temps par séquences d'activations semblent aussi possibles pour des délais de l'ordre des millisecondes aux centaines de millisecondes (Hahnloser, Kozhevnikov, & Fee, 2002; Buonomano, 2003) et pourraient émerger naturellement dans une population de neurones (Buonomano, 2005). Cependant, ce type de représentations semble moins réaliste pour des durées de l'ordre des secondes. Il concorde mieux aux données comportementales sous des intervalles de très courtes durées qu'aux données pour les intervalles de l'ordre des secondes (Buonomano, 2005; Karmarkar & Buonomano, 2007).

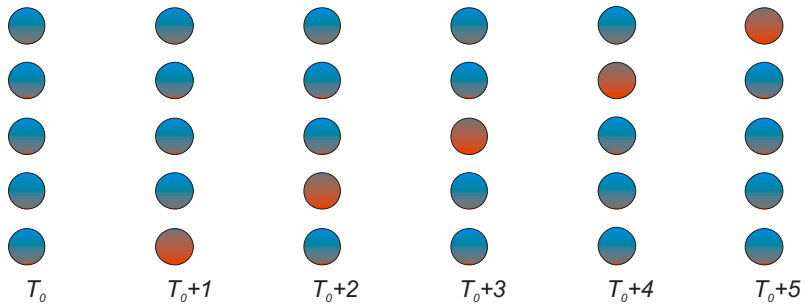


Figure 3-8 : Représentation neuronale du temps par une série d'états neuronaux distincts.

Une autre idée plus récente est qu'il n'y ait pas de mécanisme spécifique pour le temps à l'échelle des secondes ou minutes, mais plutôt que la représentation puisse être apprise à partir de mécanismes d'apprentissage généraux (Dragoi et al., 2003; Hopson, 2003). Cette idée ne remet cependant pas en questions les trois représentations précédentes, mais encourage plutôt à ne pas forcer la représentation dans la construction d'un modèle. Pour une discussion récente sur le sujet, voir (Ivry & Schlerf, 2008). Pour une revue des modèles du temps, voir aussi (Hopson, 2003).

3.3 Anatomie fonctionnelle de l'apprentissage dans le cerveau

Le cerveau est composé de plusieurs structures anatomiques distinctes, les plus petites étant souvent appelées des noyaux (ou ganglions). Le néocortex, ou plus simplement le cortex, compose la majeure partie du cerveau. Mais d'autres parties, moins volumineuses, sont aussi connues pour jouer un rôle important dans les différentes formes d'apprentissage. Bien qu'il n'y ait généralement pas de relation un à un entre les formes d'apprentissage et les structures du cerveau, cette section fait un survol des différentes structures du cerveau connues pour jouer un rôle important dans l'apprentissage. Chacun des principaux systèmes d'apprentissage dont le cortex, les ganglions de la base, l'hippocampe et le cervelet y sont brièvement décrits, ainsi que leurs fonctions et leur modèle au besoin.

3.3.1 Le cortex

Le cortex est la principale masse du cerveau (Figure 3-9). Certains l'ont considéré comme équipotent (Mountcastle, 1978), c'est-à-dire que chaque partie pourrait réaliser le même traitement d'information si elle recevait les mêmes entrées. Bref, le cortex visuel n'aurait de visuel que le fait que sa principale source d'entrées provienne de l'oeil. Chaque région du cortex est en fait une pile de six couches de neurones organisées différemment. Certaines couches servent principalement d'entrées, d'autres de sorties, d'autres de traitement quelconque. D'une région à l'autre du cerveau, l'épaisseur de ces couches varie. Bien que l'idée d'équipotentialité soit peu réaliste et exagérée (Mountcastle, 2003), elle suggère qu'un algorithme d'apprentissage modélisant bien une région corticale puisse servir de point de départ pour en modéliser une autre. Parmi les différences importantes entre les régions du cortex, le cortex frontal est beaucoup plus innervé par le système dopaminergique que la majorité des autres régions corticales. La partie frontale du cortex est celle dont le volume est beaucoup plus grand chez l'humain que chez le singe comparativement aux autres régions. Quant aux neurones dopaminergiques, ils font partie des neuromodulateurs. Ils sont situés dans les ganglions de la base qui jouent un rôle important dans l'apprentissage procédural et le conditionnement opérant (voir section 3.3.2 et Chapitre 4).

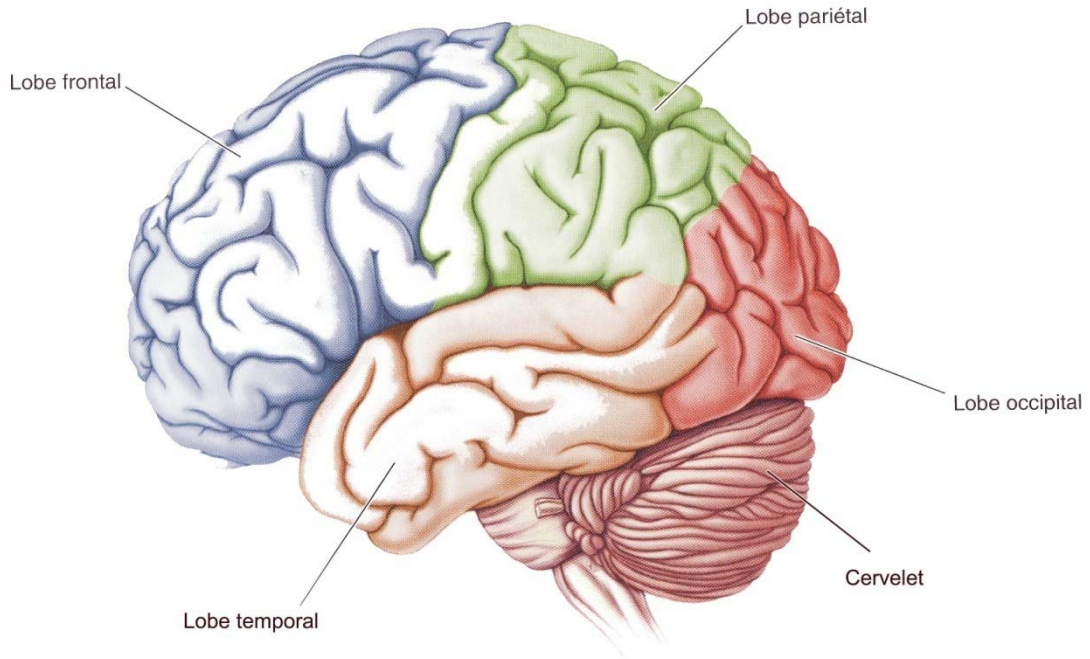


Figure 3-9 : Vue latérale du cerveau. On y voit principalement quatre lobes du cortex ainsi que le cervelet, la petite structure située en bas à droite. Adapté avec permission de (Bear, Connors, & Paradiso, 2002, p. 215).

Plus on s'éloigne à partir d'une aire sensorielle primaire (l'aire visuelle V1 par exemple, dans le lobe occipital, Figure 3-9) vers le cortex frontal, plus les neurones que l'on y retrouve répondent à des stimuli complexes. Par exemple, dans l'aire visuelle primaire, les neurones ne répondent qu'aux stimuli dans une toute petite zone du champ visuel, et principalement qu'à des contrastes importants tels que des lignes. Plus loin, dans le cortex pariétal, certains neurones répondront à la présence de stimulus à certaines positions dans l'espace autour de soi, plutôt qu'à un point de l'image sur la rétine. Les neurones de cette région peuvent aussi être modulés par l'importance des stimuli, la direction de l'attention, le contexte de la tâche et bien d'autres facteurs. Finalement, dans le cortex frontal, on peut retrouver des neurones qui ne répondent qu'à des combinaisons très complexes de signaux ou des neurones qui semblent jouer un rôle important dans la mémoire de travail, c'est-à-dire le maintien actif de certaines informations utiles pour la tâche en cours (par exemple, Romo et al., 1999), information qui n'est généralement plus visible (observable). Par exemple, dans une tâche où l'on donne une information ou une instruction par un stimulus, suivi d'un délai sans stimulus avant que l'information ne puisse être utilisée,

certains neurones peuvent être activés par le premier stimulus et restés actifs pendant la période de délai, représentant la mémoire active de l'information ou de l'instruction reçue (Romo et al., 1999; Brody et al., 2003). On peut aussi trouver des neurones dont l'activité augmente pendant l'écoulement d'un délai entre le stimulus initial et l'action suivante dans plusieurs régions corticales, incluant les cortex temporal, pariétal et frontal (Niki & Watanabe, 1979; Funahashi et al., 1989; Romo et al., 1999; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Reutimann et al., 2004; Lucchetti et al., 2005; Lebedev et al., 2008), un peu à l'instar de l'idée de neurones intégrateurs (voir section 3.2.3, Figure 3-6). Finalement, on

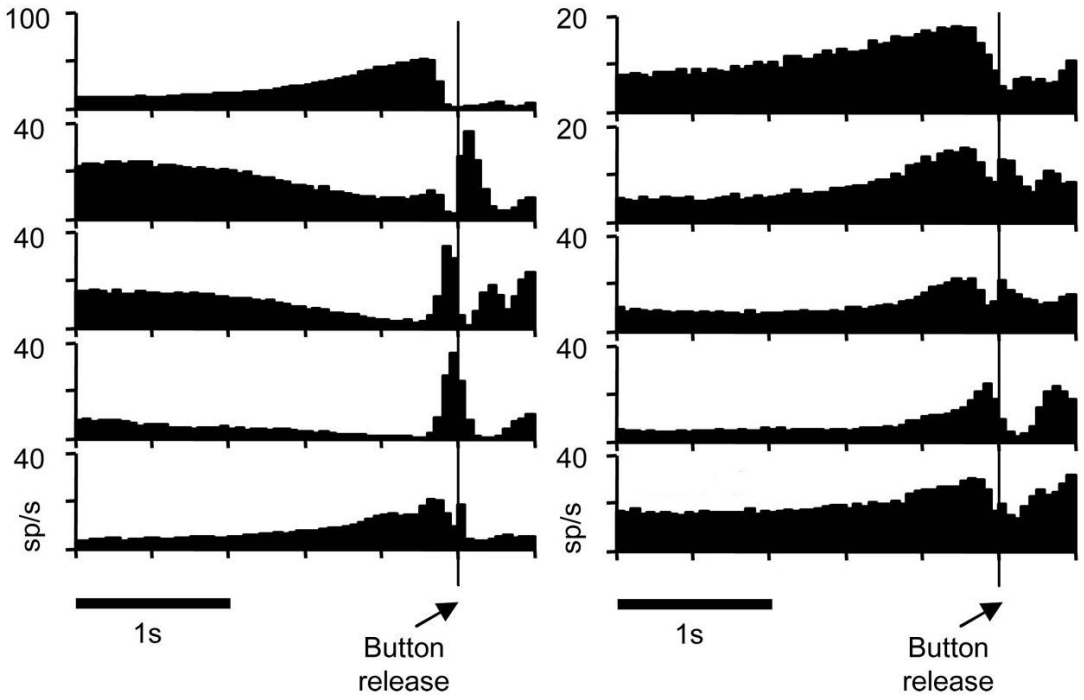


Figure 3-10 : Exemple d'activités de neurones corticaux pendant le passage du temps avant que l'animal ne relâche le bouton. Dans cette tâche, l'animal doit maintenir le bouton pendant au moins 2.5 sec et pas plus de 4.5 sec pour obtenir une récompense, sans information externe pour mesurer le temps. Adapté avec permission de (Lebedev et al., 2008, Figure 11).

retrouve dans le cortex frontal (et le cortex pariétal) les neurones miroirs. Ces neurones répondent généralement à une action sur un objet, que celle-ci soit effectuée par l'animal lui-même ou par un autre, mais pas à l'imitation du geste (Rizzolatti & Craighero, 2004). Le même neurone peut répondre, que l'action soit effectuée par le singe, ou simplement observée visuellement ou auditivement, comme déchirer du

papier (Kohler et al., 2002). Encore plus impressionnant, certains neurones pourraient répondre lorsque l'action peut-être inférée sans être directement observée (Umilta et al., 2001), une situation semblable à l'observabilité partielle (Figure 2-14). Par exemple, supposons un neurone qui répond à l'action de prendre un objet. On place un panneau devant l'objet, puis l'expérimentateur tend le bras derrière le panneau pour prendre l'objet. Le neurone répondra au moment où l'expérimentateur devrait atteindre l'objet en fonction de la partie visible du bras de l'expérimentateur. Bref, plus on s'éloigne des régions sensorielles primaires vers le cortex frontal, plus le niveau d'abstraction est généralement élevé.

D'un point de vue mathématique, le cortex est souvent comparé à un algorithme d'apprentissage non supervisé (section 2.2) (Rao & Ballard, 1999; Doya, 2000; Doi et al., 2003; Hawkins & Blakeslee, 2004; Hyvarinen et al., 2005) : c'est-à-dire qu'il apprend sa représentation des entrées simplement à partir de ces dernières, sans signal d'erreur externe. Par exemple, on réussit à modéliser le champ de sensibilité des neurones du cortex visuel à partir d'un algorithme non supervisé tel que ICA (Doi et al., 2003; Hyvarinen et al., 2005) et les modèles d'O'Reilly (O'Reilly, 1998; O'Reilly & Rudy, 2000; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly & Frank, 2006). Hawkins (Hawkins & Blakeslee, 2004) quant à lui propose que le cortex serve principalement à prédire les prochaines entrées similairement à l'autosupervision (section 2.2.5). Il y a cependant peu de modélisation de l'apprentissage dans le cortex en comparaison à ses nombreuses régions et sa grande diversité.

Le cortex pourrait donc être décrit par un (ou plusieurs) algorithme(s) d'apprentissage et serait entre autres, le siège des représentations abstraites (structurales).

3.3.2 Ganglions de la base et système dopaminergique

Les ganglions de la base forment l'un des systèmes d'apprentissages les mieux modélisés, résultats de recherches des quinze dernières années. La principale raison de cet avancement est la découverte de ressemblances entre les algorithmes

d'apprentissage machine par renforcement tel que TD et l'activité de certains neurones, en particulier, les neurones dopaminergiques.

Il est important de décrire les principales composantes des ganglions de la base et du système dopaminergique (suivre sur la Figure 3-11 ci-dessous). La porte d'entrée des ganglions de la base est le striatum (en orange). Celui-ci est composé de deux sections principales, le striosome et le matrisome. Le striosome (en vert), dans le modèle acteur-critique (Barto, 1995) d'apprentissage par différence temporelle (TD), représente l'apprentissage de la récompense espérée; c'est-à-dire que cette section tente de prédire la récompense à venir. Ce signal est dirigé vers deux autres noyaux (en bleu), la SNc (substance noire *pars compacta*) et le VTA (aire tegmentale ventrale ou *ventral tegmental area*). Ces deux zones sont la source du signal dopaminergique (DA, en bleu), considéré comme équivalent au signal d'erreur sur la récompense prédite δ dans TD (Équation 3-4, voir aussi sections 2.3.5 et 3.2.1) (Montague, Dayan, & Sejnowski, 1996; Schultz et al., 1997). C'est ce même signal qui est envoyé au cortex frontal. Ce signal est aussi retourné au striatum servant à moduler le changement synaptique, c'est-à-dire à indiquer si les synapses (poids entre les neurones) doivent changer (Reynolds & Wickens, 2002). Quant au matrisome (en orange), l'autre partie du striatum, il projette dans d'autres noyaux, pour finalement retourner vers le cortex frontal. Il représente possiblement la partie acteur du système acteur-critique (Barto, 1995), c'est-à-dire qu'il contrôlerait avec le cortex les actions à prendre. Le Chapitre 4 contient une revue plus complète de l'anatomie, de l'électrophysiologie, de la plasticité synaptique et des modèles de l'apprentissage des ganglions de la base et du système dopaminergique.

Finalement, le signal dopaminergique semble avoir un lien avec la motivation intrinsèque, le contrôle du comportement sexuel, les effets de dépendance, etc. Les ganglions de la base sont aussi perçus comme un élément potentiellement important de l'apprentissage procédural et de l'automatisation. Le signal dopaminergique a aussi été démontré comme étant un élément nécessaire dans l'effet de récompense de la nouveauté (Bevins et al., 2002). Les ganglions de la base et le système

dopaminergique forment donc une partie importante de ce projet et seront décrits plus en profondeur dans le prochain chapitre.

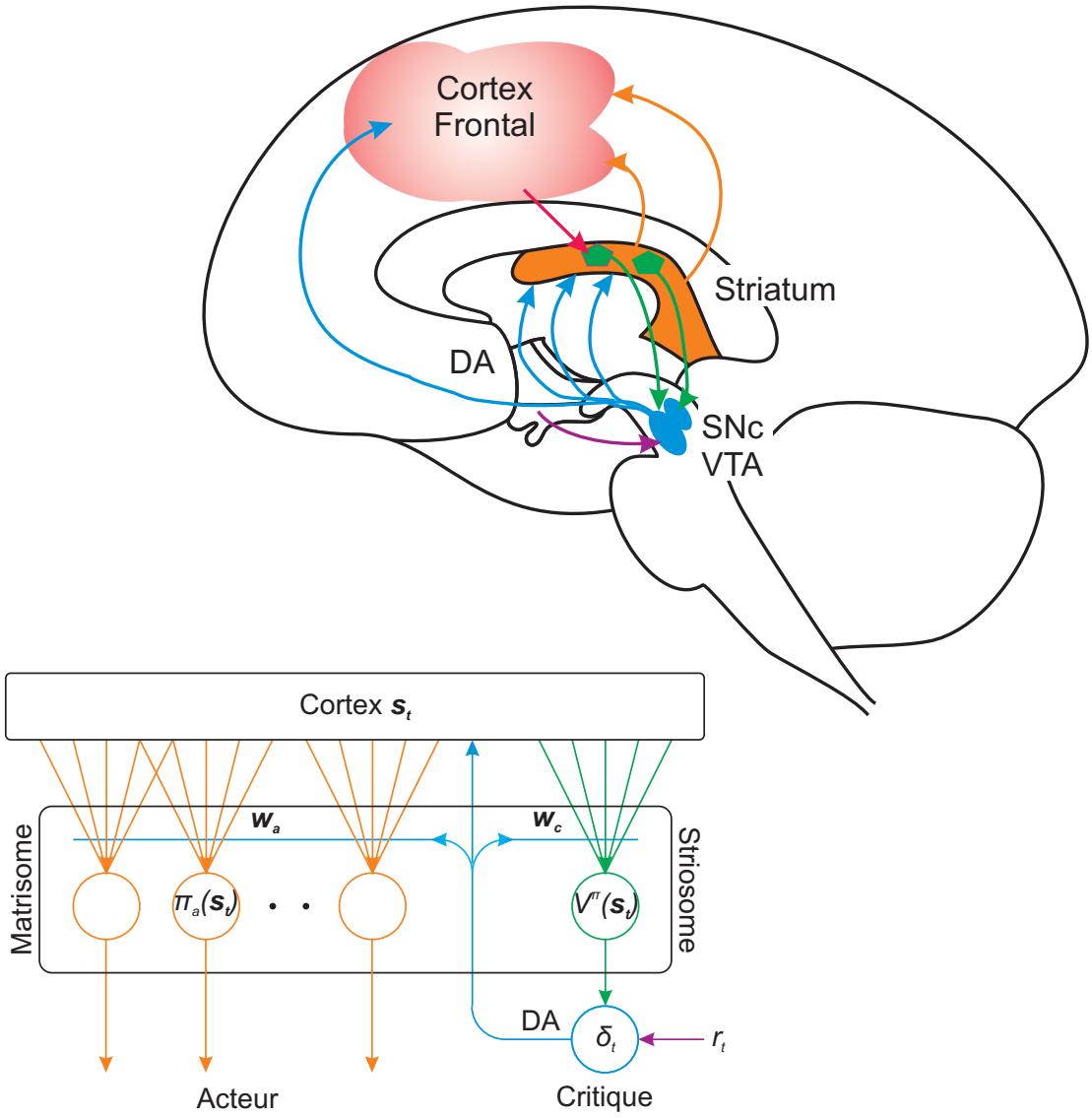


Figure 3-11 : Schéma simplifié du cortex et des ganglions de la base. À gauche, vue médiane du cerveau. À droite, modèle acteur critique.

3.3.3 L'hippocampe

L'hippocampe est au cœur de la mémoire explicite. Son rôle dans l'apprentissage procédural ou les autres formes d'apprentissage implicite reste toutefois moins clair.

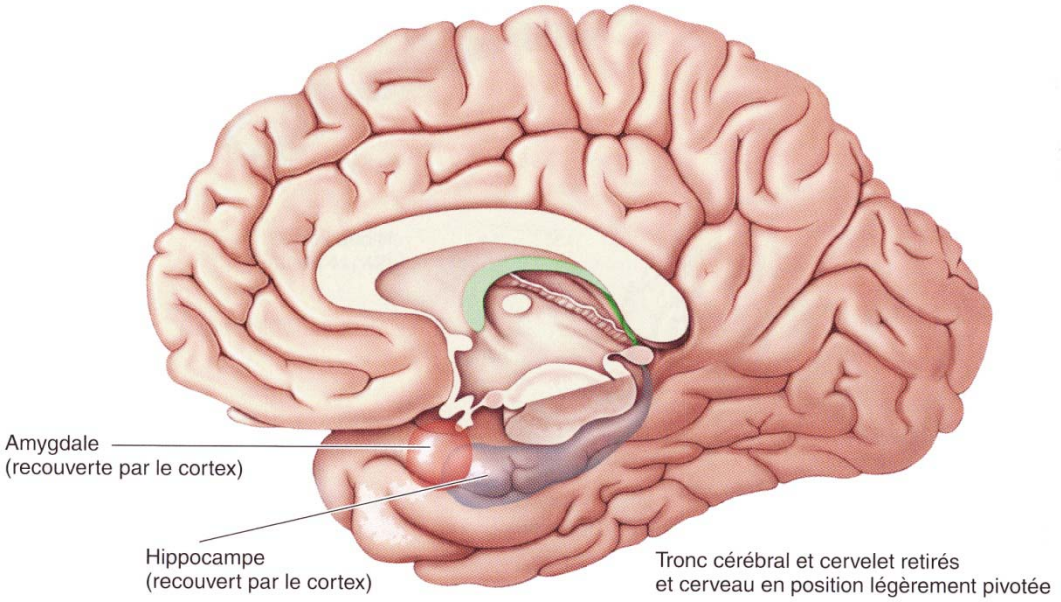


Figure 3-12 : Vue médiane (du milieu) du cerveau. On peut entre autres y apercevoir l'hippocampe et l'amygdale. Reproduit avec permission de (Bear et al., 2002, p. 218).

Un patient, du nom de H. M., a subi un enlèvement neurochirurgical bilatéral de l'hippocampe et des régions avoisinantes pour soulager une épilepsie grave. Ce patient a perdu sa mémoire antérograde, c'est-à-dire la capacité à former de nouveaux souvenirs. Tout ce qu'il apprenait après son opération, il l'oubliait dès que son esprit pensait à autre chose. « Bonjour, je me présente, je m'appelle François, comment allez-vous ? » « Très bien, merci François. » « Comptez d'un à dix svp. » « Un, deux, …, neuf, dix. » « Quel est mon nom? » « Je ne sais pas, vous ne me l'avez pas dit. »[3] Cependant, à des tâches manuelles, et avec de la pratique quotidienne, H. M. pouvait s'améliorer significativement, sans jamais se rappeler avoir pratiqué la tâche auparavant (Milner, Corkin, & Teuber, 1968).

Bien que l'hippocampe ne semble pas nécessaire à l'apprentissage procédural ni à un grand nombre de conditionnements, il semble toutefois accélérer l'apprentissage et devient requis lorsqu'il faut associer de façon complexe plusieurs stimuli (Gluck & Myers, 2001). Par exemple, il semble que le conditionnement de trace (section 3.1.4, Figure 3-2, en haut à droite) requière l'hippocampe lorsque le

[3] Texte inventé pour les besoins de l'exemple, mais similaire en contenu à la réalité.

délai interstimuli est supérieur à 2 sec (Clark & Squire, 1998; Beylin et al., 2001). Il est aussi possible que ce dernier joue un rôle important dans la détection de nouveauté (Sirois & Mareschal, 2004; Lisman & Grace, 2005).

L'hippocampe n'est pas non plus le siège des souvenirs en soi, car H. M. n'avait pas perdu les connaissances explicites qu'il avait longtemps avant son opération (Milner et al., 1968). L'hippocampe semble, entre autres, jouer le rôle d'une forme de mémoire tampon, entre la mémoire de travail, ce que l'on tient activement en mémoire pendant que l'on pense à quelque chose, et la mémoire à long terme. Bien que son rôle et son importance y soient discutés, l'hippocampe n'est pas l'objet de modélisation dans cette thèse.

3.3.4 Le cervelet

Le cervelet (Figure 3-9) représente une composante importante de l'apprentissage moteur. Bien qu'on ait longtemps cru que son rôle soit principalement moteur, il semble aussi jouer un rôle cognitif, entre autres au niveau du langage et de la notion de temps (voir Chapitre 5).

Lorsque vous attrapez un objet et qu'il est plus léger ou pesant que prévu, vous ajustez très rapidement la force exercée. Le cervelet joue un rôle important dans ce type d'adaptation (Lang & Bastian, 1999). Des troubles au cervelet entrainent principalement des difficultés d'anticipation musculaire, de coarticulations musculaires, d'adaptations musculaires. Il y a aussi une grande littérature sur le cervelet et le conditionnement aversif. La réponse non conditionnée est souvent musculaire, par exemple, cligner des yeux suite à un souffle d'air dans les yeux.

Il existe plusieurs modèles du cervelet dont le modèle Marr-Albus (Marr, 1969; Albus, 1971) récemment revu (Ito, 2000). Il serait entre autres possible de modéliser le cervelet comme une forme d'apprentissage supervisé (Doya, 1999, 2000), bien qu'une composante non supervisée puisse être de la partie (Schweighofer, Doya, & Lay, 2001). Il y a aussi des travaux très intéressants joignant cervelet et ganglions de la base en contrôle moteur adaptatif (Doya, Kimura, & Miyamura, 2001).

Le cervelet joue clairement un rôle sur le contrôle moteur et les sensations qui lui sont associées dans un monde où l'espace, le temps et les forces sont des variables continues. Dans la mesure où l'on se limite à des tâches n'ayant que des actions discrètes dans un environnement stable, on peut généralement éviter de modéliser le cervelet. La majorité des travaux présentés dans cette thèse n'ont pas d'actions du tout et le cervelet n'y sera pas modélisé. Pour plus de justifications, voir le Chapitre 5.

3.3.5 Autres régions cérébrales

Il serait trop long et peu utile pour ce projet de mentionner toutes les régions du système nerveux reliées à l'apprentissage. De chaque senseur, en passant par la moelle épinière, au cerveau, puis de nouveau à la moelle, et jusqu'à chaque muscle, tous ces éléments sur le parcours de l'observation à l'action, possèdent un certain niveau d'adaptabilité. Un autre noyau mérite cependant d'être mentionné, l'amygdale (Figure 3-12). Ce noyau semble jouer un rôle important dans certaines formes de conditionnement aversif et la mémoire émotionnelle, ainsi que dans le phénomène de la peur. Cependant, cette thèse porte principalement sur l'apprentissage avec récompense, et non avec punitions, l'amygdale ne sera donc pas discutée.

3.3.6 Hypothèses sur l'apprentissage animal dans le cerveau

L'objectif de cette thèse est de tenter de mieux comprendre le développement de représentations abstraites dans le cerveau dans un contexte d'apprentissage par renforcement. Les conditionnements classique et opérant sont des formes d'apprentissage par renforcement (section 3.1.5) et peuvent être modélisés par des algorithmes de cette famille (section 3.2.1). Les ganglions de la base et le système dopaminergique semblent jouer un rôle important dans cet apprentissage (section 3.3.2) et une revue plus en profondeur de cette région du cerveau et de ses différents modèles sera donc présentée au Chapitre 4. Finalement, le cortex semble un candidat idéal où pourraient se développer des représentations abstraites (section 3.3.1), incluant des représentations du temps (sections 3.2.3 et 3.3.1). Comment le temps est-il représenté dans le cerveau, cette représentation est-elle acquise et quelles en sont les bases neurobiologiques, restent le sujet de nombreux débats (section 3.2.3).

Chapitre 4. Ganglions de la base et système dopaminergique

Plusieurs évidences montrent que les ganglions de la base et les neurones dopaminergiques du mésencéphale jouent un rôle majeur dans l'apprentissage par renforcement, l'apprentissage de séquence et les phénomènes de dépendance (drogue, jeux pathologiques, etc.) (Montague, Hyman, & Cohen, 2004). Ils sont au coeur d'une famille de modèles informatiques basés sur l'apprentissage par renforcement (TD) où la réponse phasique des neurones dopaminergiques semble corréler de façon stupéfiante avec le signal d'erreur de prédiction de récompense (δ). Ce chapitre est donc consacré à cette région du cerveau et aux modèles en lien avec cette thèse.

Ce chapitre se divise en quatre sections, en commençant par une brève revue de l'anatomie des ganglions de la base et des neurones dopaminergiques et quelques simplifications. La deuxième partie consiste en une revue des effets principaux retrouvés dans les enregistrements extracellulaires *in vivo.* Ceci consiste en une revue de l'activité des neurones en fonction des tâches effectuées par des animaux. La troisième section est une courte révision de la plasticité synaptique à l'intersection des afférences corticales et dopaminergiques à l'entrée des ganglions de la base et leur lien possible avec l'apprentissage. Finalement, le chapitre se termine sur une revue critique des modèles informatiques d'apprentissage les plus importants de cette région du cerveau.

4.1 Anatomie

Les ganglions de la base (Figure 4-1) sont situés profondément au centre du cerveau (à sa base) et comprennent le *striatum*, le *globus pallidus* (substance pâle) et une partie de la *substance noire*. Les neurones dopaminergiques sont situés dans le mésencéphale et forment une partie de l'*aire tegmentale* et de la substance noire. Le mésencéphale est juste sous les ganglions de la base[4].

[4] Dans la littérature scientifique francophone du siècle précédent ont parlait aussi de *noyaux gris centraux*, mais cette notation détermine des régions du cerveau légèrement différentes de la dénomination anglo-saxonne des *ganglions de la base* pour des raisons anatomiques plutôt que fonctionnelle (Yelnik, 2002). Comme Yelnik, on utilisera ici la dénomination anglo-saxonne de ganglions de la base qui convient mieux ici et qui est généralement acceptée.

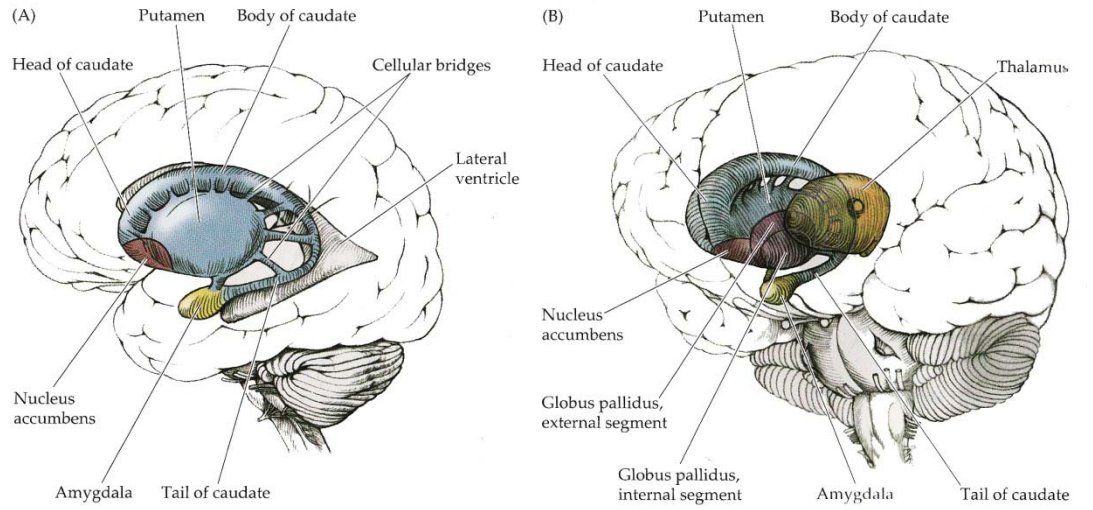


Figure 4-1: Les ganglions de la base. En bleu le putamen et le noyau caudé (tête, corps et queue). En rouge leur jonction dans le striatum ventral : le noyau accumbens. En mauve le globus pallidus, segment interne et externe. Reproduit avec permission de (Blumenfeld, 2002, Figure 16-1, p.690).

4.1.1 La porte d'entrée : le striatum

Le striatum est la principale porte d'entrée de l'information des ganglions de la base. Cette structure peut-être divisée de différentes façons (Figure 4-1). On reconnait visuellement le *putamen* (PT) au milieu rattaché au *noyau caudé* (CD, *caudate nucleus*) qui fait comme une corne autour. Il y a aussi le *noyau accumbens* (NAc, *nucleus accumbens*) situé à la jonction de la tête du noyau caudé et du striatum dans la partie *ventrale* (du bas). On appelle aussi *striatum ventrale* la partie inférieure du putamen et du noyau caudé, incluant le noyau accumbens. Le NAc est aussi parfois séparé en cœur (*core,* la partie *dorsale*, du haut) et coquille (*shell*, la partie ventrale), mais cette distinction n'est pas claire chez le primate (Joel & Weiner, 2000). Dans ce travail, nous utiliserons tout simplement le terme striatum et supposerons une certaine uniformité de celui-ci pour le moment.

Le striatum reçoit principalement trois types d'entrées (afférences). La principale source d'entrées provient d'afférences excitatrices glutamatergiques en provenance du cortex, du système limbique, dont l'hippocampe et l'amygdale, et du thalamus, le principal relais entre les sens et le cortex. Ces entrées sont relativement ségrégées et dépendent de la position où l'on se situe dans le striatum. Une revue de

l'organisation de ces entrées au striatum apparaît en Figure 4-2 (Voorn, Vanderschuren, Groenewegen, Robbins, & Pennartz, 2004). La seconde source d'entrées la plus importante dans le cadre de cette recherche est le système dopaminergique. Il n'est cependant toujours pas clair si ces entrées jouent un rôle d'inhibition et d'excitation (selon le type de récepteurs, voir section 4.1.3), de gain (Montague et al., 2004) ou de plasticité (section 4.3), ou une combinaison de tout ceci. Encore une fois, il pourrait y avoir une certaine topographie entre la localisation des projections dopaminergiques dans le striatum et leurs sources (Haber, 2003). Finalement, la troisième source d'entrées dans le striatum provient des neurones sérotoninergiques des *noyaux du Raphé*. Les noyaux du Raphé font partie du système sérotoninergique qui diffuse dans presque tout le cerveau et qui joue un rôle dans les cycles éveil-sommeil, l'état de vigilance, l'humeur et le comportement émotionnel (Bear et al., 2002). Ces derniers ne seront pas traités ici.

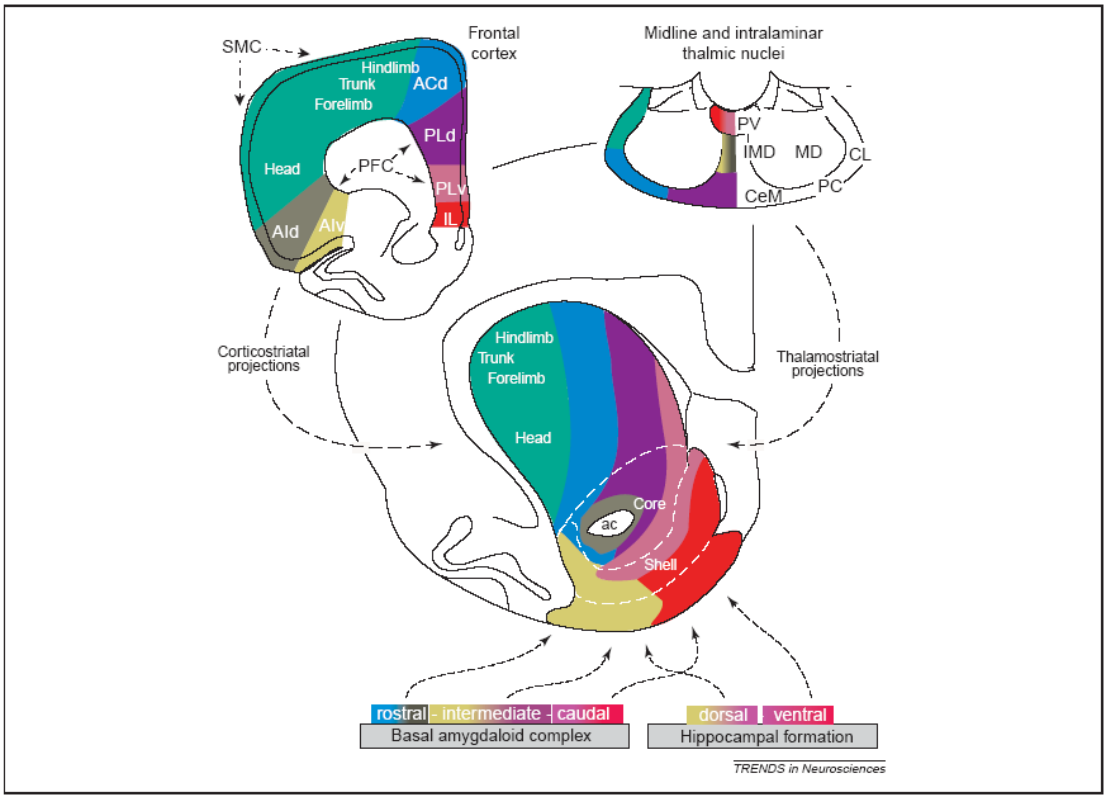


Figure 4-2: Entrées glutamatergiques du striatum en fonction de la localisation dans le plan coronal. Reproduit de *Trends in Neurosciences*, 27, Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. A., Putting a spin on the dorsal–ventral divide of the striatum, Pages 468-474, Copyright (2004), avec permission d'Elsevier.

Le striatum est composé d'au moins deux types de neurones : les neurones de projections épineux moyens (MSN, de l'anglais *medium spiny neuron*) et les interneurones actifs toniquement (TAN, de l'anglais *tonically active neuron*). Les MSN représentent environ 95 % des neurones du striatum. Ce sont des neurones inhibiteurs qui projettent principalement sur le globus pallidus et la substance noire (voir Figure 4-3). Les TAN sont de gros interneurones que l'on croit être cholinergiques et dont l'activité est parfois corrélée au signal dopaminergique (Apicella, 2007). Cette recherche ne descendra pas au niveau de ces interneurones, cependant, leur position et leur activité laissent croire qu'ils pourraient aussi jouer un rôle dans la plasticité hétérosynaptique des synapses corticostriales sous le contrôle de la dopamine (section 4.3).

4.1.2 La porte de sortie : GPi/SNr

Le *segment interne* du globus pallidus (GPi) et la substance noire *pars reticulata* (SNr) forment la principale voie de sortie des ganglions de la base vers le thalamus, un relais important vers le cortex. Comme pour le striatum, nous considérerons simplement que ces deux structures sont semblables et qu'elles diffèrent principalement par leur site de projection. Comme la majorité des entrées sensorielles et des sorties du cervelet, les sorties des ganglions de la base passent par le thalamus avant d'aller au cortex. En agissant sur le thalamus, ces neurones ont donc des projections indirectes vers le cortex ainsi que vers les principales voies motrices descendantes. Un schéma des connexions des ganglions de la base est présenté en Figure 4-3.

4.1.3 Voies directe et indirecte

Il fut longtemps accepté qu'il y a deux principales voies du striatum vers les sorties GPi/SNr (Figure 4-3): les voies directes et indirectes (Albin, Young, & Penney, 1989). La voie directe correspondrait principalement aux MSN qui projettent directement vers la porte de sortie SNr ou GPi. La voie indirecte correspondrait principalement aux MSN qui projettent d'abord dans le *segment externe* du globus pallidus (GPe), qui projette dans le *noyau sous-thalamique* (STN), qui lui projette dans les noyaux de sorties (GPi/SNr). La connectivité réelle est cependant beaucoup

plus complexe. Ces deux voies ont aussi été différentiées par leur type de récepteurs dopaminergiques (Gerfen et al., 1990), l'un ayant un effet excitateur et l'autre inhibiteur sur les MSN. Des études récentes remettent cependant cette ségrégation en question (Aizman et al., 2000; Levesque & Parent, 2005). Pour simplifier la modélisation, ces voies de sortie seront considérées comme une seule voie. Le traitement des connectivités complètes de ces voies va bien au-delà du présent projet.

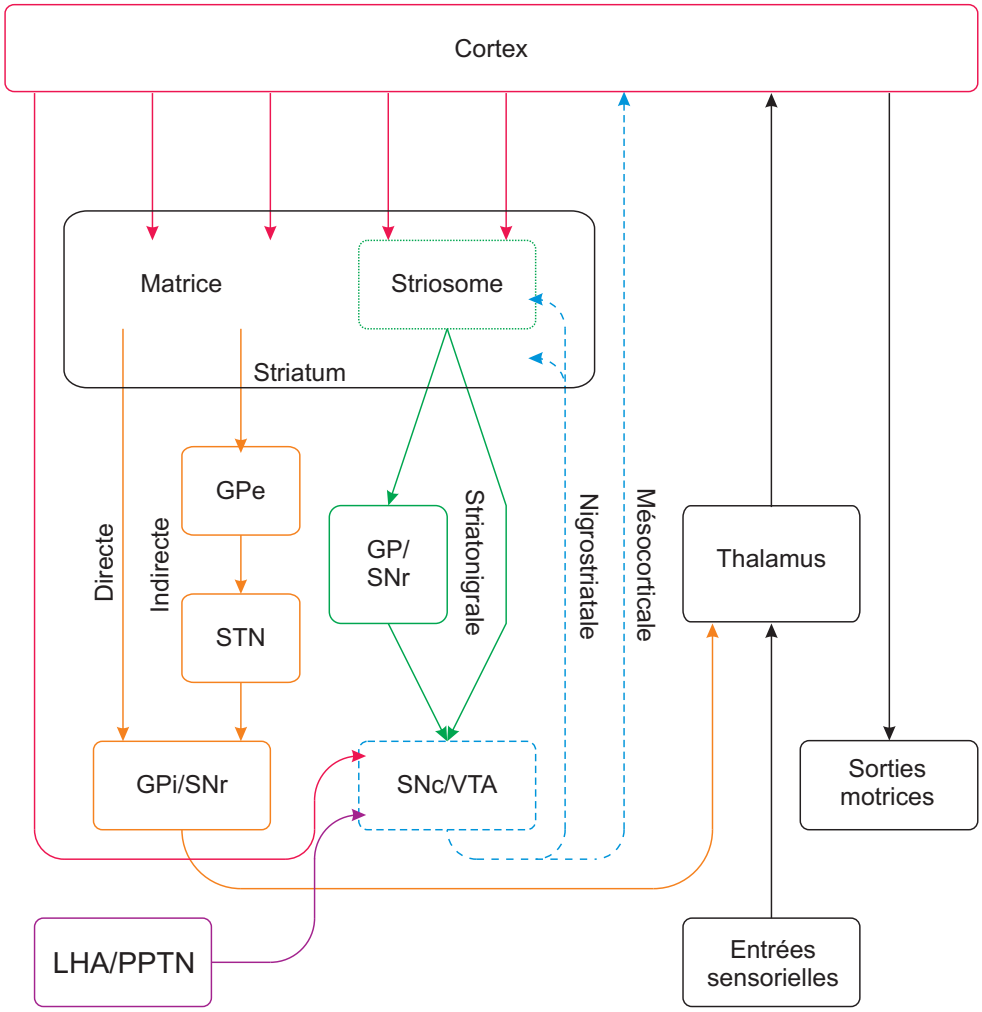


Figure 4-3 : Schéma des connexions principales dans les ganglions de la base. À l'intérieur : striatum (le striosome est en pointillé), globus pallidus (GP), noyau sous-thalamique (STN), substance noire (SN), et aire tegmentale ventrale (VTA). Les voies directe, indirecte et striatonigrale y sont aussi représentées. Les projections dopaminergiques sont en ligne tiretée. À l'extérieur : hypothalamus latéral (LHA), noyau pédonculopontine (PPTN), thalamus et cortex. Les connexions avec le système limbique ne sont pas montrées. Les mêmes couleurs qu'en Figure 3-11 sont utilisées.

4.1.4 Voie striatonigrale

Le striatum peut aussi être divisé en deux parties à partir de certains marqueurs chimiques (Gerfen, 1984) : le *striosome* (ou *patch*) et le *matrisome* (ou *matrice*). Cette division est particulièrement visible dans la partie dorsale du striatum. Le striatum, principalement du matrisome, est comme tacheté de striosome. Certaines études suggèrent que les neurones du striosome feraient contact sur les neurones dopaminergiques de la substance noire, alors que les neurones de la matrice conduiraient plutôt aux voies directes et indirectes de sorties (SNr, GP) (Jimenez-Castellanos & Graybiel, 1989; Gerfen, 1992). Est-ce que cette distinction existe vraiment? Est-ce que certains neurones du striatum dorsal projettent directement sur les neurones dopaminergiques, ou indirectement via le SNr? Le débat reste ouvert (Levesque & Parent, 2005). La partie ventrale du striatum pourrait aussi avoir des projections directes (Gerfen, 1985; Gerfen, 1992) ou indirectes (Groenewegen, Berendse, & Haber, 1993), via le pallidum ventral, sur les neurones dopaminergiques. Certains détails varient aussi selon l'espèce animale étudiée. L'existence d'une projection directe ou indirecte de neurones striataux vers des neurones dopaminergiques semble toutefois bien réelle; c'est le seul élément qui sera retenu ici.

4.1.5 Voies parallèles ségrégées

Des études anatomiques de connectivité sont à la base de l'idée que les ganglions de la base, des entrées striatales aux sorties vers le thalamus, forment des boucles parallèles plus ou moins indépendantes (Alexander, DeLong, & Strick, 1986; Strick, Dum, & Picard, 1995). Par exemple, le traitement des informations relatives au système oculomoteur en provenance des aires visuelles et des aires visuelles supplémentaires par les ganglions de la base, jusqu'aux sorties vers les centres oculomoteurs, forment l'une de ces boucles. Ces boucles utilisent toutes une mécanique semblable, mais chaque région du striatum se concentre sur une fonctionnalité différente selon ses entrées et sorties.

4.1.6 Système dopaminergique mésencéphalique

Il y a trois principales sources de neurones dopaminergiques : les aires A8 ou l'*aire rétrorubrale*, A9 ou la substance noire *pars compacta* (SNc) et A10 ou l'aire

tegmentale ventrale (VTA) (Joel & Weiner, 2000). Comme mentionné dans la section précédente, il semble aussi y avoir une certaine topographie entre les neurones dopaminergiques et leurs cibles, mais ce système est beaucoup plus diffus.

D'une façon générale, on parle de trois projections dopaminergiques. La projection nigrostriatale origine principalement du SNc et projette dans le striatum dorsal (noyau caudé et putamen) et ailleurs dans les ganglions de la base. La projection mésolimbique origine surtout de la VTA et projette vers le système limbique dont l'hippocampe et l'amygdale ainsi que vers la partie plus ventrale du striatum. Finalement, la projection mésocorticale, aussi issue de la VTA, projette dans tout le cortex, particulièrement dans le cortex frontal. Bien que générale, cette définition des projections dopaminergiques reste la plus simple et la mieux établie (Kandel et al., 2000). Les projections mésocorticale et mésolimbique sont parfois réunies dans l'appellation mésocorticolimbique.

Les neurones dopaminergiques reçoivent entre autres des entrées excitatrices de l'*hypothalamus latéral* (LHA), du *noyau pédonculopontine* (PPTN), de l'*habénula* et de tout le cortex (Geisler, Derst, Veh, & Zahm, 2007). Le LHA et le PPTN pourraient signaler les récompenses. Les neurones dopaminergiques reçoivent aussi possiblement des entrées inhibitrices du striatum (section 4.1.4) et d'interneurones locaux (Hebb & Robertson, 2000).

Encore une fois, nous ne nous attarderons pas trop à la localisation des sources dopaminergiques puisque de toute façon la plupart des enregistrements cellulaires de l'activité des neurones dopaminergiques dont il sera question ici sont simplement identifiés comme venant de la grande région d'A8, A9 et A10. Il est donc difficile d'identifier si ces signaux sont différents en fonction de leur position morphologique ou de leurs projections. Par contre, nous prendrons en considération les cibles majeures, dont le striatum et le cortex (voir Figure 4-3).

4.1.7 En résumé

Il va de soit que le résumé présenté ici et en Figure 4-3 n'est qu'un survol des éléments clefs de l'anatomie des ganglions de la base et des neurones dopaminergiques pour permettre une revue des modèles de cette région du cerveau.

Plusieurs autres noyaux, interneurones et projections ont été volontairement omis pour ne pas alourdir inutilement cette section. De plus, les modèles développés dans cette thèse seront encore plus simplifiés et seront beaucoup plus près de l'anatomie rapportée en Figure 3-11. Pour un traitement plus approfondi de la connectivité des ganglions de la base et des débats actuels sur le sujet, voir (Parent & Hazrati, 1995a; Parent & Hazrati, 1995b; Graybiel, Canales, & Capper-Loup, 2000; Joel & Weiner, 2000; Yelnik, 2002; Haber, 2003; Voorn et al., 2004).

4.2 Électrophysiologie

L'enregistrement de l'activité électrique des neurones pendant qu'un animal effectue une tâche précise permet d'estimer l'information signalée par ces neurones. La fréquence de décharge est souvent l'une des mesures les plus appropriées pour faire cette estimation. Par exemple, le neurone est-il plus actif, c'est-à-dire a-t-il une plus haute fréquence de décharge, en présence d'un stimulus visuel ou d'un stimulus auditif? Sa fréquence encode-t-elle une propriété de ce stimulus? Le neurone réagit-il à la nouveauté ou à la disparition d'objet? Et ainsi de suite.

4.2.1 Neurones dopaminergiques

Les neurones dopaminergiques des aires A8, A9, et A10 ont une fréquence de décharge de base sous les 10 Hz. Cependant, lors de la présence de stimuli associés à une récompense, il peut y avoir une brève réponse quelque 150 ms plus tard sous forme d'une bouffée de décharges à 20 Hz ou plus (Hyland, Reynolds, Hay, Perk, & Miller, 2002). On considère généralement ces deux signaux, la fréquence de base et la réponse brève au stimulus, comme étant deux signaux importants et distincts. Le premier est le signal tonique, c'est-à-dire la tonalité de base, et le second est le signal phasique, c'est-à-dire le signal en phase avec l'évènement. C'est ce dernier qui est considéré comme signalant une erreur de prédiction de récompense (Montague et al., 1996; Schultz et al., 1997) et qui sera étudié ici. Il y a toute une série d'effets classiques qui peuvent être observés dans des neurones dopaminergiques.

4.2.1.a) *Conditionnement*

Les neurones dopaminergiques semblent coder l'erreur de prédiction de récompense δ du modèle TD (Équation 4-1). Par exemple, un grand nombre de neurones dopaminergiques répondent à l'arrivée inattendue d'une récompense. On peut observer une transition pendant le conditionnement classique (Figure 4-4, à gauche). Avant le conditionnement, les neurones ne répondent qu'à la récompense imprévue (US). Rapidement, ils commencent à répondre au stimulus (CS) précédant la récompense. Éventuellement, une grande partie des neurones ne répondent plus qu'au stimulus conditionné (CS) (Ljungberg, Apicella, & Schultz, 1992; Schultz, Apicella, & Ljungberg, 1993; Pan, Schmidt, Wickens, & Hyland, 2005). Après le conditionnement, la majorité de ces neurones dopaminergiques répondent donc positivement au stimulus imprévu (CS) et répondront moins à la récompense prévisible (US) qui suit. Ils continueront toutefois de répondre à une récompense imprévue (Figure 4-4, signal 1 à droite). Si l'on omet la récompense après le stimulus conditionné, les neurones dopaminergiques répondront négativement par une réduction de leur fréquence sous la fréquence de base en l'absence de récompense (Figure 4-4, signal 3 à droite) (Ljungberg et al., 1992; Schultz et al., 1993; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004; Pan et al., 2005).

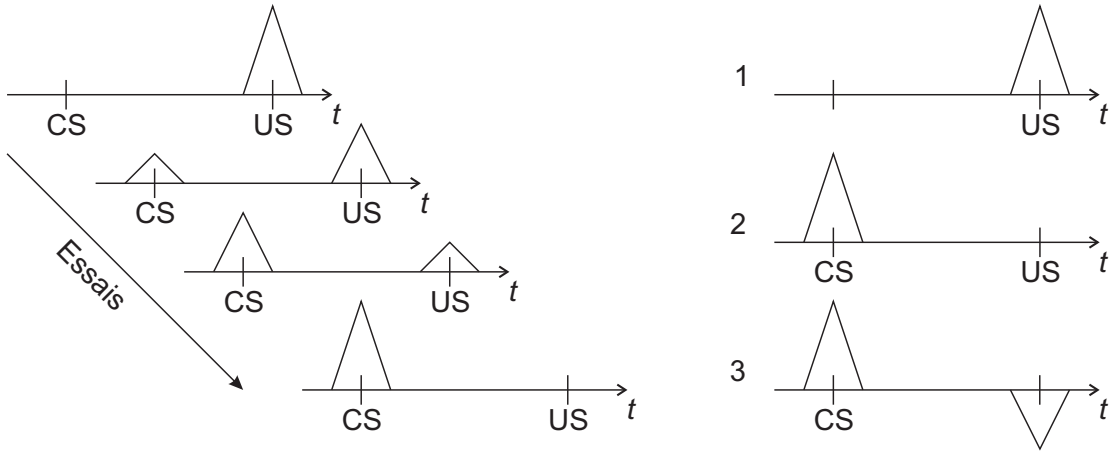


Figure 4-4 : Réponse des neurones dopaminergiques sous conditionnement. À gauche, évolution de la réponse au stimulus conditionné (CS) et à la récompense (US) pendant le conditionnement. À droite, réponse à des essais après conditionnement : 1- Récompense sans stimulus conditionné (récompense imprévue); 2- Essai normal; 3-Stimulus conditionné sans récompense (récompense prévue manquante).

4.2.1.b) *Séquence*

Si l'on allonge la séquence en présentant d'abord un stimulus B, puis du stimulus A (le CS), puis la récompense (séquence B→A→US, section 3.2.1), les neurones dopaminergiques répondront éventuellement à A, puis à B, plutôt qu'à la récompense (Schultz et al., 1993; Pan et al., 2005). Si, après l'entrainement, on omet A uniquement, on obtient une dépression au temps prévu de A et une activité à la récompense maintenant inattendue (Pan et al., 2005).

Dans le modèle TD (section 3.2.1), la valeur de la récompense serait associée à A et B de sorte que $V(B) = V(A) = 1$ et le signal dopaminergique correspondrait au signal d'erreur de prédiction de récompense

$$\delta_{t+1} = r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}}) - V^{\pi}(\boldsymbol{s_t}), \qquad \textbf{Équation 4-1}$$

pour les états observés $\boldsymbol{s}_t$ et $\boldsymbol{s}_{t+1}$ aux temps t et $t+1$ et une récompense donnée r_{t+1}, Lorsque B apparaît, $V()$ passe soudainement de 0 à 1, expliquant la réponse dopaminergique positive lors de son apparition. Lorsque B disparaît et que A n'apparaît pas, alors l'estimation $V()$ retombe soudainement à 0. Expliquant la réponse négative. Comme il n'y a plus de récompense prévue, les neurones dopaminergiques ont alors une réponse positive à la récompense.

4.2.1.c) *Masque et inhibition*

Les neurones dopaminergiques concordent aussi avec TD pour le masque et l'inhibition (section 3.1.4). En effet, si l'on conditionne l'animal d'abord sur le stimulus A (CS), puis sur la paire de stimuli AB présentés simultanément, les neurones dopaminergiques ne développeront pas de réponse importante à B présenté seul, mais répondront à la récompense qui suivra (effet de masque, Waelti, Dickinson, & Schultz, 2001). À l'opposé, si l'on conditionne d'abord l'animal sur le stimulus A, puis qu'on ajoute la paire de stimuli AB présentés simultanément sans récompense, les neurones dopaminergiques développeront une réponse positive au stimulus A, puis perdront leur réponse à la paire AB. Testés sur B sans récompense, ils répondront de façon plus modérée à B, voire même négativement. Cependant, testés

sur B avec récompense, ils répondront à la récompense imprévue (inhibition de conditionnement, Tobler, Dickinson, & Schultz, 2003).

Probabilité et magnitude : D'une façon plus générale, les neurones dopaminergiques peuvent coder la probabilité d'une récompense et sa grandeur. Par exemple, supposons que l'on assigne à cinq stimuli différents des probabilités de récompense distinctes. Après conditionnement, les neurones dopaminergiques répondront plus fortement aux stimuli ayant les plus fortes probabilités de récompense. À l'opposé, lorsque la récompense arrivera, ils répondront moins fortement aux récompenses les plus probables selon le stimulus présenté juste avant. Si la récompense n'est pas présentée, alors la dépression sera proportionnelle à la probabilité de récompense associée au stimulus (Fiorillo, Tobler, & Schultz, 2003). Il en va de même pour la magnitude de la récompense. Si le stimulus prédit une récompense plus grande, la réponse dopaminergique au stimulus conditionné sera plus grande. Si un même stimulus est suivi la moitié du temps par une grande récompense, la moitié du temps par une petite récompense, alors la réponse à la récompense sera semblable à la différence entre la récompense moyenne et celle obtenue, c'est-à-dire une dépression pour la récompense plus petite et une activation pour la récompense la plus grande (Tobler, Fiorillo, & Schultz, 2005). Donc, les neurones semblent coder la différence entre la récompense espérée (moyenne) et la récompense reçue.

Deux phénomènes semblent toutefois déroger à cette règle. Premièrement, lorsque la probabilité de récompense ne permet pas de très bien faire une prédiction telle qu'une probabilité de récompense de 50%. Dans cette situation, certains neurones dopaminergiques montrent une activité croissante ou soutenue durant le délai entre le stimulus et la récompense. Cette activité est généralement proportionnelle à l'incertitude (une probabilité de 50% ayant une grande incertitude de récompense alors qu'une probabilité de 0% ou de 100% a une petite incertitude) et à la grandeur de la récompense attendue (Fiorillo et al., 2003). Deuxièmement, l'encodage de la magnitude de la récompense dépend de l'échelle des récompenses utilisées. Même si les récompenses possibles pour un stimulus A sont dix fois plus

grandes que pour un autre stimulus B, les réponses dopaminergiques aux différentes récompenses respectives seront relativement les même pour A et B. Les neurones dopaminergiques pourraient coder la différence entre la récompense reçue et espérée en fonction de la magnitude des récompenses pouvant être associées au stimulus présenté (Tobler et al., 2005).

4.2.1.d) *Délai fixe et passage du temps*

Les neurones dopaminergiques offrent aussi le même genre de réponses proportionnelles en fonction du délai entre le stimulus et la récompense. Si quatre stimuli sont associés à quatre délais différents, alors la réponse au CS après conditionnement sera plus grande pour le stimulus ayant le plus court délai, et inversement, la réponse à la récompense sera la plus grande pour le stimulus ayant le plus long délai (Fiorillo, Newsome, & Schultz, 2008; Kobayashi & Schultz, 2008). Si le délai est tiré d'une distribution uniforme, la réponse à la récompense sera la plus faible au délai moyen, et plus grande si la récompense est plus tôt ou plus tard (Fiorillo et al., 2008). Lorsque le délai est fixe, certains neurones dopaminergiques reflètent aussi la connaissance du temps écoulé depuis le début de l'essai et du délai habituel. En effet, lorsqu'une récompense est omise, la dépression de l'activité dopaminergique a lieu juste après le moment ou la récompense est généralement reçue. Si la récompense est reçue en avance, il n'y a qu'une activation en réponse à la récompense, il n'y a pas de réaction à l'absence de récompense au moment habituel. Si la récompense est en retard, il y a une dépression au moment attendu de la récompense, suivie d'une activation au moment de la récompense (Hollerman & Schultz, 1998). Bien que la magnitude et le sens de la réponse dopaminergique au stimulus conditionné ou à la récompense indiquent la connaissance de l'intervalle de temps entre les deux, l'activité dopaminergique entre les deux évènements, elle, ne varie pas en fonction du temps écoulé (à moins d'une grande incertitude sur la récompense, section précédente) (Fiorillo et al., 2003). Les neurones dopaminergiques ne semblent donc pas coder le passage du temps lui-même.

4.2.1.e) *Autre information cachée*

Dans les situations où toute l'information n'est pas directement observable, tel que le passage du temps, les neurones dopaminergiques offrent parfois un signal démontrant une grande connaissance de l'environnement. Par exemple, dans une étude de Nakahara (2004), la récompense donnée lors d'un essai dépend des essais précédents. Les essais précédents ne sont pas directement observables. Pour prédire la récompense lors d'un essai, l'animal doit avoir appris à retenir en mémoire les essais précédents et à découvrir la règle qui dicte la récompense en fonction des essais qui précèdent. Après suffisamment d'entrainement, la réponse des neurones dopaminergiques montre que l'animal a appris à retenir cette information sur les essais précédents et en tient compte dans sa prédiction de récompense. La quantité d'entrainement devient cependant un facteur important dans l'apparition de cette réponse. Dans l'étude de Nakahara (2004), les singes ont dû être entrainés pendant plus de cinq mois et faire plus de 36 000 essais, probablement le temps que les autres régions appropriées du cerveau puissent acquérir la représentation nécessaire.

4.2.1.f) *Nouveauté*

Les neurones dopaminergiques peuvent aussi répondre aux nouveaux stimuli, mais cette réponse disparaît rapidement au fur et à mesure que l'animal s'habitue au stimulus, à moins bien sûr que le stimulus permette de prédire une éventuelle récompense (Ljungberg et al., 1992; Schultz, 1998).

4.2.1.g) *Aversion*

Les neurones dopaminergiques répondent principalement aux récompenses positives et semblent beaucoup moins sensibles aux punitions ou aux stimuli aversifs (Mirenowicz & Schultz, 1996). Pour une revue de ces rares exceptions où les neurones dopaminergiques répondent aux stimulus aversifs, voir (Horvitz, 2000).

4.2.1.h) *Revues*

Pour une revue à jour et succincte de la réponse phasique des neurones dopaminergiques, voir (Schultz, 2007). Pour une revue moins complète, mais plus en contexte avec la neurobiologie, voir (Schultz, 1998).

4.2.1.i) *En résumé*

Bien que le modèle TD explique plusieurs des phénomènes dopaminergiques ci-dessus, le modèle TD à lui seul ne peut expliquer toutes ces données. Il ne peut pas, par exemple, expliquer comment les neurones dopaminergiques gèrent différentes échelles de récompenses. Il ne peut pas non plus expliquer la réponse dopaminergique à la récompense en fonction du temps écoulé depuis le stimulus conditionné, ni acquérir une connaissance directe du délai habituel, sans une forme de représentation du temps. Similairement, le modèle TD ne peut pas reproduire l'activité dopaminergique dans des tâches où toute l'information n'est pas directement observable, sans être augmenté d'un modèle de l'information cachée à laquelle les vrais neurones dopaminergiques doivent avoir accès pour avoir le comportement qu'ils ont. L'objectif de cette thèse est de mieux comprendre comment, en conjonction au modèle TD, des représentations de cette information peuvent être acquises à l'aide de modèle d'apprentissage. La section 4.4 fait une revue des modèles dopaminergiques TD actuels.

4.2.2 Neurones striataux

La littérature utilise différentes appellations pour les neurones du striatum. Entre autres, les neurones de projections étudiés chez les primates sont souvent les PAN (de l'anglais *phasically active neuron*) par opposition aux TAN (*tonically active neuron*). Quant à la littérature chez les rongeurs, elle ne semble pas utiliser l'abréviation TAN.

Les neurones PAN sont actifs pendant différentes phases des tâches sur lesquelles les animaux sont entrainés. Par exemple, supposons que la tâche consiste à parcourir un labyrinthe en T, et qu'à mi-chemin entre le départ et la fourche, un stimulus indique le coté de la jonction qui sera récompensée. Dans cette tâche, il y a

des neurones sélectifs pour une ou plusieurs de ces étapes : départ, stimulus, fourche et récompense (Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999; Schultz, Tremblay, & Hollerman, 2003; Lau & Glimcher, 2007). Dans une tâche de saccade oculaire avec mémoire, on présente d'abord à l'animal la direction dans laquelle il devra effectuer une saccade oculaire pour avoir une récompense (Kawagoe, Takikawa, & Hikosaka, 1998). Ensuite, après un court laps de temps sans aucun stimulus, on indique à l'animal qu'il peut maintenant faire la saccade. L'animal doit faire la saccade dans la direction affichée quelques secondes plus tôt pour obtenir sa récompense. Si, d'un bloc d'essais à l'autre, on change la récompense en fonction de la direction indiquée, on retrouve des neurones sélectifs uniquement à la direction la plus fortement récompensée pour le bloc d'essais en cours, ainsi que leurs opposés, des neurones qui ne répondent qu'aux autres directions. Dans l'étude de Kawagoe et ses collaborateurs (1998), chaque bloc comportait 60 essais, soit 15 pour chacune des quatre directions possibles. Plusieurs neurones se sont adaptés à la nouvelle direction récompensée en une quinzaine d'essais (soit environ quatre essais dans chaque direction). De telles réponses sélectives à une partie de la tâche et à un stimulus et dépendant de la récompense ont aussi été observées à d'autres reprises (Hassani, Cromwell, & Schultz, 2001). Certains neurones peuvent même être sélectifs pour une grandeur de récompense lorsque des récompenses de différentes magnitudes sont offertes (Cromwell & Schultz, 2003). Finalement, les neurones de projection du striatum pourraient également encoder la récompense espérée associée à une action sans pour autant coder l'action choisie (Samejima, Ueda, Doya, & Kimura, 2005). Les chercheurs qui ont fait cette découverte ont utilisé des séries de blocs où ils changeaient les probabilités de récompense associées à chaque direction de façon à pouvoir faire varier la récompense espérée et la meilleure action (celle menant à la plus grande récompense espérée) de façon indépendante.

L'activité des neurones du striatum évolue de différentes façons avec l'apprentissage. Dans le cas du labyrinthe en T, il y a, en début d'entrainement, des neurones actifs pour chaque étape de la tâche. Par contre, après l'entrainement, la population devient surtout active au début et à la fin de la tâche, mais relativement

moins au parcours intermédiaire comme à la fourche (Jog et al., 1999). Si on enlève la récompense (extinction), alors l'activité générale réapparaît pendant la tâche. Si l'on réinsère la récompense (ré-acquisition), alors les neurones redeviennent plus spécifiques comme elles étaient avant l'extinction (Barnes, Kubota, Hu, Jin, & Graybiel, 2005). La réception de la récompense peut amener plus qu'une simple réponse des neurones du striatum. Dans une tâche à probabilité de récompense fixe pour toutes les actions, certains neurones ne semblent encoder que la présence ou l'absence de récompense au moment où celle-ci est généralement attribuée, alors que d'autres semblent sélectifs à l'action qui a eu lieu précédemment (Lau & Glimcher, 2007).

Pour un survol de l'activité des neurones de projections du striatum (MSN), voir (Schultz et al., 2003). L'article suivant en fait aussi un bon survol en introduction et discussion (Lau & Glimcher, 2007). Malgré leur rôle important dans l'apprentissage, l'activité des MSN ne sera que survolée dans cette thèse comparativement aux neurones dopaminergiques.

Les TAN semblent avoir une activité souvent corrélée aux neurones dopaminergiques. Pour une revue de l'activité des interneurones du striatum (TAN), voir (Apicella, 2007).

4.3 Plasticité

On pourrait discuter des mécanismes moléculaires et cellulaires de la plasticité synaptique en général ou plus particulièrement de celle des connexions corticostriatales. Mais ces connaissances sont souvent incomplètes. Lorsqu'on regarde une section du cerveau en particulier, ou un certain type de neurones, on ne connaît souvent que quelques éléments du casse-tête et il est difficile d'avoir une vue d'ensemble à ce niveau. Les études de costimulation et d'enregistrements cellulaires donnent une vue d'ensemble appropriée pour le sujet de cette thèse. Elles permettent de voir l'effet de la coactivation en termes de fréquence de décharge de deux neurones sur la plasticité de leur synapse commune. C'est donc principalement cette littérature qui sera traitée ici.

Le principe général se construit à partir de deux neurones A et B, le neurone présynaptique A projetant sur le neurone postsynaptique B. Premièrement, on stimule électriquement le neurone A, augmentant ainsi sa fréquence de décharge, tout en mesurant la réponse du neurone B. Ensuite, on stimule simultanément les deux neurones. Finalement, on stimule à nouveau le neurone A seulement tout en mesurant la réponse du neurone B. Lorsque la réponse du neurone B après la costimulation est plus grande qu'avant, on parle alors de LTP (*long term potentiation)*, ou de *potentialisation à long terme de l'efficacité synaptique*. Si, à l'inverse, la réponse du neurone B est plus petite après la costimulation qu'avant, on parle alors de LTD (*long term depression)*, ou *dépression à long terme de l'efficacité synaptique*.

4.3.1 Plasticité corticostriatale et dopamine

Plusieurs études ont eu lieu sur la LTP et la LTD des synapses corticostriatales, les projections du cortex vers les neurones MSN du striatum, et nigrostriatales, projection des neurones dopaminergiques du SNc vers les neurones MSN du striatum. Ces synapses sont donc à l'intersection des afférences corticales et dopaminergiques du striatum.

De façon générale (Reynolds & Wickens, 2002), la stimulation électrique d'un seul groupe de neurones, soit cortical, nigral, ou striatal, n'a aucun effet à long terme : ni habituation, ni désensibilisation. La costimulation des neurones dopaminergiques et d'un autre groupe de neurones, cortical ou striatal, n'a aussi aucun effet. Par contre, la costimulation des neurones corticaux et de leurs cibles striatales entraine généralement, mais pas exclusivement, une LTD. Le niveau de dopamine semble avoir un effet important sur le résultat de cette costimulation.

Plus précisément, des études montrent que la coactivation phasique des trois groupes de neurones (cortical, nigral, ou striatal) entraine généralement une LTP, alors qu'une coactivation des neurones corticaux et de leurs cibles striatales dans un bain de dopamine génère plutôt une LTD ou rien (Reynolds & Wickens, 2002). La théorie actuelle veut donc qu'à la base la coactivation des neurones présynaptiques (corticaux) et postsynaptiques (striataux) détermine l'éligibilité d'une synapse à être modifiée. Le signal dopaminergique phasique, qui crée une variation du niveau de

dopamine arrivant sur de petites fenêtres de temps très courte (150 ms), déterminerait la direction de cette modification : soit une potentialisation ou soit une dépression. Un faible taux de dopamine produisant une LTD, un taux moyen (actif, ~30 Hz) assurant le maintien et un taux vif (100 Hz) produisant une LTP.

Pour une revue des travaux sur la LTP et LTD corticostriatales, voir (Reynolds & Wickens, 2002). Cette information est toutefois importante puisqu'elle pourrait bien correspondre aux règles d'apprentissage (Équation 2-24 et Équation 2-25, Figure 2-13, reprises à la section 4.4.2 qui suit) des modèles TD.

4.3.2 Plasticité corticale et dopamine

Similairement, la dopamine semble jouer un rôle dans la plasticité des synapses du cortex frontal (Otani, Daniel, Roisin, & Crepel, 2003).

4.4 Modèles

Il y a probablement plus d'une centaine de modèles informatiques adaptatifs des ganglions de la base, dont la grande majorité sont, depuis 1995, basés sur l'algorithme TD. Une revue exhaustive de tous ces articles serait non seulement trop longue, mais peu utile. À la place, cette section traitera des éléments constitutifs clefs les plus importants. Après un bref historique des modèles TD des ganglions de la base suivra une section sur les différentes composantes du modèle TD et une section sur le problème de la représentation des entrées, en particulier sur la représentation du temps. Deux autres sections relatives au problème de représentation suivront : la première porte principalement sur les modèles à base de modules parallèles et la seconde sur les modèles incluant plusieurs régions du cerveau. De plus, les modèles sans liens à la dopamine ou qui par exemple se concentrent uniquement sur le chemin direct et indirect entre le striatum et la sortie GPi/SNr seront omis. Les modèles importants ou intéressants qui ne sont pas nécessairement basés sur TD seront mentionnés dans une dernière section avant la conclusion.

4.4.1 Historique

Les modèles TD (où l'on apprend à prédire la somme des récompenses à venir par différence temporelle, section 2.3.3 et 3.2.1) sont un excellent exemple de situation où l'apprentissage machine, la modélisation en psychologie de

l'apprentissage animale et les modèles informatiques des ganglions de la base se regroupent au travers d'un même cadre théorique. Sutton et Barto (1981) tentaient à l'époque de développer un modèle mathématique de style neuronal de l'apprentissage animal. Ces travaux ont été fortement inspirés des travaux de Klopf en apprentissage machine sur l'importance de la différence entre l'apprentissage supervisé et l'apprentissage par renforcement. C'est finalement vers la fin des années 80 que l'on voit apparaître TD dans sa forme actuelle comme algorithme en apprentissage machine (Sutton, 1988) et comme modèle en psychologie animale (Sutton & Barto, 1990). Pour plus de détails sur cette période, voir (Sutton & Barto, 1998).

C'est en janvier 1992 que Ljungberg et ses collaborateurs ont publié les premiers résultats (Ljungberg et al., 1992) suggérant clairement que les neurones dopaminergiques puissent encoder le signal d'erreur dans TD. Cette suggestion fut faite entre autres par Montague et ses collaborateurs à NIPS (Montague, Dayan, Nowlan, Pouget, & Sejnowski, 1993). En 1995, Houk, Davis et Beiser ont publié un livre (Houk, Davis, & Beiser, 1995) qui a grandement influencé la modélisation des ganglions de la base par la suite. C'est dans une section de quatre chapitres (Wickens & Kotter, 1995; Barto, 1995; Schultz et al., 1995; Houk, Adams, & Barto, 1995) qu'ils ébauchent certaines hypothèses sur les liens possibles entre la plasticité du striatum, TD, l'activité des neurones dopaminergiques et l'organisation du striatum. En 1996, Montague et ses collaborateurs publient un premier article (Montague et al., 1996) d'importance qui jette les bases actuelles du modèle TD des neurones dopaminergiques. Ces travaux seront suivis d'un article (Schultz et al., 1997) dans *Science* en collaboration avec un expert de l'activité des neurones dopaminergiques, Wolfram Schultz. Depuis, une bonne centaine d'articles en neuroscience ont été publiés utilisant des dérivés de ce modèle qui a cependant relativement peu évolué depuis.

En 2003, apparut l'un des premiers articles (O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003) reliant l'imagerie fonctionnelle (fMRI) au signal d'erreur de TD. Cependant, comme la dopamine peut affecter les vaisseaux sanguins, il faut

être prudent avec ces méthodes basées sur le taux d'oxygène du sang et leur interprétation à titre computationnel (Niv & Schoenbaum, 2008).

4.4.2 Modèle TD des ganglions de la base

Le modèle est principalement basé sur le fait que le signal phasique d'une majorité de neurones dopaminergiques mésencéphaliques (A8, A9 et A10) se comporte comme le signal d'erreur de prédiction de récompense δ dans les algorithmes de style TD (sections 2.3.3 et suivantes) (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). L'équation principale est rapportée ici :

$$\delta_{t+1} = r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}}) - V^{\pi}(\boldsymbol{s_t}), \qquad \textbf{Équation 4-2}$$

où $\boldsymbol{s}_t$ et $\boldsymbol{s}_{t+1}$ sont les états aux temps t et $t+1$, et r_{t+1} la récompense donnée, et γ le facteur de remise (sections 2.3.2 et 2.3.3). Tel qu'expliqué précédemment, ce signal représente la différence entre deux estimations consécutives de la somme des récompenses à venir *V()* prévue et la récompense obtenue *r*. Mais quelle variante de l'algorithme le cerveau utilise-t-il et comment cette dernière est implémentée restent incertains. Quels neurones contribuent à signaler la récompense *r*? Y a-t-il des neurones différents qui calculent une partie de la fonction de valeur *V()* ou *Q()* au temps *t* et *t-1* ou leur différence? Où les actions possibles sont-elles comparées et comment l'action à prendre est-elle déterminée?

4.4.2.a) *La récompense*

Brown et ses collaborateurs (Brown, Bullock, & Grossberg, 1999) ont proposé dans un modèle qui n'était pas basé sur TD que le signal premier de récompense provienne du LHA (Figure 4-3). Cette région est reconnue pour jouer un rôle dans l'alimentation et ces neurones répondent à la récompense première (telle que le sucre) (Nakamura & Ono, 1986). De plus, au fur et à mesure que les besoins sont satisfaits, elles ne répondent plus à la récompense (Fukuda, Ono, Nishino, & Nakamura, 1986). Le LHA projetterait vers le PPTN (Semba & Fibiger, 1992) ainsi que directement sur les neurones dopaminergiques du VTA (Fadel & Deutch, 2002). Par contre, il est important de noter que le LHA reçoit aussi des afférences du cortex, du striatum, et surtout du VTA (Duva et al., 2005) et qu'il est aussi le lieu de plasticité synaptique

sous modulation dopaminergique (Fukuda, Ono, Nakamura, & Tamura, 1990). Après conditionnement, certains neurones du LHA répondent à la récompense et au stimulus conditionné (Fukuda et al., 1986; Nakamura & Ono, 1986). Par conséquent, on peut se demander si dans certains cas particuliers, le LHA pourrait être le lieu d'un conditionnement primaire permettant à un stimulus de devenir une récompense, si une récompense y est associée.

Condé (1992) a proposé que le signal de récompense puisse passer principalement par le PPTN dont les neurones cholinergiques projettent directement sur les neurones dopaminergiques du SNc. En plus des afférences en provenance du LHA, le PPTN reçoit des afférences des sorties des ganglions de la base (Parent & Hazrati, 1995a; Parent & Hazrati, 1995b). Des études récentes montrent que non seulement certains neurones du PPTN répondent à la récompense première *r*, mais que d'autres encoderaient la prédiction *V()* d'une récompense (Kobayashi & Okada, 2007). Ceci mène nécessairement à de nouvelles hypothèses sur la façon dont le système dopaminergique calcule son erreur de récompense.

Bref, même si l'on ne cherche qu'à étudier les connexions principales et trouver la relation entre TD et les ganglions de la base, la connectivité est beaucoup plus complexe que ce qu'elle semble être dans la majorité des modèles de connectivités (voir section 4.1 et Figure 4-3). La plasticité y est potentiellement présente partout. La simple définition de récompense, pour l'animal, peut changer avec le temps et peut s'avérer bien différente de celle de l'expérimentateur. Peu de modèles, à l'exception de (Brown et al., 1999; Kobayashi & Okada, 2007), s'intéressent à la provenance du signal de récompense et au rôle du LHA et du PPTN au-delà de cette simple fonction. Il n'est pas impossible qu'en plus de jouer un rôle dans le calcul de la récompense, ces régions puissent aussi jouer un rôle dans le calcul de la prédiction *V()* ou même plus directement du signal d'erreur δ (Kobayashi & Okada, 2007).

4.4.2.b) *La prédiction*

Le striatum est souvent considéré comme une région de choix pour l'estimation de la somme des récompenses à venir (*V()* et/ou *Q()*, section 2.3.3 et

suivantes) (Houk et al., 1995; Suri & Schultz, 1999; Doya, 2000; Nakahara, Doya, & Hikosaka, 2001; Suri, 2001; Suri, Bargas, & Arbib, 2001; Suri & Schultz, 2001; Baldassarre, 2002; Suri, 2002; Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004; Samejima et al., 2005; Moustafa & Maida, 2007; Redish, Jensen, Johnson, & Kurth-Nelson, 2007; Khamassi, Mulder, Tabuchi, Douchamps, & Wiener, 2008) dans un modèle du type acteur-critique par exemple. Parmi les bonnes raisons de supposer que certains neurones du striatum puissent encoder *V()* ou *Q()*, comptons : l'activité de neurones pendant les différentes étapes entre le début d'une tâche et la récompense (section 4.2.2) dont certaines calculant clairement la prédiction associée à une action (*Q()*) (Samejima et al., 2005), des modèles ayant réussi à reproduire l'activité de ces neurones à l'aide de *V()* (Suri & Schultz, 2001; Khamassi et al., 2008), la plasticité synaptique des connexions corticostriatales qui dépend de la dopamine (section 4.3.1) (Reynolds & Wickens, 2002) et les entrées corticales susceptibles de contenir une représentation de l'état $\boldsymbol{s}$. En effet, si l'on reprend les équations en 2.3.5 pour *V(),*

$$V^{\pi}(\boldsymbol{s_t}) = \boldsymbol{w_c}^{\boldsymbol{T}}\boldsymbol{s_t}, \qquad \textbf{Équation 4-3}$$

et pour la modification des connexions $\boldsymbol{w_c}$,

$$\boldsymbol{w_c} \leftarrow \boldsymbol{w_c} + \alpha\delta_{t+1}\boldsymbol{s_t}, \qquad \textbf{Équation 4-4}$$

et que l'on considère que l'état $\boldsymbol{s}$ est représenté par les entrées corticales, alors les équations, la connectivité et la plasticité du striatum concordent. Un modèle général (extrait de, mais pas nécessairement équivalent à: Houk et al., 1995; Egelman, Person, & Montague, 1998; Suri & Schultz, 1999; Doya, 2000; Baldassarre, 2002; Suri, 2002; Nakahara et al., 2004; Khamassi, Lacheze, Girard, Berthoz, & Guillot, 2005; Khamassi et al., 2006) combinant l'acteur-critique simple du chapitre 2.3.5 à l'anatomie simplifiée de la section 4.1 se trouve en Figure 4-5. À première vue, l'anatomie et l'algorithme de base semblent bien se superposer.

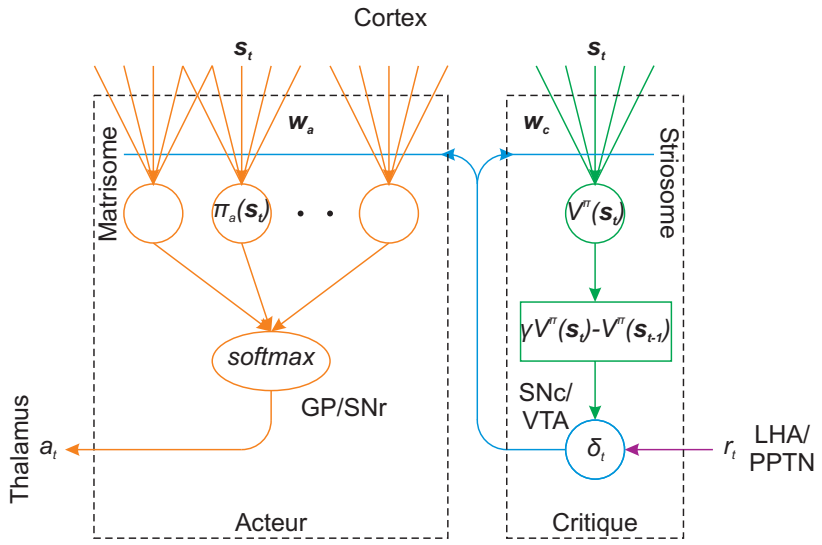


Figure 4-5 : Anatomie de la Figure 4-3 superposée au modèle acteur-critique neural simple du Chapitre 2 (Figure 2-12). Les mêmes couleurs qu'en Figure 3-11 et Figure 4-3 sont utilisées.

Le principal problème de ce modèle relève du calcul de δ par les neurones dopaminergiques (Joel, Niv, & Ruppin, 2002). Que ce soit via la séparation striosome/matrisome, ou striatum ventral/striatum dorsal (revoir section 4.1.4), la connexion monosynaptique (voie directe) du striatum sur les neurones dopaminergiques est inhibitrice et la connexion excitatrice, s'il y en a une, doit donc être polysynaptique (indirecte). Cette deuxième étant plus lente (mais de peu) que la première, si le striatum représentait *V()* ou *Q()*, alors l'équation des neurones dopaminergiques ressemblerait plutôt à

$$\delta_{t+1} = r_{t+1} - \gamma V^{\pi}(\boldsymbol{s_{t+1}}) + V^{\pi}(\boldsymbol{s_t}), \qquad \textbf{Équation 4-5}$$

au lieu de

$$\delta_{t+1} = r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}}) - V^{\pi}(\boldsymbol{s_t}). \qquad \textbf{Équation 4-6}$$

Une erreur de signes difficilement réconciliable. Cependant, peu de modélisateurs discutent des interneurones des aires dopaminergiques (section 4.1.6) et du PPTN et de leurs rôles potentiels. Tel que mentionné dans la section sur la récompense, il est possible que le PPTN puisse fournir une voie excitatrice indirecte du striatum aux neurones dopaminergiques.

Le cortex orbitofrontal ainsi que quelques autres régions du cortex limbique sont aussi souvent suggérés comme jouant un rôle dans l'estimation de la récompense

à venir. Les raisons sont principalement les mêmes que pour le striatum : des neurones qui semblent encoder la valeur de prédiction de récompense (Schultz, Tremblay, & Hollerman, 2000) associée à différentes options (Padoa-Schioppa & Assad, 2006; Padoa-Schioppa & Assad, 2008), un apport dopaminergique jumelée à une projection sur les neurones du VTA (section 4.1.6) et finalement la modélisation de certaines activités à partir de la fonction *V()* (Suri & Schultz, 2001). Cependant, malgré la relation importante entre l'activité de certains de ces neurones et la récompense, on considère plus souvent leurs projections vers le striatum que celles vers les neurones dopaminergique. Leur rôle exact dans le modèle reste donc nébuleux. Il est possible qu'il soit relié à la question de la représentation (voir section suivante 4.4.3) et des préférences de l'animal. Toutefois, il serait aussi possible qu'une région corticale comme le cortex orbitofrontal fournisse la composante positive de la prédiction de récompense et le striatum la composante négative; ou qu'ils puissent se compléter d'une quelconque autre façon dans le calcul de *V()*.

Plus récemment, l'habénula, projetant sur les neurones dopaminergiques, a été proposé comme source d'inhibition possible relié à la prédiction de récompense. Certains de ses neurones semblent en effet être activés par des stimuli qui ne sont pas associés à une récompense et inhibés par des stimuli associés à une récompense (Matsumoto & Hikosaka, 2007). Cette nouvelle piste de recherche semble une avenue prometteuse pour déterminer comment les neurones dopaminergiques arriveraient à calculer le signal d'erreur δ.

Mais pourquoi est-ce si difficile de déterminer les neurones qui prédisent la récompense alors que les neurones dopaminergiques indiquent clairement l'erreur de prédiction? En fait, ce qui rend le signal dopaminergique si clair, c'est qu'une majorité de neurones donne la même réponse. Bref, le signal dopaminergique est un signal d'erreur précis, mais, il contient peu d'information sur le contexte de l'erreur. Ce signal unique (scalaire) est diffusé dans plusieurs régions du système nerveux. Par contre, la représentation de la récompense semble plutôt distribuée et contextuelle. Par exemple, Suri et Schultz ont modélisé *V()* en utilisant un neurone de prédiction pour chaque stimulus (Suri & Schultz, 1998, 1999; Pan et al., 2005); ces neurones

sont donc sélectifs au stimulus présent. Dans un second modèle (Suri, 2001; Suri & Schultz, 2001), ils ont utilisé un neurone *V()* pour prédire chaque stimulus comme s'il était une récompense. Dans cet article, un neurone artificiel associé à un stimulus répond donc de façon anticipatoire au stimulus ou à l'action sur le point d'être exécuté, si l'on considère une action comme un autre stimulus à prédire (Suri et al., 2001). D'autres (Baldassarre, 2002; Khamassi et al., 2005; Khamassi et al., 2006; Bertin, Schweighofer, & Doya, 2007) ont utilisé une forme de mélange d'experts où chaque neurone $V()_i$ contribue à *V()* selon sa *responsabilité.* Dans le cerveau, les neurones qui semblent prédire une récompense semblent être sélectifs aux stimuli, aux contextes ou aux récompenses mêmes. Par conséquent, il est difficile de dissocier la représentation de l'état de la représentation de sa valeur potentielle en termes de récompense à venir.

4.4.2.c) *L'action*

Si le striatum est une région de choix pour la prédiction de récompense, il l'est encore plus pour le processus de sélection d'une action (l'acteur). Non seulement tous les éléments qui le suggéraient comme critique s'appliquent aussi pour le rôle d'acteur, mais en plus, les problèmes de projections du striatum vers les neurones dopaminergiques ne s'appliquent pas à l'acteur puisque ce dernier n'a pas obligatoirement besoin de projections vers les neurones dopaminergiques. Si l'on regarde la règle de correction de l'acteur

$$\pi_a(\boldsymbol{s_t}) = \boldsymbol{w_a}^{\boldsymbol{T}}\boldsymbol{s_t} \qquad \textbf{Équation 4-7}$$

pour chaque action $a \in A$,

$$\boldsymbol{w_{a_t}} \leftarrow \boldsymbol{w_{a_t}} + \alpha\delta_{t+1}\boldsymbol{s_t}, \qquad \textbf{Équation 4-8}$$

où a_t est l'action gagnante au temps *t*, comme pour le critique, tous les éléments s'y trouvent : l'entrée corticale, la modulation de la plasticité par le critique, c'est-à-dire le signal dopaminergique, ainsi que la projection vers le thalamus pour le contrôle moteur. Cependant, la plupart des modèles TD ayant un acteur (Houk et al., 1995; Egelman et al., 1998; Suri & Schultz, 1998, 1999; Doya, 2000; Nakahara et al., 2001; Suri, 2002; McClure, Daw, & Montague, 2003; Montague et al., 2004; Moustafa &

Maida, 2007) s'arrêtent là et ne vont pas plus en détails sur l'anatomie des voies de sorties, ni sur le rôle de chacune des stations sur le parcours (à l'exception de Suri et al., 2001; Baldassarre, 2002).

Les ganglions de la base et spécialement le striatum sont reconnus depuis longtemps pour jouer un rôle important dans le contrôle moteur comme en font foi les maladies de Parkinson et de Huntington (Kandel et al., 2000). La maladie de Parkinson est marquée par la mort de neurones dopaminergiques et donc par la diminution du taux de dopamine dans le striatum et les voies de sortie des ganglions de la base auxquelles est associée une difficulté à initier ou à faire certains mouvements. La maladie de Huntington est quant à elle marquée au début par la mort des neurones du striatum et conduit entre autres à une surgénération de mouvements indésirables.

Que le striatum représente l'acteur, dans un modèle acteur-critique, ou les prédictions *Q()* pour chaque action, dans un modèle de style SARSA ou *Q*-Learning (Redish, 2004; Redish et al., 2007), il n'en reste pas moins qu'une action doit être sélectionnée. Les voies de sortie striatopallidothalamiques (voir Figure 4-3) font partie des éléments considérés dans les modèles des ganglions de la base comme jouant un rôle dans la sélection de l'action. Une série de modèles basés sur cette idée, indépendamment de l'apprentissage par renforcement, ont été produits sur le sujet (Gurney, Prescott, & Redgrave, 2001a; Gurney, Prescott, & Redgrave, 2001b; Gurney, Humphries, Wood, Prescott, & Redgrave, 2004).

4.4.2.d) *Le modèle*

Il reste à déterminer si les ganglions de la base sont plus susceptibles de ressembler à un algorithme de style acteur-critique, où le choix de l'action se fait indépendamment du calcul *V()* de prédiction de la récompense, ou du style *Q-Learning*/SARSA, où le choix de l'action dépend de l'évaluation *Q()* des paires état-action possibles.

Le modèle acteur-critique a longtemps été le modèle privilégié (Houk et al., 1995; Egelman et al., 1998; Suri & Schultz, 1998, 1999; Doya, 2000; Nakahara et al., 2001; Suri, 2002; McClure et al., 2003; Montague et al., 2004; Moustafa & Maida,

2007). Dans ce modèle, la fonction *V()* tend vers une estimation moyenne des récompenses en fonction des probabilités des différentes actions que l'acteur peut choisir (section 2.3.3). Le calcul de l'erreur δ dépend donc de *V()*, une fonction calculée indépendamment de l'action qui sera choisie :

$$\delta_{t+1} = r_{t+1} + \gamma V^{\pi}(\boldsymbol{s_{t+1}}) - V^{\pi}(\boldsymbol{s_t}).$$ **Équation 4-9**

Récemment, des chercheurs ont observé des réponses dopaminergiques proportionnelles à la prédiction des récompenses *Q* spécifique à l'action choisie (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006), suggérant plutôt un modèle du type SARSA (section 2.3.3) :

$$\delta_{t+1} = r_{t+1} + \gamma Q^{\pi}(\boldsymbol{s_{t+1}}, a_{t+1}) - Q^{\pi}(\boldsymbol{s_t}, a_t).$$ **Équation 4-10**

D'autres chercheurs ont observé des réponses dopaminergiques proportionnelles à la récompense qu'amènerait la meilleure action (Roesch, Calu, & Schoenbaum, 2007), suggérant plutôt une règle d'apprentissage comme *Q-Learning* (section 2.3.4) :

$$\delta_{t+1} = r_{t+1} + \gamma \max_{a'}\{Q^{*}(\boldsymbol{s_{t+1}}, a')\} - Q^{*}(\boldsymbol{s_t}, a_t)$$ **Équation 4-11**

La découverte de neurones codant la prédiction de récompense pour chaque action possible dans le striatum (*Q()*) (Samejima et al., 2005) favorise aussi ces deux alternatives basées sur l'apprentissage d'une fonction *Q()* des paires d'états-action $(\boldsymbol{s_t}, a_t)$. Anatomiquement, une ségrégation des neurones du striatum projetant vers les neurones dopaminergiques de ceux projetant vers les voies de sorties des ganglions de la base (voir section 4.1.4) suggèrerait plutôt un modèle acteur-critique. Par contre, la récente découverte de neurones du striatum avec des projections vers chaque relais des voies de sorties (Levesque & Parent, 2005) et donc possiblement aussi vers les neurones dopaminergiques, porte à penser que finalement, une architecture du type *Q-Learning* ou SARSA est tout aussi probable, sinon plus, qu'une architecture du type purement acteur-critique.

4.4.3 Le problème de la représentation

La modélisation du signal dopaminergique par le signal d'erreur δ des modèles TD n'est cependant pas le simple résultat de l'Équation 4-2. Elle dépend entièrement de la forme de la fonction *V()* (ou *Q()*) et surtout de la structure de l'espace d'états *S*, c'est-à-dire de la représentation des états ou des entrées. Il y a

principalement deux façons de décrire *V()* selon la structure de *S*. Si *S* est un espace d'états distincts dont un seul est VRAI à la fois, comme c'est le cas dans les modèles d'automates ou de Markov, alors une valeur *V(s)* peut être apprise pour chaque état *s*. Si par contre $\boldsymbol{s}$ est un vecteur dans $\Re^N$ ($S \subseteq \Re^N$) où plusieurs dimensions peuvent prendre une valeur différente de 0 en même temps, alors nous obtenons une représentation distribuée (un vecteur) dans un continuum de valeurs possibles (les réels). La première forme peut être réécrite sous forme vectorielle en utilisant une dimension pour chaque état $s \in S$ de sorte que pour chaque vecteur $\boldsymbol{s}$, il n'y aura toujours qu'une seule dimension portant la valeur « 1 » (l'état courant) et les autres seront à zéro. De cette façon, il est possible de discuter des différentes approches en conservant une seule formulation, l'Équation 4-3. En considérant que le striatum forme principalement une seule couche et que ses entrées sont principalement des neurones corticaux, on peut représenter l'activité corticale sous forme vectorielle et *V()* comme une simple combinaison linéaire du vecteur d'état $\boldsymbol{s}$ (voir Équation 4-3). Cette section porte principalement sur la représentation des états et entrées $\boldsymbol{s} \in \Re^N$.

Supposons un modèle avec deux entrées ($\boldsymbol{s} \in \Re^2$), chacune représentant la présence (1) ou l'absence (0) d'un stimulus (la première A et la deuxième B; voir Tableau 4-I à gauche). Un tel modèle peut très bien apprendre à prédire que le stimulus A prédit une récompense et que le stimulus B prédit une récompense en apprenant un poids approprié pour chacune de ces entrées. Cependant, un tel modèle ne peut pas, en plus, apprendre que la paire de stimuli AB ne mène à aucune récompense, même pour TD (revoir section 3.2.1). *V()*, dans cet espace, n'a pas une capacité calculatoire suffisante. C'est le problème du *OU exclusif* (XOR). Pour modéliser l'apprentissage du OU exclusif, Pearce et d'autres (Pearce, 1994) ont dû trouver une façon de représenter cet état (A et B présent simultanément) différemment que par le simple marquage de la présence de A et de B indépendamment. Par exemple, en utilisant une troisième entrée pour marquer la présence de A et de B simultanément (voir Tableau 4-I à droite). On peut observer de façon indirecte le développement d'une telle représentation chez le pigeon, où l'on voit très bien l'inaptitude de l'oiseau au début de l'entrainement à discriminer A ou B

de AB (Redhead & Pearce, 1995, dans une version à trois stimuli). Il n'est cependant pas raisonnable de penser que toutes les combinaisons possibles de stimuli possibles soient associées à un neurone. La seule solution est donc la combinaison d'une représentation distribuée et d'une fonction *V()* non linéaire par rapport aux entrées brutes. Si *V()* est linéaire, alors la représentation distribuée entre *V()* et les entrées brutes doit être non linéaire.

Tableau 4-I : Représentation des entrées pour le *OU exclusif*. À gauche, une représentation à deux dimensions qui ne permet pas son apprentissage par une fonction linéaire de la forme de l'Équation 4-3. À droite une représentation augmentée qui le permet (*w = (1,1,-2)*).

Stimuli	Représentation	Récompense
∅	(0,0)	0
A	(1,0)	1
B	(0,1)	1
AB	(1,1)	0

Stimuli	Représentation	Récompense
∅	(0,0,0)	0
A	(1,0,0)	1
B	(0,1,0)	1
AB	(1,1,1)	0

4.4.3.a) *Lignes de délai et représentations du passage temps*

Dans toutes modélisations de type TD, la relation temporelle entre les évènements est un élément important. Dans plusieurs situations, il est suffisant de décomposer la simulation selon les évènements (O'Reilly & Frank, 2006). Par exemple, pour représenter une séquence d'évènements ∅ → A → B → US, on pourrait avoir : $\boldsymbol{s_0} = (0,0)$, $r_1 = 0$, $\boldsymbol{s_1} = (1,0)$, $r_2 = 0$, $\boldsymbol{s_2} = (0,1)$, $r_3 = 1$. Mais lorsque la durée de chacun de ces évènements devient importante, alors le passage du temps lui-même doit pouvoir être représenté.

La première approche est celle des lignes de délai (aussi appelée *Complete-Serial-Coumpound,* Sutton & Barto, 1990). Cette méthode fournit au modèle une connaissance parfaite du passage du temps du début d'un stimulus jusqu'à une durée maximale (le nombre de lignes). En fait, c'est le même principe que pour les réseaux à lignes de délai TDNN (*time-delay neural networks*, et non *temporal-difference*, revoir section 2.1.3). On utilise une composante différente du vecteur d'entrée $\boldsymbol{s}$ pour représenter chaque délai possible depuis l'apparition du stimulus, et ce, pour chaque entrée possible (Figure 4-6). Cette méthode requiert une série de lignes de délai pour chaque entrée originale. Bien que de telles lignes de délai existent dans certaines régions sous-corticales pour des délais de l'ordre des millisecondes et des

microsecondes (Carr & Konishi, 1990), ça semble physiologiquement irréaliste pour des délais de l'ordre des secondes (voir section 3.2.3). On peut aussi imaginer une série de neurones actifs les uns après les autres dans un ordre précis à partir de l'apparition du stimulus, semblable à une série de mémoire tampon. Là aussi, bien que certains neurones corticaux (Buonomano, 2003) et sous-corticaux (Hahnloser et al., 2002) semblent pouvoir produire de telles séquences, par exemple pendant la génération d'actions de courtes durées (< 1 s) (Hahnloser et al., 2002), une telle représentation prédéfinie pour toutes les entrées reste peu probable pour couvrir des délais de l'ordre des secondes (Buonomano, 2005; Karmarkar & Buonomano, 2007). Néanmoins, c'est la représentation utilisée dans les premiers modèles TD et elle l'est encore aujourd'hui (Sutton & Barto, 1990; Montague et al., 1996; Schultz et al., 1997; Suri, 2001; Suri et al., 2001; Suri & Schultz, 2001; Daw & Touretzky, 2002; Pan et al., 2005; Bertin et al., 2007).

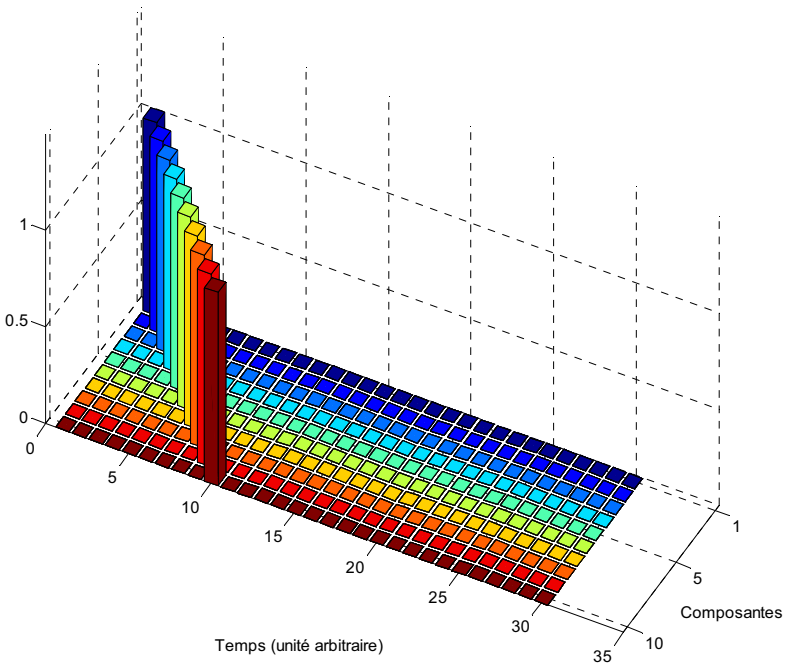


Figure 4-6 : Représentation d'un seul stimulus apparaissant au temps *t = 0* et utilisant neufs lignes de délai.

Pour modéliser des résultats dopaminergiques où le délai a de l'importance, Suri et Schultz (Suri & Schultz, 1998, 1999) ont fabriqué une représentation

temporelle plus complexe basée sur une série de mémoires de travail pour chaque stimulus. Chaque mémoire de travail de la série couvre une durée différente, son activité est inversement proportionnelle à sa durée et elle augmente très légèrement avec le passage du temps (Figure 4-7). Comme la représentation précédente, cette représentation exige la préexistence d'une série de mémoire pour chaque stimulus possible et pour chaque délai requis avec une distribution de délais en fonction de la précision temporelle désirée. A priori, elle n'est pas très différente de la précédente puisqu'il ne suffit que de soustraire l'activité de la composante mémoire *k-1* de celle de *k* pour obtenir l'équivalent de la ligne de délai *k* de la représentation précédente.

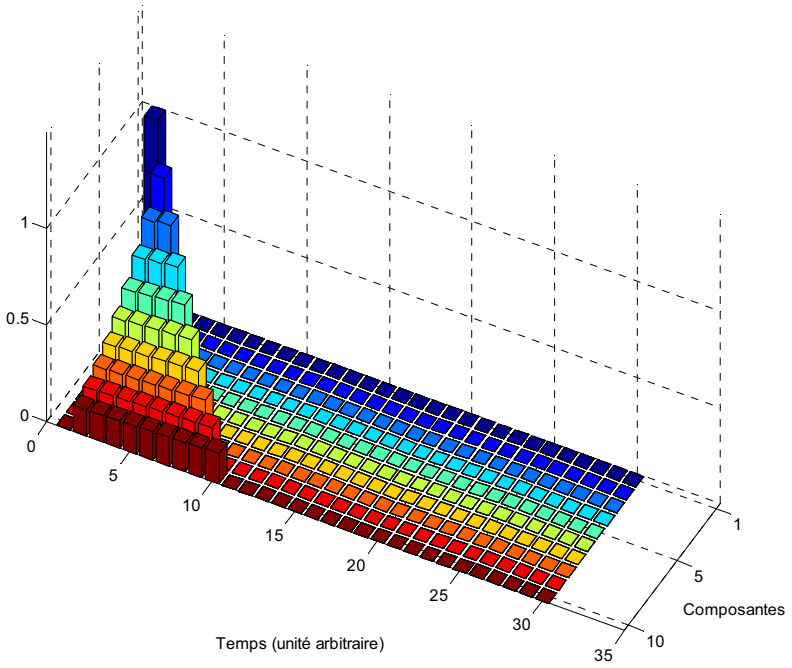


Figure 4-7 : Représentation de l'apparition d'un seul stimulus au temps *t = 0* et utilisant neuf lignes de mémoires soutenues.

Plus récemment, une version plus proche des modèles du passage du temps dans le cerveau a été utilisée avec TD (Ludvig, Sutton, & Kehoe, 2008; Ludvig, Sutton, Verbeek, & Kehoe, 2009). Plutôt qu'une série de mémoires de travail, on utilise une série d'intégrateurs temporels, chacun atteignant un sommet à un délai différent (Figure 4-8). En fait, c'est une version plus lisse des lignes de délai et dont les propriétés mathématiques sont plus près des données empiriques sur le passage du

temps (Machado, 1997; Fiorillo et al., 2008). Pour différencier le conditionnement de trace du conditionnement classique (Figure 3-2), une entrée supplémentaire peut être utilisée pour représenter la présence du stimulus (composante 1, bande bleue foncée de $t = 0$ à $t = 30$, Figure 4-8) (Ludvig et al., 2009). Cependant, à certains niveaux, ce modèle possède les mêmes problèmes que les précédents. Il suppose l'existence, pour chaque stimulus, d'une population complète d'intégrateurs temporels couvrant tous les délais nécessaires.

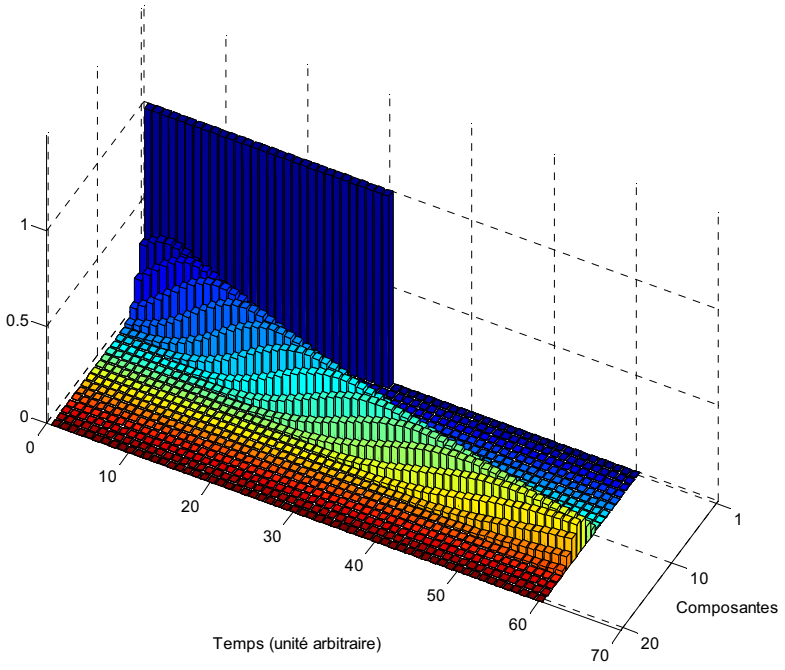


Figure 4-8 : Représentation d'un seul stimulus apparaissant au temps *t = 0,* présent pour 30 unités de temps et utilisant 19 micro-stimuli.

Finalement, il est possible de représenter le passage du temps à l'aide d'une population d'oscillateurs ayant chacun une fréquence différente (Figure 4-9). Le plus récent et le plus complet de ces modèles est appelé SBF (de l'anglais, *striatal beat frequency*) (Matell & Meck, 2004). Il a l'avantage de reproduire la loi de Weber pour le temps (section 3.1.5), c'est-à-dire que sa précision temporelle est proportionnelle à la durée du délai. Dans ce modèle, chaque délai possible est représenté par un vecteur unique d'activités des oscillateurs (tant que le délai est plus petit que l'inverse du plus petit commun multiple des fréquences des oscillateurs). Comme pour les modèles

précédents, si le passage du temps doit être mesuré pour plusieurs stimuli indépendants, alors plusieurs populations d'oscillateurs sont nécessaires. Le modèle suppose donc une population d'oscillateurs précis résidant dans le cortex dans la bande de fréquences alpha (8 à 13 Hz) et démarrant tous à l'instant où le CS apparaît.

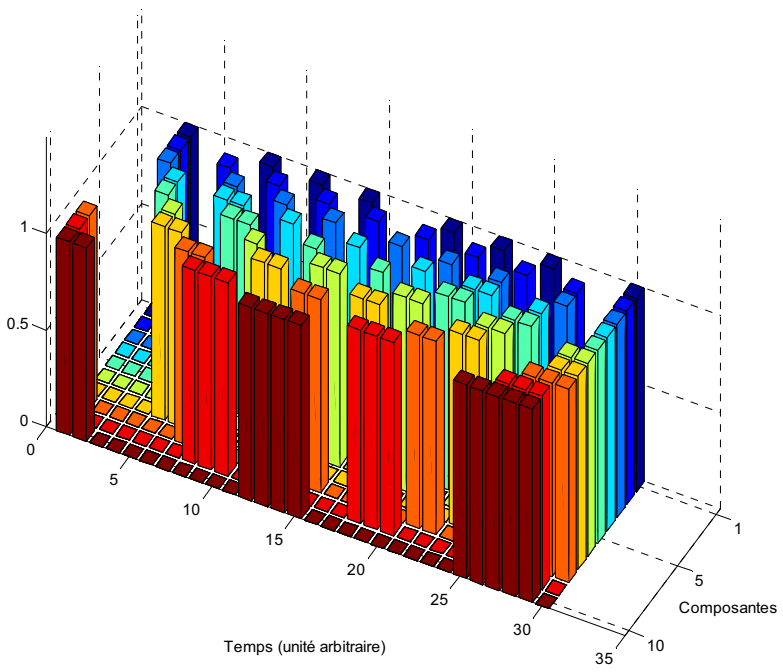


Figure 4-9 : Représentation de l'apparition d'un seul stimulus au temps *t = 0* et utilisant dix oscillateurs de fréquences différentes et démarrés simultanément.

Cette représentation a certaines similarités avec les lignes de délai (Figure 4-6). Si l'on suppose que ces dernières ont une activité périodique avec une période d'environ quatre fois leur délai de base, on obtient une représentation semblable. Mais, au-delà de la plus longue ligne de délai, c'est le motif spécifique d'activités de tous les oscillateurs combinés qui encodent le temps écoulé. Le modèle SFB suggère le striatum comme région où cette reconnaissance du délai serait effectuée. Les auteurs suggèrent que la dopamine puisse servir à démarrer les oscillateurs, par son activité à l'apparition du CS, et à apprendre à reconnaître le motif représentant le bon délai, par son activité à l'apparition de la récompense. Bien qu'aucune règle précise d'apprentissage n'y soit proposée, une telle règle pourrait être développée par les méthodes habituelles (Chapitre 2). De plus, même si ce modèle n'en est pas un du

signal dopaminergique, il serait possible d'utiliser cette représentation temporelle dans les modèles TD et d'en vérifier les prédictions pour le signal dopaminergique (sur les données de Fiorillo et al., 2008, par exemple). L'article (Matell & Meck, 2004) ne fournit cependant aucune description mathématique précise du modèle bien que les résultats de plusieurs simulations y soient présentés.

Tous ces modèles comportent trois problèmes importants. Le premier, qui a déjà été mentionné, est qu'il suppose la préexistence de populations complètes de neurones représentant le temps écoulé depuis l'arrivée du stimulus, et ce, pour chaque stimulus possible. Cette représentation est peu probable pour des intervalles de l'ordre des secondes (Karmarkar & Buonomano, 2007). Deuxièmement, lorsqu'une mémoire de travail est nécessaire, elle est simplement ajoutée au modèle. Pour ce faire, on ajoute une composante représentant la mémoire du stimulus pendant l'intervalle de temps (comme la représentation du stimulus en Figure 4-8). Dans un tel modèle, il n'y a plus de différence entre le conditionnement classique et le conditionnement de trace. La mémoire de travail n'a pas de rôle actif; elle n'a pas besoin d'apprendre quel stimulus doit être gardé en mémoire. Troisièmement, une fois la représentation temporelle démarrée, celle-ci ne peut plus être arrêtée à moins d'ajouter d'autres modifications *ad hoc* au modèle (par exemple, Suri & Schultz, 1999). Bref, tous ces modèles ont une représentation totalement statique et construite avec la connaissance parfaite de la tâche par l'expérimentateur.

Dans les faits, il y a de vrais neurones qui semblent représenter une intégration temporelle de l'ordre des secondes et certaines s'adaptent au délai de la tâche lorsqu'il change (Niki & Watanabe, 1979; Funahashi et al., 1989; Komura et al., 2001; Romo et al., 1999; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Reutimann et al., 2004; Lucchetti et al., 2005; Lebedev et al., 2008). Ces données portent à croire que plusieurs régions du cerveau, par un mécanisme général d'apprentissage, puissent apprendre à représenter le passage du temps, lorsque nécessaire (Dragoi et al., 2003; Reutimann et al., 2004; Hopson, 2003; Ivry & Schlerf, 2008). Selon cette hypothèse, plutôt que d'avoir des populations entières d'intégrateurs avec des délais prédéterminés, certains neurones pourraient prendre ce

rôle et s'adapter automatiquement au délai de la situation. De plus, au lieu de devoir relier les entrées à un mécanisme de calcul du temps, les neurones qui reçoivent les entrées pourraient directement évaluer le passage du temps, lorsque ce dernier est constant. Cependant, ces modèles n'en sont encore qu'à leur balbutiement.

4.4.3.b) *Modèles internes*

Pour résoudre les problèmes des lignes de délai, Daw et ses collaborateurs (2003, 2006) ont remplacé ces dernières par un modèle semi-markovien interne de l'environnement. Ils supposent qu'avec le temps, l'animal a appris quelque part dans le cortex, une représentation de la tâche incluant la relation temporelle entre les différents éléments (stimulus et récompense) et que c'est le signal de ces neurones représentant l'état supposé du monde qui sert d'entrée à TD. En conditionnement, deux états sont possibles, être dans un essai en attente de la récompense, ou être entre deux essais. L'arrivée du stimulus ou de la récompense amène le système directement dans le bon état. De plus, la règle de TD est changée pour directement tenir compte du temps passé dans l'état. Ils proposent aussi une façon purement mathématique d'apprendre ce modèle interne en provenance d'articles précédents, sans plus de détails. Cependant, ce modèle ne fournit aucune représentation neuronale du temps ou de son apprentissage.

Dans une tâche légèrement plus complexe, Nakahara et ses collaborateurs (2004) ont montré que le signal dopaminergique pouvait même témoigner de l'utilisation d'autres informations contextuelles par l'animal, telle que l'information sur les essais précédents. Pour modéliser ces résultats, ils ont simplement ajouté des dimensions au vecteur d'entrée de TD pour représenter cette information supplémentaire. C'est en fait ce qui arrive dans la majorité des modèles TD et autres. Il n'y a qu'une infime adaptation dans le modèle, souvent parce que de toute façon, on cherche à modéliser l'apprentissage d'animaux surentrainés chez lesquels l'adaptation réelle est minime et rapide lors de l'acquisition des données. Relativement peu de recherches portent sur le développement des réponses neurales au cours de l'apprentissage car de tels enregistrements sont très difficiles à effectuer et à interpréter. C'est pourquoi l'apprentissage des représentations reste méconnu.

4.4.4 Modules parallèles et apprentissage de représentations

Un certain nombre de modèles sont basés sur l'idée qu'il puisse y avoir plusieurs *modules* TD en parallèle (Nakahara et al., 2001; Baldassarre, 2002; Redish, 2004; Khamassi et al., 2005; Khamassi et al., 2006; Bertin et al., 2007; Khamassi et al., 2008; Moustafa & Maida, 2007). Par exemple, Khamassi (Khamassi et al., 2008) a utilisé trois modules, chacun ayant une représentation partielle des entrées.

Dans un modèle acteur-critique, Nakahara (Nakahara et al., 2001) utilise deux boucles d'acteurs en parallèle, l'un fonctionnant en coordonnées visuelles et l'autre en coordonnées motrices, selon le modèle des voies parallèles ségrégées (section 4.1.5) (Alexander et al., 1986). Bien que les deux systèmes n'apprennent pas à la même vitesse, ils partagent le même critique et toutes les représentations et transformations de coordonnés font partie du modèle et ne sont pas acquises. Similairement, Moustafa et Maida (2007) utilisent deux acteurs en parallèle et un seul critique pour simuler les tâches ayant une période de délai entre la consigne et la réponse demandée. Le premier acteur doit apprendre à maintenir la représentation de la consigne en mémoire, alors que le deuxième effectue la réponse demandée lorsque le signal est donné. Comme dans la plupart des modèles TD, seules sont apprises les connexions corticostriatales, c'est-à-dire les poids à l'entrée des acteurs et du critique.

Parmi les modèles les plus intéressants, mentionnons (Baldassarre, 2002). Bien qu'il n'y ait pas de représentation des entrées apprises, ce modèle utilise un *mélange d'experts* pour représenter *V()* et une autre série de réseaux pour les acteurs. Dans ce modèle, la récompense dépend de l'objectif de l'agent. Par exemple, si l'agent a faim, alors trouver de la nourriture est récompensant. L'objectif faisant partie des entrées, les experts pour *V()* ont tendance à se spécialiser en fonction de l'objectif. Par exemple, un expert peut se spécialiser à prédire la récompense quand l'agent a faim. Similairement, les acteurs sont représentés par un double mécanisme de sélection. Le premier mécanisme de sélection représente les ganglions de la base et sélectionne l'acteur pouvant agir. Une fois l'acteur sélectionné, ce dernier peut choisir une action, c'est le deuxième niveau de sélection. Chaque niveau apprend selon la règle de TD. Contrairement aux experts du critique, les acteurs ne se spécialisent pas

par objectif. Tout comme il est sensé de prédire la récompense en fonction de l'objectif, il est aussi sensé d'apprendre des actions qui seront utiles à plusieurs objectifs. Khamassi (Khamassi et al., 2005; Khamassi et al., 2006) a aussi utilisé différentes variantes de mélange d'experts pour *V()* combiné au modèle de sélection d'actions de Gurney (Gurney et al., 2001a; Gurney et al., 2001b).

Cependant, les experts dans le modèle de Khamassi (Khamassi et al., 2005) ne se répartissaient pas très bien pour la tâche utilisée. Pour améliorer la répartition des experts, ils ont préentrainé une carte autoorganisée (SOM, de l'anglais *self-organizing map*), une forme de réseaux de neurones de catégorisation non supervisée. L'entrainement est effectué en promenant l'agent de façon aléatoire dans son environnement, avant l'ajout du modèle TD. La carte ainsi apprise crée différentes catégories dans lesquelles elle classe les états perçus. Dans cette nouvelle version (Khamassi et al., 2006), la carte sert à déterminer à chaque moment quel expert est responsable de *V()*. Redish (Redish et al., 2007) a aussi expérimenté l'apprentissage de représentations en utilisant un réseau compétitif non supervisé dont la sortie est utilisée directement comme représentation pour *V()*. Cet article étudie aussi les interactions possibles entre la dopamine et le réseau. Cependant, l'article ne porte pas sur l'apprentissage de représentations, mais plutôt sur l'effet de cet apprentissage sur le comportement humain comme le jeu compulsif. Finalement, on peut aussi mentionner le modèle de Nakahara et ses collaborateurs (Nakahara, Amari, & Hikosaka, 2002) qui, sans être basé sur TD, réussit à reproduire la sélectivité adaptative aux situations récompensées des neurones du striatum (section 4.2.2, Kawagoe et al., 1998). Ce modèle utilise une règle hebbienne et un signal dopaminergique théorique pour moduler l'activité des neurones du striatum.

Parmi les modèles TD ayant une forme d'apprentissage de représentations, les modèles de Suri et Schultz (Suri, 2001; Suri et al., 2001; Suri & Schultz, 2001) demeurent parmi les plus généraux. Ils utilisent TD de façon généralisée comme prédiction de n'importe quel stimulus comme si c'était une récompense. Par conséquent, le modèle apprend d'une certaine façon à prédire la dynamique de l'environnement. Cependant, il n'est pas clair s'il est préférable de prédire la somme

des évènements à venir (TD) ou plus simplement le prochain évènement (prédiction de séries temporelles, section 2.2.5). Évidemment, dans certaines situations, un stimulus peut annoncer un évènement à venir beaucoup plus tard ou dans un délai indéterminé. Quoi qu'il en soit, ces modèles de Suri & Schultz dépendent entièrement de la représentation du temps parfaite que produisent les lignes de délai.

4.4.5 Modèles multiples et apprentissage de représentations

Doya (1999, 2000) a développé une théorie générale sur l'apprentissage dans le cerveau dont le présent ouvrage, quoique bien différent, est grandement inspiré. Dans cette théorie, les régions anatomiques importantes sont considérées comme effectuant différents types d'apprentissage. Par exemple, le cortex pourrait faire de l'apprentissage non supervisé, les ganglions de la base de l'apprentissage par renforcement et le cervelet de l'apprentissage supervisé (revoir les Chapitre 2 et Chapitre 3). Quelques modèles ont regardé l'interaction de ses composantes, mais relativement peu a été fait sur le développement de représentations pendant l'apprentissage par renforcement en neuroscience.

O'Reilly (O'Reilly, 1998; O'Reilly & Munakata, 2000) a développé une théorie générale basée sur six principes de base : le réalisme biologique, la représentation distribuée, la compétition inhibitrice (*softmax*), la récurrence, l'apprentissage par erreur et l'apprentissage hebbien. En utilisant des paramètres différents pour les ganglions de la base, le cortex et l'hippocampe, il a modélisé l'apprentissage simultané de plusieurs régions du cerveau et a réussi à reproduire plusieurs résultats dans un seul et même cadre (O'Reilly & Munakata, 2000; Rougier et al., 2005; O'Reilly & Frank, 2006). Ces simulations suivent une ligne d'idée similaire à cette thèse. Cependant, la capacité réelle de ses algorithmes à créer de nouvelles représentations reste à démontrer. Par exemple, il discrétise toujours le temps selon l'arrivée des évènements dans les simulations et il ne s'occupe pas du passage du temps (voir section 4.4.3). O'Reilly utilise aussi un modèle légèrement différent de TD pour les ganglions de la base et le signal dopaminergique (O'Reilly & Frank, 2006).

4.4.6 Autres modèles

Un élément important qui ne vient pas de TD mais qui peut y être combiné est relié au concept de la *saillance incitative* (traduction libre de l'anglais *incentive saliency*) (Berridge & Robinson, 1998). La question générale est toute simple : quel est le rôle de la dopamine? La dopamine a-t-elle simplement un rôle d'*apprentissage*? Sert-elle plus généralement à indiquer la présence de quelque chose que l'on *aime* (ayant l'effet d'une récompense)? Ou sert-elle à motiver la recherche de ce quelque chose (le *vouloir*)? Par exemple, un rat placé dans un labyrinthe peut avoir appris qu'il y a une récompense à l'autre bout, mais si l'on bloque le signal dopaminergique, il ne partira pas à sa recherche (Ikemoto & Panksepp, 1996). Ce comportement est a priori incompatible avec le rôle de la dopamine comme simple signal d'erreur dans les modèles TD où elle ne sert qu'à modifier les synapses. La façon la plus simple de réconcilier le pouvoir motivationnel de la dopamine au modèle TD, est de faire jouer un second rôle au signal δ en lui permettant d'influencer immédiatement les probabilités de prendre une action. Par exemple, dans un acteur de type *softmax* (sections 2.1.2 et 2.3.5) où il y aurait aussi une probabilité de ne pas agir, le signal pourrait augmenter la probabilité des actions par rapport à l'inaction (Montague et al., 1996; Egelman et al., 1998; McClure et al., 2003). Le modèle de Redish (2004) utilise aussi une variante de ce concept. Dans les modèles de Schultz (Schultz et al., 1997) et d'Alexander (Alexander & Sporns, 2006; Alexander, 2007), δ joue un rôle similaire sur le *changement d'action* et le *changement d'attention* respectivement. Pour une discussion approfondie sur la saillance incitative et TD, voir (Berridge, 2007).

Un autre concept important et relativement peu analysé est le concept de la nouveauté. La nouveauté peut faire office de récompense, et lorsque c'est le cas, elle dépend du système dopaminergique pour son conditionnement (Bevins et al., 2002). Kakade et Dayan (2002) ont proposé deux modèles d'utilisation de la nouveauté pour expliquer la réponse dopaminergique à celle-ci et le rôle de la nouveauté sur l'exploration. Le premier suppose un signal direct de nouveauté faisant office de récompense. Lisman et Grace (2005) ont proposé des voies anatomiques d'acheminement d'un tel signal en provenance possible de l'hippocampe. Le second

modèle plus mathématique est basé sur la surévaluation de tout nouvel état, favorisant ainsi leur exploration. Mentionnons les résultats récents d'enregistrements dans le striatum (Barnes et al., 2005) où celui-ci semble partiellement se désactiver avec le surentrainement (exploitation), mais reprend une activité plus variée pendant l'extinction (exploration). Redgrave et Gurney (2006) discutent aussi du rôle possible du signal de nouveauté dopaminergique dans la découverte de nouvelles actions. Le lien possible entre les ganglions de la base, la dopamine, la nouveauté et les actes, bien que peu discuté dans cette thèse, pourrait s'avérer très important pour le développement de représentations. Une grande part de notre apprentissage se faisant par les jeux et l'exploration, cette exploration, sous un contrôle possible des ganglions de la base, pourrait fournir au cortex les informations nécessaires à la construction d'abstractions.

Finalement, il y a un certain nombre d'autres modèles mathématiques des ganglions de la base, seuls ou en incorporant d'autres régions du cerveau, qui ne sont pas basés sur TD. Il y a tout d'abord le modèle de (Beiser & Houk, 1998) dans lequel il n'y a aucun apprentissage, le modèle de (Dominey & Arbib, 1992; Arbib & Dominey, 1995) dans lequel il y a plusieurs régions du cerveau, mais un seul ensemble de poids adaptatifs (corticostriataux) sur la base du renforcement, les modèles de (Brown et al., 1999; Contreras-Vidal & Schultz, 1999) qui reproduisent des données dopaminergiques en utilisant une forme de lignes de délai et sans apprentissage de représentations, et les modèles de (Berns & Sejnowski, 1998; Gurney et al., 2001a; Gurney et al., 2001b; Gurney et al., 2004) concentrés sur les voies de sorties et la sélection d'action.

4.4.7 Littérature, conclusion et direction

Parmi les revues de littérature sur le sujet, mentionnons (Brown et al., 1999; Gillies & Arbuthnott, 2000) qui font une bonne revue des modèles de l'époque, ainsi que (Joel et al., 2002; Suri, 2002). Horgan et Cummins (2006) comparent les modèles à représentation temporelle (Suri, 2001) et les modèles à représentation abstraite (Daw, Courville, & Touretzky, 2006). Finalement, pour une revue des modèles des ganglions de la base, en particulier sur l'apprentissage de séquences et la relation

entre ces modèles et les connaissances biophysiques des ganglions de la base, voir (Worgotter & Porr, 2005). Pour une bonne introduction au modèle TD des ganglions de la base à tous les niveaux, voir (Montague et al., 2004). Pour un résumé de l'état actuel des modèles, voir (Daw & Doya, 2006). Pour une liste des questions et réponses fréquentes sur le sujet, voir (Niv & Schoenbaum, 2008).

En conclusion, la modélisation de l'apprentissage dans les ganglions de la base a connu un essor considérable avec l'arrivée des modèles TD. Une grande partie des données dopaminergiques a été reproduite d'une façon ou d'une autre à l'aide de ces modèles. Cependant, le problème de l'apprentissage d'abstractions structurales, c'est-à-dire de représentations, n'est pas résolu par les ganglions de la base seuls. Ceux-ci y jouent pourtant un rôle important : soit par leur interaction directe avec le cortex frontal, par les projections dopaminergiques ou par leur rôle dans le contrôle moteur et l'exploration. Rares sont les modèles neurophysiologiques (pas nécessairement biophysiques) s'attaquant au problème de l'apprentissage intégré entre le cortex et les ganglions de la base. En fait, bien peu de modèles regardent réellement le développement d'abstractions. La grande majorité des modèles fonctionnent grâce aux représentations implémentées par les expérimentateurs. De plus, si le cerveau développe une représentation de la dynamique de l'environnement, alors il doit pouvoir apprendre à gérer le passage du temps. Aucun des modèles revus n'apprend de représentations du temps.

L'objectif de cette thèse est de mieux comprendre à l'aide de modèles comment les ganglions de la base et le cortex permettent le développement d'abstractions, plus particulièrement de représentations temporelles, dans le contexte de l'apprentissage par renforcement. Dans les manuscrits qui suivent la Méthodologie (Chapitre 5), un nouveau modèle combinant TD à des réseaux de neurones artificiels non supervisés représentant le cortex est élaboré. Le défi est d'arriver à construire un modèle capable de développer de façon autonome sa propre représentation de la tâche tout en reproduisant les données dopaminergiques avec un minimum d'ajustements et dont la représentation temporelle pourra être comparée à l'activité de vrais neurones. Enfin, ce modèle devra permettre d'expliquer les données observées.

Chapitre 5. Méthodologie

Dans cette thèse, la modélisation est l'outil principal de recherche utilisé pour mieux comprendre le développement de représentations dans le contexte de l'apprentissage par renforcement, en particulier l'interaction entre le cortex et les ganglions de la base. Cette section décrit et justifie l'approche ainsi que le choix des tâches et des données sélectionnées pour cette étude.

Dans la littérature scientifique, plusieurs régions du cerveau font l'objet de modélisation individuelle à un niveau relativement fin (modélisation du potentiel de membranes et du potentiel d'action pour un ou quelques neurones, modélisation de la fréquence de décharge d'un neurone représentant une population, ou de toute une population, etc.), parfois même en y incluant la plasticité synaptique. Il y a aussi quelques modèles d'envergure plus globaux qui incluent plusieurs régions du cerveau. Mais, ceux-ci incluent rarement de la plasticité pour plus d'une structure. À une autre échelle, il y a les modèles psychologiques aux principes généraux abstraits. Mais, ils tardent à être reliés à des processus neurologiques précis, en particulier les modèles connexionnistes. Ces modèles, qui se veulent plus *neuronaux*, gagneraient beaucoup à être plus collés aux données neurophysiologiques existantes.

Or, il arrive un moment où l'étude de l'apprentissage dans le cerveau ne peut se limiter à une seule région de ce dernier. Le cerveau est un amalgame de différents systèmes travaillant ensemble à apprendre une nouvelle tâche. Lorsqu'on ne modélise qu'une région, on ne peut étudier qu'une petite partie du processus d'apprentissage, et ce, généralement en insérant *à la main* toutes les autres connaissances nécessaires développées par les autres régions du cerveau non incluses dans le modèle. C'est le cas de la majorité des modèles présentés au chapitre précédent (section 4.4). Dans le cerveau, l'acquisition de ces connaissances possiblement distinctes, mais certainement complémentaires, est souvent le fruit d'une collaboration entre plusieurs structures. Il est même fort possible que les signaux d'erreur d'une région du cerveau puissent servir de guide à une autre. Par exemple, le signal dopaminergique issu des ganglions de la base est souvent modélisé comme un signal d'erreur de prédiction de récompense. Ce même signal, en plus d'être utilisé localement par les ganglions de la

base dans leur apprentissage, est aussi envoyé dans le cortex, lieu probable du développement d'abstractions (voir section 3.3.1). Or le cortex est aussi l'une des sources d'entrées principales des ganglions de la base (voir section 4.1). De telles boucles d'interaction se retrouvent partout dans le cerveau. Si l'apprentissage dans une région du cerveau est dépendant de l'apprentissage dans une autre et de certains signaux relatifs à cet apprentissage, alors un modèle de l'apprentissage incluant les deux régions devrait nous permettre d'étudier des interactions importantes de cet apprentissage que deux modèles individuels, et donc partiels, ne peuvent nous donner.

Dans cette thèse, l'objectif est d'essayer de mieux comprendre le développement de représentations dans le contexte de l'apprentissage par renforcement (par récompense). L'hypothèse de travail utilisée ici est que différentes régions du cerveau adressent différents problèmes d'apprentissage et qu'une partie de la solution réside dans leurs interactions. C'est donc un modèle de ces systèmes qui sera construit afin de mieux les étudier. Le niveau de modélisation a été choisi suffisamment proche des neurones pour pouvoir tirer parti des données neurophysiologiques tout en gardant un niveau d'abstraction suffisamment élevé pour conserver le potentiel explicatif du modèle, c'est-à-dire, qu'il nous aide à mieux comprendre l'apprentissage dans le cerveau. L'approche proposée ici est de mettre ensemble les modèles plastiques des ganglions de la base, du système dopaminergique et du cortex, afin de pouvoir étudier leurs dynamiques et leurs interactions dans l'apprentissage.

Étudier plusieurs systèmes plastiques en interaction devient rapidement très compliqué. Il faut donc limiter le nombre de structures cérébrales à étudier. Dans le domaine de l'apprentissage par renforcement avec récompenses et sans punitions, les ganglions de la base et le système dopaminergique semblent les structures les plus importantes. Ils ont d'ailleurs des modèles mathématiques bien établis qui concordent avec de nombreux enregistrements électrophysiologiques en conditionnement classique et opérant appétitif (voir Chapitre 4). Le cortex est quant à lui le complément le plus intéressant puisqu'il est le siège de nombreuses représentations (voir section 3.3.1) et qu'il est une source importante d'information pour les

ganglions de la base (voir section 4.1). Il ne semble cependant pas y avoir de modèle d'apprentissage général du cortex aussi fortement établi que pour les ganglions de la base et le système dopaminergique.

Dans un premier temps, il est important d'évaluer si le modèle TD des ganglions de la base et les idées d'apprentissage non supervisé du cortex peuvent effectivement fonctionner ensemble et permettre l'apprentissage par renforcement de tâches intéressantes. Pour ce faire, on peut construire un petit modèle incluant un modèle TD acteur-critique des ganglions de la base combiné à un modèle d'apprentissage non supervisé pour une région corticale. Ensuite, on peut mesurer les capacités d'apprentissage de différentes variations du modèle sur des tâches typiques de la littérature de l'apprentissage machine par renforcement et les comparer à ce qui se fait dans ce domaine. Par exemple, on peut utiliser le modèle pour voir comment la rétroaction dopaminergique pourrait influencer positivement l'apprentissage dans le cortex. Les résultats positifs de cet exercice sont publiés dans le premier article de cette thèse (Chapitre 6) (Rivest et al., 2005).

Pour vérifier si un apprentissage non supervisé du cortex est suffisant pour apprendre les représentations nécessaires à expliquer les résultats dopaminergiques lors de l'apprentissage par renforcement, il est important de choisir une tâche d'apprentissage simple nécessitant probablement ces deux régions du cerveau et pour lesquelles il y a suffisamment de données électrophysiologiques. Dans la situation simple de conditionnement appétitif de trace à délai fixe, les modèles TD sont bien établis, mais demeurent incomplets. Par exemple, aucun n'explique très bien l'acquisition du délai et de la différence entre la période interstimuli et la période interessais, pourtant bien visible dans les enregistrements des neurones dopaminergiques de Hollerman & Schultz (1998) (voir section 4.4.3). Comment les intervalles de temps dans l'ordre des secondes sont représentés ou comptés dans le cerveau reste encore aujourd'hui la source de débats, pour ne pas dire un mystère. Mais, il y a de bonnes raisons de croire que cette fonctionnalité est en partie acquise ailleurs dans le cortex (Niki & Watanabe, 1979; Funahashi et al., 1989; Romo et al., 1999; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Reutimann

et al., 2004; Lucchetti et al., 2005; Lebedev et al., 2008). Le conditionnent appétitif de trace à délai fixe est donc une tâche idéale pour ce projet de modélisation de l'apprentissage dans le cerveau. De plus, cette tâche a l'avantage de permettre de remettre à plus tard l'introduction de la partie action de la boucle striatothalamocorticale dans le modèle. Plus il y a de boucles, plus il y a d'interactions et donc plus l'analyse peut être compliquée. Outre la réponse à la réception de la récompense, il n'y a pas d'action comme telle, et donc peu d'influences sur les entrées sensorielles. On peut donc simplement dire que l'animal reçoit la récompense. De plus, il y a des enregistrements du signal dopaminergique précis pour cette tâche avant, pendant et après le conditionnement, ainsi que sur des essais tests variés, montrant son acquisition (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz & Schultz, 1994; Hollerman & Schultz, 1998; Pan et al., 2005). Il est donc possible de directement modéliser la réponse de ces neurones plutôt que le comportement animal. Étant donné que la partie acteur des ganglions de la base projette aussi au cortex, éliminer temporairement la partie action permet de réduire le nombre d'interactions entre le cortex et les ganglions de la base au minimum.

Le Chapitre 7 (Rivest et al., 2010a) utilise donc le conditionnement appétitif de trace pour étudier le développement de représentations temporelles dans le cortex. Le modèle cortical du Chapitre 6 doit cependant être légèrement modifié pour permettre l'apprentissage de telles représentations temporelles. Il est remplacé par un réseau de longue mémoire à court terme (LSTM, voir section 2.1.4) entraîné à prédire les prochaines entrées (voir section 2.2.5). Les représentations apprises par ce modèle du cortex servent d'entrées au modèle TD des ganglions de la base et du système dopaminergique. En plus de fournir de nouvelles explications possibles à l'apprentissage de représentations temporelles, ces simulations permettent de vérifier si les activités dopaminergiques enregistrées peuvent être le résultat de représentations développées dans le cortex. Le modèle permet aussi d'évaluer si la rétroaction dopaminergique vers le cortex pourrait aider le développement de représentations par ce dernier. Le conditionnement appétitif de trace est au cœur des

simulations effectuées sur le modèle dans le deuxième article de cette thèse (Chapitre 7).

En élargissant les simulations à différentes variantes de conditionnement, il est possible d'utiliser le modèle pour prédire l'activité dopaminergique et corticale dans différentes situations jamais évaluées auparavant. En variant l'information disponible au modèle pendant l'apprentissage, comme la présence constante du stimulus jusqu'à l'arrivée de la récompense (conditionnement classique) ou sa disparition un certain temps avant son arrivée (conditionnement de trace), il est aussi possible de vérifier si les réseaux sont faciles à analyser et s'ils permettent une explication claire des façons dont le cortex pourrait utiliser l'information pour créer une représentation de la tâche. Ceci permet aussi de vérifier si l'activité dopaminergique devrait être sensible à de telles variations ou non, selon le modèle. Bref, cette approche permet d'évaluer le pouvoir explicatif et prédictif du modèle en plus d'apporter de nouvelles hypothèses de recherches. C'est le cœur du troisième article de cette thèse (Chapitre 8) (Rivest et al., 2010b).

Pourquoi pas le cervelet? Bien qu'indéniable, le rôle du cervelet dans l'évaluation du passage du temps semble principalement se situer sous la barre des deux secondes ou dans l'ordre des millisecondes ainsi que dans les tâches demandant une réponse motrice (Ivry & Keele, 1989; Ivry & Spencer, 2004; Buhusi & Meck, 2005). Il pourrait toutefois ne pas jouer un rôle aussi important dans l'apprentissage d'intervalles de l'ordre des secondes (Breukelaar & Dalrymple-Alford, 1999; Harrington, Lee, Boyd, Rapcsak, & Knight, 2004; Livesey, Wall, & Smith, 2007). Cependant, ne pas modéliser le cervelet rend plus difficiles les liens avec une vaste littérature sur le conditionnement classique (et le temps) basé sur le conditionnement aversif (décharges électriques, bouffées d'air dans les yeux, etc.). Alors qu'il est possible de mesurer le conditionnement appétitif en partie par le signal dopaminergique, le conditionnement aversif a une composante motrice dans laquelle le cervelet semble jouer un rôle clef (Ivry & Spencer, 2004). De plus, le système dopaminergique semble moins sensible aux stimulus aversif (Mirenowicz & Schultz, 1996) (l'aversion, ou la peur, pourrait passer par l'amygdale). Bref, l'importance de

chaque structure cérébrale semble différente selon que le conditionnement soit aversif (par punition) ou par récompense. Cependant, cela n'exclut pas que le cervelet puisse jouer un rôle important pendant l'apprentissage du délai interstimuli lorsqu'il est constant même en conditionnement appétitif. Le cervelet n'est toutefois pas modélisé dans cette recherche.

Pourquoi pas l'hippocampe? *A priori*, l'hippocampe ne semble pas avoir de rôle important dans l'intégration du temps, et il n'y en a pas vraiment de documenté. Cependant, la tâche utilisée ici, le conditionnement de trace avec une période d'attente sans aucun stimulus (Figure 3-2, en haut à droite), pourrait requérir l'hippocampe pour établir le lien entre les deux stimuli non contigus dans le temps. Effectivement sans hippocampe, le conditionnement de trace s'avère parfois beaucoup plus difficile (Clark & Squire, 1998; Beylin et al., 2001), mais la littérature comporte aussi des données contradictoires (Thibaudeau, Potvin, Allen, Dore, & Goulet, 2007). L'hippocampe n'est pas non plus modélisé dans cette recherche.

Bref, le conditionnement appétitif a été choisi afin de minimiser les structures cérébrales nécessaires à son apprentissage au cortex, aux ganglions de la base et au système dopaminergique. En plus de permettre d'éliminer la boucle striatothalamocorticale d'action, les enregistrements corticaux et dopaminergiques fournissent toute l'information nécessaire permettant de croire qu'une représentation temporelle de la tâche est acquise dans le cortex. Il est clair que plusieurs autres structures cérébrales jouent un rôle dans la réponse comportementale complète observée chez les animaux. Toutefois, cette thèse portant sur le développement de représentations dans le contexte de l'apprentissage par renforcement se limite à ces deux structures cérébrales.

Chapitre 6. Brain Inspired Reinforcement Learning

This paper (Rivest et al., 2005) by François Rivest, Yoshua Bengio, & John Francis Kalaska was published in 2005 in the *Advances in Neural Information Processing Systems **17***, the proceeding of the *2004 Neural Information Processing Systems Conference (NIPS 2004)*. Minor precisions were added in this chapter.

Most of this paper is F.R.'s work and ideas. The *e-gradient* algorithm was Y.B.'s suggestions; Y.B. also provided extensive direction, feedback, ideas, and comments throughout this work from the original ideas to the final manuscript. J.K. contributed as advisor, providing feedback and suggestions at all stages of the work.

The original goal of this paper was to find new solutions to the reinforcement machine learning problem inspired from recent findings about learning in the brain, and it does. But it appears that in fact, even learning in the brain is not so well understood. In trying to understand learning in the brain, this paper confirms the hypothesis that reinforcement learning in the basal ganglia could profit from unsupervised learning in the cortex, and that at the same time, unsupervised learning in the cortex could profit, at least via the dopaminergic signal, from reinforcement-based learning in the basal ganglia. This paper is a *proof of concept* opening the path to the next manuscript.

Abstract

Successful application of reinforcement learning algorithms often involves considerable hand-crafting of the necessary non-linear features to reduce the complexity of the value functions and hence to promote convergence of the algorithm. In contrast, the human brain readily and autonomously finds the complex features when provided with sufficient training. Recent work in machine learning and neurophysiology has demonstrated the role of the basal ganglia and the frontal cortex in mammalian reinforcement learning. This paper develops and explores new reinforcement learning algorithms inspired by neurological evidence that provides potential new approaches to the feature construction problem. The algorithms are compared and evaluated on the Acrobot task.

6.1 Introduction

Reinforcement learning algorithms often face the problem of finding useful complex non-linear features (Foster & Dayan, 2002). Reinforcement learning with non-linear function approximators like backpropagation networks attempt to address this problem, but in many cases have been demonstrated to be non-convergent (Tsitsiklis & VanRoy, 1996). The major challenge faced by these algorithms is that they must learn a value function instead of learning the policy, motivating an interest in algorithms directly modifying the policy (Sutton, McAllester, Singh, & Mansour, 2000).

In parallel, recent work in neurophysiology shows that the basal ganglia can be modeled by an actor-critic version of temporal difference (TD) learning (Barto, 1995; Suri & Schultz, 1999, 2001), a well-known reinforcement learning algorithm. However, the basal ganglia do not, by themselves, solve the problem of finding complex features. But the frontal cortex, which is known to play an important role in planning and decision-making, is tightly linked with the basal ganglia. The nature of their interaction is still poorly understood, and is generating a growing interest in neurophysiology.

This paper presents new algorithms based on current neurophysiological evidence about brain functional organization. It tries to devise biologically plausible algorithms that may help overcome existing difficulties in machine reinforcement learning. The algorithms are tested and compared on the Acrobot task. They are also compared to TD using standard backpropagation as function approximator.

6.2 Biological Background

The mammalian brain has multiple learning subsystems. Major learning components include the neocortex, the hippocampal formation (explicit memory storage system), the cerebellum (adaptive control system) and the basal ganglia (reinforcement learning, also known as instrumental conditioning).

The cortex can be argued to be equipotent, meaning that, given the same input, any region can learn to perform the same computation. Nevertheless, the frontal lobe differs by receiving a particularly prominent innervation of a specific

type of neurotransmitter, namely dopamine. The large frontal lobe in primates, and especially in humans, distinguishes them from lower mammals. Other regions of the cortex have been modeled using unsupervised learning methods such as ICA (Doi et al., 2003), but models of learning in the frontal cortex are only beginning to emerge.

The frontal dopaminergic input arises in a part of the basal ganglia called ventral tegmental area (VTA) and the substantia nigra (SN). The signal generated by dopaminergic (DA) neurons resembles the effective reinforcement signal of temporal difference (TD) learning algorithms (Sutton & Barto, 1998; Suri & Schultz, 1999). Another important part of the basal ganglia is the striatum. This structure is made of two parts, the matrisome and the striosome. Both receive input from the cortex (mostly frontal) and from the DA neurons, but the striosome projects principally to DA neurons in VTA and SN. The striosome is hypothesized to act as a reward predictor, allowing the DA signal to compute the difference between the expected and received reward. The matrisome projects back to the frontal lobe (for example, to the motor cortex). Its hypothesized role is therefore in action selection (Barto, 1995; Suri & Schultz, 1999, 2001).

Although there have been several attempts to model the interactions between the frontal cortex and basal ganglia, little work has been done on learning in the frontal cortex. In (Doya, 1999), an adaptive learning system based on the cerebellum and the basal ganglia is proposed. In (Foster, Morris, & Dayan, 2000), a reinforcement learning model of the hippocampus is presented. In this paper, we do not attempt to model neurophysiological data per se, but rather to develop, from current neurophysiological knowledge, new and efficient biologically plausible reinforcement learning algorithms.

6.3 The Model

All models developed here follow the architecture depicted in Figure 6-1. The first layer (I) is the input layer, where activation represents the current state. The second layer, the hidden layer (H), is responsible for finding the non-linear features necessary to solve the task. Learning in this layer will vary from model to model. Both the input and the hidden layer feed the parallel actor-critic layers (A and V)

which are the computational analogs of the striatal matrisome and striosome, respectively. They represent a linear actor-critic implementation of TD.

The neurological literature reports an uplink from V and the reward to DA neurons which sends back the effective reinforcement signal *e* (dashed lines) to A, V and H. The A action units usually feed into the motor cortex, which controls muscle activation. Here, A's are considered to represent the possible actions. The basal ganglia receive input mainly from the frontal cortex and the dopaminergic signal (*e*). They also receive some input from parietal cortex (which, as opposed to the frontal cortex, does not receive DA input, and hence, may be unsupervised). H will represent frontal cortex when given *e* and non-frontal cortex when not. The weights *W,* $\boldsymbol{v}$ and *U* correspond to weights into the layers A, V and H respectively (*e* is not weighted).

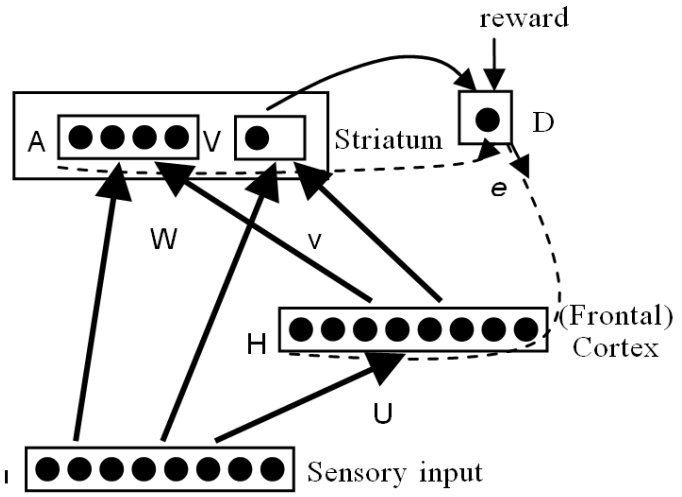


Figure 6-1 : Architecture of the models.

Let $\boldsymbol{x}_t$ be the vector of the input layer activations based on the state of the environment at time *t*. Let *f* be the sigmoidal activation function of hidden units in H. Then $\boldsymbol{y}_t = [f(\boldsymbol{u}_1\boldsymbol{x}_t), \ldots, f(\boldsymbol{u}_n\boldsymbol{x}_t)]^T$, the vector of activations of the hidden layer at time *t*, and where $\boldsymbol{u}_i$ is a row of the weight matrix *U*. Let $\boldsymbol{z}_t = [\boldsymbol{x}_t^T \; \boldsymbol{y}_t^T]^T$ be the state description formed by the layers I and H at time *t*.

6.3.1 Actor-critic

The actor-critic model of the basal ganglia developed here is derived from (Barto, 1995). It is very similar to the basal ganglia model in (Suri & Schultz, 1999)

which has been used to simulate neurophysiological data recorded while monkeys were learning a task (Suri & Schultz, 2001). All units are linear weighted sums of activity from the previous layers. The actor units behave under a winner-take-all rule. The winner's activity settles to 1, and the others to 0. The initial weights are all equal and non-negative in order to obtain an initial optimist policy. Beginning with an overestimate of the expected reward leads every action to be negatively corrected, one after the other until the best one remains. This usually favors exploration.

Then $V(\boldsymbol{z_t}) = \boldsymbol{v}^T\boldsymbol{z_t}$. Let $\boldsymbol{b_t} = W\boldsymbol{z_t}$ be the vector of activation of the actor layer before the winner take all processing. Let $a_t = argmax(b_{t,i})$ be the winning action index at time t, and let the vector $\boldsymbol{c_t}$ be the activation of the layer A after the winner take all processing such that $c_{t,a} = 1$ if $a = a_t$, 0 otherwise.

6.3.1.a) *Formal description*

TD learns a function V of the state that should converge to the expected total discounted reward. In order to do so, it updates V such that

$$V(z_{t-1}) \rightarrow E[r_t + \gamma V(z_t)] \qquad \textbf{Equation 6-1}$$

where r_t is the reward at time t and γ the discount factor. A simple way to achieve that is to transform the problem into an optimization problem where the goal is to minimize:

$$E = [V(z_{t-1}) - r_t - \gamma V(z_t)]^2 \qquad \textbf{Equation 6-2}$$

It is also useful at this point, to introduce the TD effective reinforcement signal, equivalent to the dopaminergic signal (Suri & Schultz, 1999):

$$e_t = r_t + \gamma V(z_t) - V(z_{t-1}) \qquad \textbf{Equation 6-3}$$

Thus: $E = e_t^{\,2}$.

A learning rule for the weights $\boldsymbol{v}$ of V can then be devised by finding the gradient of E with respect to the weights $\boldsymbol{v}$. Here, V is the weighted sum of the activity of I and H. Thus, the gradient is given by

$$\frac{\partial E}{\partial v} = 2e_t[\gamma z_t - z_{t-1}] \qquad \textbf{Equation 6-4}$$

Adding a learning rate and negating the gradient for minimization gives the update:

$$\Delta v = \alpha e_t \left[z_{t-1} - \gamma z_t \right]$$ **Equation 6-5**

Developing a learning rule for the actor units and their weights W using a cost function is a bit more complex. One approach is to use the tri-hebbian rule

$$\Delta W = \alpha e_t c_{t-1} z_{t-1}^{T}$$ **Equation 6-6**

Remark that only the row vector of weights of the winning action is modified.

This rule was first introduced, but not simulated, in (Barto, 1995). It associates the error e to the last selected action. If the reward is higher than expected ($e > 0$), then the action units activated by the previous state should be reinforced. Conversely, if it is less than expected ($e < 0$), then the winning actor unit activity should be reduced for that state. This is exactly what this tri-hebbian rule does.

6.3.1.b) *Biological justification*

Barto (1995) presented the first description of an actor-critic architecture based on data from the basal ganglia that resembles the one here. The major difference is that the V update rule did not use the complete gradient information.

A similar version was also developed in (Suri & Schultz, 1999), but with little mathematical justification for the update rule. The model presented here is simpler and the critic update rule is basically the same, but justified neurologically. Our model also has a more realistic actor update rule consistent with neurological knowledge of plasticity in the corticostriatal synapses (Wickens & Kotter, 1995) (H to V weights). The main purpose of the model presented in (Suri & Schultz, 1999) was to simulate dopaminergic activity for which V is the most important factor, and in this respect, it was very successful (Suri & Schultz, 2001).

6.3.2 Hidden Layer

Because the reinforcement learning layer is linear, the hidden layer must learn the necessary non-linearity to solve the task. The rules below are attempts at neurologically plausible learning rules for the cortex, assuming it has no clear supervision signal other than the DA signal for the frontal cortex. All hidden units weight vectors are initialized randomly and scaled to norm 1 after each update.

6.3.2.a) *Fixed random*

This is the baseline model to which the other algorithms will be compared. The hidden layer is composed of randomly generated hidden units that are not trained.

6.3.2.b) *ICA*

In (Doi et al., 2003), the visual cortex was modeled by an ICA learning rule. If the non-frontal cortex is equipotent, then any region of the cortex could be successfully modeled using such a generic rule. The idea of combining unsupervised learning with reinforcement learning has already proven useful (Foster & Dayan, 2002), but the unsupervised features were trained prior to the reinforcement training. On the other hand, Whiteson and Stone (2003) have shown that different systems of this sort could learn concurrently. Here, the ICA rule from (Amari, 1999) will be used as the hidden layer:

$$\Delta u_i = -\alpha_t \left[\varphi(u_i x_t) x_t - u_i x_t \varphi(u_i x_t) u_i \right] \quad \textbf{Equation 6-7}$$

where

$$\varphi(a) = -\frac{d \log q(a)}{da} \quad \textbf{Equation 6-8}$$

for probability distribution *q()* of $\boldsymbol{u_i x_t}$. This learning rule is the gradient of the cost function:

$$-E\left[\sum_{i=1}^{n} \log q(u_i x_t) \right] \quad \textbf{Equation 6-9}$$

This means that the hidden units are learning to reproduce the independent variables at the origin of the observed inputs.

6.3.2.c) *Adaptive ICA (e-ICA)*

If H represents the frontal cortex, then an interesting variation of ICA is to multiply its update term by the DA signal *e*. The size of *e* may act as an adaptive learning rate whose source is the reinforcement learning system critic. Also, if the reward is less than expected ($e < 0$), the features learned by the ICA unit may be more counterproductive than helpful, and *e* pushes the learning away from those features.

6.3.2.d) *e-gradient method*

Another possible approach is to base the update rule on the derivative of the objective function E applied to the hidden layer weights U, but constraining the update rule to only use information available locally. Let f' be the derivative of f, then the gradient of E with respect to U is approximated by:

$$\frac{\partial E}{\partial u_i} = 2e_t\left[\gamma v_i f'(u_i x_t)x_t - v_i f'(u_i x_{t-1})x_{t-1}\right] \quad \textbf{Equation 6-10}$$

Negating the gradient for minimization, adding a learning rate and removing the non-local weight information, gives the weight update rule:

$$\Delta u_i = \alpha e_t\left[f'(u_i x_{t-1})x_{t-1} - \gamma f'(u_i x_t)x_t\right] \quad \textbf{Equation 6-11}$$

Using the value of the weights $\boldsymbol{v}$ would lead to a rule that use non-local information. The cortex is unlikely to have this and might consider all the weights in $\boldsymbol{v}$ to be equal to some constant.

To avoid neurons all moving in the same direction uniformly, we encourage the units on the hidden layer to minimize their covariance. This can be achieved by adding an inhibitory neuron. Let q_t be the average activity of the hidden units at time t, i.e., the inhibitory neuron activity. Let $\bar{q}_t$ be the moving exponential average of q_t. Since

$$Var[q_t] = \frac{1}{n^2}\sum_{i,j}\mathrm{cov}\left(y_{t,i}, y_{t,j}\right) \cong TimeAverage\left((q_t - \bar{q}_t)^2\right) \quad \textbf{Equation 6-12}$$

and ignoring the f's non-linearity , the gradient of the $Var[q_t]$ with respect to the weights U is approximated by:

$$\frac{\partial Var[q_t]}{\partial u_i} = 2(q_t - \bar{q}_t)x_t \quad \textbf{Equation 6-13}$$

Combined with the previous equation, this results in a new update rule:

$$\Delta u_i = \alpha e_t\left[f'(u_i x_{t-1})x_{t-1} - \gamma f'(u_i x_t)x_t\right] + \alpha[\bar{q}_t - q_t]x_t \quad \textbf{Equation 6-14}$$

When allowing the discount factor to be different on the hidden layer, we found that $\gamma = 0$ gave much better results (e-gradient(0)).

6.4 Simulations & Results

All models of section 6.3 were run on the Acrobot task (Sutton & Barto, 1998). This task consists of a two-link pendulum with torque on the middle joint. The goal is to bring the tip of the second pole in a totally upright position.

6.4.1 The task: Acrobot

The input was coded using 12 equidistant radial basis functions for each angle and 13 equidistant radial basis functions for each angular velocity, for a total of 50 non-negative inputs. This somewhat simulates the input from joint-angle receptors. A reward of 1 was given only when the final state was reached (in all other cases, the reward of an action was 0). Only 3 actions were available (3 actor units), either -1, 0 or 1 unit of torque. The details can be found in (Sutton & Barto, 1998).

Fifty networks with different random initializations were run for all models for 100 episodes (an episode is the sequence of steps the network performs to achieve the goal from the start position). Episodes were limited to 10000 steps. A number of learning rate values were tried for each model (actor-critic layer learning rate, and hidden layer learning rate). The selected parameters were the ones for which the average number of steps per episode plus its standard deviation was the lowest. All hidden layer models got a learning rate of 0.1.

6.4.2 Results

Figure 6-2 displays the learning curves of every model evaluated. Three variables were compared: overall learning performance (in number of steps to success per episode), final performance (number of steps on the last episode), and early learning performance (number of steps for the first episode).

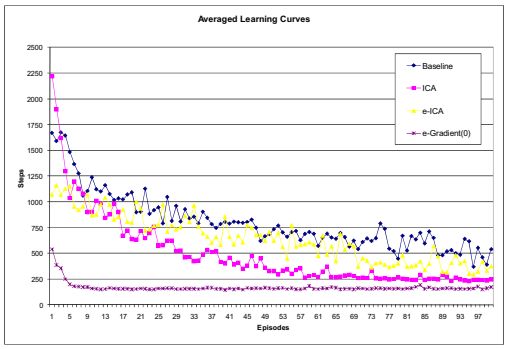


Figure 6-2 : Learning curves of the models.

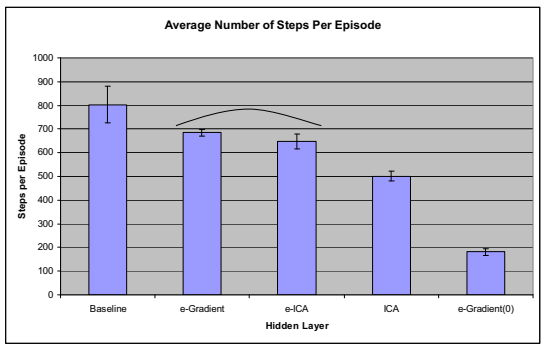


Figure 6-3 : Average number of steps per episode with 95% confidence interval.

6.4.2.a) *Space under the learning curve*

Figure 6-3 shows the average steps per episode for each model in decreasing order. All models needed fewer steps on average than baseline (which has no training at the hidden layer). In order to assess the performance of the models, an ANOVA analysis of the average number of steps per episode over the 100 episodes was performed. Scheffé post-hoc analysis revealed that the performance of every model was significantly different from every other, except for *e*-gradient and *e*-ICA (which are not significantly different from each other).

6.4.2.b) *Final performance*

ANOVA analysis was also used to determine the final performance of the models, by comparing the number of steps on the last episode. Scheffé test results showed that all but *e*-ICA are significantly better than the baseline. Figure 6-4 shows the results on the last episode in increasing order. The curved lines on top show the homogeneous subsets.

6.4.2.c) *Early learning*

Figure 6-2 shows that the models also differed in their initial learning. To assess how different those curves are, an ANOVA was run on the number of steps on the very first episode. Under this measure, *e*-gradient(0) and *e*-ICA were significantly faster than the baseline and ICA was significantly slower (Figure 6-5).

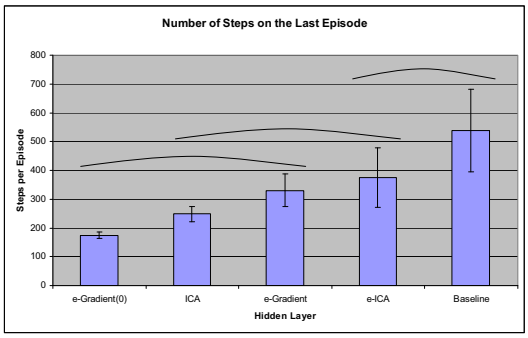


Figure 6-4 : Number of steps on the last episode with 95% confidence interval.

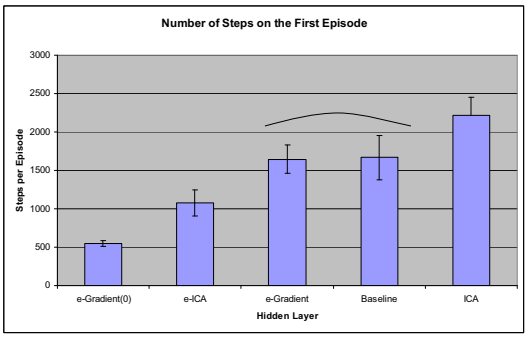


Figure 6-5 : Number of steps on the first episode with 95% confidence interval.

It makes sense for ICA to be slower at the beginning, since it first has to stabilize for the RL system to be able to learn from its input. Until the ICA has stabilized, the RL system has moving inputs, and hence cannot learn effectively. Interestingly, *e*-ICA was protected against this effect, having a start-up significantly faster than the baseline. This implies that the *e* signal could control the ICA learning to move synergistically with the reinforcement learning system.

6.4.3 External comparison

Acrobot was also run using standard backpropagation with TD and ε-Greedy policy. In this setup, a neural network of 50 inputs, 50 hidden sigmoidal units, and 1 linear output was used as function approximator for *V*. The network had cross-connections and its weights were initialized as in section 6.3 such that both architectures closely matched in terms of power. In this method, the RHS of the TD equation is used as a constant target value for the LHS. A single gradient was applied to minimize the squared error after the result of each action. Although not different from the baseline on the first episode, it was significantly worse on overall and final performance, unable to constantly improve. This is a common problem when using backpropagation networks in RL without handcrafting the necessary complex features. We also tried SARSA (see section 2.3.3) (using one network per action), but results were worse than TD.

The best result we found in the literature on the exact same task are from (Sutton & Barto, 1998). They used SARSA(λ) with a linear combination of tiles. Tile coding discretized the input space into small hyper-cubes and few overlapping tilings

were used. From available reports, their first trial could be slower than *e*-gradient(0) but they could reach better final performance after more than 100 episodes with a final average of 75 steps (after 500 episodes). On the other hand, their function had about 75000 weights while all our models used 2900 weights.

6.5 Discussion

In this paper we explored a new family of biologically plausible reinforcement learning algorithms inspired by models of the basal ganglia and the cortex. They use a linear actor-critic model of the basal ganglia and were extended with a variety of unsupervised and partially supervised learning algorithms inspired by brain structures. The results showed that pure unsupervised learning was slowing down learning and that a simple quasi-local rule at the hidden layer greatly improved performance. Results also demonstrated the advantage of such a simple system over the use of function approximators such as backpropagation. Empirical results indicate a strong potential for some of the combinations presented here. It remains to test them on further tasks, and to compare them to more reinforcement learning algorithms. Possible loops from the actor units to the hidden layer are also to be considered.

Acknowledgments

This research was supported by a New Emerging Team grant to John Kalaska and Yoshua Bengio from the CIHR. We thank Doina Precup for helpful discussions.

Chapitre 7. Alternative Time Representation in Dopamine Models

This paper (Rivest et al., 2010a) by François Rivest, John Francis Kalaska, & Yoshua Bengio appeared on-line in *Journal of Computational Neuroscience* on October 22nd 2009 (doi: 10.1007/s10827-009-0191-1). Minor precisions were added in this chapter.

Most of this paper is F.R.'s work and ideas. The idea to use the LSTM networks is a suggestion of Y.B. & Douglas Eck; J.K. provided extensive direction, feedback, ideas, and comments throughout this work from the original ideas to the final manuscript. Y.B. contributed as advisor, providing feedback and suggestions at all stages of the work.

To better understand how the cortex and the dopaminergic system interact in learning a representation of the environment, it is useful to first look at the simplest possible task for which it is known that some representation must be learned and for which there is reproducible data to model. There is dopaminergic data suggesting that even under simple trace conditioning with fixed interstimulus interval, the brain learns an internal model of the task and of its timing (Hollerman & Schultz, 1998; Daw et al., 2006). Although this dopaminergic data has been modeled several times using TD, these models were all provided with a complete representation of the task for free (see section 4.4.3). But timing representation could well be learned by the cortex when necessary (Dragoi et al., 2003; Hopson, 2003). This paper (Rivest et al., 2010a) shows that it is effectively possible to learn without supervision the necessary representation that TD requires to reproduce the dopaminergic data on this task. An unsupervised learning algorithm representing the cortex learns the environment dynamics and its internal representation is used as input to a temporal-difference learning model of the dopaminergic system. Not only does the model reproduce successfully the dopaminergic data, but the model of the cortex also successfully reproduces activity often observed in similar tasks in the cortex.

Abstract

Dopaminergic neuron activity has been modeled during learning and appetitive behavior, most commonly using the temporal-difference (TD) algorithm. However, a proper representation of elapsed time and of the exact task is usually required for the model to work. Most models use timing elements such as delay-line representations of time that are not biologically realistic for intervals in the range of seconds. The interval-timing literature provides several alternatives. One of them is that timing could emerge from general network dynamics, instead of coming from a dedicated circuit. Here, we present a general rate-based learning model based on long short-term memory (LSTM) networks that learns a time representation when needed. Using a naïve network learning its environment in conjunction with TD, we reproduce dopamine activity in appetitive trace conditioning with a constant CS-US interval, including probe trials with unexpected delays. The proposed model learns a representation of the environment dynamics in an adaptive biologically plausible framework, without recourse to delay lines or other special-purpose circuits. Instead, the model predicts that the task-dependent representation of time is learned by experience, is encoded in ramp-like changes in single-neuron activity distributed across small neural networks, and reflects a temporal integration mechanism resulting from the inherent dynamics of recurrent loops within the network. The model also reproduces the known finding that trace conditioning is more difficult than delay conditioning and that the learned representation of the task can be highly dependent on the types of trials experienced during training. Finally, it suggests that the phasic dopaminergic signal could facilitate learning in the cortex.

7.1 Introduction

To interact successfully with the world, an individual must acquire knowledge about its temporal dynamics and causal temporal relationships. For instance, feeding is vital. The ability to learn something about its food supply dynamics can give an animal a major advantage in survival. For example, a predator may see its prey disappear into a hole. To know that the prey is likely to exit its burrow after a while once he has moved away, might be very useful for the predator to learn to improve its

hunting success. But how can such knowledge about predictable relationships between events separated in time be learned?

The phasic activity of mesencephalic dopaminergic (DA) neurons in many conditioning studies is believed to represent a reward prediction error signal that could play an important role in learning under positive reinforcement situations. This activity is often modeled using the temporal-difference (TD) learning algorithm (Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Daw, Courville, & Touretzky, 2003, 2006; Pan et al., 2005; Bertin et al., 2007; Ludvig et al., 2008). However, TD models only give half the answer. Given a proper representation, TD solves only the credit assignment problem, i.e., how to associate the resulting reward to the right state, stimulus, or action. Understanding that even if the prey has disappeared, it is still there, estimating how long to wait for it or to let it go, all these elements require something else. Remembering that the prey is hiding requires some form of short-term memory. Estimating elapsed time requires some form of time perception. This information is not directly observable by the predator, but refers to an *internal representation* of the situation that is developed by the animal.

Many studies and models of DA neuron activity used a paradigm similar to trace conditioning (Figure 7-1), with a fixed delay between conditioned stimulus (CS) offset and unconditioned stimulus (US, or reward) onset. In this form of conditioning, there is a period of time between the CS and the US during which no unique informative input arrives from the environment, making trials and intertrial intervals indistinguishable over short periods of time without some form of memory. Dopaminergic neuron data in such situations shows that these neurons have some information about the fact that the CS is a predictor of the US, and about the time when the US should appear (Schultz et al., 1993; Montague et al., 1996; Hollerman & Schultz, 1998; Morris et al., 2004). However, most TD models (Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Daw et al., 2003, 2006; Pan et al., 2005; Bertin et al., 2007; Ludvig et al., 2008) do not address how the necessary representation is formed. Instead, most models of DA activity use

pre-set delay lines or temporal basis functions to provide to TD a history of past events (short term memory) and a built-in temporal representation of them (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Pan et al., 2005; Bertin et al., 2007; Ludvig et al., 2008). Delay-line representations require multiple predefined lines with specific parameters to accommodate all the possible timing information for all possible stimuli of a given task, similar to having axons of various lengths or diameters or a set of poly-synaptic connections. This seems physiologically unrealistic for delays in the order of seconds.

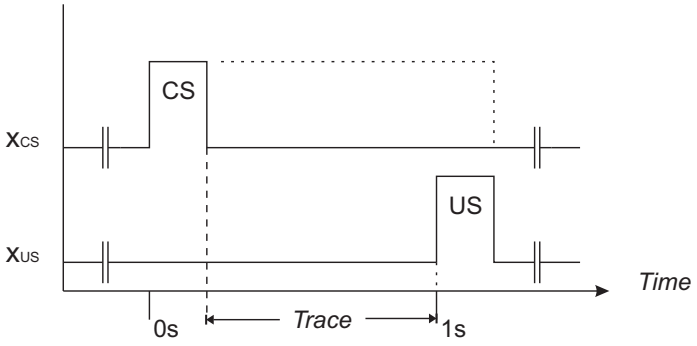


Figure 7-1: Schema of the sequence of stimulus events in appetitive trace conditioning with a constant CS-US interval. In classical conditioning, *trace* means that the CS and US do not overlap in time, i.e., there is a temporal gap between CS offset and US onset. The delay period between the CS and US requires learning to associate the US with a CS that had occurred in the past, and not with any current sensory input. Delay conditioning, in which the CS stays ON for the whole interval duration or longer, is shown using the dotted x_{cs} line. The time interval between the CS and US onsets is usually constant, unless stated otherwise.

How the brain represents and processes temporal information is a deep problem that impacts on many aspects of brain function (Ivry & Schlerf, 2008). Interval timing theory commonly uses a combination of an internal pacemaker or clock, an accumulator, and a memory to model timing. This is the basis of scalar timing theory (Gallistel & Gibbon, 2000; Church, 2003). However, many timing models do not solve the credit assignment problem and instead use explicit *start* and *stop* signals that provide explicit a priori knowledge of the temporal sequence of salient events in the task.

Alternatively, it has been proposed that a general learning mechanism that builds a representation of the world and that learns a time representation when needed might exist (Hopson, 2003; Dragoi et al., 2003). This is consistent with models of distributed time representations that arise in the cerebral cortex (Lewis, 2002;

Durstewitz, 2004) as a result of the intrinsic dynamical properties of the neural circuits implicated in learning the temporal structure of a task (Ivry & Schlerf, 2008).

Consistent with the latter perspective, the model presented here supports the hypothesis that the brain can learn a rich representation of tasks, including timing, that serves as input to neurons building an estimate of future rewards (Doya, 2000; Daw et al., 2003, 2006). As an alternative to dedicated delay-line and clock models of time, we use long short-term memory networks (LSTM) (Gers et al., 2002), a general learning algorithm, to provide an adaptive input representation to a TD(λ) model of the DA system (Pan et al., 2005). LSTM is a recurrent neural network that learns structure and relationships in sequential data and which contains some general short-term memory-like ability. In our model, an LSTM network representing the cerebral cortex builds its own representation of the world while experiencing a continuous stream of sensory events without explicit start and stop signals. We hypothesized that the LSTM learns to predict the timing and relationships of sensory events in the environment, hence building a representation for the environment dynamics, while TD builds an estimate of future rewards and computes a reward error signal. We found that once the LSTM has learned to identify the CS and to predict the timing of the US in an appetitive trace conditioning task with a constant CS-US interval, the necessary representation of the task, including time-dependent activity similar to that found in the cortex (Funahashi et al., 1989; Leon & Shadlen, 2003; Lebedev et al., 2008), emerges within the LSTM network. It allows TD to learn the proper context-dependent reward estimate, thereby still reproducing experimental findings on DA neuron activity (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz & Schultz, 1994; Pan et al., 2005), including probe trials of unexpected delay intervals (Hollerman & Schultz, 1998) without an ad hoc experimenter-built representation of time or task structure.

The *Methods* section first describes the model and the simulation set-up. After a brief summary of the training results, the *Results* contains two main sections. The first shows that our TD model using a learned representation still reproduces

dopaminergic phasic activity in a number of conditions. The second section reveals how the learned task dynamics are represented by the LSTM networks.

7.2 Methods

7.2.1 The model

The rate-code model consists of two interconnected networks (Figure 7-2): one representing a cortical area (frontal or parietal) that learns the environment via unsupervised learning mechanisms, and one representing the basal ganglia and the dopaminergic system responsible for upcoming reward estimates and reward estimation errors. The cortical area was modeled using LSTM networks (Hochreiter & Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2002) to learn its environment. LSTM is a general neural network learning algorithm used in a wide range of machine learning applications (Eck & Schmidhuber, 2002; Bakker, 2002) that implements working memory in an intuitive way using gated recurrent loop mechanisms. The basal ganglia (BG) and the dopaminergic system were modeled using the TD(λ) algorithm (Sutton & Barto, 1998; Pan et al., 2005) to calculate reward estimates and errors. TD is a learning algorithm for reinforcement learning that predicts the sum of all future rewards and then computes an error signal from the difference between two successive estimates and the actual experienced reward. It has been used in numerous dopaminergic models (Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Daw et al., 2003, 2006; Pan et al., 2005; Bertin et al., 2007; Ludvig et al., 2008). The pseudocode for the model is provided in *Annexe II, Supplemental Pseudocode*. Parameter values are given for 200ms time steps.

7.2.1.a) *LSTM model of the cortex*

Let $x_{CS,t}$ and $x_{US,t}$ be the sensory inputs CS and US respectively at time *t.* The cortical network consists of an LSTM network trained to predict its next input. Since the CS always occurs after random delays, only the US ($x_{US,t}$) can be predicted, so the network has a single output $y_{US,t}$ whose target value is $x_{US,t+1}$, the US signal at the next time step. Therefore, when stimuli reach the cortex, they fulfill two functions: first, to

update the network weights to improve the prediction it made (its output $y_{US,t-1}$ at the previous time step); and second, to be processed to predict $x_{US,t+1}$ at the next time step (its output $y_{US,t}$). Mathematically, the network is updated to minimize its squared prediction error

$$(y_{US,t} - x_{US,t+1})^2 \quad \textbf{Equation 7-1}$$

using a method similar to backpropagation. The error is computed by the node

$$e_{US,t} = (y_{US,t-1} - x_{US,t}) \quad \textbf{Equation 7-2}$$

in Figure 7-2 and the signal is fed back (as part of the gradient of the error function) into the LSTM for weight updates.

The model uses the full form of the LSTM network that can be found in (Hochreiter & Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2002). An LSTM network consists of a set of inputs, memory neurons (called *state cells* in the LSTM literature), gates and a bank of outputs. The gates control the signals that can enter the memory cells, their rate of forgetting or build-up, and their link to the outputs. Memory cells and their gates form a memory block (Figure 7-3). In each time step, input patterns are reproduced at the input layer of the LSTM, memory blocks are processed (using current input and the output from the memory block in the previous time step through the recurrent link), and their output is passed through to the output layer of the LSTM.

An LSTM network may have multiple memory blocks in parallel acting like a hidden layer in a multilayer network with possible recurrence from the output of each memory block to the input of all memory blocks (Figure 7-2). The choice of the number of memory blocks is somewhat arbitrary. We were initially concerned that using only one memory block might artificially constrain the range of potential solutions that the networks could learn. Therefore, the current model presented here has two memory blocks with two memory cells each. We chose two memory blocks because there are two critical problems to solve in this task: First, is the environment currently in an intertrial interval or a trial? Since we are not providing the network with any form of delay lines or history, the network receives no external information about recent events during the temporal interval between the CS offset and the US

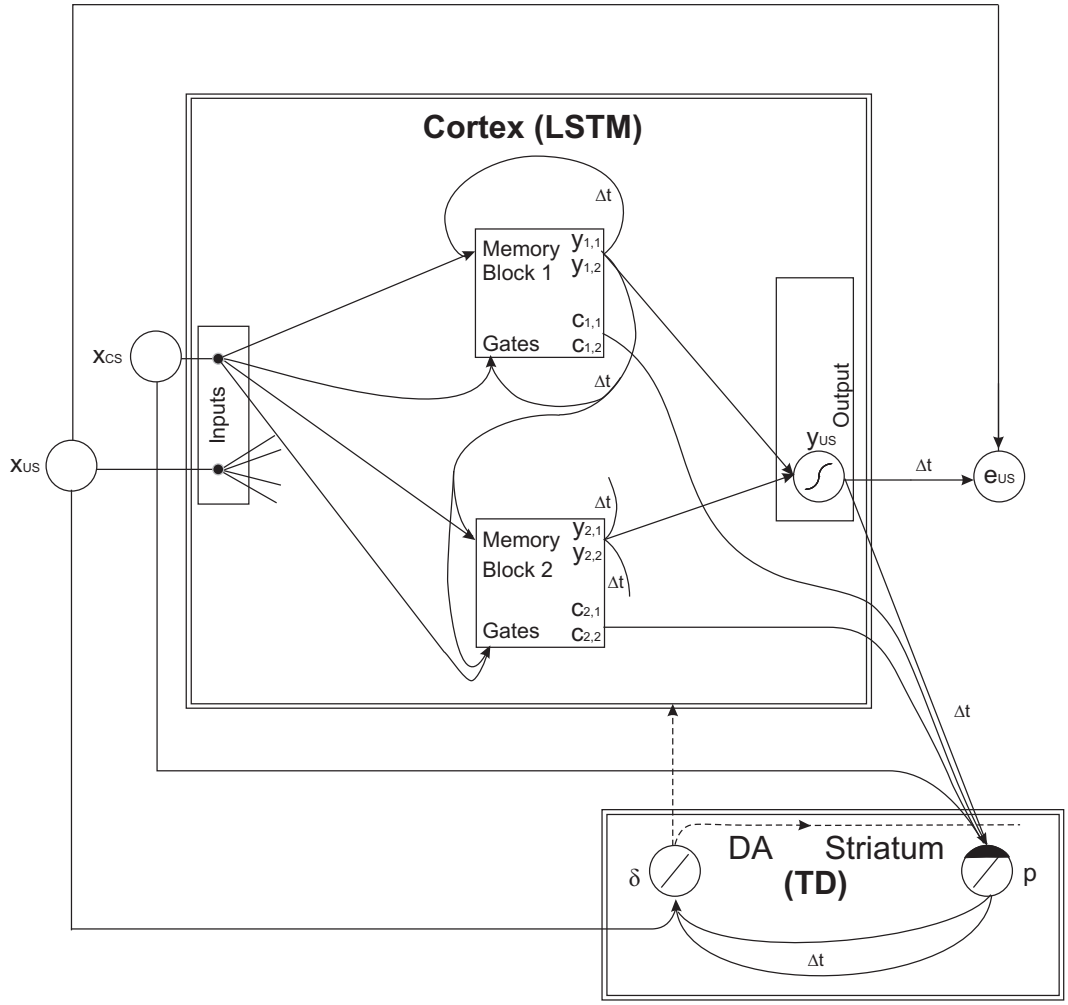


Figure 7-2: Schematic diagram of the model. An initial representation of the stimuli (x_{CS} and x_{US}) projects to the cortex (upper left box) modeled by a long short-term memory network (LSTM) whose output (y_{US}) learns to predict the next stimulus (x_{US} at time $t+1$) by minimizing the squared prediction error (e_{US}^2). The LSTM network is made of 2 *memory blocks* and an output layer. A memory block receives as input the stimuli, as well as recurrent connections from themselves and from each other. Projections from the second input and the second memory block are only partially drawn for clarity, but are similar to those of the first memory block. The memory blocks mainly project to the output layer from which an error signal can be computed. Some of the memory blocks' internal neurons also have extra-cortical projections ($c_{1,1}$, $c_{1,2}$, $c_{2,1}$, and $c_{2,2}$); the memory block's internal architecture is depicted in Figure 7-3. Δt indicates that the signal will be used in computation over the next time step (recurrent links for example). Sigmoidal and linear response curves indicate the activation functions the neurons apply to their weighted sum of inputs. The initial representation of the stimuli, the LSTM outputs, and the memory blocks' extra-cortical projections also project to a second region (lower right box), the striatum or mesencephalic dopaminergic circuits, modeled by the temporal-difference (TD) learning network. These afferent connections are used by TD *p* neurons to make predictions about future rewards. The TD error signal δ is the correlate of the phasic signal in dopaminergic neurons' activity, and plays an important role in learning. Dark boundary at the point of contact of inputs onto neuron *p* represent eligibility traces. Dashed lines from δ represent diffuse DA signals used in learning only. The dashed line pointing to the LSTM box represents the mesocortical projection in our full model.

onset shown in Figure 7-1. Second, if the environment is currently in a trial interval, how long has it been since the CS or how imminent is the US? Again, without a built-in time representation and no external cue about elapsed time, the network must build one by itself. Subsequent tests showed that more memory blocks did not improve learning performance, and that one block may have been sufficient but at some cost in performance (see *Results: LSTM Representation*).

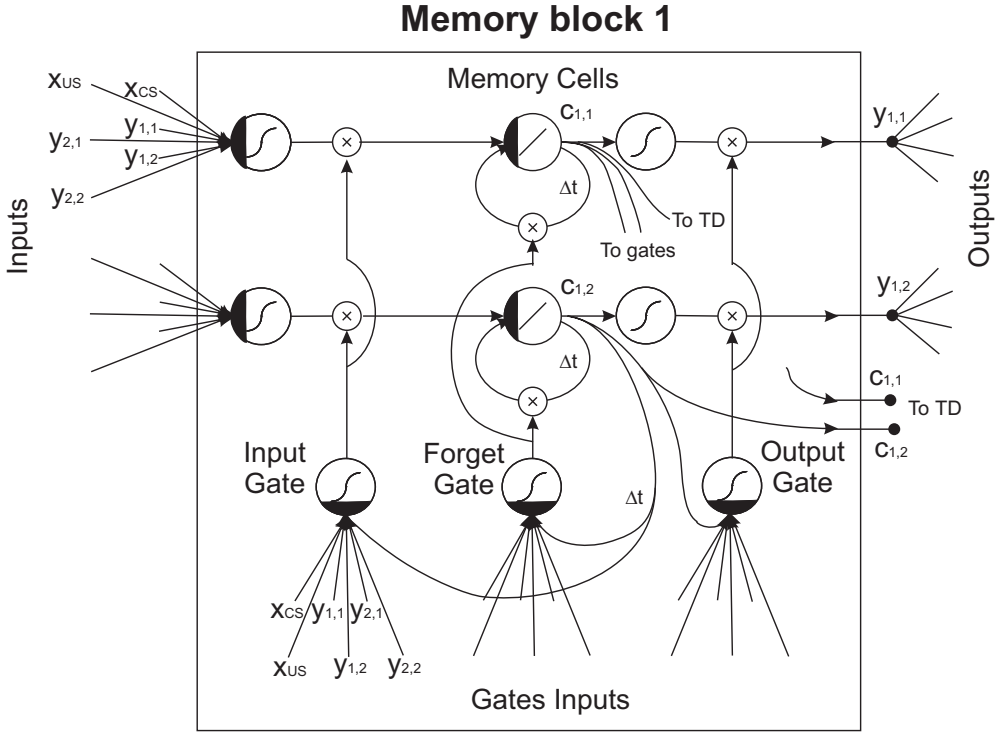


Figure 7-3: LSTM memory block connectivity. The input gate controls the gain of the stimuli or recurrent input to the memory cells. The forget gate controls the gain on the memory (state) cell recurrence (self-feedback). The output gate controls the gain on the memory block outputs. Input and forget gates receive memory cell content from previous time step (as well as stimuli and recurrent inputs), while the output gate receives the current time step content of the memory cell (as well as stimuli and recurrent input). Δt indicates that the signal will have reached its target by the next time step. Sigmoidal and linear response curves indicate the activation functions the neurons apply to their weighted sum of inputs. x represents signal multiplication. Dark boundaries at the point of contact of inputs onto neurons represent eligibility traces (see *Methods: The model: LSTM model of the cortex* for details).

In a memory block (Figure 7-3), the weighted sum of the block input is first computed. Then, this signal is multiplied by the activity of the *input* gate. This signal is then added into a linear recurrent memory cell, whose recurrence rate (leak or build-up) is controlled by the *forget* gate. Finally, the memory block output is determined by the memory cell activity multiplied by the *output* gate. Within a block,

multiple memory cells share the same gates. Each gate has its own weighted sum of the memory block inputs. Gates also have access (as input) to the memory cell activity. We used sigmoidal activation functions for the gates as well as for the memory cell inputs and outputs and for the LSTM outputs (represented by the sigmoid response curves in Figure 7-2 and Figure 7-3). It is important to note that all those input and gate weights are adaptive and must be appropriately learned. Although the LSTM memory blocks can be seen as a form of short-term memory, the network must learn for any given task to properly transform the inputs, to properly manipulate the gates, and to properly combine the memory blocks outputs.

The present model's learning was enhanced using eligibility traces within the recurrence outside and inside the memory blocks. These are represented by a dark boundary at the point of contact of the input to the neurons in Figure 7-3. The eligibility trace of a variable is a small exponentially decaying memory of that variable value. In the present simulation, the eligibility traces drop below 40% in the first second, and to about 10% by two seconds. For example if, x_i is an input variable, than its eligibility trace is given by

$$u_{i,t} = \lambda u_{i,t-1} + x_{i,t}, \ \lambda < 1. \qquad \textbf{Equation 7-3}$$

Eligibility traces replace the neuron's synaptic activity by a decreasing trace in the computation of the learning gradient, making an active synapse eligible for plasticity for a slightly longer time window. Mathematical eligibility traces could be considered as indicators of how eligible a synaptic connection is for correction, based on recent past pre-synaptic activity. This signal could be embedded in neurons through second messengers or calcium traces of synaptic co-activation (pre and post-synaptic) (Wickens & Kotter, 1995). It allows a functional connection to be made between events that happened in the immediate past and the current event. Although common in TD learning (Sutton & Barto, 1998) and shown relevant in TD models of dopamine (Pan et al., 2005), this is the first time that eligibility traces have been used in LSTM networks and may help the LSTM find a possible sequence of events. In the experiment presented here, it seemed to speed-up learning by a factor of two, but this was not thoroughly tested. The effectiveness of these modifications to LSTM

networks deserve to be studied independently in more tasks, but this is outside the scope of this paper.

All LSTM weights are initialized with a small random value in the range [-0.1, 0.1]. To update the weights and minimize the LSTM prediction error, an estimator of the gradient of this error with respect to the memory block weights must be computed. The mathematical details of this gradient are beyond the scope of this paper, but can be found in (Hochreiter & Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2002). We used a learning rate of $\alpha_{LSTM} = 0.5$ and an eligibility trace-decay parameter of $\lambda_{LSTM} = 0.8$. These parameters were manually adjusted to maximize LSTM success rate on the main training set-up (see *Simulations & training*). The cortical network projects its output prediction $y_{US,t}$ and its memory cell activities $c_{1,1,t}$, $c_{1,2,t}$, $c_{2,1,t}$, and $c_{2,2,t}$ to the BG network (Figure 7-2, LSTM to TD arrows).

LSTM networks were chosen as an abstract but relatively neural-like model of the cortex for a number of reasons. First, it was recently suggested in the timing literature that the ability to learn temporal relationships and timing could emerge from a general (i.e., non-dedicated) learning algorithm (Hopson, 2003; Dragoi et al., 2003). LSTM is a general on-line learning algorithm originally developed in the machine-learning field. Moreover, LSTM intuitively implements working memory and gating mechanisms, and by being trained to predict its next inputs, it becomes unsupervised. Doya (1999; 2000) has proposed that neural circuits in the cerebral cortex are specialized for efficient unsupervised learning of the state of the environment and of the behaving system, as is required in trace classical conditioning. Finally, and most importantly, while LSTM networks are abstract mathematical constructs and are not a biophysically realistic simulation of cerebral cortical neural circuits, they are still among the most powerful and the most biologically plausible of the recurrent neural-network learning algorithms. LSTM is a more powerful learner than backpropagation through time (BPTT) and real-time recurrent learning (RTRL) (for specific technical reasons, see Hochreiter and Schmidhuber, 1997). The latter algorithms also require either a complete history of past activities and gradients (BPTT) or a matrix of the partial derivatives of each neuron with respect to every weight in the network (RTRL)

making a biological implementation unrealistic. In contrast, the LSTM weight updates (synaptic changes) depend on signals that are mostly local in topology and time. Except for the output weights and error signals that must be fed back, a memory block contains all the information needed to do the weight update, and requires only the signals from the current and previous time step. It does not need any extra information about the other memory blocks, or about signals at even earlier time steps (*t-k, k>1*) beyond what is locally available. Although we do not expect to find an exact LSTM architecture in the cortex, nor do we claim that any particular element of the LSTM circuit corresponds to any specific cortical neuron, most of the elements from which LSTM is constructed are much more plausible than LSTM predecessors (such as BPTT or RTRL) or than multi-second delay lines. Moreover, LSTM is the only recurrent neural network that can learn in continuous mode without an explicit reset between trials (Gers et al., 2000) as in the current training setup (see *Simulations & training*).

7.2.1.b) *TD model of the dopaminergic system*

The basal ganglia and dopaminergic network are modeled by a standard TD(λ) algorithm (temporal-difference learning with eligibility traces) as in (Pan et al., 2005). TD builds an estimate

$$p(s_t) \approx \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad \textbf{Equation 7-4}$$

(where $\gamma < 1$) of the sum of future rewards (r_{t+k+1}) for the current state (s_t). γ is called the discount factor, it makes rewards expected further away in time seem less valuable (we used $\gamma = 0.98$). The estimate of expected reward p_t is computed by a linear weighted sum ($\boldsymbol{w}_p$) of the inputs ($x_{CS,t}$, $y_{US,t-1}$, $c_{1,1,t-1}$, $c_{1,2,t-1}$, $c_{2,1,t-1}$, $c_{2,2,t-1}$),

$$p_t = \boldsymbol{w}_p (x_{CS,t}, y_{US,t-1}, c_{1,1,t-1}, c_{1,2,t-1}, c_{2,1,t-1}, c_{2,2,t-1}), \qquad \textbf{Equation 7-5}$$

the inputs ($c_{1,1,t-1}$, $c_{1,2,t-1}$, $c_{2,1,t-1}$, $c_{2,2,t-1}$) are bounded to the range [0, 1]. The estimate is refined by computing an error δ_t (called the effective reinforcement signal) using the standard TD rule

$$\delta_t = r_t + \gamma p_t - p_{t-1} \qquad \textbf{Equation 7-6}$$

(i.e., δ_t is the difference between the previous and current expected reward estimate and the actual reward received r_t). δ_t is usually considered the correlate of the DA neuron phasic signal in TD models (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Daw et al., 2003, 2006; Pan et al., 2005; Ludvig et al., 2008). The weight vector $\boldsymbol{w}_p$ is updated using a three-way synaptic rule

$$\Delta \boldsymbol{w}_p = \alpha_{TD} \delta_t \boldsymbol{u}_t \qquad \textbf{Equation 7-7}$$

where $\boldsymbol{u}_t$ is the eligibility trace vector for the inputs of *p,*

$$u_i = \lambda_{TD} u_{i,t-1} + in_{i,t-1} \qquad \textbf{Equation 7-8}$$

(where $in_{i,t-1}$ is the i^{th} input signal to p_{t-1} at time *t-1* and $\lambda_{TD} = 0.9$) and bounded to the range [0, 1]. Eligibility traces usually apply to individual states (one per time step). Here they apply to a vector of units' activity, and these units may be active on numerous consecutive time steps, hence the need to limit them. The learning rate was $\alpha_{TD} = 0.1$. Weights for the TD expected reward estimate neuron *p* are initialized with a small positive value of 0.1. Considering that signals coming from LSTM may be delayed by processing within the cerebral cortex, the model uses LSTM signals from the previous time step (200 ms). Hence, the basal ganglia inputs are $x_{CS,t}$, $y_{US,t-1}$, $c_{i,j,t-1}$ as stimuli (corticostriatal connection) and r_t as the hedonic reward signal. Since there are no actions in the tasks simulated, there are no actor units here.

The last signal r_t is not an input to the prediction portion of TD learning circuit in which the striatal neurons could be computing the prediction (Houk et al., 1995), but a real hedonic reward signal that may come from the lateral hypothalamus (LHA) possibly via the pedunculopontine tegmental nucleus (PPTN or PPTg) (Fukuda et al., 1986; Dormont, Conde, & Farin, 1998; Brown et al., 1999). The hedonic value may also be the result of a more primitive association of the unconditioned stimulus (such as taste) to something having a hedonic value (e.g., a slower vegetative effect related to homeostatis such as changes in blood glucose or amino acid levels after ingestion of rewarding food or juice). Such association could be learned in LHA (Fukuda et al., 1990), but this is outside the scope of this model.

The TD error signal δ_t is now widely used to model the dopaminergic phasic signal. Striatal neurons could be computing *p* (Houk et al., 1995; Schultz, Apicella,

Romo, & Scarnati, 1995; Samejima et al., 2005). Some striatal neurons project directly or indirectly to dopaminergic neurons (Joel & Weiner, 2000) and some back to the cortex through the globus pallidus and thalamus. The latter are usually considered to be actors in TD models, but in this study, the model has no action to take, so there are no actor units feeding back to LSTM.

Finally, the TD learning rule for the neurons computing p is a heterosynaptic rule that combines neural pre- and post-synaptic co-activation with a delayed phasic dopaminergic signal. Biological mechanisms have been proposed for such rules (Houk et al., 1995; Wickens & Kotter, 1995). A review of relevant corticostriatal dopamine-dependent plasticity (Reynolds & Wickens, 2002) also seems to agree with the learning rule required for p, if striatum neurons implement it. It was also shown mathematically that such a three-way rule could work even at the level of spike timing (Florian, 2007; Izhikevich, 2007; Roberts, Santiago, & Lafferriere, 2008; Kolodziejski, Porr, & Worgotter, 2009; Potjans, Morrison, & Diesmann, 2009).

7.2.1.c) *Model of the mesocortical projection*

In trying to improve learning in the LSTM networks, we investigated how the mesocortical projection could play a role in learning in frontal cortex. In the main set of experiments, we added a link between the dopamine signal computed by the TD model and the cortex. Our hypothesis was that dopamine, which is a learning factor in TD and which could be considered a potential factor of the corticostriatal plasticity (Reynolds & Wickens, 2002), could also have a learning role in the cortex (Otani et al., 2003). An alternative hypothesis is that in the cortex, phasic dopaminergic inputs could also play an attentional role, by signaling time steps of greater importance or sensory events of greater salience, or by modulating the memory cell sensitivity to new inputs (gating) (Montague et al., 2004).

We implemented the first idea of a learning role by using the TD signal δ_t to modulate learning in the LSTM through the mesocortical projection (dashed arrow from TD to LSTM, Figure 7-2) by adding the absolute value $|\delta_t|$ to the LSTM basic constant learning rate α_{LSTM} :

$\alpha_{LSTM,t} = \alpha_{LSTM} + \beta\,|\delta_t|,$ **Equation 7-9**

where β is a constant controlling the level of contribution of $|\delta_t|$ to the LSTM learning rate (we used $\alpha_{LSTM} = 0.5$ and $\beta = 0.5$). Since LSTM learning requires the sensory input from the next time step for correction, and δ_t based on LSTM output also appears only at the next time step, the LSTM error signal and TD δ_t signal are well-aligned in time. Hence, the LSTM correction for $y_{US,t-1}$ depends on stimulus input $x_{US,t}$ and on their resulting signal δ_t.

7.2.2 Simulations & training

The simulations were done using a custom-built simulator written in Java. The simulator can be downloaded from *ModelDB* (http://senselab.med.yale.edu/modeldb/) (see *Appendix A: Supplemental Material 1: Java simulator* for details).

Let $x_{CS,t}$ and $x_{US,t}$ be the conditioned (CS) and unconditioned (US) stimulus signals at time t respectively. The CS presentation is modeled using $x_{CS,t} = 1.0$ and the US presentation by $x_{US,t} = 1.0$ for 1 time step, otherwise, these signals are zero (Figure 7-1). There is no noise in the signal. Time is incremented in 200ms time step intervals.

The networks were trained for 1000 4-minute (simulated time) *training block*s (i.e. a total of 40,000 conditioning trials) of alternating trials and intertrial intervals. Trials were appetitive classical trace conditioning with a fixed delay of 1s between stimulus onsets (Figure 7-1), while intertrial intervals were random delays between 4s and 6s. Every 10 training blocks, a *test* block was run that included 5 probe trials that had a delay of either 0.4s or 1.4s (early- and late-probe respectively) between the CS and US onsets. Individual early- or late-probe trials were randomly chosen (not balanced) in each test block. Training began and ended with a test block.

A training block was considered successful when the LSTM network was able to properly predict its next input (absolute error $|y_{US,t} - x_{US,t+1}| < 0.5$) for 300 consecutive time steps, i.e. for about 10 trials (or 1min), within it. A network was considered successful if it maintained this performance criterion in the last (1000^{th}) training block. In no case was the DA signal or any other signal from the TD model used to assess network success. The sole performance criterion was the ability of the

LSTM network to recognize the CS and predict the US in the task. We continued to train new naïve networks until we had 30 successful networks, according to our criteria, which we then used for analysis. LSTM parameters such as learning rate and eligibility trace-decay were roughly adjusted to maximize LSTM success without the mesocortical projection (modifying its learning rate on-line) in this setup and were then kept fixed for all other simulations. Data from the last 20 training and 3 test blocks of those 30 networks were recorded in an MS Access database that can be downloaded from the *Université de Montréal* institutional digital repository *Papyrus* (see *Appendix A: Supplemental Material 2: Simulations data* for details).

Some analyses required an assessment of the network responses to unpaired random CS and US presentations before and after training. For that, we used a 4-minute *control* block in which CS and US were randomly presented for a single time step (200ms). Each presentation was separated by a random delay (uniform between 5s and 7s). These blocks had the same number of single stimulus events as the training blocks' number of trials, but there was no pairing of stimuli and no "trials" per se. These blocks were also recorded (*Appendix A: Supplemental Material 2: Simulations data*).

7.3 Results

7.3.1 Successful training

The criterion for LSTM learning was correct prediction of network output for 300 successive time steps (approximately 10 sequential trials), indicating that the LSTM had properly learned the CS and US temporal relationship and that it could predict precisely the US arrival from the CS presentation. (see *Methods: Simulations & training*). In a first simulation, we did not use the mesocortical projection from TD to the LSTM. Without this connection, the LSTM is totally independent of TD; information flows only from the LSTM to TD. Only 52% of those LSTM networks had some successful blocks, taking 533 training blocks on average to achieve first success, and only half of those remained successful in the last training block (Table 7-I).

In a second simulation, we added the mesocortical projection from TD to the LSTM in a new set of naïve networks. Most (69%) of the networks achieved criterion in at least one training block, after 233 training blocks (≈10,000 trials) on average, which was more than twice as fast to first success as the networks without the mesocortical projection (see Table 7-I). Two-thirds of those networks remained correct until the end of 1000 40-trial training blocks and so were considered successful. While the first model resulted in a high variability of δ amplitudes from network to network, this second model also had a much smaller range of δ amplitudes (see Supplemental Table 7-II for details). We trained new naïve networks until we had 30 successful networks for analysis. Unless stated otherwise, analyses were done on this set of networks.

Table 7-I: Total number of networks trained (col 2), number of networks showing some successful blocks (a training block with 300 consecutive time steps (10 trials) of correct LSTM prediction) (col 3), number of networks that remained successful in the last (1000th) block (i.e., successful networks kept for analysis) (col 4), and average (mean and standard deviation) number of training blocks before the first successful training block (col 5).

Model	Total trained	With some successful blocks	Successful networks	First successful block
Without mesocortical projection (trace paradigm)	104	53 (52%)	28 (27%)	533±131
Full model (trace paradigm)	67	46 (69%)	30 (45%)	233±133
Full model (delay paradigm)	37	37 (100%)	30 (81%)	20±4

In a third simulation with new naïve networks and the mesocortical projection from TD to the LSTM, we used a delay (non-trace) task in which the CS remained ON for the whole interval, terminating at the same time as the US (dotted line, Figure 7-1). All the networks achieved criterion in at least one training block, after only 20 training blocks (≈800 trials) on average. These simulations are consistent with the finding that delay conditioning should be much easier to learn than trace conditioning.

7.3.2 Dopamine neuron responses

The LSTM network was the sole source of information about timing in our TD model. Further evidence that the LSTM provided a sufficiently rich representation of the task environment came from an analysis of the neurons in TD after training. Consistent with previous studies (Montague et al., 1996; Schultz et al., 1997; Suri &

Schultz, 1998, 1999; Pan et al., 2005; Bertin et al., 2007; Ludvig et al., 2008), the TD δ neurons in our model showed most of the characteristics of dopaminergic phasic signals (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz & Schultz, 1994; Pan et al., 2005).

Real dopamine neuron activity reported in the literature has been documented in three different contexts: activity before learning, transfer from US to CS during acquisition, and activity after learning. We verified that our combined LSTM-TD model could reproduce the known response properties of dopamine neurons during classical conditioning, within the context of a completely naïve network that must learn the task dynamics while building a representation for it. That is, the network must first learn that there is a relationship between CS and US, then it must learn to maintain a memory of a recent CS event (or code for the gap period between CS and US), and finally it must learn a proper time representation for the CS-US timing relationship.

For instance, before learning a task, real dopamine neurons usually have a brisk response to unexpected rewards (US) or to novel non-rewarding stimuli (CS) (Ljungberg et al., 1992; Pan et al., 2005), but the response to unexpected novel non-rewarding CS stimuli typically ceases after a few presentations (Ljungberg et al., 1992). To test whether our model would reproduce this behavior, we ran 30 untrained networks on a *control* block with only random unpaired CS or US presentation every 6 seconds on average. A typical δ signal is shown in Figure 7-4A. The TD δ (DA) neurons initially responded to rewarding stimuli and to the novel non-rewarding stimulus (the latter is caused by the non-zero weights initialisation in TD, as in (Suri & Schultz, 1998)), but the response to the latter stimulus rapidly decreased, fitting well the experimental data in animals. A *paired-samples t-test* (*one-tailed*) between the responses to the first and last stimulus presentation in the block showed a significant mean decrease of 81% in magnitude to the CS ($P < 1.0E\text{-}31$, $df=29$). Therefore, the δ response to a novel non-rewarding stimulus habituates rapidly after a few repetitions when it is not predictively coupled with reward.

During learning, and once well conditioned, real dopamine neurons usually show a transfer of response from a main response to the US at the beginning, to a response to the CS (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz & Schultz, 1994; Pan et al., 2005) by the end. Some neurons even stop responding to the US (about 2/3), while for other neurons, the response may persist. No real DA neurons maintain a sustained activity during the fixed delay (Pan et al., 2005; Schultz et al., 1993) unless reward is uncertain in the task (Fiorillo et al., 2003), which is not the case here.

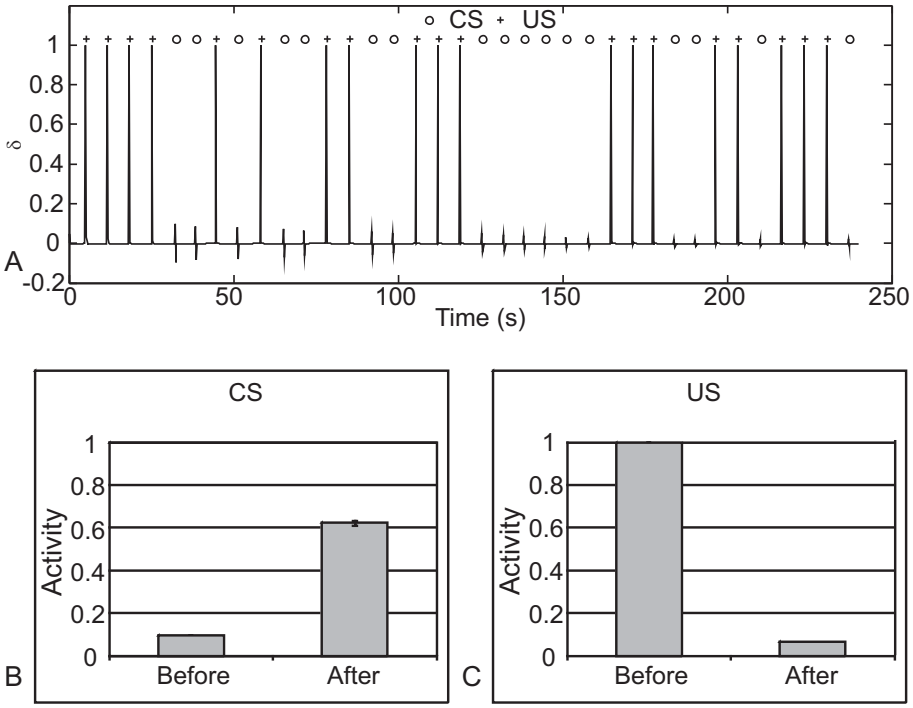


Figure 7-4: Simulated dopamine neuron responses. A. Typical untrained network δ response on random presentation of arbitrary non-rewarding stimulus or reward in a control block. Reward presentations are marked by a plus sign and non-rewarding stimulus by a circle. B. Mean and standard deviation of mean networks δ responses to the CS before and after a single training block. C. Mean of mean networks δ responses to the US before and after complete training (1000 blocks).

We recorded the first training block for a set of 30 untrained networks and compared the δ response of the first and last trials. The mean response to the CS increased significantly by a factor of 5 in magnitude (*one-tailed, paired-samples t-test, $P < 1.0E-51$, $df=29$*), showing a rapid acquisition of the CS response (Figure 7-4B). We then compared the δ response to the US from the control block of

untrained networks to those of the last training block of the 30 successful networks. For that, we calculated for each network the mean δ signal at reward presentation over the whole block. The δ response to the US presentation decreased significantly by 94% of its magnitude from before to after learning *(two-sample t-test, one-tailed, unequal variance, $P < 0.001$)*. Of the 30 networks, only 9 stopped responding to reward (neurons' response to the US was less than a third of its response to the CS). This is about half as frequent as reported in real neural populations, although this may vary depending how "stop responding" was defined in different studies, and depending on the training situation. From those 9 networks, 4 responded with a depression (neuron's response was lower than -0.1). This is much more than in real dopaminergic neural populations; neural depression to expected reward is rare. Mean δ response including all networks (not just those that stop responding to the US) to normal US, before and after training is shown in Figure 7-4C. No network δ neurons showed evidence of sustained activity during the delay period of normal trials in the last training block, indicating that time is not encoded explicitly in δ neuron activity. On average, the model showed the characteristics of the transfer of the DA signal from the US to the CS and the lack of explicit coding of time in δ, that are both seen in real DA responses in animals.

Finally, real dopamine neurons have typical responses after training. Among them, they continue to respond to unexpected rewards (Mirenowicz & Schultz, 1994; Pan et al., 2005) and show depression at the expected time when a reward is omitted (Schultz et al., 1993; Morris et al., 2004; Pan et al., 2005). To test that the TD error signal δ would still respond to unexpected reward after conditioning, we ran a post-training control block on the successful networks with only random US or CS presentation (no pairing) every 6 seconds or so (as in Figure 7-4A). The averaged δ responses of the networks to unexpected rewards were around 0.98 ± 0.02 (mean and standard deviation). To test that the δ signals were depressed by an omitted reward, we looked at the late probe trials of the last 3 test blocks (Figure 7-5). The mean response at the expected time of the US was significantly lower than 0 by an average

of -0.5 (*one-sample t-test, two-tailed, df=29, $P \leq 0.005$*), again fitting well real dopaminergic neuron data.

7.3.3 Dopamine and timing

Real DA neurons differentiate expected from unexpected CS-US timing (Hollerman & Schultz, 1998). We presented the networks with *test* blocks containing a few early (0.4s) and late (1.4s) interstimulus *probe* trials. More than half of the networks showed typical early and late probe-trial δ response curves shown in Supplemental Figure 7-9 (Hollerman & Schultz, 1998) (for details, see Supplemental Table 7-III). In early trials, the δ signal averaged over all networks rose at the arrival of the early US and was weak at the expected US time (Figure 7-5, see Supplemental Figure 7-10 for results on the model without the mesocortical projection). On late trials, in contrast to early trials, δ dropped at the expected US time, and then rose in response to the unexpected late arrival of the US.

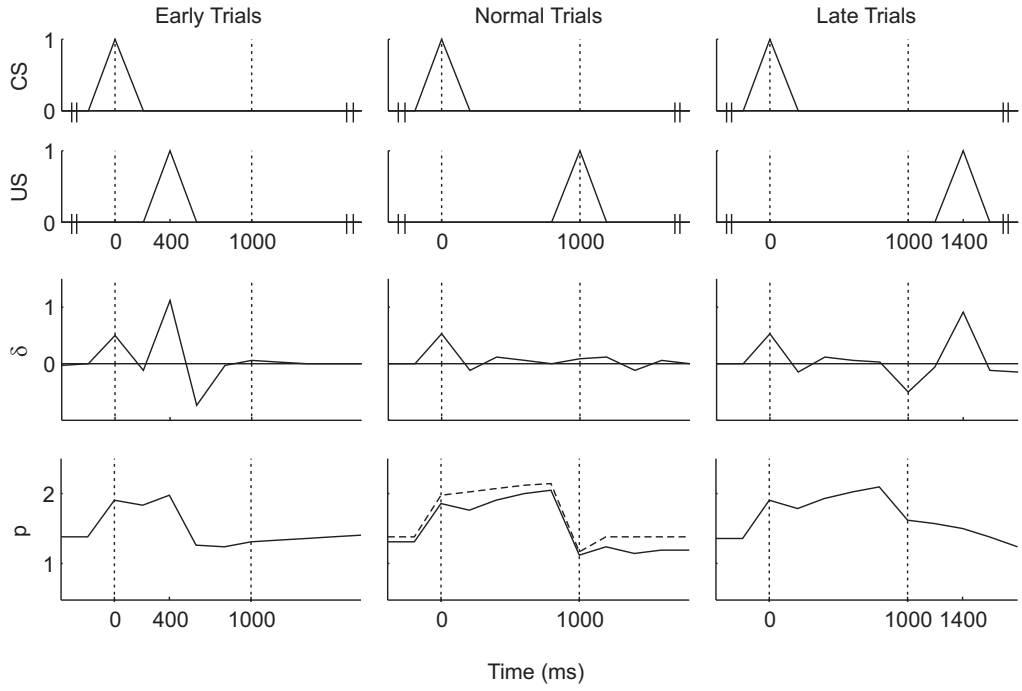


Figure 7-5: δ and p signals from trained TD networks. From left to right, each column shows signals for early (short 400ms probe trials), normal (1000ms trials), and late (long 1400ms probe trials) trials. Vertical dashed lines indicate the standard CS-US training interval of 1000ms. The top rows show the CS and US signal for each type of trial. The bottom rows show network population average δ_t and p_t signals. Early- and late-probe trials are averaged over the last 3 test blocks, normal trials are averaged over the last training block only. Dashed line (bottom centre figure) is the theoretical p value for normal trials (see *Appendix B: Theoretical p signal*).

Recently, Fiorillo et al. (2008) reported that the DA response to an early US presentation may more closely resemble the response to the normal US than to an unpredictable US. Using a control block of unpaired CS and US, our results seem closer to the original data (Hollerman & Schultz, 1998) with a δ response in early trials more similar to unpredicted US than to normal trials. The difference in findings between the two studies may be related to various factors such as the animals' training history or to differences in the relative acquired timing precision of the animals in their tasks. Also, Hollerman & Schultz (1998) illustrated their findings with a "typical" single-neuron response which may not be representative of every neuron, whereas Fiorillo et al. (2008) presented their findings with a population average. The distribution of probe trials in our simulations is closer to Hollerman's experiment than to Fiorillo's experiment that has about 20% of early probe trials. But Fiorillo argues that the difference in responses to early US does not seem to be caused by this difference in probe frequency. Furthermore, our networks were also trained to precisely predict the time of the US arrival while Fiorillo's data suggests that the animals' expectation may be more diffuse over time. More data would be necessary to resolve this issue.

Furthermore, the absence of response at the usual delay in early trials could not be captured by delay-line models unless some ad hoc reset was added (Suri & Schultz, 1999). The absence of a depression at the expected time of reward in early trials was inferred to imply that internally, the early arrival of the reward terminated the trial (Suri & Schultz, 1999; Daw et al., 2006), as it also does in our model. A recent paper successfully reproduced the result without such an assumption by adding a temporal representation to the reward that cancels out the residual CS temporal representation after the early US (Ludvig et al., 2008). More experimental data are required to determine which approach is more valid.

7.3.4 Reward-predictive Neurons

We also analyzed the activity of the TD reward-predictive neurons that learned an approximation of expected future reward $p(s_t)$. With a few assumptions, a theoretical p signal can be computed (see *Appendix B: Theoretical p signal*) to assess

how well TD learned its reward estimate *p* using the LSTM representation. Overall, the population average from the trained networks (Figure 7-5, solid line) was very close to the theoretical estimate (dashed line).

We examined the TD expectation of *p* in early and late probe trials. In early trials, the expectation returned to its intertrial baseline value once a reward was received (Figure 7-5). More interestingly, in late trials, the networks maintained some reward expectation beyond the normal 1s delay. This could indicate that the networks had learned to expect occasional late probe trials. Since they experienced a test block every 10 training blocks, about 0.62% of the trials were late trials. This late trial signal (Figure 7-5) also resembles the equivalent signal on omission trials in (Ludvig et al., 2008) when looking only at the contribution of the temporal representation of the CS provided to TD in their model. *p* signals also resemble striatal recordings in similar conditions (Schultz, Apicella, Scarnati, & Ljungberg, 1992).

7.3.5 LSTM Representation

An LSTM network consists of a set of inputs, memory neurons, gates and a bank of outputs (Figure 7-2). The gates control the signals that can enter the memory cells, their rate of forgetting or build-up, and their link to the outputs. Memory cells and their gates form a memory block (Figure 7-3). Within a block, multiple memory cells share the same gates. Each gate has its own weighted sum of the memory block inputs. We used sigmoidal activation functions for the gates. Gates also have access (as a recurrent input for the input and forget gates) to the memory cell activity. We used two memory blocks, each with two memory cells, to give the networks enough flexibility and computational power to differentiate trials from intertrials and to keep track of time. Subsequent experiments with a single memory block and a single memory cell lead to the same representation, but with much lower success rate on training. These networks will not be described here (see *Appendix C: Memory block responses* for details).

In each time step, patterns are processed through the input layer to compute the weighted sum of the memory block inputs. This signal is multiplied by the activity of the *input* gate. This signal is then added into a linear recurrent memory-cell

unit, whose recurrence rate (leak or build-up) is controlled by the *forget* gate. The memory blocks and gates also receive the previous memory blocks' output via the recurrent link. The memory block output depends on the memory cell activity multiplied by the *output* gate, and is passed through the output layer.

To successfully predict the US, the LSTM network must make an association between the CS and the US and must learn to monitor elapsed time since the last CS presentation. These elements must appear in one way or another in the memory blocks. All the memory block inputs and gates and network output weights are learned by the network. These are not set by the experimenter, leaving the network with potentially many different solutions.

7.3.5.a) *The general pattern*

Typical response patterns emerged for each element in most successful networks (17/60 memory blocks did not contribute directly to LSTM output in different networks and were removed from this analysis, leaving 43 memory blocks: see *Appendix C: Memory block responses* for details.)

Input gates: Most input gates (41/43) responded to the CS and about half also responded to the US (Figure 7-6A,B). Thus, the input gate allows the CS to enter the memory cells, and often also allows the US to enter, probably to inhibit memory cell activity once the US has arrived.

Forget gates: About two-thirds of forget gates responded to the US. Although their baseline activity during the intertrial interval may have varied, the gate was usually fully open between CS and US, and fully closed at US and sometimes the next time step (Figure 7-6C,D). When the forget gate was open, signal flowed recurrently into the memory cell, when it closed, the memory cell content depended only on its current input. Thus, the forget gate allows maintenance of the memory content during the trial, and triggers memory clearance (forgetting) when the US arrives.

Memory cells: About two-thirds of the memory cells showed a clear sustained response above intertrial baseline between CS and US. At US they returned to baseline or even show depression (Figure 7-6E).

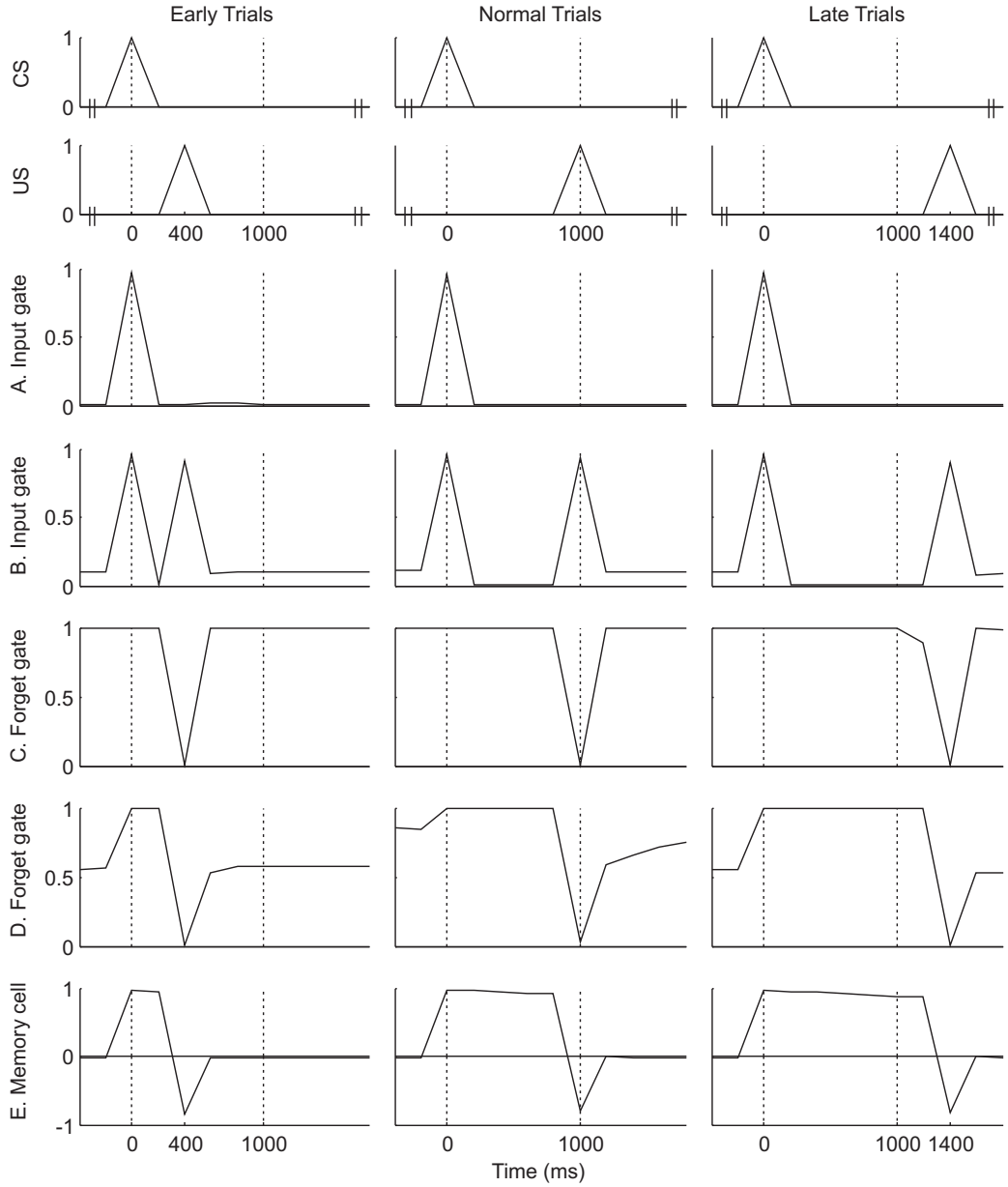


Figure 7-6: Typical LSTM responses shown using the same format as Figure 7-5. A & B. Typical input gate responses: A. an input gate responding only to the CS. B. an input gate responding to both CS and US. C & D. Typical forget gate responses: C. a forget gate always maximally open, except at US presentation when it fully closes. D. a forget gate with variable baseline activity during the intertrial interval (more than half open), but that is also maximally open only during the CS-US interval, and fully closes on US prior to returning to its baseline. E. Typical memory cell response.

Output gates: About two-thirds of the output gates fully closed on CS and then showed a slow ramp-like progressive opening during the memory cells' sustained period to reach their maximum by 800ms, i.e., one time step before the expected end of the CS-US interval. After the US presentation, they usually closed or returned to baseline. Intertrial baseline activation varied from fully closed to fully open for different output gates (Figure 7-7A,B).

Block outputs: 80% of memory-block outputs resembled the output gates, except that they turned off or showed depression at US. The block output was computed by multiplying the memory cell activity by the output gate. Thus, memory block outputs reached their maximal response at 800ms, thereby predicting the expected arrival of US in the next time step (Figure 7-7C). This was a robust representation of the temporal structure of the task.

Overall, we found that 26/30 successfully trained networks from the original pool had at least one memory block in which all 5 neuron classes (*input gates, forget gates, output gates, memory cells,* and *memory-block outputs*) displayed the typical responses described above. In the other 4 networks, 2 had a memory block with a linear build-up in the memory cells, and 2 did not maintain their US prediction past 800ms (see below).

7.3.5.b) *Time computation by the intrinsic dynamics of recurrent loops*

The above signals suggest that the CS closes the output gate, which is then slightly opened by the memory cell activity or its bias. This allows feedback to leak into the output gate via the recurrence external to the memory block (Figure 7-2), so that the output gate opens approximately quadratically during the CS-US interval, peaking by 800ms. The depression of memory cells at US presentation passes through the memory-block output and the external recurrence to shut down the output gate at the next time step. The essence of this mechanism is that the memory block output and the output gate are coupled in such a way that they act as an accumulator that estimates elapsed time since the CS. Estimation of the correct predicted CS-US

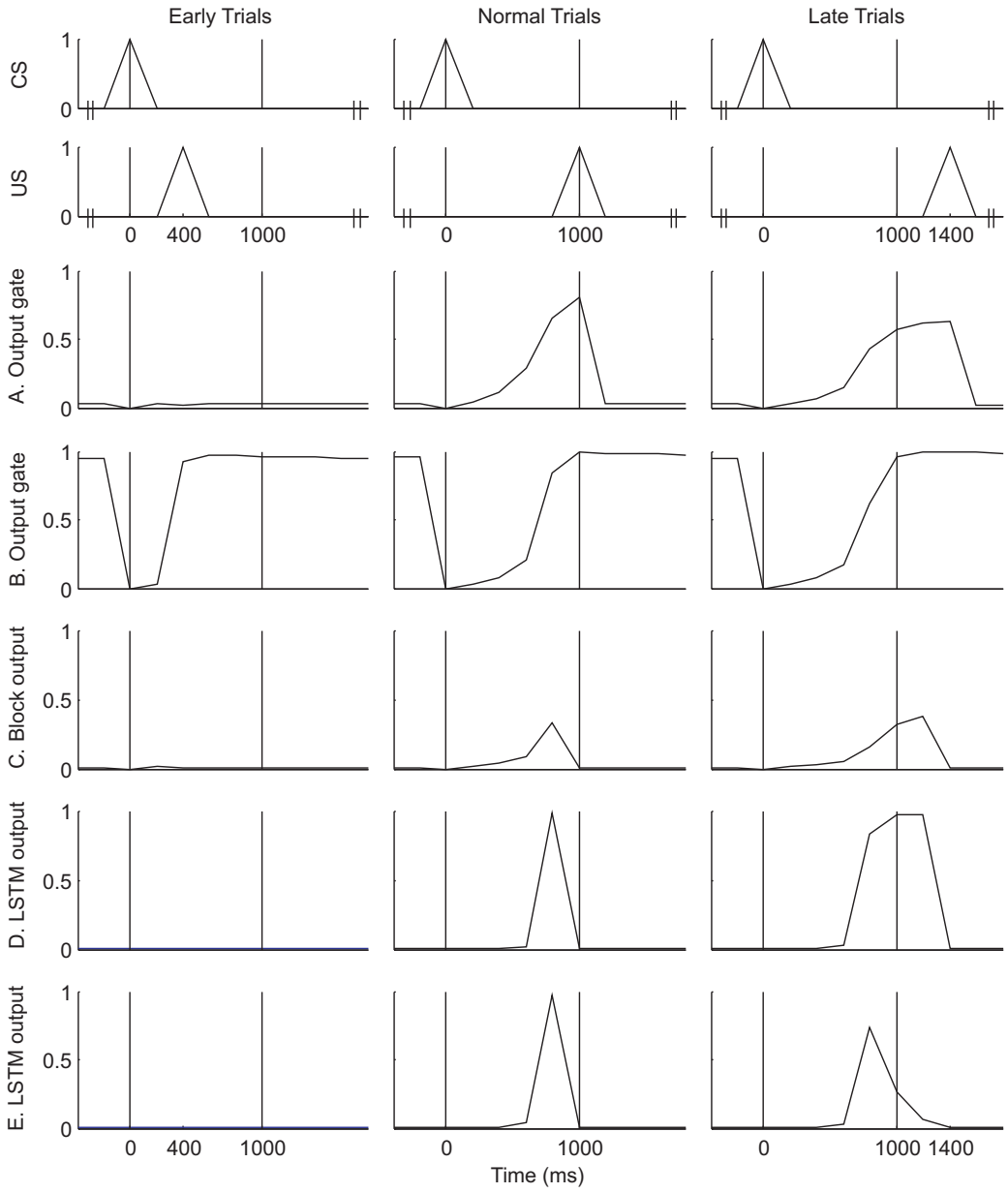


Figure 7-7: Typical LSTM responses shown using the same format as Figure 7-5. A & B. Typical output gate responses: A. an output gate with near fully closed baseline during the intertrial interval. B. an output gate with a near fully open baseline during the intertrial interval. Nevertheless, the signals from the two output gates during the CS-US interval are very similar. C. Typical memory block output response. D & E. LSTM output responses: D. the typical response. E. the response for a network that does not assume US will come once the expected delay is passed.

interval depends on learning the appropriate gain of this recurrent feedback accumulator circuit. In summary, the networks autonomously develop a single multi-

component integrator comprised of several different elements (memory cells, gates, and recurrent loops) that collectively has the appropriate integration constant for the given delay and also learn which external signal should start and stop the integration (timing) process.

7.3.6 Prediction in late probe trials

Most networks kept expecting the US in late probe trials until it appeared (Figure 7-7D). This is a reasonable assumption for the networks to make since throughout their learning, they always experienced a US after a CS, even if it was sometimes late (about 3 out of every 400 trials). Only 2 networks stopped responding after 800ms, indicating that they stopped expecting the US on late trials (Figure 7-7E).

To test whether this was an environmental effect, we trained another set of 30 successful networks but replaced the late probe trials of all but the last 3 test blocks by no-reward trials (CS only). These infrequent no-reward trials (0.62% of trials) were sufficient to drastically change the results. Only 3 networks kept expecting a US in late probe trials in the last 3 test blocks, 21 stopped responding after the expected time, and the remaining 6 showed more variable behavior, sometimes waiting briefly for a reward, and sometimes not. The model is therefore very sensitive to the observed data and predicts very reliable differences in expectations and behaviors in probe trials as a result of small but important differences in the history of observed trials in their past experience.

7.3.7 Memory cells with linear build-ups

A few memory cells showed a continuous linear build-up of activity between the CS and US (Supplemental Figure 7-11A,C). The four memory blocks containing them had a rather different signalling organization. Instead of responding to the CS and the US only, their input gate showed a build-up or sustained activity from CS to US inclusively (Supplemental Figure 7-11B,D). Their forget gates, output gates, and memory block outputs, in contrast, did not seem to be different from the other memory blocks (Figure 7-6C,D, Figure 7-7A,B, and Figure 7-7C).

The main difference between these few memory blocks and the majority resided in the input gate responses. They remained open during the delay period, providing a continuous flow of input that was linearly accumulated in the memory cells. The forget gate reset assures that the signal accumulates only during the CS-US interval, and not during the intertrial interval. Those memory blocks were probably not dependent on the external recurrent link between the block outputs and the output gates since both can be computed directly from the build-up level in the memory cell. This is similar to the kind of neural code suggested by Durstewitz (2004), i.e., a neural integrator in the form of a self-feedback loop. These linear build-ups could also look like a continuous implementation of the accumulators of scalar timing theory (Church, 2003). Why this is not the LSTM's preferred solution, how closely it resembles scalar timing theory, and how this connectionist representation could affect the model's behaviours, all remain to be investigated.

7.3.8 Intertrial timing and speed of new learning

The intertrial interval was fairly long compared to the CS-US interval in our standard simulations, and was variable in duration. Most networks showed little evidence of signals that gradually anticipated the arrival of the CS during the intertrial interval. Nevertheless, a few networks also seemed to keep track of time during the intertrial interval, as demonstrated by an increasing activation of memory cells or p signals (Supplemental Figure 7-12).

Finally, a few early or late trials occasionally occurred in succession in the randomized trial sequence of test blocks. This was sufficient for the LSTM of some overtrained networks to adapt to the extra delay, also affecting the average p and δ results on test trials (Figure 7-8). Such rapid adaption to change in delays has already been observed in real neurons (Komura et al., 2001; Brody et al., 2003). This rapid adaptation and the small number of probes also explain why in Figure 7-7D,E the averaged activity of the LSTM output at 800ms is different in late trials from normal trials. The probe trials present during training may play a role in the development of this rapid adaptation ability. More experiments will be needed to study this ability.

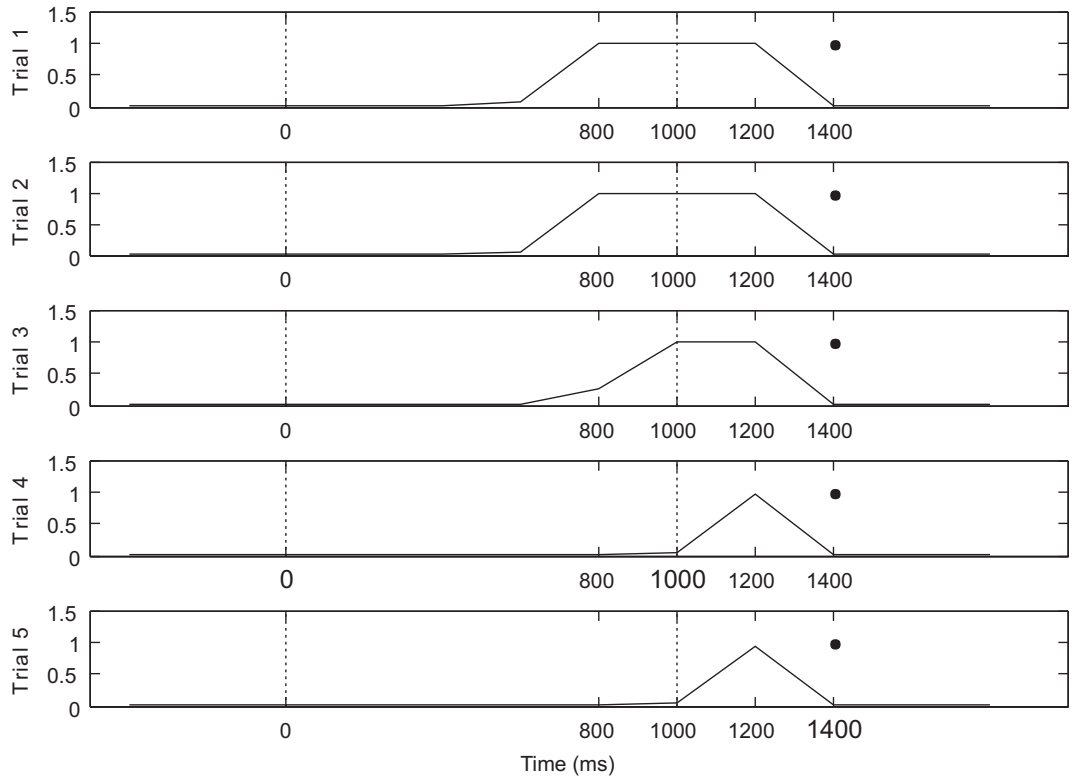


Figure 7-8: LSTM output signal for a network showing rapid acquisition of the late probe trial delay when encountering five such trials in a row in a test block. The large dot represents the late US presentation at 1400ms. As shown, on the first late trial, the network responded at 800ms (1 time step before the expected US timing). By the fourth trial, the network responded at 1200ms (1 time step before the late US). Vertical dashed lines indicate the standard CS-US training interval of 1000ms.

7.4 Discussion

The LSTMs in our simulations clearly and autonomously learned the environment dynamics as well as the necessary representation to discriminate trials from intertrial intervals and elapsed time within the trial interval between CS and US. They recognized the CS and started counting from the CS to the expected US arrival 1s later. The networks continued to expect US beyond 1s after CS onset if they experienced only occasional late-probe trials, or stopped expecting the US at 1s if they experienced only occasional no-reward trials. They also stopped expecting the US when the reward was given early. Many LSTM memory block input gates responded to the CS alone, while some others responded to both CS and US. In contrast, most forget gates responded selectively to the US. These events signaled the beginning and end of the 'trial' period, and segregated the continuous stream of

sensory events into alternating 'trial' and 'intertrial' segments. The predictive association between the CS and US was captured by the sustained activation of memory cells and forget gates during the CS-US interval. Elapsed time was encoded by the gradual ramp-like changes in memory block output gates throughout the CS-US interval and by the associated predictive activations of memory block outputs and LSTM outputs just prior to the expected arrival of the US. Strikingly, most of the successfully trained naïve LSTM networks, with no a priori knowledge about the structure of the task, found very similar solutions to both computational problems. Finally, the sustained activity from the memory cells $c_{i,j}$ and the LSTM predictive outputs y_{US} provided to TD all the necessary state information for it to learn a correct estimate of future rewards.

These modeling results suggest that time can be encoded in a context-dependent manner as ramp-like changes in single-neuron activity by a temporal integration mechanism within recurrent loops whose parameters are learned by general-purpose learning mechanisms and implemented by a small distributed population of cortical neurons, rather than within dedicated special-purpose timing circuits (Ivry & Schlerf, 2008).

7.4.1 Relation to previous work

Not all studies of DA phasic activity used trace conditioning. Some used delay conditioning and others used operant conditioning (Hollerman & Schultz, 1998). Nevertheless, the effects are generally considered similar across conditions (Daw et al., 2006). There have been five previous successful simulations of the Hollerman & Schultz (1998) data on DA activity under variable delay intervals (Brown et al., 1999; Suri & Schultz, 1999; Daw et al., 2003, 2006; Bertin et al., 2007; Ludvig et al., 2008). Four (Brown et al., 1999; Suri & Schultz, 1999; Bertin et al., 2007; Ludvig et al., 2008) used a form of delay-lines or a set of temporal basis functions to represent the stimulus over time, but one of those (Brown et al., 1999) was not TD-based.

Delay-lines are equivalent to having in memory an exact copy of the inputs for each of the $k < K$ preceding time steps, each of these copies moving from one memory buffer to the next at each time step, and with K properly chosen to span the

interstimulus interval, but not the intertrial interval. With delay-lines, TD simply has to find which stimulus at which of the *k* preceding time steps properly predicts the US. More complex temporal basis functions have a similar flavor, but the further away in time the input (higher *k*), the more diffuse the representation of the stimulus at *t-k.* Both representations usually require a built-in reset on reward to reproduce DA results on early trials, otherwise, TD *p* neurons continue to expect the US at the usual time, even after receiving it early. The reset can also be replaced by adding a similar temporal representation for the reward stimulus on the input side (Ludvig et al., 2008), but this would be closer to an expectation inhibition than to a reset or a shift of the time representation to the intertrial interval epoch. Current DA data do not allow us to validate or reject this possibility, but our model seems to learn a reset, consistent with most other TD models. The representation learned by the LSTM can be viewed as learning one of temporal basis functions and to associate it to the relevant stimulus, while previous models need to be given a whole set of fixed basis functions covering the necessary temporal space for every possible stimulus.

The successful simulation of the data that did not use delay-lines (Daw et al., 2003, 2006) is related to the present model in that it used an internal semi-Markov model of the task as a rich representation of the current state for TD. Those states resembled some of the present LSTM units, but there was no mechanism to learn state structure. The internal model of the task was inferred a priori and given to TD, while we started with a fully naïve system that learned its own model of the task. Also, while the semi-Markov model of Daw et al. (2003, 2006) assumes that the animal has a discrete probabilistic state space model and a separate clock mechanism, the present model embeds several distinct processes (credit assignment, world state estimation, and time representation) within the continuous time-varying state of an adaptive dynamical representation generated by the LSTM networks, without probabilistic inference or a separate clock/timing mechanism.

An innovative aspect of the present model is that it proposes an integrated approach to learning both the environment (LSTM) and reward expectations (TD). Furthermore, previous models of DA activity did not incorporate new concepts in

time interval modeling. For example, as an alternative to delay lines, Durstewitz (2004) proposed a progressive build-up of single-neuron activity or a progressive recruitment of neurons into the active population as neural codes for time intervals. The interval timing literature also suggests mechanisms for the representation of time based on general learning algorithms (Hopson, 2003; Dragoi et al., 2003). This latter approach is taken here.

7.4.2 Model's representation of the task

Our model presents an integrated adaptive framework that extends TD models of DA signals (Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Daw et al., 2003, 2006; Pan et al., 2005; Ludvig et al., 2008) with a model of the task (Doya, 2000; Daw et al., 2003, 2006) learned conjointly by a general-purpose LSTM neural network. Instead of special-purpose timing circuits such as unrealistically long delay lines, pacemakers or clocks, time could be represented by time-dependent increases or decreases in either the activity level of single neurons or by the number of active neurons (Durstewitz, 2004), or by less intuitive coding mechanisms such as transitions from one arbitrary network state to another over time (Karmarkar & Buonomano, 2007). Our model predicts that time is encoded by time-dependent ramp changes in single-neuron activity by temporal integration within feedback loops in recurrent network architectures. The LSTM generated ramp (Figure 7-7A) and sustained (Figure 7-6E) activity that closely resembles real cortical neurons in many tasks (Funahashi et al., 1989; Lucchetti & Bon, 2001; Leon & Shadlen, 2003; Lucchetti et al., 2005; Lebedev et al., 2008), including prefrontal neurons implicated in short-term working memory processes during delay periods (Romo et al., 1999; Brody et al., 2003). Our model predicts that this kind of cortical activity may appear even in very simple tasks such as trace conditioning. Exactly how time is represented in the brain is still controversial (Ivry & Schlerf, 2008), but the present results suggest some common computational principles between the LSTM and a possible distributed representation of time in the cerebral cortex (Lewis, 2002; Durstewitz, 2004) rather than in dedicated circuits.

The DA activity in early- and late-probe trials (Hollerman & Schultz, 1998) suggests two inferences about the brain's internal model of the task. First, once a reward has occurred, no further reward is expected until the next CS because the DA neural response does not decrease at the time of expected reward in early-probe trials. This can be inferred by the LSTM since it never encountered two consecutive US (i.e. consecutive rewards) while training. The arrival of US therefore switches the LSTM to intertrial mode. Second, the brain learns the task's time structure, since DA responses are weak when the US occurs at the expected time but are enhanced when it arrives early or late. The TD δ signal shows that pattern (Figure 7-5). We also found that the LSTM expected the reward until it finally occurred in late-probe trials (Figure 7-7D). This reflected the LSTM's training experience, in which a reward occurred in every trial. Networks trained with infrequent no-reward trials developed a different representation which typically stopped expecting the US if it did not arrive on time. Therefore, the model predicts very different cortical neural responses to probe trials depending on the training history of previously observed trials. This may also be closer to what animals experienced in the Hollerman & Schultz task (1998) in which animals had to indicate which of two stimuli was associated with reward. To study learning, stimuli were frequently replaced by novel ones, for which the animals did not know the right choice. As a result, they often experienced CSs without a US while learning. Nevertheless, whether LSTM networks kept waiting for a US or not had little effect on the TD-δ signal, possibly because LSTM can only predict the delay for normal trials, but not for the infrequent late trials. Training networks using no-US probe trials and keeping the early and late probe trials only for the last three test blocks, but not the others, also yield similar TD δ signals in early and late trials.

7.4.3 Contributions

This is the first completely naïve model of the environment and of DA activity in appetitive trace conditioning to reproduce the Hollerman and Schultz data (1998). Starting with no ad-hoc memory or temporal representation of the input, no prewired reset mechanisms and no a priori knowledge of task structure, it autonomously learns the environment dynamics while forming a representation of it. It received no

indication of which signal was the predictor (the CS), nor was it trained on discontinuous trials with a clear start and stop which would provide implicit cues about the temporal structure of salient events. In contrast, our model learned the task by experiencing the environment in continuous mode. Moreover, our simulation can run in both trace and delay modes. Previous models usually did not make such a distinction by directly coding the CS onset in the simulation (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998, 1999; Bertin et al., 2007) or the working memory process (Brown et al., 1999), except for (Ludvig et al., 2009). Our model also correctly reflects that learning is easier in delay conditioning than in trace conditioning, as it is usually for animals (Gallistel & Gibbon, 2000; Beylin et al., 2001).

Performance was assessed only by how well the LSTM circuits predicted the US. The successful TD δ simulations of DA phasic responses were not biased by any criterion that drove the network in that direction, but emerged while the linked LSTM-TD networks learned a general model of the task, supporting our hypothesis.

Beyond reproducing dopaminergic activity and cortical sustained activity and build-ups, our model makes a number of predictions. First, since we hypothesized that the dopaminergic signal depends on the representation learned in the cortex, one should be able to find cortical neurons with sustained activity (Figure 7-6E) as well as time-dependent build-ups of activity during the interstimulus interval (Figure 7-7A,B) even in simple trace conditioning with a constant CS-US interval. Second, depending on the training history, if the animal encountered mainly no-reward probe trials, then cortical ramping neurons should stop their activity after the usual delay even in absence of reward. In contrast, if the animal encountered mainly late probe trials while training, the activity of the ramping neurons should persist for a period of time after the usual CS-US interval when probed with a no-reward trial. Finally, the model also predicts that blocking the dopaminergic phasic signal should slow down learning in the frontal cortex.

7.4.4 LSTM success rate and current limitations

The major weakness of the model at the moment is the large amount of training it requires. In the best cases, some networks learned the task in about 4000 trials (but within 800 trials in average for all non-trace networks, see Table 7-I) while animals seems to learn timing imprecisely as fast as the conditioning itself within tens of trials (Balsam, Drew, & Yang, 2002) in trace conditioning. However, this slow acquisition of the task must be viewed in the context of the model itself. First, for the LSTM to be considered successful, we required that it not only learn the predictive association between the CS and US, but also to predict the timing of the US with a temporal precision of 1 time step. In contrast, in many animal studies, conditioning is usually considered successful as soon as the subjects learn the predictive CS-US association, while the precision of timing is only assessed later or not at all. Since TD learns a response to the CS within few tens of trials (see *Results: Dopamine neuron responses*), the basic CS-US association, but not the timing, could theoretically be learned rapidly in our model.

In appetitive interval timing literature, animals are often trained for about 1000 trials on a fixed delay of the order of 30s before being probed on long CS only probes to evaluate the temporal profile of their response to the stimulus (Buhusi & Meck, 2000). In trace variations, the CS can be there for as long as the gap (Figure 7-1) between the CS and the US (Buhusi & Meck, 2000; Thibaudeau et al., 2007). For example, the CS could be presented for 15s followed by a gap of 15s before the reward. The animal response curve with respect to elapsed time has a Gaussian shape and its width is usually proportional to the fixed interval used with a response peak (the Gaussian mean) at about the trained delay. That is, the timing precision is usually proportional to the timed duration. In the 1s interval presented here, Hollerman & Schultz (1998) DA data suggests the existence of an internal width in the order of our simulation resolution (200ms for a 1s delay) while Fiorillo et al. (2008) DA data suggests a much larger width (~1s for a 2s delay). Note that the LSTM is trained to learn the exact delay (high temporal precision), and that it tries to minimize the error on its prediction for each time step, not the temporal error of its prediction. This

means that for the LSTM, it is preferable to predict nothing, than to predict a US arrival that is too early or too late. Allowing LSTM to have a temporally smoother error calculation and softening the success criterion may be necessary for it to scale up well to longer delays and to be compared fairly with animal precision. It may also be necessary to reproduce the findings of the Fiorillo et al. (2008) study showing the relationship between dopaminergic response, trial interval length, and peak expectation under Weber's law.

Ljungberg et al. (1992) considered 30,000 repetitions a criterion for overtraining, which is similar to our simulation. Nevertheless, more than 50% of our networks failed to learn and retain the task for all 40,000 repetitions. We also tried a somewhat more relaxed success criterion of 80% successful trials within a block instead of 10 consecutive correct trials, but global success rates were similar. Furthermore, when learning tasks with memory components such as in delayed response tasks, animals are often first trained using very short inter-stimulus or inter-event delays. As they learn the association, the memory interval is introduced and gradually extended. These are the kind of tasks in which neural activities similar to the ones in the LSTM presented here are usually observed (Funahashi et al., 1989; Romo et al., 1999; Brody et al., 2003; Leon & Shadlen, 2003). In contrast, our networks began totally naïve and with the full CS-US delay of 1 second (i.e., 5 time steps in our simulation). Fiorillo and collaborators (2008) report dopaminergic activity showing some knowledge of timing when recording after about 600 trials of pretraining in a classical conditioning (non-trace, no-memory) set-up. This seems on a comparable scale with our non-trace networks (800 trials on average, Table 7-I).

Also, all of the networks were truly naïve at the beginning of training, with no prior experience of time intervals whatsoever nor any a priori knowledge either of the nature of the task or of the environment built into the networks. Learning in TD models with delay lines (or similar temporal representations) is a much easier problem, since each stimulus has a complete temporal representation to which TD only has to select the right one. Here, the LSTM has to build the temporal representation once it has found that the CS may be a predictor of the US. On the

other hand, a learned temporal representation could be re-used in many ways. For example, it could adapt to a different delay (Komura et al., 2001; Brody et al., 2003) or it could be associated to a new stimulus (Ivry & Schlerf, 2008) as shown in animals. This is certainly not something delay lines could reproduce, since they need a distinct temporal representation for each stimulus. In contrast, LSTM networks could technically transfer an acquired time representation to a new stimulus by connecting it to the appropriate memory block. Whether LSTM could transfer an acquired time representation to a new situation remains to be fully tested, but the speed at which some networks acquired new delays (a few consecutive late trials were sometimes sufficient for the networks to adapt to it, Figure 7-8), shows that once some timing mechanism has emerged, adapting it to a new situation (a new delay in this case) can be quite rapid.

The gap between the two stimuli without a sensory event (trace conditioning) in our simulation is also definitely a major challenge for the model. Delay conditioning, in which CS and US overlaps and turn off together (Figure 7-1, dotted x_{CS} line), is much easier than trace conditioning (Gallistel & Gibbon, 2000; Beylin et al., 2001). This difficulty comes from the partial observability of the task. During the gap, the environment provides no cues to the animal (or model) to differentiate CS-US trial intervals (when there is an imminent reward ahead) from intertrial intervals. While delay-line models provide an internal memory of all recent past events, our network must first learn to maintain the CS in memory in order to associate it to the US and to differentiate trials from intertrials. Human data also suggest that similar awareness is needed in trace conditioning, at least in eye-blink conditioning (Clark & Squire, 1998). For example, half of human subjects may not develop awareness of the CS-US relationship, and acquiring a trace-conditioned response is highly correlated to this awareness. In contrast, delay conditioning is acquired independently of the subject's awareness of the CS-US relationship. These results are very similar to ours where half of our networks failed to learn the trace task while all networks trained on delay conditioning succeeded (Table 7-I). The sustained presence of the CS in delay conditioning is undoubtedly one of the main reasons why our preliminary studies

suggest that the learning rate and success rate of our networks are substantially enhanced in delay conditioning rather than trace conditioning mode. Results show that our networks also learn significantly faster in non-trace conditions, learning the task nearly 10 times faster (20 training blocks instead of 233 on average, Table 7-I). This seems consistent with the data, even though the networks require many more examples.

Another important factor to consider is the very limited scope of the networks in our simulation. Through evolution, the central nervous system of higher organisms has undoubtedly become richly endowed with circuits that seek out causal relationships between events, especially sensory events or motor actions that lead to favourable outcomes for the organism. One such mechanism is episodic memory. The lack of episodic memory (e.g., hippocampus) in our model could also be a crucial factor contributing to the slow learning rate in our model. The acquisition of eye-blink trace conditioning without a hippocampus is very difficult, sometimes almost impossible even for 1s or 2s delays (Clark & Squire, 1998; Beylin et al., 2001). The difficulty for LSTM is that, until it learns to recognize and maintain the CS in working memory using sustained activity, it has little or no information about the past CS events in memory when the US arrives, beyond the eligibility trace, to form a link between them. Animals are probably helped by an episodic memory system such as the hippocampus, which, even if it does not store precise timing information, could provide a history of recent events to facilitate the establishment of an initial association between them. Surprisingly, data for appetitive trace conditioning on hippocampus-lesioned rats on 2s interval seems contradictory (Thibaudeau et al., 2007). More hippocampus lesion studies of appetitive trace conditioning, especially with longer delays, are certainly needed.

Nevertheless, our current model clearly cannot accommodate the rapid acquisition of timing found in animals (Balsam et al., 2002). It is possible that more timing-specific mechanisms in such structures as the cerebellum, could be involved in this fast learning, and may then be used to teach timing to the cortex. More work is certainly needed to better understand the role of each subsystem in classical and trace

conditioning as well as to learn which sub-system is necessary in which condition. Conditioning responses may also appear well before the cortex has learned a complete representation of the task. Faster time learning mechanisms or more complete brain models will be needed to account for time learning data in animals.

7.4.5 Mesocortical projections speed up learning

Our results demonstrate the utility of the computational models to assess possible roles of mesocortical DA projections to the frontal cortex. Extensive testing on a large range of learning rates (α_{LSTM}) and DA contributions (β) scales (*Methods: The model: Model of the mesocortical projection*) showed that the DA signals generally improve the learning performance (see *Appendix D: Mesocortical performance*). Even using near-optimal parameters for both model (with and without the mesocortical projection), the mesocortical DA projection model was nearly two times faster to reach a first successful block in the trace conditioning task than the other model. The mesocortical feedback also seems to help synchronize LSTM and TD learning, maintaining TD δ (and hence p) signals in a more reasonable range (see Supplemental Table 7-II).

But how does the mesocortical feedback help learning in the model presented here? The δ_t (DA) signal in recorded blocks is near 0 most of the time. Reward prediction changes when either US arrives, or when some variable in the LSTM network that TD found to be a reliable indicator of reward prediction indicates a possible change in reward expectation. δ neuron activity varies mostly on CS and US presentation, on which it can easily be 1, tripling the LSTM base learning rate (a timely increase of 200%) for its output prediction at those specific time steps when the LSTM inputs to TD seem to cause TD to make errors in reward prediction. On the other hand, the LSTM is likely to predict nothing most of the time (an easy solution to error minimization) and is therefore likely to have near zero weight changes except on US presentation. However, our results suggest that simply increasing the LSTM learning rate without using the δ_t signal cannot completely account for the extra performance that this signal provides. The first and easiest solution in this task for an error-minimization learner without a pre-existing temporal representation is to always

predict nothing. This reduces the error to the single and relatively rare US events. When TD δ_t feedback increases the LSTM learning rate parameter on those events, it acts as if it was increasing the error-value of those time steps relative to other time steps. It becomes 2 or 3 times more important to predict the US on time, than to falsely predict a US elsewhere (a bit earlier or later for example). This may be one way the δ_t signal helps the LSTM to learn. The δ_t signal does not provide information about the general goodness of the LSTM solution, but only when it leads TD to make reward prediction errors.

Our present hypothesis is that the mesocortical dopaminergic signal could mark events of importance in reward prediction and facilitates the learning in the cortex. Others have also used DA signals to learn better working memory representations (O'Reilly & Frank, 2006) or explicitly to learn a gating policy for working memory (Todd, Niv, & Cohen, 2009). Dopamine is a powerful modulator of prefrontal cortex plasticity (Otani et al., 2003) and activity (Montague et al., 2004). We thus suspect that its role in reality is of even greater importance. Since the δ signal responds to the CS early in learning (in the very first training block), it should help the LSTM network recognize CS as a predictor of US. According to the gating hypothesis (Braver & Cohen, 2000; Montague et al., 2004; Rougier et al., 2005), PFC sensitivity to inputs could be directly controlled by the DA phasic signal. The initial response of δ neurons to the CS could thus be used to control the LSTM input gate to memory, facilitating the association between the CS and US and helping prevent irrelevant distracters from entering working memory. Because of its gate and memory architecture, the LSTM network is particularly suitable for studying such hypotheses. The present work is only a first step in studying how the DA signal could facilitate learning in the cortex. Connecting the mesocortical projection to those LSTM neurons (such as the output gate in this case) may also explain some of the effects of dopaminergic drugs on processing of temporal information such as its effect on the subjective perception of the duration of time intervals (Buhusi & Meck, 2005).

Combining models of different brain structures also allows a whole new set of pre-experimental computational studies that could lead to a better understanding of

how learning occurs in the cerebral cortex and the basal ganglia. Computational models can be used to simulate complex experiments to answer questions that would seem untestable (Samejima et al., 2005; Daw & Doya, 2006). When learning a novel task, "Which region learns first?" and "Which region learns what?" are still wide open questions (Laubach, 2005). Having a model permitting us to pre-test some ideas could help in the design of animal experiments that would yield useful insights. By including adaptive models of multiple brain structures, we open the door to such exploration when dealing with the distributed plasticity of the brain in learning.

7.4.6 Adding actions, and the scalar property of time

Adding actions to TD is a fairly straight forward step that was not necessary in our simulations (Barto, 1995; Suri & Schultz, 1998, 1999; Ludvig et al., 2009). However, learning and reproducing data under acting conditions is much more complex. First, there is no guarantee that under an adaptive representation of the inputs, as in the model presented here, the system will converge (although see Rivest et al., 2005 for positive results using mesocortical feedback to control the adaptation rate of the input representation), even under a fixed policy. Second, one must determine which kind of policy learning the basal ganglia are more likely to implement (Samejima et al., 2005; Morris et al., 2006; Roesch et al., 2007). Finally, if you include the important question of when to respond, then a good temporal model needs to be developed in the first place.

This last item is in part the basis we tried to establish here, using information from dopaminergic and cortical neurons activities as actionless measures of timing in the brain. There are a number of relevant phenomena for which there are data that the model still has to fit, to validate it as an alternative mechanism in interval timing. A good model of timing needs the scalar property of time (Gallistel & Gibbon, 2000). Fixed-interval (FI) and peak-interval (PI) tasks are examples of tasks in which animals are rewarded for actions they make only after a certain delay set by the experimenter and where the animal's precision can be evaluated. Data from these experiments shows that the animals learn to self-pace the delay from task onset. Reproducing these data will require an action component in the model. There is also

some recent relevant dopaminergic data in classical conditioning that relates dopaminergic response, trial interval length, and peak expectation (Fiorillo et al., 2008) and that do not require actions.

Adding the capacity for action to the model would open up a wide range of other electrophysiological rate-codes and behavioural data, such as striatal neuron activity, operant conditioning, and the role of different components of the cortico-striato-cortical network in such functions as appetitive conditioning, motor skill acquisition and motor habit formation.

7.4.7 Conclusion

The LSTM learns to predict the timing and relationships of salient sensory events, while TD builds an estimate of future rewards and computes an error signal that replicates many of the task-related properties of mesencephalic DA neuron phasic activity. The proposed model seems a promising alternative timing model via a general learning mechanism. It does not require predefined delay-lines, it is not clock-based, it autonomously learns the environment dynamics, and it develops ramp-like activity that reflects temporal estimates as seen in various cortical areas as well as sustained working-memory activity. To our knowledge, this is the first time that such a recurrent neural network has been used in conjunction with TD to model the DA signal in time interval learning. The model also reflects the fact that trace conditioning is much harder than delay conditioning, that the learned representation of the task is highly dependent on the types of trials experienced during training and that removing the phasic dopaminergic signal to the frontal cortex could slow down its learning. It also allows the study of new problems such as cortico-BG learning interactions.

Appendix A

Supplemental Material 1: Java simulator

This program is a Java package containing the original simulation software used to perform the present simulations and can be downloaded from *ModelDB* (http://senselab.med.yale.edu/modeldb/) at http://senselab.med.yale.edu/ModelDB/ShowModel.asp?model=124329. No publication or

commercial derivatives can be made without the author's written consent. On the other hand, researchers who would like to have more options to try to make predictions and plan animal experiments are welcome to contact the author [FR]; model options, environmental or tasks options, as well as options to include other models can be discussed.

Supplemental Material 2: Simulations data

Three MS Access .mdb files containing all recorded data for the present experiment can be downloaded from *Université de Montréal* institutional digital repository *Papyrus* (https://papyrus.bib.umontreal.ca/jspui/) at http://hdl.handle.net/1866/3073. We highly recommend contacting the author [FR] before making any inferences about the data.

The *Master* database contains the last 20 training blocks with the test blocks before, in between, and after them (there was one test block every 10 training blocks). These are then followed by an extra training block and a final control block. The *PreTraining* database contains 2 control blocks run on 30 untrained networks. The *InitialTraining* database contains 2 training blocks run on 30 untrained networks.

Appendix B

Theoretical p signal

A theoretical *p* signal can be computed to evaluate the internal model the networks have learned. We make few assumptions about the networks' internal model of the task, compute a theoretical *p* signal for that model using TD, and compare it to the networks' *p* signal.

We will assume that the networks are unable to keep track of time during the intertrial interval between the arrival of a US and the presentation of the next CS and that they have perfect knowledge of the task within a trial. We modeled this using a 7-state Markov model: the model begins in intertrial state; when the CS appears, it shift into a CS state followed by 5 other states representing 200ms, 400ms, 600ms, 800ms and 100ms after CS onset; then the model automatically comes back in the intertrial state again. The *p* neurons learn an estimate of the sum of discounted future rewards, i.e.

$p \approx \sum_{k=0,\infty} \gamma^k r_{t+k+1}$, **Equation 7-10**

$\gamma = 0.98$. To compute the theoretical p signal implied by the Markov model, we simply reran the simulations (without test blocks) using the same TD component as in the paper, but replacing its input by the 7-states model described above. We also decreased the TD learning rate on each training block (1% smaller) to ensure convergence.

This theoretical p signal converged to an expected value of 1.4 during intertrial intervals, increased from 2.0 to 2.1 from the CS onset to usual US arrival (Figure 7-5, bottom row, middle panel, dashed line), and goes down to 1.2 on the US (1000ms) state. This means that in general, the networks could expect their sum of future rewards in a block, without more information, to be about 1.4. Once a trial has started, signalled by the arrival of CS, a more precise value that includes the imminent reward of value 1 can be computed, thus the sudden rise in the p estimate. The p value slowly increases as the expected reward gets closer in time due to discounting. Finally on US presentation, networks can expect less than in intertrial state since the US is never closely followed by a reward. Intertrial state estimate is less precise since all intertrial steps share the same estimate. In TD, the error is always pushed backward in time, since the only unpredictable event is the arrival of the CS, this is where most of the error signal (δ) appears, i.e. at CS presentation.

As shown on Figure 7-5, the theoretical p signal and the averaged networks' p signal are relatively close to each other.

Appendix C

Memory block responses

Some of the memory blocks in the 30 successful networks contained elements with unphysiological response patterns. In particular, the activity of some memory cells increased throughout the duration of entire training blocks, mainly because of sustained ramp increases during the intertrial intervals (Supplemental Figure 7-13). This increase probably would have continued indefinitely if the neurons' activity was not reset between each training block.

A deeper analysis revealed that memory blocks with such memory cell behaviours did not contribute directly to the network output at time *t*, even though they had some phasic responses to the task's signals. Of the 60 memory blocks (2 per network), 19 had one of the above patterns. No successful network had two memory blocks with such signals, indicating that it is probably impossible to learn the task with such signals only. Using a correlation measure between the memory block outputs and the LSTM outputs to evaluate their contribution to network function, 17 of the 19 memory blocks with unusual responses were removed from the LSTM representation analysis (memory blocks with absolute correlations < 0.03 where removed from the analysis), leaving a total of 43 memory blocks (out of 60) analyzed. None of the 41 memory blocks considered *normal* were rejected by the correlation test. The finding that a number of networks learned the task even though one of their memory blocks had unphysiological response patterns and did not contribute significantly to network output indicates that one memory block is probably sufficient to learn this classical conditioning task.

In an extra experiment post-analysis, we ran a small number of networks using a single memory block and a single memory cell. Although the success rate of these networks was much lower (~6%), their final solution matched the one found in *Results: LSTM Representation*. This also suggests that blocks removed from the analysis were not contributing to the final solution and that the final solution probably stands within a single memory block most of the time.

Appendix D

Mesocortical performance

In order to verify that the mesocortical speed-up was not simply due to the addition of two constant learning rates, we trained 50 networks for each pair of hyper-parameters values in the range $\{0.0, 0.0032, 0.01, 0.0316, 0.1, 0.3162, 1.0, 3.1623\}^2$ (powers of square root of 10). We than reduce the space of hyper-parameters by looking only at the number of successful networks (networks still successful on the last training block, Supplemental Figure 7-14). With $\beta = 0$, there was at best 34% of successful networks ($\alpha_{LSTM} = 1$). The averaged first successful epoch (using the 1000

limit for the 13 that fails to have a successful block) was 480 blocks. Higher success rate with the mesocortical model (from 50% to 60%) were obtained within the range $(\alpha_{LSTM}, \beta) \varepsilon \{\{0.0316, 0.1, 0.3162\} \times \{0.3162, 1.0\}\}$. In this range, the averaged first successful block of the fastest networks ($\alpha_{LSTM} = 0.3162$ and $\beta = 1.0$) was 260 (only 6 networks failed to have a successful block and had a value of 1000). Finally, we also looked whether networks could learn only using DA as learning rate (with no basic intrinsic learning rate, $\alpha_{LSTM} = 0.0$). $\beta = 0.3162$ and $\beta = 1.0$ gave the best success rates (30% and 34% respectively), the fastest networks ($\beta = 1.0$) had an averaged first successful block of 616 meaning that in this task, DA alone, without a basic intrinsic learning rate, seemed sufficient to learn. We performed a one-way ANOVA on these 3 sets of networks ($\alpha_{LSTM} = 1.0$ and $\beta = 0.0$, $\alpha_{LSTM} = 0.3162$ and $\beta = 1.0$, $\alpha_{LSTM} = 0.0$ and $\beta = 1.0$) and found a significant difference ($P < 1.0E\text{-}6$). A post-hoc Scheffe test showed that the mesocortical model using a basic learning was significantly faster than the two other groups.

Acknowledgements

We are grateful to Douglas Eck, Aaron Courville, Doina Precup, and many others for discussion in the development of the present work. This manuscript also profited from the comments of Pascal Fortier-Poisson and Elliot Ludvig, as well as from the anonymous reviewers.

F.R. was supported by doctoral studentships from the New Emerging Team Grant in Computational Neuroscience (CIHR) and from the Groupe de recherche sur le système nerveux central (FRSQ). Y.B and J.K. were supported by the CIHR New Emerging Team Grant in Computational Neuroscience and an infrastructure grant from the FRSQ.

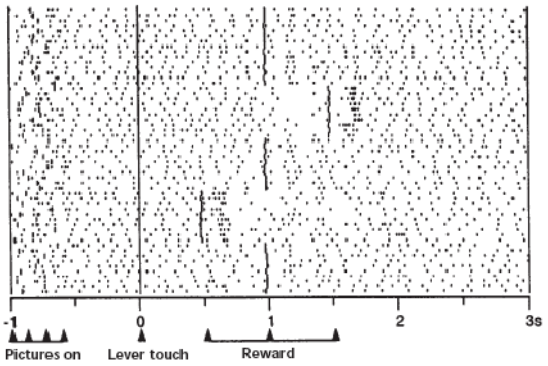
Supplemental Tables and Figures

Supplemental Table 7-II: Means and standard deviations of the minimum, median, and maximum values of the raw δ signal taken from networks over the recorded training blocks. One network was removed from this analysis in the last column because its extremes were a few orders of magnitude higher. Absolute values higher than one may be caused by unstable TD inputs. To converge, TD requires stable inputs. But while LSTM is learning, similar situations can lead to different LSTM activities and hence to different TD inputs. The full model seems to show relatively stable behaviours while the model without the error signal feedback control on LSTM learning rate seems more affected, as shown by its optimums high means and standard deviations.

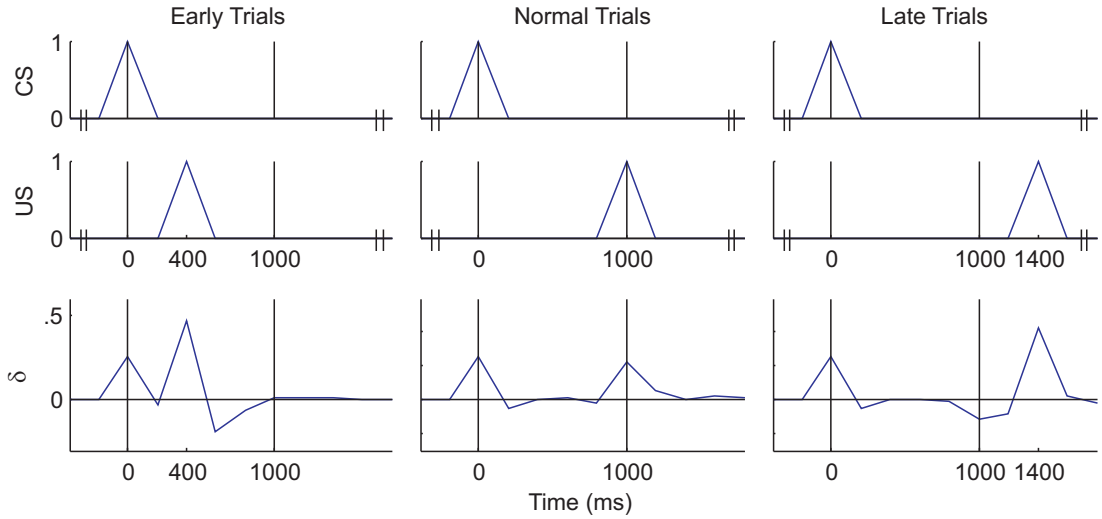
	Full model	Without mesocortical projection
Minimums	-2.6±1.9	-5.2±9.5
Medians	-0.02±0.02	-0.02±0.07
Maximums	2.3±1.3	3.4±5.0

Supplemental Table 7-III: Mean δ response properties for each network. Column 1 (CS acquisition): On normal trials, $\delta > 0.1$ at CS presentation. Column 2 (Early trials response) : On early trials, $\delta > 0.1$ at CS and US presentation and $-0.1 < \delta < 0.1$ at expected US presentation. Column 3 (Missing US response) : On late trials, $\delta < -0.1$ at expected US presentation. Column 4 (Late trials response): On late trials, $\delta > 0.1$ at CS and US presentation and $-\delta < -0.1$ at expected US presentation. Column 4 (US extinction): On normal trials, δ response at US presentation is < $\delta/3$ response at CS presentation.

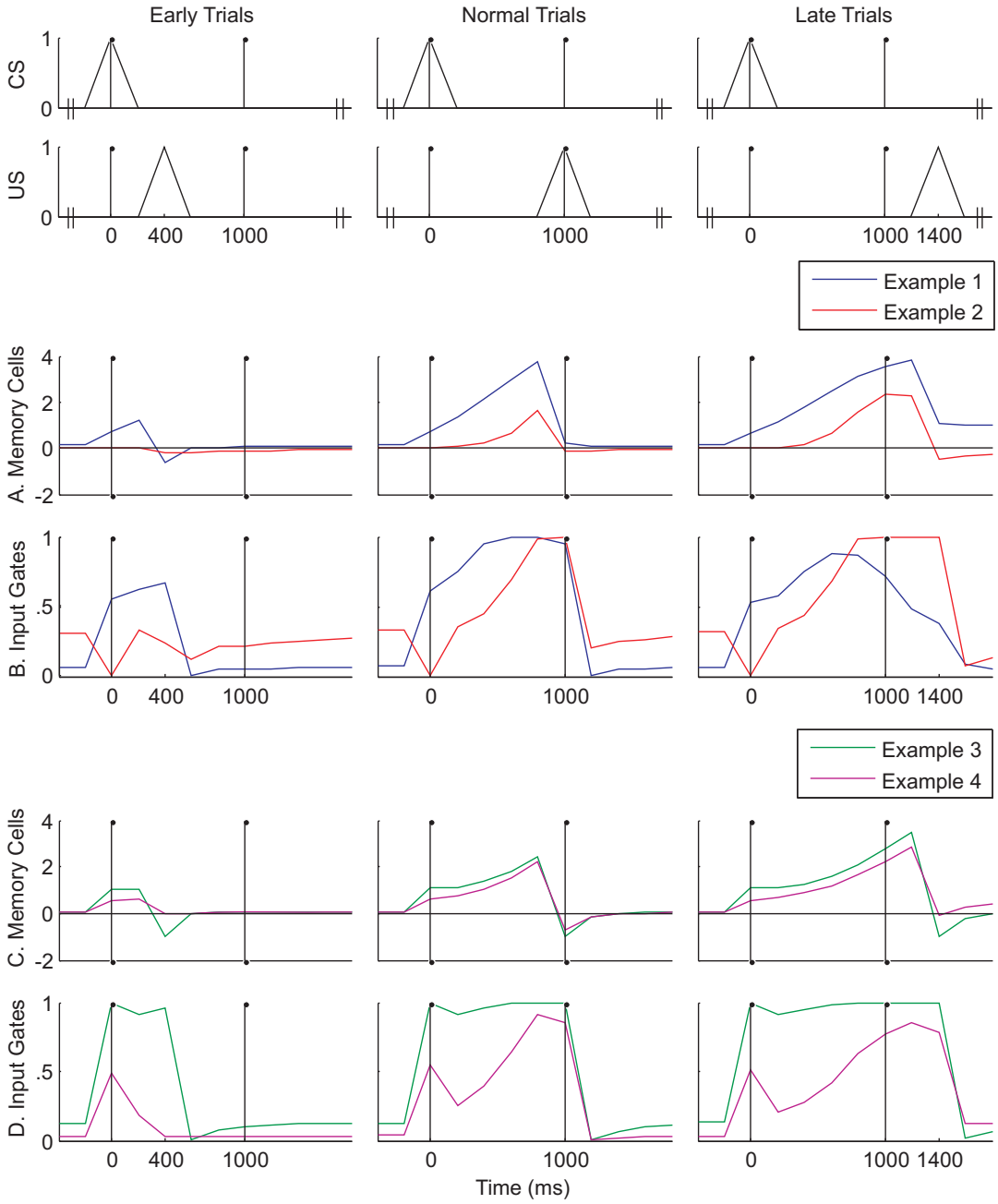
Network	1-CS	2-Early	3-Miss	4-Late	5-US
1	X	X	X	X	X
2	X	X	X	X	X
3	X	X	X	X	X
4	X	X	X	X	X
5	X	X	X	X	X
6	X	X	X	X	X
7	X	X	X	X	
8	X	X	X	X	
9	X	X	X	X	
10	X	X	X	X	
11	X	X	X	X	
12	X	X	X	X	
13	X	X	X	X	
14	X	X	X	X	
15	X	X	X	X	
16	X	X	X	X	
17	X	X			
18	X	X			
19	X	X			
20	X		X	X	X
21	X		X	X	X
22	X		X	X	X
23	X		X	X	
24	X		X	X	
25	X		X	X	
26	X				
27	X				
28					
29					
30					



Supplemental Figure 7-9: Response of dopamine neurons on probe trials with different delays. Top, middle and bottom sections show DA activity on normal (1s) trials. The second section shows DA activity on late trials, when the delay is longer than usual (1.5s). The fourth section shows DA activity on early trials, when the delay is shorter than usual (.5s). On late trials, there is a depression at the expected time of reward (1s) and a burst of activity when reward is finally received. On early trials, there is only a burst when reward is unexpectedly received. Reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience (Hollerman & Schultz, 1998, Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat.Neurosci.* 1:304-309), copyright (1998).



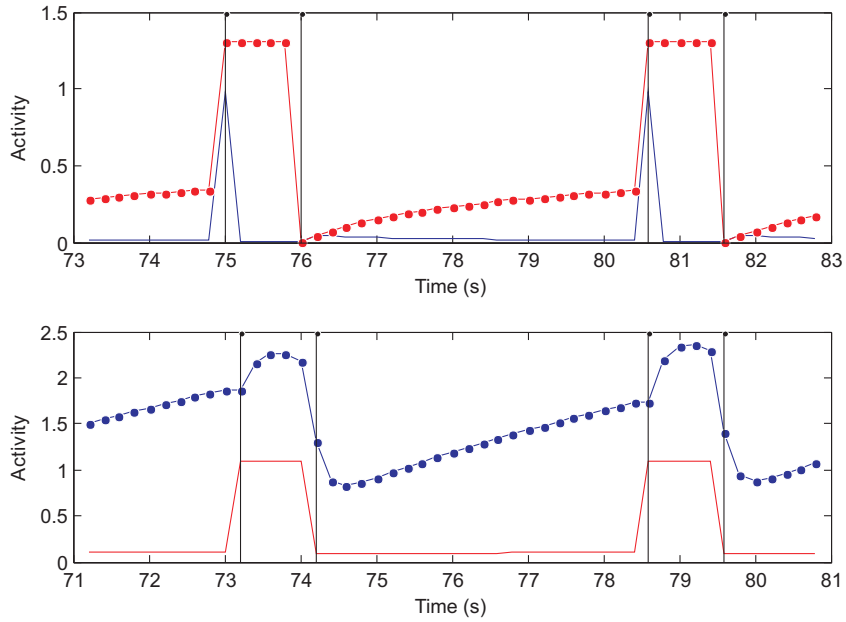
Supplemental Figure 7-10: Normalized δ signal from trained TD networks without mesocortical projection using the same format as Figure 5. Although learning was twice as slow as in the full model, the final normalized δ responses are nearly identical. Since the TD *p* neuron, and hence δ (DA), are unbounded, the amplitude of the individual signals vary greatly from network to network (see Supplemental Table 7-II). When looking at the averaged δ temporal profile, it is important not to let an individual network with a very high amplitude bias the population temporal profile. This is especially important in the model without the mesocortical feedback where some networks can have a DA signal an order of magnitude higher than others. Since all networks are trained under the same experimental conditions, and since the reward size is always 1, amplitude differences between networks are variables of little interest. Therefore, in this figure, the δ signal of each trained network was normalized globally, prior to analysis, such that their intertrial baseline activity (the median of the signal) over all recorded blocks is 0 and the maximum amplitude of their signal is 1. There is a single median and a single absolute maximum per network applied uniformly on all of its blocks. This procedure does not affect the temporal profile of individual networks.



Supplemental Figure 7-11: Four memory cells and input gate pairs from different networks (Examples 1-4). Memory cells (A & C) show linear build-up during the CS-US interval of a trial. Their corresponding input gates (B & D) show ramp or sustained responses during the CS-US interval. Vertical dashed lines indicate the standard CS-US training interval of 1000ms.

A.

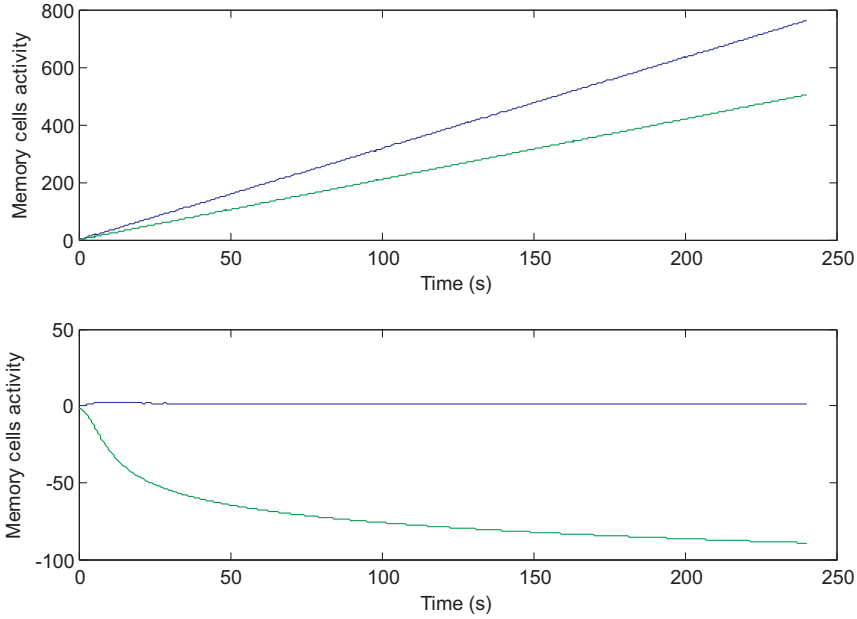
B.



Supplemental Figure 7-12: A. Example of a memory cell (red) that not only shows sustained activity during the conditioning trial, but that also shows build-up of activity during the intertrial interval. The other signal (blue) is the corresponding input gate of the memory cell. B. Example of a TD prediction neuron *p* (blue) that shows an increase in reward expectancy during the intertrial delay and that adjusts its prediction on CS presentation. The other signal (red) is a corresponding memory cell which shows normal sustained activity with no build-up during the intertrial interval of the conditioning trial. Vertical dashed lines indicate the standard CS-US training interval of 1s.

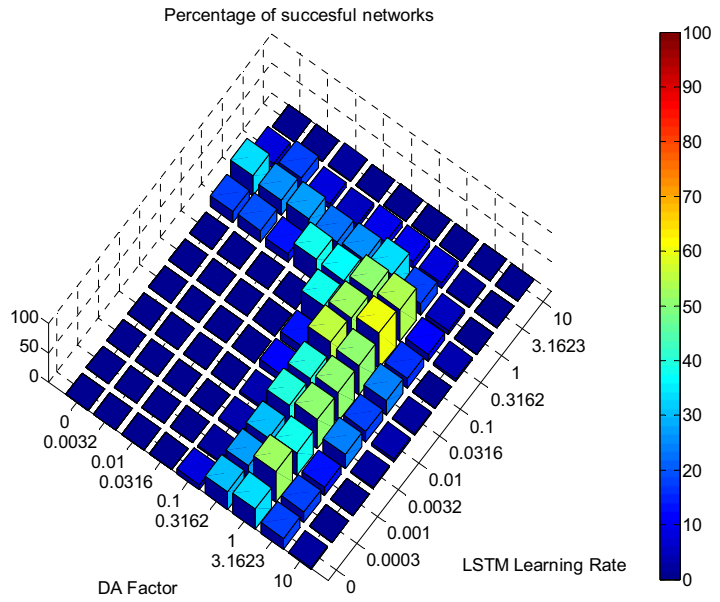
A.

B.

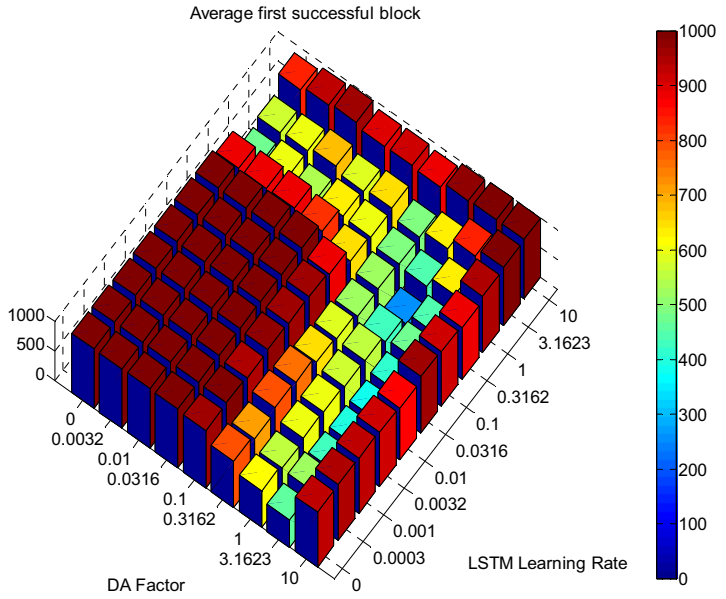


Supplemental Figure 7-13: A. Example of a memory block with two memory cells that increment continuously throughout the training block. B Example of a memory block that has one cell that seems to stabilize (in blue), and one cell that while possibly bounded, nevertheless decreases significantly during the training block.

A.



B.



Supplemental Figure 7-14: Results when varying the LSTM fixed learning rate and the DA factor used to modify the LSTM learning rate on-line (mesocortical projection, see *Methods : The Model : Model of the mesocortical projection*). A. Percentage of successful networks (with last training block successful). B. Average first successful block (using 1000 for networks that had no successful blocks).

Chapitre 8. Conditioning and Time Representation in the Long Short-term Memory Networks

This paper (Rivest et al., 2010b) by François Rivest, John Francis Kalaska, & Yoshua Bengio is near submission. Most of this paper is F.R.'s work and ideas. J.K. provided extensive direction, feedback, ideas, and comments throughout this work. Y.B. contributed as advisor, providing feedback and suggestions at all stages of the work.

The main objective of this paper is to evaluate ability of the model developed in the preceding chapter (*Chapitre 7*) (Rivest et al., 2010a) to reproduce and explain relevant data, as well as its ability to make interesting predictions and to identify some of its major limitations. This is done through a set of useful experiments. These experiments compare the time representation learned by the model under different conditioning paradigms such as classical and trace conditioning. It attempts to test whether the dopaminergic phasic signal should be similar under these different paradigms and it also makes predictions about neural activity when testing under a given paradigm (say, trace conditioning) is preceded by training under a different paradigm (such as delay conditioning). Finally, it looks at how the time representation evolves with respect to the timed interval length in order to evaluate the potential of the model as an interval timing model. The results are then used to discuss the model's validity, its explanatory power, as well as its current limitations.

Abstract

Dopaminergic models based on the temporal-difference algorithm (TD) usually do not differentiate trace from delay conditioning. Instead, they use a complete time representation from conditioned stimulus onset. Recently, a new dopaminergic model was proposed in which timing is learned within a modeled patch of cortex. In this model, time learning takes place in a long short-term memory artificial neural network. The objective of this paper is to evaluate the model's ability to reproduce and explain relevant data, as well as its ability to make interesting predictions and to isolate elements that can be improved. The model can differentiate

trace and delay conditioning at the cortical and dopaminergic level. Therefore, we evaluate how trace versus delay conditioning affect the time representation learned by the model. We found that compared to trace conditioning, the networks use a strikingly different time representation in delay conditioning, using the continuous stream of inputs provided by the conditioned stimulus. We tested whether the model predicts similar or distinct DA responses to probes of unexpected interstimulus intervals between delay and trace conditioning. As often assumed by modelers, the model predicts no important difference in DA responses between those two conditions. We use the model to generate predictions about animals' expectations and dopaminergic responses when trained on one conditioning paradigm and tested on the other. The model predicts that in trace conditioning, timing should start with the conditioned stimulus offset as opposed to its onset. In classical conditioning, it predicts that if the conditioned stimulus does not disappear after the reward, the animal may expect a second reward. These two predictions appeared to have relevant precedents in the literature. Finally, we studied how time is represented with respect to the interval length to be timed. We found that the build-up of activity of many neurons adapted to new delays by adjustment to the rate of integration. Some important limitations of the model were also revealed by these experiments.

8.1 Introduction

The ability to learn the temporal dynamics of the environment can be extremely useful. For example, a hunter that can learn its prey's behavioral dynamics, such as its escape habits, is more likely to catch it. More generally, temporal associations between events in the environment can help to establish causal relations. The temporal-difference (TD) learning algorithm (Sutton, 1988) has been widely used to model the subcortical dopaminergic (DA) system that processes reward-related information (Montague et al., 1996; Schultz et al., 1997). However, little is known about how an *'internal model'* of the environment is developed in the brain, especially for the timing of events on time scales on the order of seconds (Buhusi & Meck, 2005; Ivry & Schlerf, 2008).

In the interval timing literature, most models are abstract. The most common model is scalar expectancy theory (SET) that assumes a central clock, an accumulator and a memory (Gallistel & Gibbon, 2000). The accumulator is initiated at a *start* signal and on the arrival of a *stop* signal the accumulator value is saved into memory. While this allows learning of the timing between two events, the start and stop signals must be known *a priori*. A second model is the learning to time theory (LET). This model is based on the idea of a diffusion of the start stimulus through a sequence of reservoirs (Machado, 1997). This could be considered a diffused form of delay lines. A set of adaptive weights is placed between each action node and the reservoir's values. The proper actions are then learned by modifying the weights based on the reservoir's values and the reward. The *striatal beat frequency* (SBF) model is closer to the neurophysiology (Matell & Meck, 2004). In this model, the delay lines are replaced by a population of oscillators with distinct fixed frequencies. Using the appropriate weighted sum of the oscillators' outputs, the system can recognized the specific pattern of activity associated to a particular time interval. In all these models, timing is learned by properly using a given time representation.

There are a few more adaptive neural representations of learning to time. The first one uses a small oscillator made from two interacting units (Dragoi et al., 2003) whose frequency adapts to the interval to time. A second similar model uses a pair of leaky integrators whose integration and leak rates can adapt to fit experienced delays (Reutimann et al., 2004). These models are very limited as to what they can learn, but one useful property is that they do not have to be instantiated with a preset array of time-scales, but rather they adapt to the environmental dynamics. However, they also must be provided *a priori* with pre-defined *start* and *stop* signals.

In the TD-DA modeling literature, is it common to use delay-lines to represent events at previous time steps (Montague et al., 1996; Schultz et al., 1997; Pan et al., 2005). This is similar to representing each time step by a different state representing distinct elapsed time since last stimulus onset or offset. More advanced timing models use diffused delay lines similar in some respect to using the LET state flow as input to the TD model (Suri & Schultz, 1999; Ludvig et al., 2008). In those models, the

temporal representation is not learned. Instead, a representation of the full distribution of temporal intervals is provided *a priori* to TD for the model to match the data. Other models have used a semi-Markov state representation of the environment (Daw et al., 2006). Once again, a model of the environment dynamics was provided to the TD model, so it was not learned.

Recently, a new connectionist model of working memory, based on the long short-term memory network (LSTM, Gers et al., 2002), was developed to learn the environment dynamics and to provide an internal model of the world to a TD model of the dopaminergic system (Rivest et al., 2010a). This was the first model that could autonomously learn the task, without a built-in set of temporal representations provided by the experimenter. In the trace conditioning task used, the LSTM, trained to predict its next input, developed an interesting representation of time and the activity of units in the model showed similarities with real cortical neurons that have either sustained or ramp-like changing activity in memory-based fixed-delay tasks (Funahashi et al., 1989; Lucchetti & Bon, 2001; Romo et al., 1999; Brody et al., 2003; Leon & Shadlen, 2003; Lucchetti et al., 2005; Lebedev et al., 2008).

In this study, we explored more deeply how time is represented in LSTM networks and what it can tell us about learning under different conditioning paradigms. This is done through a set of scientific questions. In particular, does it represent time the same way in trace and delay conditioning? Dopaminergic neuron activity is often assumed to be similar under these two distinct paradigms (Daw et al., 2006): does the combined LSTM-TD network predict a difference? What could the model reveal about the trace versus delay conditioning paradigms? How does the LSTM time representation change depending on the interstimulus interval length? Are there other parallels between the model unit activity and behavioral and neurophysiologic data in the timing literature? What are the LSTM limitations as a neural model of timing in the brain? Could it make a good model of timing with working memory and of dopamine phasic activity?

The first experiment compares the LSTM time representation under different conditioning procedures. The second experiment evaluates whether the model

predicts similar or different DA responses to unexpectedly shorter or longer interstimulus intervals for these procedures. The third experiment uses the LSTM-TD networks to predict how cortical and dopamine neurons of animals trained on one form of conditioning would respond when tested on a different one. The last experiment looks at how LSTM time representation changes with respect to time interval length.

8.2 The Model

The model is a rate-code model made of two interconnected artificial neural networks. The first network is a long short-term memory (LSTM) network (Gers et al., 2002) representing the cortex and trained to predict its next input, i.e. to learn the task dynamics. The second network is a very simple temporal-difference (TD) learning algorithm (Sutton & Barto, 1998) trained to estimate the sum of future rewards. The LSTM network learns the task dynamics and provides states and temporal representation to TD as they develop. TD uses these inputs as well as direct stimuli inputs to predict the sum of future reward. The error in reward predictions is the correlate of the phasic dopaminergic signal (Montague et al., 1996; Schultz et al., 1997). The combined LSTM-TD model used here is the same as in (Rivest et al., 2010a).

The LSTM network (Figure 8-1) has two inputs ($x_{CS,t}$ and $x_{US,t}$), representing the observable conditioned (CS) and unconditioned (US, reward) stimuli at time t, and one sigmoid output ($y_{US,t}$) in the range [0, 1], representing the LSTM prediction about the US for the next time step. There is no corresponding predictive CS output since the CS is not predictable in the task used here. Between the input and output layers lie a number of memory blocks with outputs $y_1, \ldots, y_M$, where M is the number of memory blocks. Each memory block i (Figure 8-2) contains one linear and recurrent memory cell c_i and three multiplicative gates (input gate $y_{in,i}$, forget gate $y_{fgt,i}$, and output gate $y_{out,i}$) that control the flow of information around the memory cell. The output of each gate is a weighted sum of the memory block inputs

$$\boldsymbol{x_t} = [b, x_{CS,t}, x_{US,t}, y_{1,t-1}, \ldots, y_{M,t-1}] \quad \textbf{Equation 8-1}$$

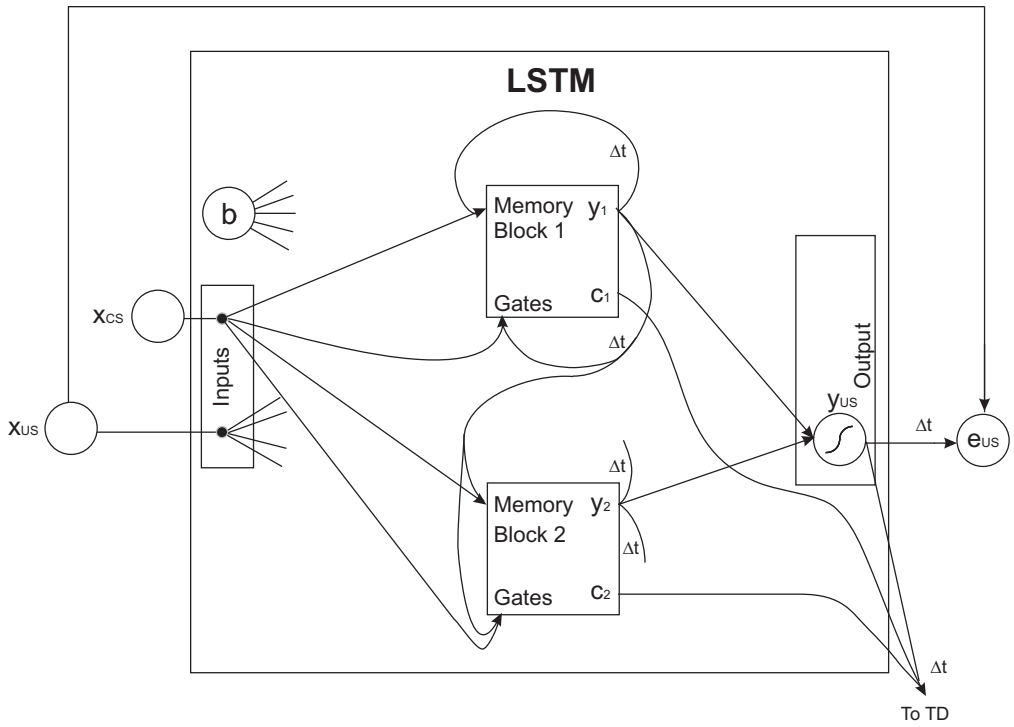


Figure 8-1: LSTM network.

and the memory cell activation c_i squashed by a sigmoid in the range [0, 1] (*asig()*):

$y_{in,i,t} = asig(\boldsymbol{w}_{in,i,t} \cdot [\boldsymbol{x}_t \; c_{i,t-1}])$, **Equation 8-2**

$y_{fgt,i,t} = asig(\boldsymbol{w}_{fgt,i,t} \cdot [\boldsymbol{x}_t \; c_{i,t-1}])$, and **Equation 8-3**

$y_{out,i,t} = asig(\boldsymbol{w}_{out,i,t} \cdot [\boldsymbol{x}_t \; c_{i,t}])$ **Equation 8-4**

where $\boldsymbol{w}_{g,i,t}$ is the weight vector for gate g (*in*, *fgt*, or *out*) of memory block i and *[]* is the concatenation operator. The activity of a memory cell is given by the weighted sum of its memory block inputs squashed through a sigmoid in range [-1, 1] (*sig()*) and multiplied by its input gate plus its previous value multiplied by its forget gate:

$c_{i,t} = y_{in,i,t} \; sig(\boldsymbol{w}_{i,t} \cdot \boldsymbol{x}_t) + y_{fgt,i,t} \; c_{i,t-1}$ **Equation 8-5**

where $\boldsymbol{w}_{i,t}$ is the input weight vector of memory block i. The result is then multiplied by the input gate activation and squashed in range [0, 1] again to generate the memory block output

$y_{i,t} = y_{out,i,t} \; sig(c_{i,t})$. **Equation 8-6**

Note that the memory cells c_i also feed into their gates (Equation 8-2, Equation 8-3, and Equation 8-4). Within the LSTM, memory blocks are fully recurrent to each

other. There is also a bias unit ($b = 1$) feeding the memory blocks and the output layer. The network output is given by the squashed weighted sum

$$y_{US,t} = asig(\boldsymbol{w}_{US,t} \cdot [b, x_{CS,t}, x_{US,t}, y_{1,t}, \ldots, y_{M,t}]), \quad \textbf{Equation 8-7}$$

where $\boldsymbol{w}_{US,t}$ is the weight vector for the LSTM output. All LSTM weights are initialized with small random values in range [-0.1, 0.1].

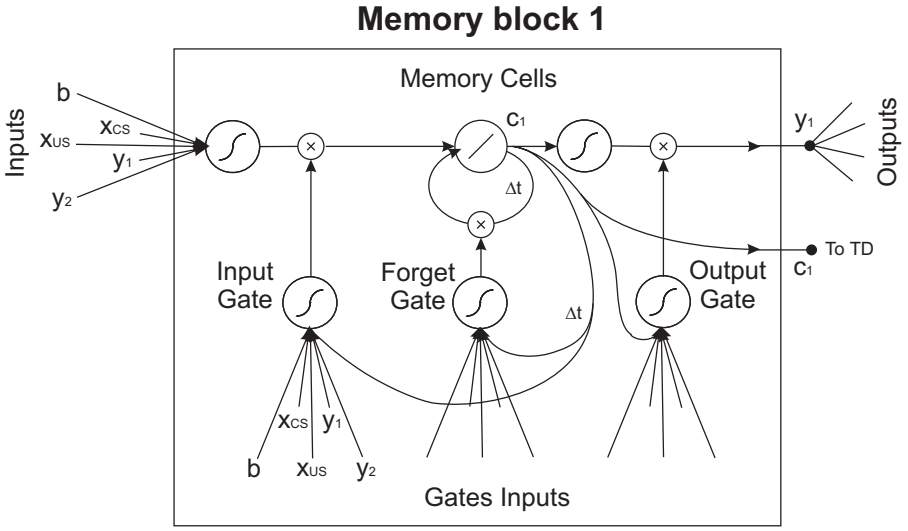


Figure 8-2: LSTM memory block.

LSTM are trained online using a backpropagation learning algorithm to minimize the next input prediction squared error

$$(y_{US,t} - x_{US,t+1})^2. \quad \textbf{Equation 8-8}$$

A useful property of the LSTM network is that learning computation time and memory requirements are independent of the training sequence length. At each time step, the gradient with respect to a memory block weight can be computed without any information from the other blocks except their output and without any information from previous time steps except the immediately preceding one. This makes LSTM computationally simpler than standard backpropagation algorithms. Finally, the LSTM learning rate ($\alpha_{LSTM,t}$) in this model is modulated additively by the TD error signal used to represent dopamine modulation of cortical plasticity (Otani et al., 2003)

$$\alpha_{LSTM,t} = \alpha_{LSTM} + \beta\,|\delta_t|. \quad \textbf{Equation 8-9}$$

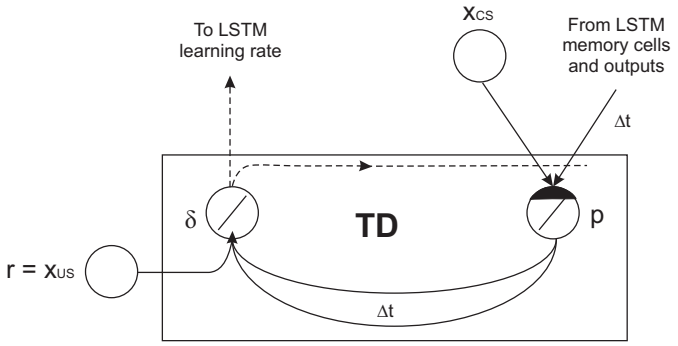


Figure 8-3: TD(λ) model.

The TD(λ) model (Figure 8-3) is made of a single prediction node (p_t) that computes a weighted ($\boldsymbol{w_p}$) sum of TD inputs

$p_t = w_{\boldsymbol{p},t} \cdot [x_{CS,t}, y_{US,t-1}, c_{1,t-1}, \ldots, c_{M,t-1}])$, **Equation 8-10**

which are the CS ($x_{CS,t}$) and the LSTM memory cells' activation bounded in the range [0, 1] and network outputs at the previous time step ($y_{US,t-1}$). The difference between the two consecutive predictions and received reward (r_t) is then used to compute the TD error

$\delta_t = r_t - \gamma p_t - p_{t-1}$ **Equation 8-11**

where $0 < \gamma < 1$ is the discount factor. Weights are updated using the rule

$\Delta \boldsymbol{w}_{p,t} = \alpha_{TD} \delta_t \boldsymbol{u}_{t-1}$, where **Equation 8-12**

$u_{j,t} = \lambda u_{j,t-1} + x_{j,t}$ **Equation 8-13**

for TD j^{th} inputs $x_{j,t}$. and where $0 < \lambda < 1$ is the eligibility trace decay factor. TD traces are also bounded in the range [0, 1]. $\boldsymbol{w}_p$ is initialized with small positive values (0.1). For a list of differences between the current model and (Rivest et al., 2010a), see *Appendix A: Changes from (Rivest et al., 2010a)*.

8.3 Tasks

Three variations of classical conditioning were used: *trace conditioning*, *delay conditioning*, and *extended conditioning*. In these tasks, the interstimulus interval (ISI), i.e. the time interval between the conditioned stimulus (CS) and the unconditioned stimulus (US, the reward) onsets, is always fixed. The difference

between the three tasks resides in the CS duration. The signals for all three tasks are depicted in Figure 8-4.

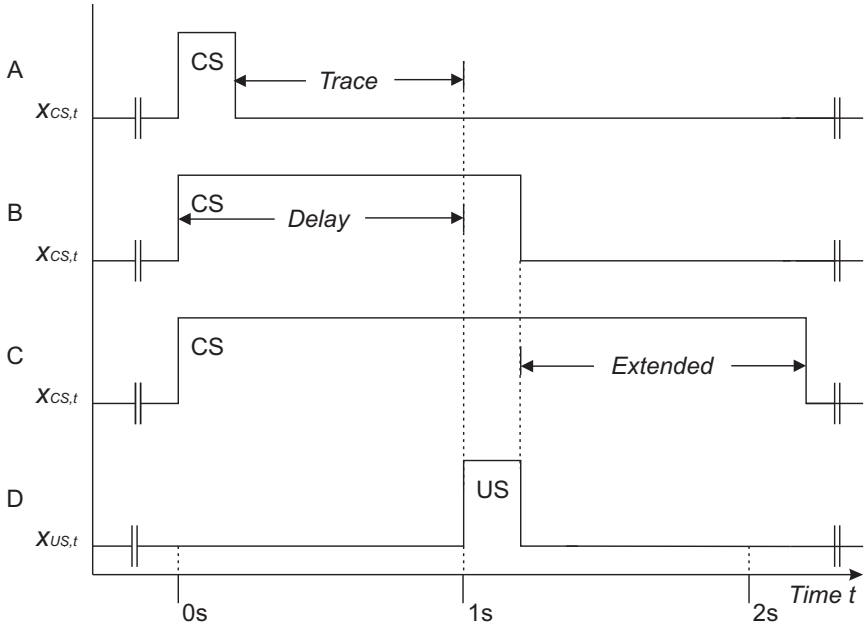


Figure 8-4: Three fixed-delay conditioning variations. In *trace conditioning* (A), there is a gap ('trace') between the CS offset and US onset. In *delay conditioning* (B), the CS and US offsets coincide in time. In *extended conditioning* (C), the CS offset occurs after the US offset. In the paradigms used in this study, the US signal (D) was the same for all three types of conditioning, using the same fixed time interval of 1s between CS onset and US onset for all standard training trials.

8.4 Experiment 1: Time representation with respect to conditioning paradigm

Recently, the LSTM network was proposed as a model of how a working memory in the cortex could learn timing between the CS and the US under trace conditioning (Rivest et al., 2010a). The representation found in the model was similar to neural activities found in the cortex during delayed response tasks or other similar tasks (Funahashi et al., 1989; Romo et al., 1999; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Lucchetti et al., 2005; Lebedev et al., 2008). We first attempt to reproduce the simulations at a higher simulation temporal resolution (10Hz instead of 5Hz), with a single memory block (instead of two) and a single memory cell (instead of two) and no probe trials of various duration during training. Having multiple memory blocks in the previous simulations seemed to complicate the

analysis of the representation in the networks. We would like to know if with a single memory block, all the networks will find the same representation. Networks with a single memory block should be easier to analyse. We then expand the study to see whether the LSTM would use the same representation under delay and extended delay conditioning. While in trace conditioning the network keeps in memory whether it is in a trial or intertrial (Rivest et al., 2010a), in delay conditioning, the network should not need to specifically code for the trial since the presence or absence of stimulus provides this information directly. In extended conditioning, the CS presentation goes beyond the US. To differentiate the pre-US period (for which there is a US to predict) from the post-US period (for which there is no US to predict), we expect the network to code whether the US is passed or not.

8.4.1 Methods

Randomly initialized networks with a single memory block were trained ($M = 1$, $\alpha_{LSTM} = 0.01$, $\beta = 1.0$, $\alpha_{TD} = 0.1$, $\lambda = 0.9$, $\gamma = 0.98$) on 4-minute (simulated time, 2400 time steps at 10Hz) *training block*s of alternating trials and intertrials intervals (i.e., a sequence of alternating CS and US signals, Figure 8-5). Three groups of networks were trained, one for each conditioning paradigm: trace, delay, and extended delay conditioning (Figure 8-4). Each group was trained using an interstimulus onset interval of 1s between the CS and US onsets. A trial terminates when both the CS and US are off, marking the beginning of an intertrial interval. Intertrial intervals were random delays between 4s and 6s without any stimuli.

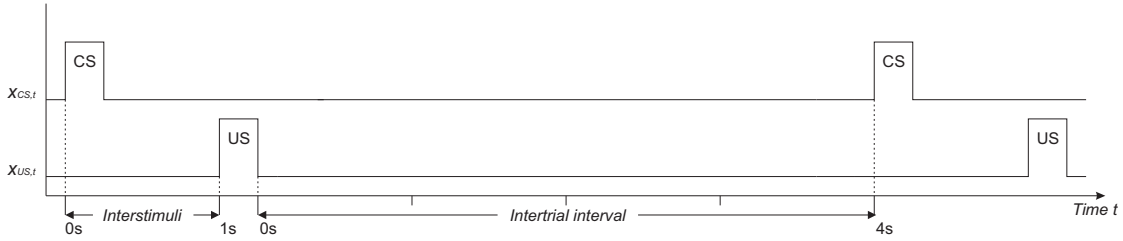


Figure 8-5: Zoom on a trace conditioning training block.

A training block was considered successful when the LSTM network was able to properly predict its next input (absolute error $|y_{US,t} - x_{US,t+1}| < 0.5$) for 600 consecutive time steps (i.e. for about 10 consecutive trials or 1 minute of simulated

time). A network was considered successful when 16 of the last 20 training blocks were successful (80%). Training was limited to 5000 blocks in trace conditioning and to 500 blocks in the two other paradigms. Trace conditioning was shown to be much harder for LSTM networks (Rivest et al., 2010a). We continued to train new random networks until we had 30 successful networks in each task, according to the above criteria, which we then used for analysis.

8.4.2 Results

8.4.2.a) *Time representation in trace conditioning*

Most LSTM networks (27/30) trained under trace conditioning paradigm developed a representation similar to the one in Figure 8-6. This representation is significantly different from the one found in (Rivest et al., 2010a) for the same model under similar (but slightly different) conditions (see *Appendix A: Changes from (Rivest et al., 2010a)*). In the original experiments, the LSTM networks used the input gate on CS to generate a sustained activity in the memory cell during the interstimulus interval, and the forget gate (or input gate) to shut it down on US (Supplemental Figure 8-17A-B,D). The elapsed time build-up occurred in the memory block output and output gate loop, at the output of the memory block (Supplemental Figure 8-17C,E). In contrast, in the present simulation, the memory cell plays both roles, marking intertrial versus interstimulus interval as well as pacing time during the interstimulus interval (Figure 8-6D). To do that, the memory cell encodes the intertrial period using a negative steady-state value. (Note that since memory cells and weights are unbounded, some memory cell activity and memory block output were the inverted image of those shown here. This does not affect the interpretation given here.) On CS, the forget gate closes, forcing a reset of the memory cell to 0 (Figure 8-6B). From that point in time, a positive build-up is created in the memory cell using memory cell to input gate feedback loop (Figure 8-6A,D, see also Figure 8-2). This allows an increasing flow of input into the memory cell probably coming from the bias unit. Finally, on US, the forget gate together with the input gate, resets the memory cell to its negative baseline activity (Figure 8-6A-B,D).

The forget gate brings it back to 0 by clearing the memory cell content while the input gate probably allows the US to re-activate the negative steady-state of the memory cell. The LSTM output (Figure 8-6F) depends directly on the memory cell build-up with little or no use of the output gate that is always open (Figure 8-6C). In summary, the networks trained with a trace paradigm developed an explicit internal representation of the environmental state (trial versus intertrial) and elapsed time within a single memory cell.

Further experiments revealed that a number of factors could influence the time representation within the networks. Among those, the value of the hyper-parameters (α_{LSTM}, β, λ_{LSTM}, α_{TD}, and λ_{TD}) and the simulation resolution (10Hz vs 5Hz) seem the most important ones. More memory blocks (M) only seemed to increase variability in representations, but does not seem to change the most frequent representation. The number of memory cells within each memory block did not seem to impact the final representation. Finally, and surprisingly, probe trials during training seemed to have little effect on the representation of normal trials. Differences between the present simulations and previous ones can be found in *Appendix A: Changes from (Rivest et al., 2010a)*. It is important to assess the impact of the hyper-parameters on the networks final representation since different results could lead to different predictions.

8.4.2.b) *Time representation in delay conditioning*

All LSTM networks trained under delay conditioning developed the same representation. As expected, the representation, shown in Figure 8-7, was very different from the one found in trace conditioning. Under this paradigm, the network does not need to remember whether it is in a trial or an intertrial in working memory since this information is directly provided by the presence or absence of the CS at its input node. Like in trace conditioning, the memory cell has a negative intertrial baseline (Figure 8-7D). When the CS appears, it opens the input gate slightly (Figure 8-7A), allowing the CS signal to level off the memory cell activity. This releases the inhibition the memory cell is applying on the input gate, allowing it to fully open (in

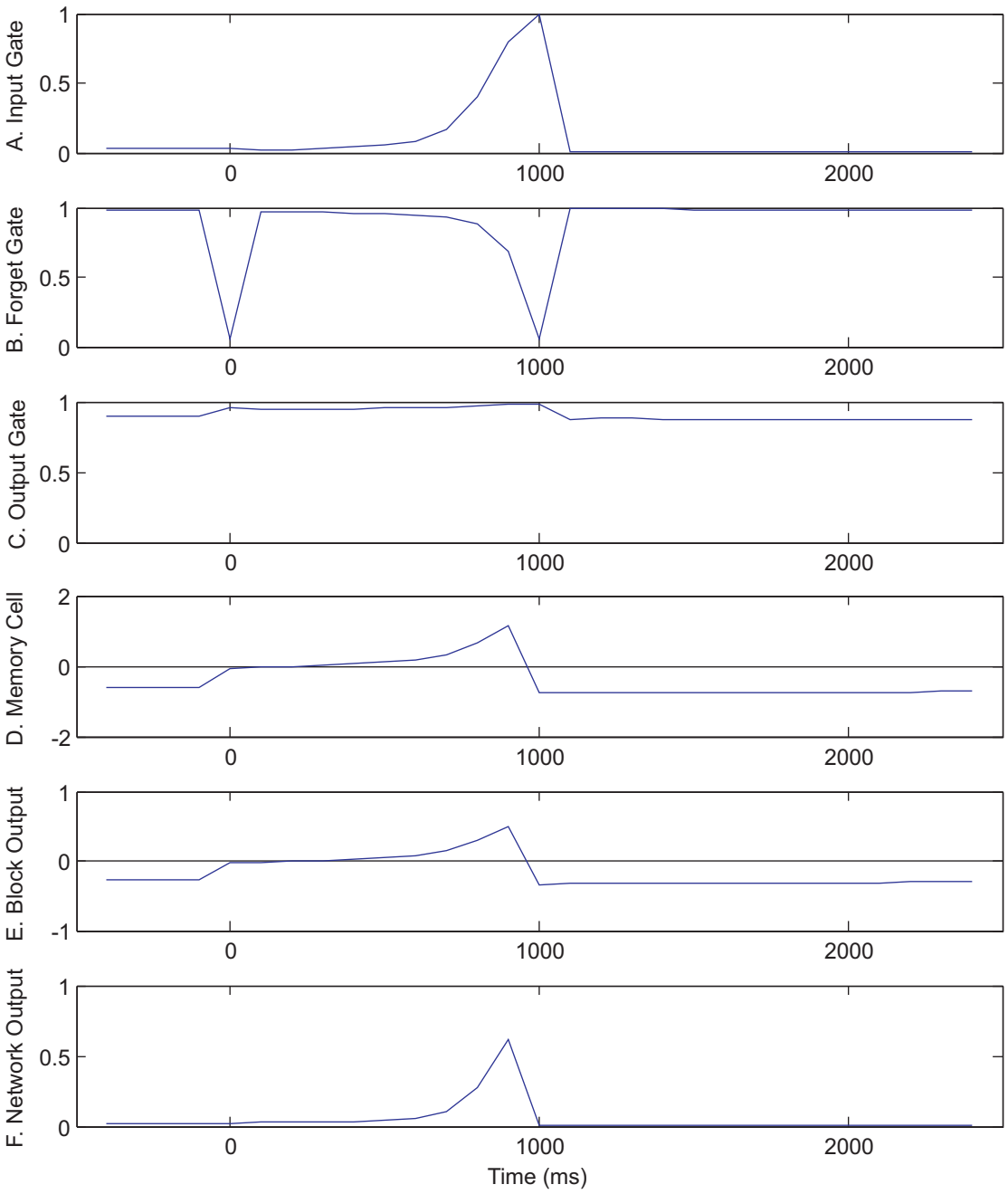


Figure 8-6: Typical time representation within an LSTM network under trace conditioning.

about 400ms). During the trial, the continuous flow of activity from the CS allows the memory cell to climb almost linearly until the specific delay is reached (Figure 8-7D). When appropriate, the output gate begins to open (Figure 8-7C), probably allowing a more precise LSTM output than if the networks were only using the memory cell activity as output, as they do under the trace paradigm. On US, the forget gate resets

the memory cell into a negative state (Figure 8-7B,D). Although it does not seem to inhibit the input gate enough to close it immediately (Figure 8-7A), the absence of CS flowing into the memory cell lets the latter shunt back to its negative steady-state from which it fully closes the input gate (Figure 8-7A,D). We found no signal in the LSTM network that could discriminate whether the network is in a trial or an intertrial, meaning that the network is not coding this information which it receives as input. Trace networks used the sign of the memory cell activity to code this information (Figure 8-6D). In summary, the evaluation of elapsed time in delay networks (the memory cell build-up) is dependent on the presence or absence of the observable CS. When the CS is present, the memory cell increases its activity over time; when it is not, it decreases to its baseline. This leads us to another question: what would append if the CS stays on beyond the US?

8.4.2.c) *Time representation in extended delay conditioning*

LSTM networks trained under extended delay conditioning also developed their own representation of the task (Figure 8-8). Under this paradigm, the networks have a finer control of the CS inflows to the memory cell build-up than in delay conditioning (Figure 8-8A), probably because they have to deal with the persistence of the CS beyond the US presentation. While the time representation may look similar to that in delay conditioning, instead of a linear build-up only within the memory cell, both the memory cell and the input gate show a build-up activity pattern (Figure 8-8A,D) similar to one in trace networks (Figure 8-6A,D). In contrast to delay networks, the forget gate combined to the input of the US allows a full reset of the memory cell in a single time step (Figure 8-8A-B,D) as in trace conditioning. This reset seems to push the memory cell to a lower value than its negative steady-state outside trial. But the activity seems to rise again under the persistent CS (from 1100ms to 2000ms, Figure 8-8D). This suggests that the network does not encode distinctively the pre- and post-US period except for few tenths of a second following the US. This could be similar to resetting the clock to a small negative value, just enough so that it does not reach the necessary accumulator value to expect a second

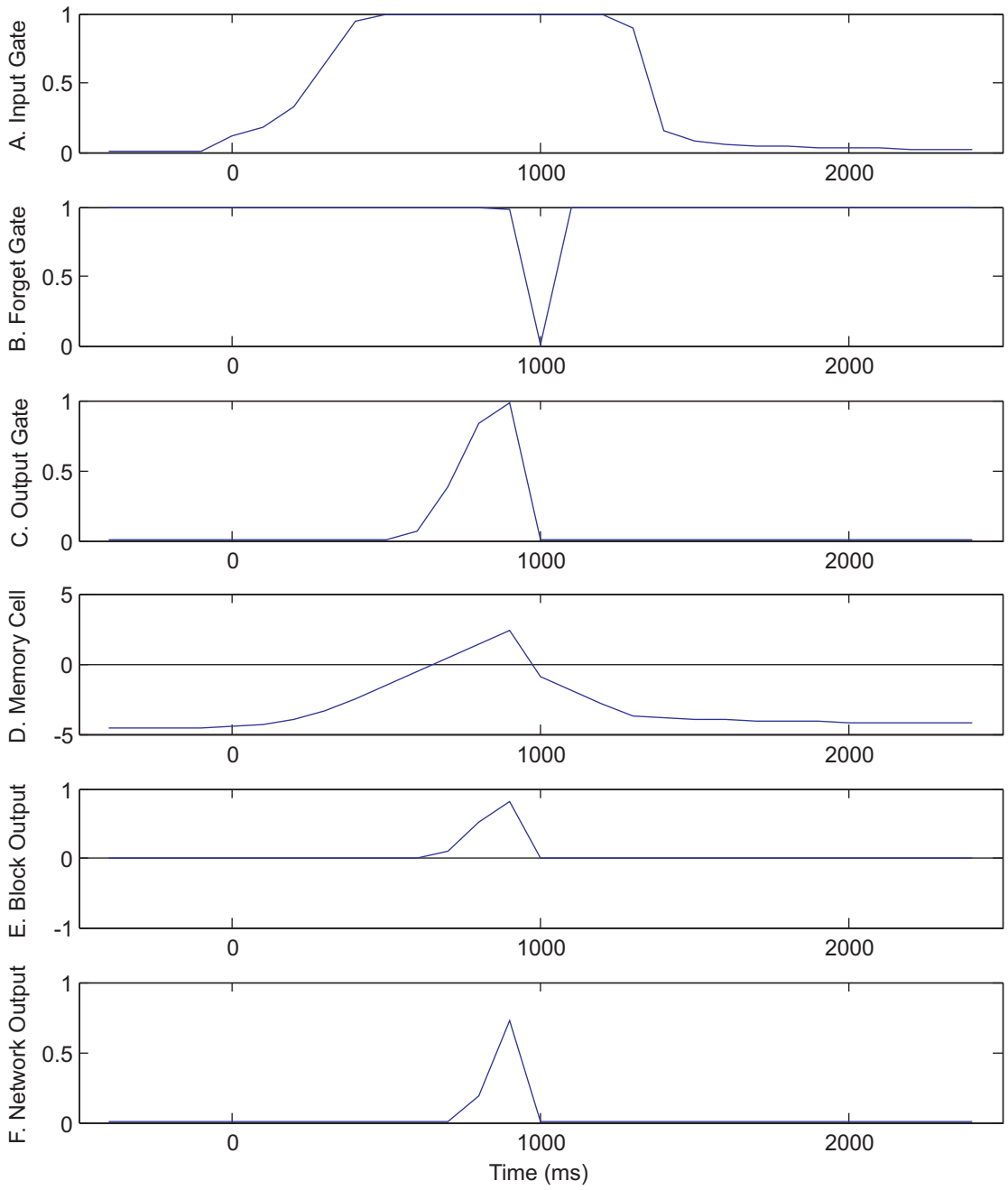


Figure 8-7: Typical time representation within an LSTM network under delay conditioning.

US during the persistence of the CS. It may well forget about the US if the CS is presented long enough. Except for the output gate, all networks had the same representation. Some extended delay networks (9/30, not shown here) seemed to use the output gate to increase the network output precision as in delay networks (Figure 8-7C). Note that as opposed to trace conditioning, in delay and extended conditioning

the memory cell activity does not shift in sign on CS appearance. In summary, extended networks seem to use every gate available to have a finer control on the CS inflows and their output prediction and to compensate for the persistent CS after the US offset. They do not seem to code for pre- versus post-CS period for a very long time.

8.4.2.d) *Time representation summary*

Experiment 1 shows that the type of conditioning protocol has a significant effect on the LSTM network representation of the task. Although the memory cells always seemed to code elapsed time in the present simulations, the networks allotted their resources according to the demands of the task. In trace conditioning, the network had to generate internal representations of both the state of the environment and of elapsed time. In contrast, when the CS provided relatively unambiguous information about trial/intertrial state (delay conditioning), the network allotted most resources to time estimation and there was no clear representation of state. Finally, when the CS persists beyond the US, the network seems to use all the resources to control timing more finely.

An objective of these experiments was to verify if using a single memory block with a single memory cell would result in consistent and easy to analyze representations of the task in the LSTM networks. To test that, we replicated (Rivest et al., 2010a) results using a single memory block with a single memory cell and we expanded the analysis of the LSTM representation to delay and extended conditioning. The networks were very hard to train (very low success rate, even with the best choice of hyper-parameters) and their final representations seem more sensitive to the hyper-parameters. On the other hand, the final solutions of randomly initialized networks under the same settings were extremely stable and they lead themselves to an easy analysis, an important property for a model, especially when it comes to making predictions or explaining neurophysiological data.

8.5 Experiment 2: Response to probes with unexpected interstimulus interval

Since the first dopaminergic data on unexpected delay length appeared (Hollerman & Schultz, 1998), modellers used to assume that dopaminergic response should be similar whether the data were collected under trace-like or delay-like

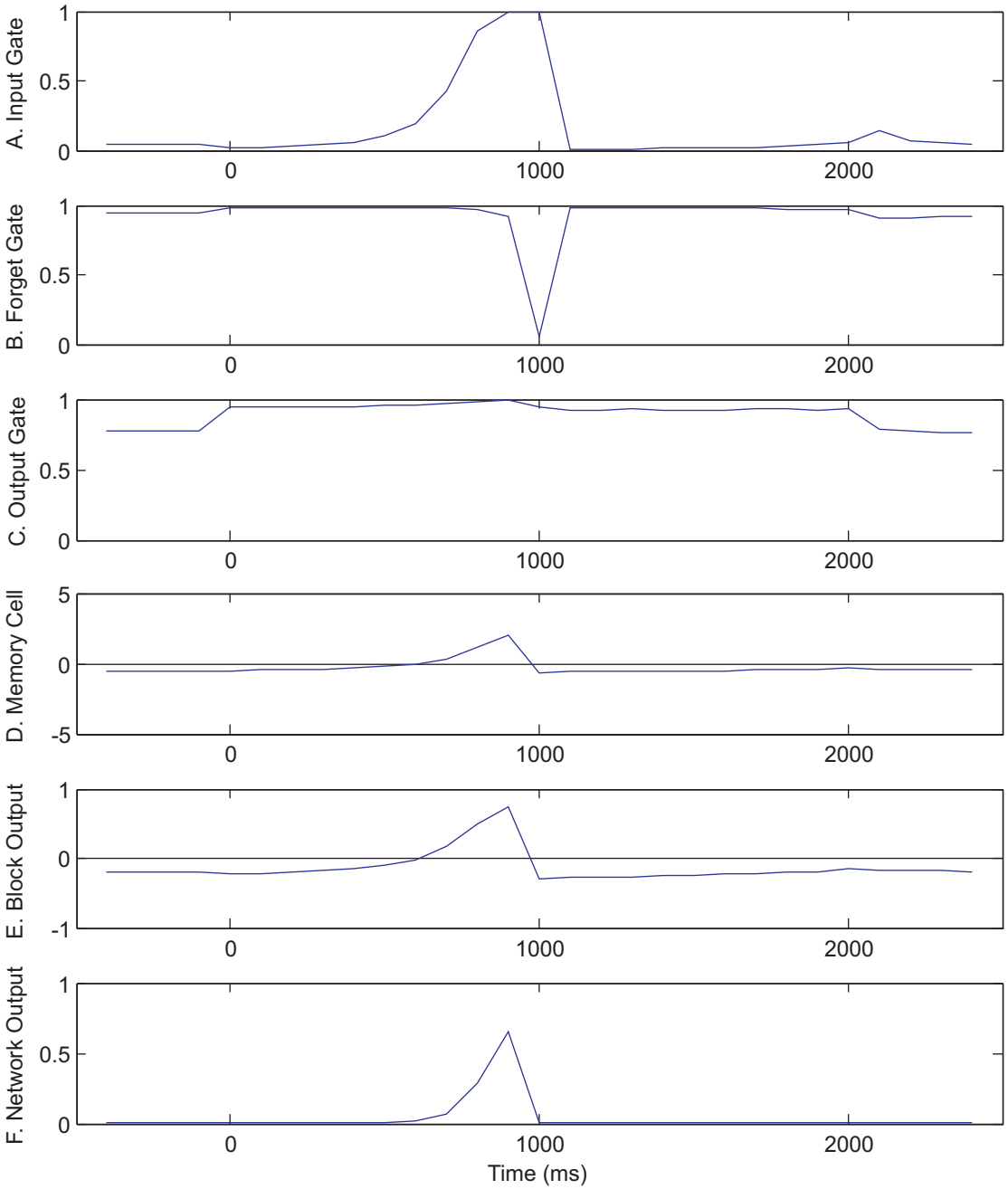


Figure 8-8: Typical time representation within an LSTM network under extended conditioning.

conditioning paradigms (Daw et al., 2006). To do that, they usually encode time from the CS onset, instead of its presence or absence (Suri & Schultz, 1999; Brown et al., 1999; Daw et al., 2006; except in Ludvig et al., 2009). But it is not clear that it is the case in the brain (Morris et al., 2004; Fiorillo et al., 2008; Ludvig et al., 2009). The model presented here develops its own time representation and is very sensitive to the presence or absence of the CS (see *Experiment 1*). For example, and similarly to animals (Gallistel & Gibbon, 2000), it learns delay conditioning much faster than trace conditioning (Rivest et al., 2010a). It is therefore useful to test whether the model would predict similar or different dopaminergic responses (modeled by the δ of TD) under the three paradigms presented here. To our knowledge, there are no dopaminergic data under situations comparable to the extended delay conditioning paradigm.

8.5.1 Methods

To test how the model would respond to unexpected interstimulus intervals we ran 90 networks trained as in *Experiment 1* (with one, two, and five memory blocks, $M \in \{1,2,5\}$) on two *time-probe* blocks. The first block alternates between normal and early trials, where early trials had an interstimulus interval of .5s instead of 1s. The second block alternates between normal and late trials where late trials had an interstimulus interval of 1.5s. Time-probe blocks have the same duration as training blocks (2400 time steps). During these probe blocks, all network learning rates are set to zero ($\alpha_{LSTM} = 0.0, \beta = 0.0, \alpha_{TD} = 0.0$) to ensure the networks are not learning the new interstimulus interval.

8.5.2 Results

8.5.2.a) *Predicted dopaminergic responses on time-probe trials*

Averaged δ signals (the correlate of the phasic dopaminergic response) for early, normal, and late trials and for each conditioning procedure are shown in Figure 8-9. With two memory blocks in the LSTM networks, there seems to be little difference in δ responses between trace (Figure 8-9A-B) and delay (Figure 8-9C-D) conditioning, both being comparable to (Hollerman & Schultz, 1998) data

(Supplemental Figure 8-18). The most visible difference is the depression following the US on early probe trials in delay conditioning (Figure 8-9D, left column) that does not appear in trace (Figure 8-9B, left column). Networks trained under extended delay conditioning (Figure 8-9E-F), however, showed two significant differences. First, they show a depression on early trials, at usual US timing (Figure 8-9F, left

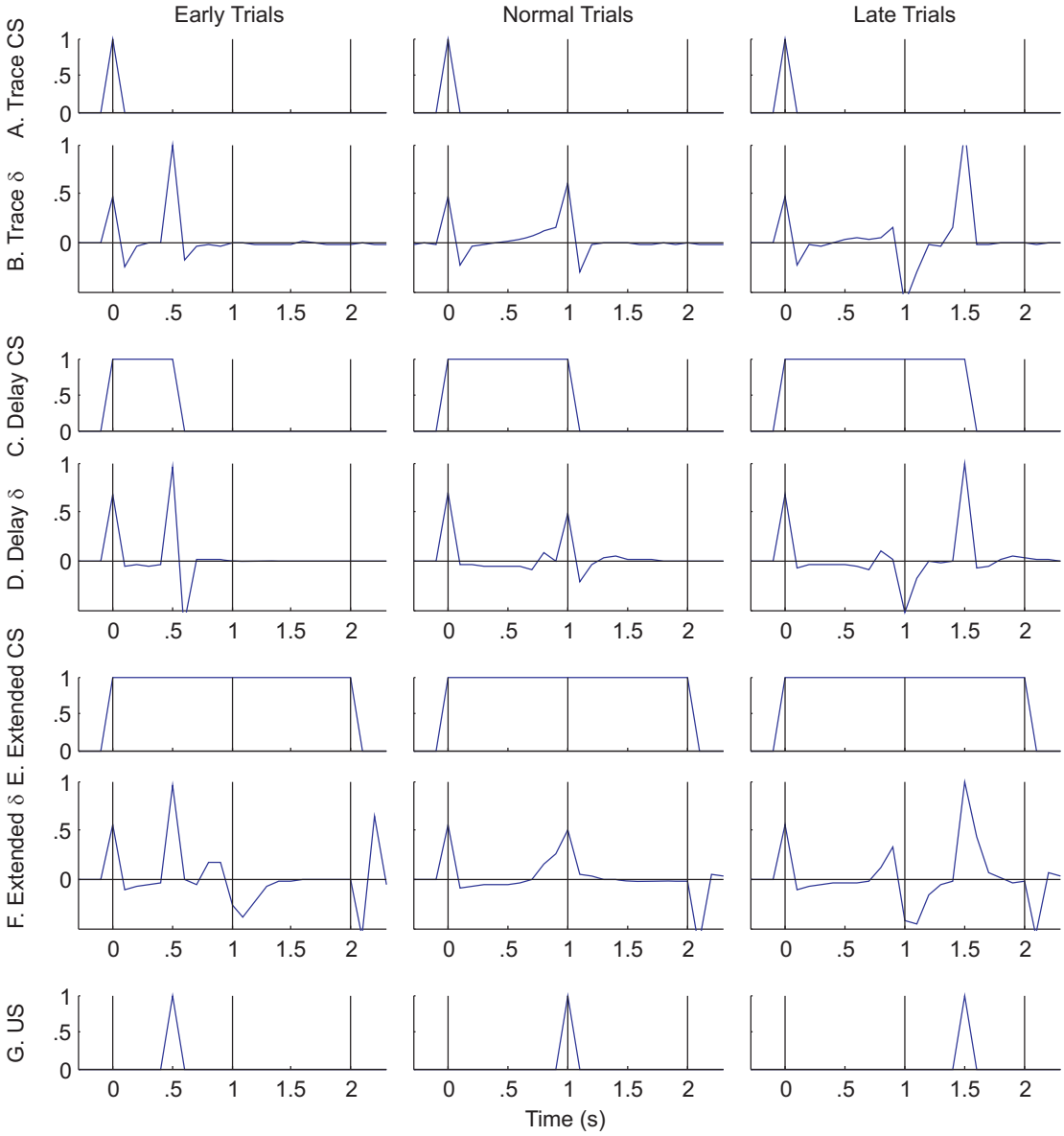


Figure 8-9: Population average δ signal on early (0.5s, left column), normal (1.0s, center column) and late trials (1.5s, right column) for each condition paradigm. For each condition (A-B trace, C-D delay, E-F extended delay) procedures, the first figure panel shows the CS signal while the second panel shows the δ signal. The last panel (G) shows the US signal (same for all procedure). LSTM networks had two memory blocks ($M = 2$).

column). While trace networks stop expecting reward (US) after its occurrence, these results suggest that the extended delay networks do not stop expecting it. Second, these networks show a δ depression whenever the CS disappeared (Figure 8-9E-F). These activities may be caused by networks not differentiating between pre-US and post-US intervals. Training the networks with a longer CS persistence or with time-probes trials throughout training could force them to acknowledge that two US in a row within the same trial is unlikely. Networks making such an inference may lead to a different δ. Similar sensitivity to probe trials during training was shown in (Rivest et al., 2010a). Finally, if the network had to learn to predict the CS, and hence its termination, such internal representation of the CS offset timing could make the CS offset depression disappear. Further simulations would be needed to make clear predictions.

Table 8-I : Number of networks showing dopaminergic responses properties on early, normal and late probe trials. Results where networks populations differ are marked by a *.

		Trace (/30)	Delay (/30)	Extended delay (/30)
Early trials	$\delta(0.0s) > 0.2$	28	30	30
	$\delta(0.5s) > 0.2$	30	30	30
	$\delta(1.0s) > -0.2$	30	30	18*
	All	**28**	**30**	**18**
Normal trials	$\delta(0.0s) > 0.2$	28	30	30
	$\delta(1.0s) > 0.0$	22*	29	28
	All	**20**	**29**	**28**
Late trials	$\delta(0.0s) > 0.2$	28	30	30
	$\delta(1.0s) < 0.0$	20	23	20
	$\delta(1.5s) > 0.0$	30	30	30
	All	**19**	**23**	**20**

We completed the analysis of the networks with two memory blocks by testing how many showed some of the dopaminergic responses reported in the literature in similar situations (Table 8-I). Two differences due to the type of training were visible. First, many extended delay networks (>1/3) showed an important δ depression at usual US timing (as showed in Figure 8-9F, left column). Second, some trace networks (<1/3) also showed a depression on US presentation in normal trials (as reported in Rivest et al., 2010a). This last result was not visible from the average population signal Figure 8-9B, center column).

Trace networks with a single memory block ($M = 1$) were unable to replicate known depression at the expected time (1s) on late trial (Figure 8-9B, right panel). Memory cell and weights in LSTM networks are not sign-constrained. The activity shown in Figure 8-6D can thus also be inverted (symmetric over time axis) and this does not affect the interpretation of how LSTM represents the task. But the LSTM to TD connection clips the memory cells signals in the range [0, 1]. Moreover, the LSTM networks under trace conditioning presented here encode elapsed time and trial versus intertrial in the same neurons using the positive and negative part of the signal differently. This may explain why the trace simulation with a single memory block does not reproduce the dopaminergic data appropriately. It does not receive the full information contained in the LSTM memory cell. All other panels from Figure 8-9 were similar for the number of memory blocks tested ($M \in \{1,2,5\}$).

8.5.2.b) *Time-probe trials summary*

Overall, the simulations properly replicate existing dopaminergic data and predict little or no differences in dopaminergic phasic activity between trace and delay conditioning, as many modellers in the past assumed. On the other hand, the simulations predict a depression at expected US timing on early trials under extended delay conditioning. The model also predicts a depression when the CS disappeared in this paradigm. But more simulations are necessary to isolate more precisely the conditions under which these depressions are likely to appear versus conditions under which they may not. To our knowledge, no physiological data are available for these conditions. Finally, the connectivity from the LSTM networks to the TD model should be revised to ensure all the information the LSTM can provide to TD is received by TD and can be used linearly (since p_t is linear, Equation 8-10). Beyond this issue, the results were consistent and independent of the number of memory blocks.

8.6 Experiment 3: Response to probes from different conditioning paradigms

The role of models is not only to provide explanations for observed phenomena, but also to make predictions about what could be observed under different situations according to the models. In particular, when different hypotheses need to be tested, simulations could help determine the best protocol to test those hypotheses. Simulations are often much faster to do than real experiments. This allows the experimenters to search through simulations the set-up that is most likely to discriminate hypotheses.

Experiment 3 aimed at making interesting predictions that could eventually be tested. An interesting question we asked is what the model would predict in terms of reward expectation and dopaminergic activity if the animal was first trained on one type of conditioning paradigm and then tested on a second one. If these predictions happened to be true, what do they indicate about the animal's inferences about the task he was trained on?

8.6.1 Methods

To test how the model responds to probes from other conditioning paradigms, we ran 90 networks trained as in *Experiment 1* (with two memory blocks, $M = 2$) on eight *conditioning-probe* blocks. Each block alternates between normal trials from the same conditioning paradigm on which the network was initially trained and probe trials consisting of normal trials from a second conditioning paradigm. There are four possible conditioning paradigms: trace, delay, extended delay, and a new very extended delay conditioning paradigm in which the CS is presented during four times the CS-US interval. There are two types of conditioning-probe blocks: one in which probe trials are rewarded as usual, and one in which probe trials are not rewarded. Conditioning-probe blocks have the same duration as training blocks (2400 time steps). As in *Experiment 2*, all networks learning rates are set to zero ($\alpha_{LSTM} = 0.0$, $\beta = 0.0$, $\alpha_{TD} = 0.0$) during probe blocks to ensure the networks are not learning the new interstimulus interval.

8.6.2 Results

8.6.2.a) *LSTM responses on cross-probe trials*

Trace networks displayed the most unexpected behavior of all. They had been trained in an environment in which the CS appeared for one time step (e.g. 100ms), followed by a US after a trace delay of 900ms. When tested in non-trace probe trials, their predictive output showed that some networks continued to predict the US 900ms after CS offset and not 1000ms after CS onset. For trace probes, this was the normal response, but the behavior persisted when the networks were presented with delay, extended and very extended trials. Few networks seemed to measure elapsed time from CS onset, but most of them started later and the population average of US prediction peaked at a little more than 900ms after the CS offset (as shown on Figure 8-10). In a different set of simulations (unpublished), where learning rates were non-zero, the results were even more striking, with almost every network consistently peaking at 900ms after CS offset on every type of trial. This behavior also seemed to be independent of whether the US appears or not, except for the delay condition, where rewarded probes usually did not lead to US prediction. Finally, when no reward is given, networks tend to wait for it for a while. When the CS is long enough, it makes sense to minimize temporal error by using the closest cue in time. Here, trace networks normally see the CS for a single time step. Nevertheless, they seem to start to count on CS offset when tested in probe trials with longer CS presentations.

Delay networks responded as expected to probe trials. They made no US predictions in trace probes, but they did expect the US on time in all other conditions (shown by a peak of activity at 900ms, Figure 8-11). Like trace networks, delay networks kept waiting for the reward to appear in unrewarded trials, but in contrast to trace networks, their US prediction stops with the disappearance of the CS. On extended rewarded trials, delay networks show an almost immediate resurgence of US prediction. In *Experiment 1*, we found that the memory cell of those networks is not fully reset by the US. It is likely that the persistence of the CS beyond the US

offset and the residual build-up in the memory cell together lead the networks to predict the US again without waiting for the usual interstimulus interval.

Extended delay networks respond similarly to delay networks in trace and delayed probe trials, but they respond very differently in both type of rewarded extended delay trials (compare Figure 8-11 and Figure 8-12 lower right panels). On rewarded normal extended delay trials, extended delay networks do not show resurgence, as they were trained to. On unrewarded very extended trials (the lowest rightmost panel in Figure 8-12), they wait a full second after the US before showing resurgence. In contrast to delay networks, this resurgence is not immediate and takes a little more than 1s to reach its peak. As we found in *Experiment 1*, these networks reset their state within a single time step in order to avoid resurgence in the extended period of 1s in training. The amount of time each network waits before predicting the US again is a measure of how long they can remember that the US is past. The results presented here show that those networks are relatively limited in terms of how long they can inhibit resurgence of prediction, confirming the results found in *Experiment 1*.

8.6.2.b) *δ (DA) responses on cross-probe trials*

We also looked at the behavior of average δ signals in each condition (similar to LSTM outputs, see Supplemental Figure 8-19, Supplemental Figure 8-20, & Supplemental Figure 8-21) to generate predictions for the dopaminergic phasic signal. First, δ signals for all conditions showed a burst on CS appearance. Second, δ signals for all conditions also showed a burst of activity on US, but the burst size varied with reward predictability. For example, delay networks showed a larger burst to US on trace probe trials than to other probe trials; for trace networks, it was the opposite. Third, δ signals for all conditions showed a depression on CS offset, unless a reward co-occurred with the CS offset. Finally, for unrewarded probes, delay and extended networks showed a depression at expected reward time (except on trace probes) and trace networks showed such depression only on unrewarded trace probes. A little depression is visible around 1s after the CS offset for trace networks on non-trace

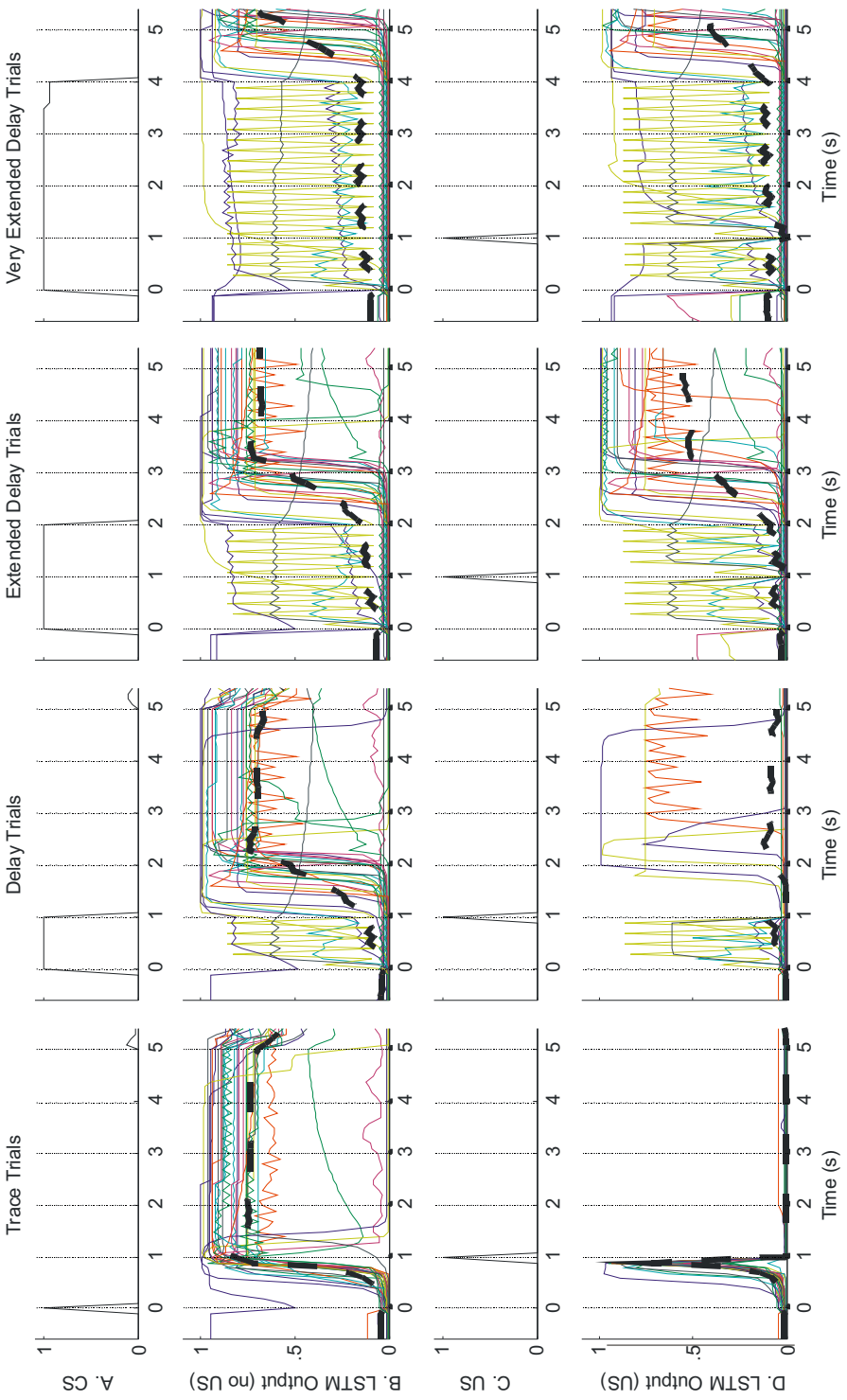


Figure 8-10: Mean LSTM output for trace networks on cross-probe trials. Each column represents a different type of probe trial. Rows are aligned in each column. The first row is the CS, the second is the LSTM output on unrewarded probes, the third is the US, and the fourth the LSTM output on rewarded probes. Each line represents a different network; the wide dashed line represents the networks population average.

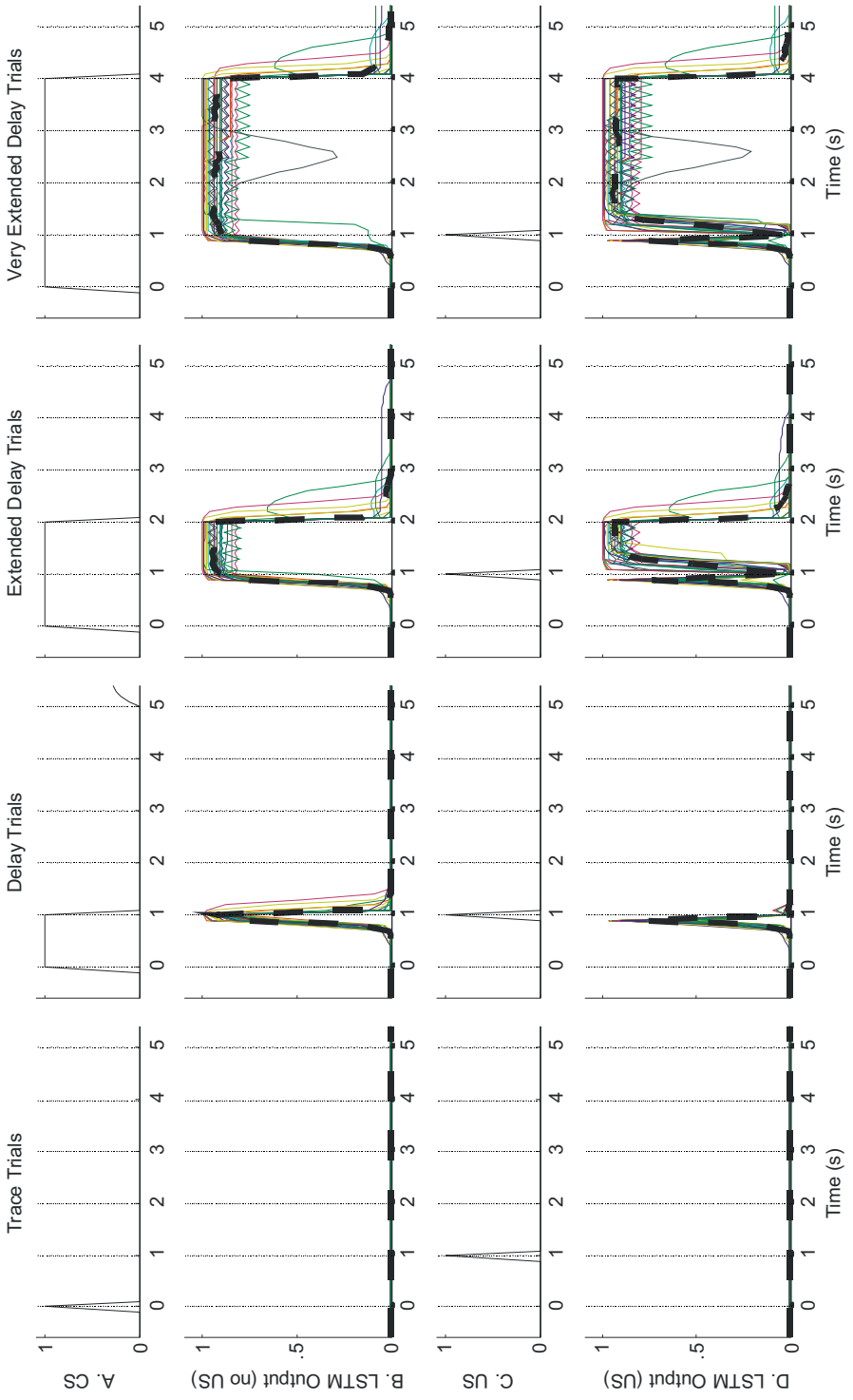


Figure 8-11: Mean LSTM output for delay networks on probe trials. Same format as in Figure 8-10.

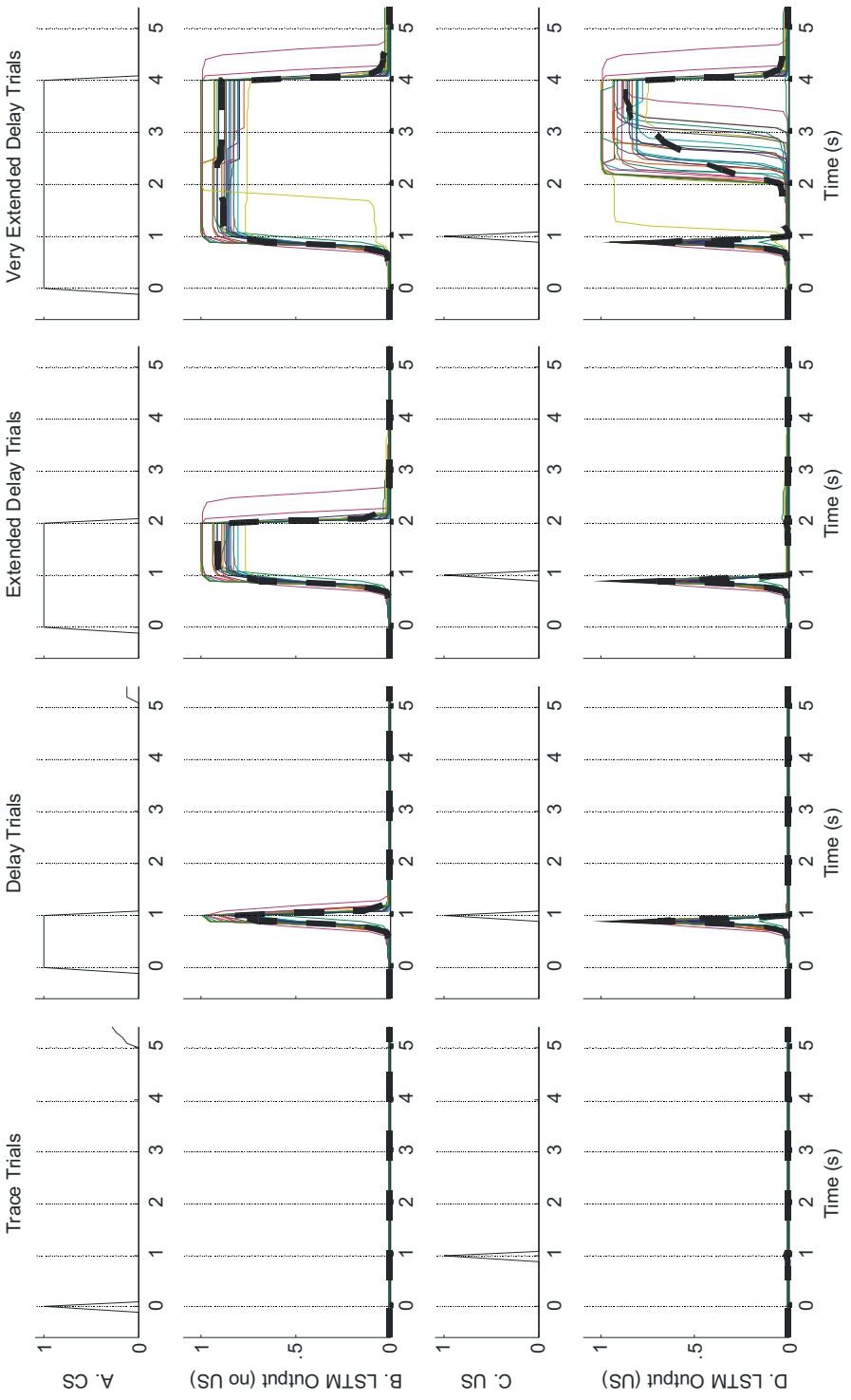


Figure 8-12: Mean LSTM output for extended delay networks on probe trials. Same format as in Figure 8-10.

probes. Although this response is not clear, it is certainly related to networks counting time from CS offset. In summary, while LSTM networks vary significantly in their representation and responses to different probes, the resulting δ is relatively stable independently of the training condition. Only expected US time varies since trace networks start counting on CS offset while other networks require the presence of the CS to count at all.

8.6.2.c) *Cross-probe trials summary*

The goal of these simulations was to show that the model could generate interesting and testable predictions. Two very interesting predictions came out of these simulations, beyond dopaminergic prediction.

First, the model predicts that animals first trained under delay conditioning may show resurgence of reward expectations under extended delay conditioning. The model explanation for this behaviour would be that the animal is basing its expectations solely on the presence of the CS (see *Experiment 1*). It is possible that if the animals encountered enough probe trials, it could understand that there are never two rewards in a row (or it could have a natural bias toward this inference). More simulation would be needed to determine under what circumstances the model predictions change. The model also predicts that even if trained under extended delay conditioning, the animal could show resurgence of expectations if the CS suddenly persists longer than usual. Remark that the model solution could have been different if it was trained on slightly longer persistent CS. Right now, the length of the post-US CS interval is such that the network has found a solution without remembering the US event by resetting its accumulator to a small negative value. Training on longer persistent CS may have allowed the model to learn to remember the US is past and may have led to a different prediction. Again, more simulations could allow isolating specific situations under which we should expect the animal to show resurgence versus conditions under which we should not.

The second prediction of the model is that the trace interstimulus interval is likely to be counted from CS offset and not from CS onset. Although the CS was

short (one time step) in the present experiment, it does make sense to start counting on CS offset on long CS duration, since it reduces the interval duration to represent and hence, increases the accuracy or ease of learning.

Overall, this experiment shows that the model is making interesting predictions for which it proposes a clear explanation, assuming timing is learned by trying to predict upcoming events in some form of working memory in the cortex. It also allows generating simulations that can help determine the best experimental situation to test the model hypothesis and explanations.

8.7 Experiment 4: Time representation with respect to interval length

Finally, an important question is how time is represented in the LSTM with respect to different time interval lengths. This question can help evaluate whether LSTM can make a good timing model and how well it fits neurophysiological data as well as provide directions for future improvements of the model.

8.7.1 Methods

In this experiment, we took the trace and delay networks from (Rivest et al., 2010a), but extended their delay to try to better understand how time is represented in each network as a function of training history. We trained networks by expanding the delay one time step at a time. When a network had a successful block, it was then trained on the same task, but under an interstimulus interval one time step longer. We limited such retraining to 250 blocks per increment. In case of failure, up to 10 training attempts per increment were run. Networks were initially retrained on a 1s delay to eliminate possible learning of the last early and late probe trials. They were then trained for up to 3s interstimulus onset intervals by increments of 0.2s. The LSTM learning rate α was set to 0.1 instead of 0.5, as this seemed to speed re-learning and the chance of success.

8.7.2 Results

8.7.2.a) *Time representation under trace conditioning*

We monitored the memory blocks with the most common temporal representation on 1s intervals (Supplemental Figure 8-17) and looked at their evolution as we extended the CS-US delay interval up to 3s (27 memory blocks from 26 distinct networks from (Rivest et al., 2010a)). The overall structure did not change much (Supplemental Figure 8-22). The input gate still opened on the CS, the forget gate still closed on the US and the memory cells still showed sustained activity from CS to US marking the trial interval. The interesting signals to look at are the output gate and memory block output whose loop integrates time. The correct network prediction depends on the memory block output, which depends on the output gate assuming relatively constant memory cell activation. We calculated the averaged output gate signal over the whole population, normalized as a function of the duration of the delay interval (Figure 8-13) (two output gates were removed from this analysis since their activity clearly did not adapt to the delay like the other output gates). Although there is a slight adjustment in curve's shape up to about 1.6s, longer delays have very similar curves. The slope seems to scale linearly with the timed interval.

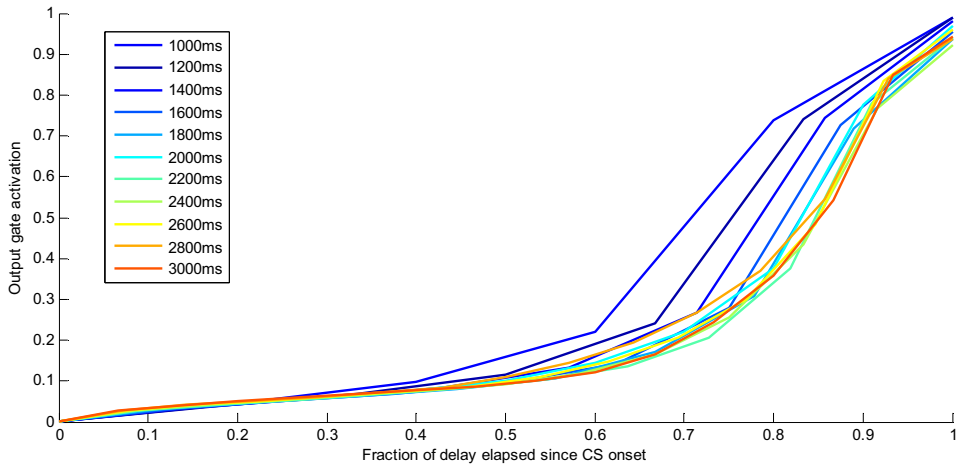


Figure 8-13: Averaged output gate activity with respect to the proportion of elapsed time between CS onset and US onset.

This seems consistent with scalar expectancy theory (Gallistel & Gibbon, 2000) that predicts that the temporal behavioral profile for different delays, when normalized for duration, should closely match. This further indicated that learning different time intervals was achieved by appropriate scaling of gate openings, feedback activations and integration rates within the memory block. However, much larger intervals will be needed to verify this possibility. The earlier increase in output gate activity appearing in the shorter intervals (between 1s and 1.6s) may be caused in part by the insufficient temporal resolution of the simulation. A minimum of 2 time steps may be needed to cross threshold and open the gate, taking a minimum of .4s (or 40% of the time interval) for networks initially trained on the 1s interval. Curves are extensively overlapping for time intervals between 1.8s and 3s. The training procedure also requires networks to reach an absolute precision of 1 time step (200ms). This condition does not allow a comparison to animals who usually demonstrate a timing precision proportional to the interstimulus interval.

8.7.2.b) *Time representation under delay conditioning*

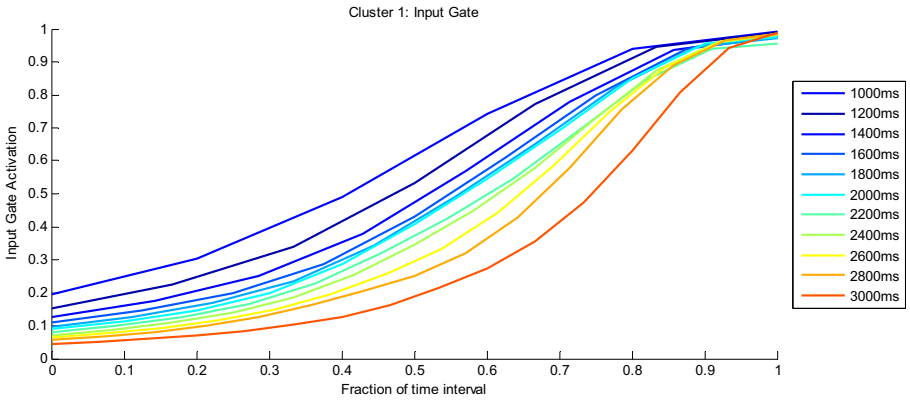
The most common representation in those networks was the same as the one shown in *Experiment 1* of this paper (Figure 8-7). Among the memory blocks with this representation (36), at least two most common classes of adaptation to larger time intervals appeared.

The first and largest group (I, 18 memory blocks), directly scaled the input gate slope behavior, adjusting the whole timing process to the new delay (see example in Supplemental Figure 8-23). Figure 8-14 shows the temporal alignment of the gate signals. Although the normalized input gate activations (Figure 8-14A) did not align perfectly, we should not forget that the LSTMs are all trained with the same absolute temporal precision, so the longer the interval to time, the more precise the LSTM has to be relative to the time interval length. Therefore it should not be surprising to see networks that adapt their input gate to the interval length showing a sharper input gate response curve. Forget gates are mostly triggered by the US and hence, do not scale with respect to time (Figure 8-14B). Output gates (Figure 8-14C)

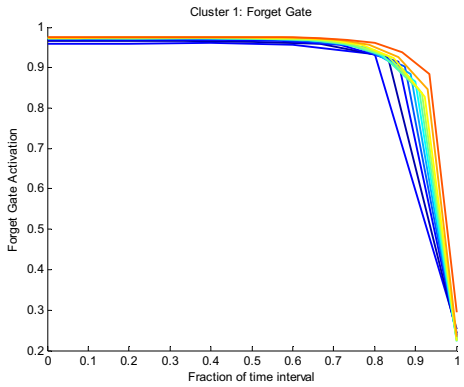
seem more difficult to evaluate. Simulations with much larger time intervals would be needed to determine whether their behaviors scale with time interval length.

The second most common adaptation (II, 13 memory blocks) was to sustain the input gate response, which resulted in a larger memory cell activation range due to the constant input that followed its opening (Supplemental Figure 8-24). While memory cells of the first group (I) showed a change in slope with the adaptation of the input gate, the second group (II) showed a relatively constant slope but an increase in memory cell activation range proportional to the interval length. Such a change then requires a change in threshold at the output gate to ensure responding at the appropriate time. Output and forget gate signals of this second group are very similar to those of the first group.

A



B



C

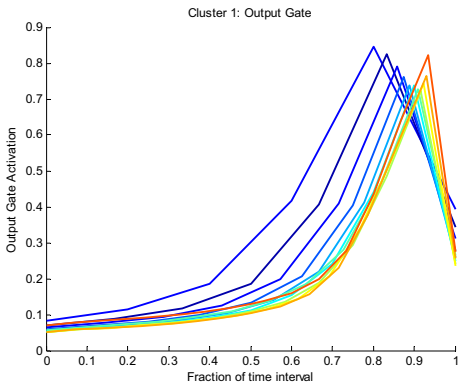


Figure 8-14: Averaged input gate (A), forget gate (B), and output gate (C) activity with respect to the proportion of elapsed time between CS onset and US onset for memory blocks of delay networks with input gate adaptation to time interval (group I).

For both groups, we performed a regression on the memory cells' activation range between the first CS onset time step and the time step immediately preceding US onset. While the memory cell activation range was mostly independent of the interval duration for group I, it was highly dependent on it for group II (Figure 8-15, left panel). A regression on the slope of the memory cell activity over the same time period (CS onset to one time step before US onset) showed that the slope of memory cells from group I was much more dependent on the interval length then for group II. Once scaled by the time interval, the slope for memory cells of group I was mostly independent of the delay, meaning that the slope scaled linearly with the timed interval (Figure 8-15, right panel). The second group (II) is only possible because LSTMs have unbounded memory cells, a clearly unphysiological feature. The fact that the slopes of the memory cells from group I scaled linearly with the timed interval and their range was bounded makes group I a much more interesting and realistic neural model of timing than group II. Group I seems a good match for similar slope-adapting-to-delay neurons in the brain. Adaptation to variable delay by adjusting slopes was also found real neurons (Komura et al., 2001; Leon & Shadlen, 2003; Reutimann et al., 2004).

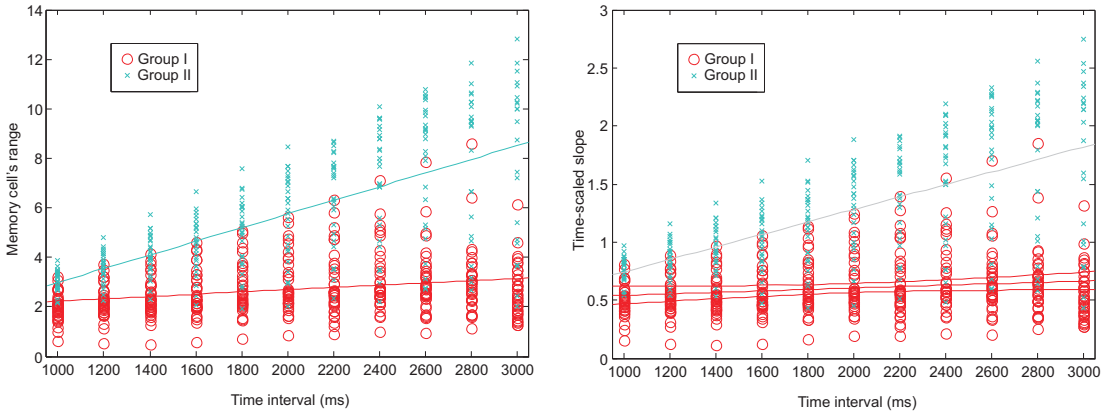


Figure 8-15: Regression on memory cell's range (left) and time-scaled slope (right) for group I and II.

8.7.3 Summary

The goal of this experiment was to study how time is represented in the LSTM with respect to different time interval lengths in order to evaluate whether the LSTM networks could make a good timing model and how well they would fit

neurophysiological data as well as to provide directions for future improvements of the model. While some gates were triggered by a stimulus, such as the CS or the US, a number of networks seem to have signals that scaled linearly with the time interval length; an important characteristic for a timing model. Under delay conditioning, some networks adapt their slope to the different interval lengths (group I). These units seem to match neurophysiological data. Others (group II) adapt their activation range. This behaviour may be less realistic, although not totally impossible for some small ranges. Modifying the model to favour slope-adaption rather than range-adaptation to new delays could be preferable.

8.8 Discussion

8.8.1 Animals vs the model, conditioning and timing

In these experiments, networks were trained like animals using blocks of trace or delay conditioning with a fixed interstimulus interval separated by random intertrial intervals. However, it is important to review major differences between the model and animal studies as well as between the simulations, conditioning and interval timing.

First, while animals are active agents that must lick to drink, the model studied here is a purely passive agent. It does not need to act to get the reward, nor does it have an overt unconditioned response to the US. The model's only unconditioned response is the reward signal r that responds to the US. In animals, a similar signal acting as internal hedonic reward value could come from lateral hypothalamus (Nakamura & Ono, 1986) when the animal licks the reward. In the conditioning situation in which we placed the model, the model never had to press a lever or lick to get a reward, it just receives it, similar to brain stimulation reward paradigms (Hernandez et al., 2006).

Second, the cortical part of the model, the LSTM, tries to predict the next stimulus with precise timing and we evaluate the model performance or training success based on this very hard criterion. In reality, animals are never trained until their cerebral cortex perfectly predicts the environment, they are trained until they

show some stable behavior we want or expect them to have. As shown with the dopaminergic data, while LSTM networks in trace conditioning can take as much as 10 000 trials or more before showing near perfect prediction, the necessary signal to learn to lick on CS (conditioning) can appear within the first tens of trials (see Rivest et al., 2010a). Moreover, animals may never learn the perfect timing of events. In interval timing, they show Gaussian-like response curves whose peak is centered at about the correct delay. These curves are averaged over multiple trials and do not represent the response of any single trial (Gallistel & Gibbon, 2000). Animals do not seem to change their response onset and offset much over training (Balci et al., 2009) and hence, may have a limited timing precision under this condition. The requirements imposed on the model to have a very small-width response curve may be equivalent to a huge amount of overtraining. One must be careful before comparing learning time in the model and in animals.

In the literature, conditioning acquisition is usually measured in terms of the rate of response to the CS. There are no such measure here, although it could be added with minimal changes to the model such as in (Ludvig et al., 2009). Moreover, learning speed for timing is barely discussed in the interval timing literature (Balsam et al., 2002; Kaiser, 2008, 2009; Balci et al., 2009), but timing could be acquired rapidly in delay conditioning (Balsam et al., 2002). Studies usually focus on the rate of acquisition of conditioning responses to the CS, but rarely for timing. The cerebellum, which is not integrated here, could also play a role in the rapid timing acquisition (Ivry & Schlerf, 2008).

Finally, the model does not have a hippocampus to support episodic memory. The necessity of an intact hippocampus in trace appetitive conditioning is still not clear (see Beylin et al., 2001; but Thibaudeau et al., 2007). We also consider this model a model of appetitive conditioning more than of aversive or fear conditioning because of its relationship with the dopaminergic system which seems to support reward-based conditioning (Schultz, 2007). On the other hand, the model does not yet integrate the idea that dopamine could play a role in gating working memory (Braver

& Cohen, 2000; Montague et al., 2004; Rougier et al., 2005). However the LSTM architecture seems well suited to explore this hypothesis.

8.8.2 Prediction of dopaminergic phasic responses under various conditions

One contribution of this paper is to validate the assumptions made by a number of other models (except Ludvig et al., 2009) that whether we model delay or trace conditioning should not have much of an effect on the final DA behavior in probe trials (*Experiment 2*). Most models either provided the equivalent of the sustained memory of the CS assuming trace conditioning (Brown et al., 1999), sustained the representation of the transient CS onset by diffusion of the activation through some form of delay lines (Suri & Schultz, 1999), or modeled the task in a sufficiently abstract manner that they did not have to simulate any stimulus gaps in the trace trials (Daw et al., 2006). Since our model considered these two cases as different by virtue of our adaptive model of the frontal cortex, we were able to examine what our model predictions were on this issue. As shown in *Experiments 2 & 3*, the DA behavior in probe trials did not vary much across the conditioning paradigms used (trace versus delay), except maybe for extended conditioning.

The model also made an interesting prediction about extended probe trials, predicting a DA depression on CS offset in the absence of an almost co-occurring reward. In trace conditioning, the model also predicts that the timing of the DA depression on unrewarded trials depends on the CS offset, and not the CS onset as it would be the case for delay conditioning.

A first limitation of the model as it stands is its need to have at least two memory blocks to reproduce the dopaminergic data under trace conditioning (depression at expected time on late/omission trials, Morris et al., 2004). This problem seems to be caused by the memory cells in the LSTM networks that can encode different information using positive and negative activities (Figure 8-6D) as well as the memory cell signal clipping in the LSTM to TD connectivity. Neurons with positive and negative outputs are physiologically unrealistic, but it is not hard to imagine a solution to this problem. For example, the signal could be split into two units representing the positive and negative components. These units could then

project to the TD prediction unit *p*. Having a fully working solution with a single memory block would lead to a clear explanation about how working memory representation of the task is used by TD to compute its sum of future reward estimation and the resulting δ signal as shown in see Supplemental Figure 8-25.

A second limitation of the model as a timing model for dopaminergic neurons is related to the more recent dopaminergic data presented by (Fiorillo et al., 2008). First, Fiorillo's data on early probe trials reveal a much smaller dopaminergic response to unexpected US, similar to the one on normal trials. But Fiorillo reports population averages while Hollerman and Schultz (1998) report only typical examples. Second, Fiorillo's (2008) data also shows dopaminergic responses when the animal is trained on randomly selected interstimulus intervals from a uniform distribution. It is not clear how LSTM could be trained on this type of problem and how this would affect *p* and δ signals in TD. Finally, with sufficient training on delay conditioning, the dopaminergic response to US on normal trials should be very small (near zero). That result does not seem to be fully captured yet by the LSTM-TD model (see Figure 8-9D) (see also, Rivest et al., 2010a).

In summary, the LSTM-TD model is doing an excellent job at reproducing some of the existing dopaminergic data and at explaining it. But some more difficult data remains to be covered by the model. This may imply moving away from LSTM to something better. Still, LSTM is among state-of-the-art algorithms to learn time-series and the basis of timing in the brain and whether and where time representations are learned is still under debate (Ivry & Schlerf, 2008). So one should not be surprised that there are still issues to be solved to model dopaminergic data.

8.8.3 Time representation

Results from *Experiment 1* suggest a very different representation between trace and delay conditioning in LSTM networks. In delay conditioning, the sustained CS provides continuous and unambiguous information about the state of the environment, such as whether or not it is currently in a trial or inter-trial interval. In trace conditioning, the system has to learn to use network resources to maintain that information, which every network did using memory cells. Finally, in extended delay

conditioning, the networks had to remember the US was past, at least for a little while.

It was intriguing to see the input-gate responses persist beyond the US in delay conditioning (*Experiment 1*, Figure 8-7A). In the interval timing literature, animals are often first trained on a mixture of normal rewarded delay intervals and unrewarded very extended delay probe trials similar to those in *Experiment 3* (Figure 8-11, top right panel). In these unrewarded probe trials, the CS is presented for an interval much longer than the usual rewarded fixed interval. This mixture of trials is called the peak-interval (PI) procedure (Buhusi & Meck, 2000; Roberts, 1981; Kirkpatrick-Steger, Miller, Betti, & Wasserman, 1996) and the probe trials are used to measure the animal timing precision. Since the context of the task presented here is different than the PI procedure (see section 8.8.4 for details), it was unexpected to see that a PI-like response curve could appear by itself, especially without noise or similar stochastic mechanism. The bell-shaped response curve is usually due to averaging and is explained by noise in SET theory. Individual trials have a relatively sharper 'low-high-low' pattern (Church, Meck, & Gibbon, 1994). It is interesting to see that this could be the result of an appropriately self-trained feed-back loop that generates the proper build-ups and leaks, and that this curve could adapt to different delays (as in *Experiment 4*, delay networks, group I). It would certainly be interesting to validate the model on the PI procedure. One could easily see how actor units receiving this input gate signal could lead to PI-type behavior.

In *Experiment 4*, the ramp activity changes in the memory blocks of trace networks and of delay networks (group I) showed relatively good scaling with time, an important property. But there were some issues with memory blocks from delay networks of group II who adapt to larger intervals by increasing their response range, a relatively unrealistic adaptation. Finding a way to limit the dynamic range of the memory cells property to get more stable and realistic representations could be necessary in the future. It is also difficult to assess the model potential as a neural model of timing under the current training procedure. LSTM networks are currently trained to reach an absolute time precision while animals are usually trained for a

fixed period of time before being evaluated for their timing precision. It would be important in the future to change the LSTM objective function to allow them to make temporal errors during training instead of forcing them to make predictions only at one specific time step. Such improvement would make it easier to compare LSTM's training time and performance to that of animals.

8.8.4 Resurgence of expectation

In *Experiment 3*, the model predicts a second rise in expectations following the reward for delay networks on extended and very extended probe trials and for extended delay networks on very extended delay probe trials. A precedent for this finding exists in the behavioral literature (Kirkpatrick-Steger et al., 1996; Sanabria & Killeen, 2007). Its causes and relation to animal timing are still under debate (Sanabria & Killeen, 2007).

Our model explanation of this behavior is simple. In delay networks, the presence or absence of the CS fully determines the timing build-up. When the CS disappears, it does not fully reset the accumulated time in the memory cell (Figure 8-7D), since this takes some extra time. This explains why, in Figure 8-11 (lower panels, two last columns), the output signals rise again rapidly without waiting for a full second. Extended delay networks have a slightly different behavior since they learn to fully reset the accumulated timing (probably to a small negative value, Figure 8-8D) sufficiently not to respond during the post-US interval equal to about 1s. But on very extended delay conditioning probe trials, they show resurgences of expectation past the usual 1s post-US interval (Figure 8-12, bottom right panel). This varies through the network population from immediate resurgence to a full 1s or more of waiting time before resurging. Again, we can suspect that the differences among networks depend on how much negative time they reset their cumulated time on US presentation.

The main experimental difference between the model predictions and the animals' data lies in the experimental protocol. In the interval timing literature, animals are often first trained on fixed interval rewarded trials (FI). Under this protocol, the animal receives a reward for the first lever press after a fixed delay from

CS onset, and this action also turns off the CS. After training, they usually begin to press the lever repeatedly before the end of the fixed interval from CS onset, hence the CS and US they observed from that point in training is similar to our delay conditioning paradigm. Animals are then trained on a mixture of 50% fixed interval trials, and 50% unrewarded very extended delay probe trials similar to those in *Experiment 3* (Figure 8-11, top right panel). In these unrewarded probe trials, the CS is presented for an interval much longer than the usual rewarded fixed interval. This mixture of trials is the basis of the peak-interval (PI) procedure (Buhusi & Meck, 2000; Roberts, 1981; Kirkpatrick-Steger et al., 1996). (Note that our model was not trained on these probes, it was only tested on them with learning rate fixed to 0, see 8.6.1.) Animal timing is then inferred by fitting a Gaussian to the response rate with respect to elapsed time from CS onset on the probe trials. From this mixture of trials, animals learn that if the reward does not come on time, then it will not come at all, and hence, learn to cease responding before the CS disappeared, once the fixed rewarded interval has elapsed on probe trials. Resurgence usually appears on very extended probes trials (Kirkpatrick-Steger et al., 1996; Sanabria & Killeen, 2007). An example of results under this protocol is shown in Figure 8-16.

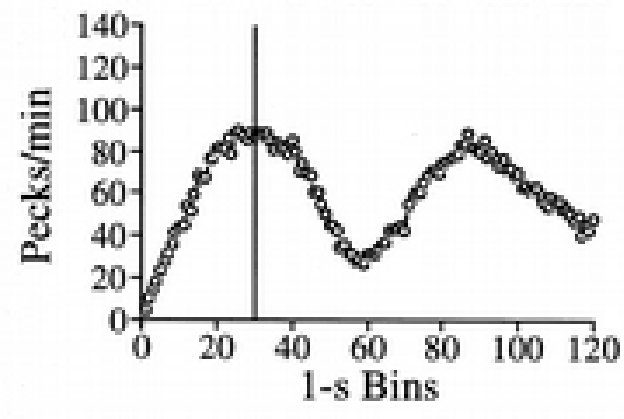


Figure 8-16 : Pigeons response rate with respect to elapsed time since CS onset on 120s unrewarded probe trials. Pigeons were initially trained on 30s FI protocol. They were then trained on a mixture of 30s FI trials and 120s unrewarded probe trials. Copyright © 1996 by the American Psychological Association. Reprinted with permission from (Kirkpatrick-Steger et al., 1996).

Pigeons learned to stop responding if not rewarded within 60s from CS onset through training, while our delay networks stop responding on rewarded very

extended delay probe trials (lower right panel, Figure 8-11 & Figure 8-12) because they actually receive the reward on time. Nevertheless, in both cases, the persistence of the CS induced a resurgence of responses (pigeons) or reward expectations (networks). When delay conditioning networks received unrewarded probe trials (Figure 8-11, second row, last two columns), similar to a probes in the PI protocol, they did not show a stop in response after the usual delay, nor did they show resurgence on very extended trials. These results should not be taken as a mere failure of the model to reproduce PI behaviors. Beyond the difference discussed in an earlier section (*8.8.1 Animals vs the model, conditioning and timing*), the networks were not trained on the usual 50% normal trials versus 50% probe trial protocol as in the PI protocol. They were only tested on probe trials with all their weights fixed (learning rate set to zero, see section 8.6.1). Getting animals to stop responding after the normal delay can be sensitive to the ratio of probes (Kaiser, 2008) and requires some training on probe trials to appear (Balci et al., 2009). More simulations are necessary to fully assess whether the model presented here could reproduce resurgence data and provide an interesting explanation.

Standard timing theory such as SET and LET cannot easely account for this type of data (see Lejeune & Wearden, 2006 for a list). The *Theory of Stochastic Counters* can account for this data (Killeen & Taylor, 2000). It does so by assuming the existence of a binary counter (a digital binary accumulator) and using the periodicity of low order bits. For a recent review of resurgence and relevant models, see (Sanabria & Killeen, 2007).

8.8.5 Timing and GAPs

The input-gate and memory-cell interaction in delay conditioning (Figure 8-11A,D) may also explain some behavioral results in gap paradigms. Gaps in timing occur when the CS is off for a brief period of time during delay conditioning. When the gap is very short, timing seems to be paused. When the gap is longer, timing seems to restart (Buhusi & Meck, 2000). Such behavior seems in agreement with the representation found in *Experiment 1* in delay conditioning. A lack of CS input for a small amount of time could have little effect on the build-up, but a long enough gap

could lead to the memory-cell to eventually reach its negative steady-state, equivalent to a full reset. More experiments would be needed to verify if the model could account for gap data. However, current results suggest that this is a situation worth testing.

8.8.6 Timing in trace depends on CS offset

Another prediction of the model appeared in *Experiment 3*. Under trace conditioning, instead of timing from CS onset, as it does for delay and extended conditioning, LSTM networks start timing on CS offset. This can be inferred from the trace network behavior on *Experiment 3* probe trials (Figure 8-10). Under delay, extended, or very extended probes trials (rewarded or not), where the CS is on for 1s, 2s, and 4s respectively (instead of 100ms for trace), the output of trace-trained LSTM networks (its US prediction) rises 800ms after the CS offset (i.e., after the end of what would be considered the trial interval). This behavior was completely unexpected. However, a precedent for this finding exists in the behavioral literature (Buhusi & Meck, 2000) in trained rats on a trace fixed interval procedure, similar to our fixed trace conditioning. They then trained them on a trace peak-interval procedure to evaluate the animals' timing ability. On the last few days of the experiment, they changed the length of the CS on the probe trials. The animals' behavior in those probes support the idea that the animals were timing the interval not from the CS onset, but mainly from the CS offset, as we see found in our LSTM trace networks.

This revealed that the LSTM network learned to recognize the most useful predictive feature of the CS across all conditioning conditions. In delay and extended paradigms, that is the onset of the CS, because the offset is either coincident with (delay) or long after (extended) the presentation of the US and so has little or no predictive power. The onset of the CS could also serve as the timing cue in trace conditioning as well. However, the LSTM network settled on a solution that used the end of the CS as the cue to begin predictive timing. This might have enhanced the precision timing by reducing the duration of the time interval to that of the trace (memorized delay) itself.

8.8.7 LSTM limitations revealed

Compared to animals, the LSTM networks were relatively slow at learning timing (Rivest et al., 2010a). In particular, it seems that smaller time steps increase the learning difficulty. An important limiting factor for the LSTM network as a model of interval timing is its current learning scheme. The LSTM network learns to predict the next stimulus by minimizing the error at each time step individually. This is why we had to train networks until they learn the precise timing of events. This also means that for a stimulus presented for a single time step, it is better for the LSTM to predict nothing, than to erroneously predict the event a time step earlier or later. Results showed that LSTM networks cannot always easily adapt to changes in the environmental dynamics such as changes in velocity. If the constant velocity of some environmental process changed, this would lead to different timing between events. Many networks did not succeed in adapting to small changes in delay (*Experiment 4*). Allowing them to have a smoother learning criterion based on the temporal profile instead of the individual steps would certainly be a major improvement to their architecture.

Finally the unbounded range of the memory cells could be a problem. Their linearity is a crucial property behind the LSTM learning ability and computational property. However, their unbounded nature caused problems in trace conditioning. With two memory blocks in (Rivest et al., 2010a), a number of memory blocks had to be removed from analysis because their memory cell activity kept increasing during the training block, making those blocks computationally ineffective. It can also lead to a less physiologically realistic representation of time (e.g., *Experiment 4*, delay networks, memory blocks group II). Finding a way to limit the range of those cells without reducing the LSTM computational learning power would be an asset for future modeling work.

8.8.8 Conclusion & future work

The goal of these experiments was to evaluate the LSTM-TD model's (Rivest et al., 2010a) explanatory and predictive power by analyzing its signals under

variations of the conditioning paradigms and to assess its validity as a model of timing and dopaminergic data. To do this, we used the model to suggest answers to a number of scientific questions.

With respect to dopamine, the model successfully reproduced dopaminergic data and made a number of interesting predictions. Among them, the model predicts that dopaminergic response should be very similar after training whether recorded under trace or delay conditioning paradigm, a hypothesis that still requires more neurophysiological experiments to be confirmed. Modifications of the LSTM to TD connectivity should also improve the model's predictive and explanatory power. But there are still dopaminergic timing data the model cannot account for.

The model's encoding of the task states in trace, delay and extended delay conditioning provides clear and interesting explanations to various behaviors. The model is easy to analyze with a single memory block. The finding that the model showed behavior related to resurgence, gap, and trace from CS offset, data that other current models have trouble explaining, reveals the predictive power that this model may have in interval timing. Moreover, the model's neural activity is very similar to activity found in the cortex during fixed delay periods (Funahashi et al., 1989; Romo et al., 1999; Komura et al., 2001; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Lucchetti et al., 2005; Lebedev et al., 2008) and also makes predictions about how this activity could vary depending on the conditioning paradigm used. The model's generality makes it an interesting candidate for intrinsic timing theory (Ivry & Schlerf, 2008), as opposed to models of timing that propose dedicated special-purpose clock mechanisms. Further investigations of the model in a more animal like set-up (with actions) is required, but is certainly worthwhile in the light of the results presented here.

Three elements could be improved in the model. First, the LSTM to TD connections can be improved to allow TD to receive and use all the information the LSTM can provide it. This should allow better dopaminergic predictions as well as a clearer explanation of the dopaminergic data. Second, the model could provide even more neural-like results if the LSTM architecture or learning algorithm were

modified to favor memory cells with more limited dynamic range. This would favor more adaptation to different delays through change in slopes and eliminate memory cell with unbounded range (Rivest et al., 2010a). The model would therefore behave in a more neurophysiologically realistic way. Finally, finding a way to train the LSTM networks in a less rigid context than single-step precision would allow much better comparison of performance between the model and animals.

Appendix A

Changes from (Rivest et al., 2010a)

The new simulation was done at 10Hz instead of 5Hz, did not use probe trials during training (only probe trials on frozen networks after training), and used different α_{LSTM}, β, λ_{LSTM}, and α_{TD} hyper-parameters values.

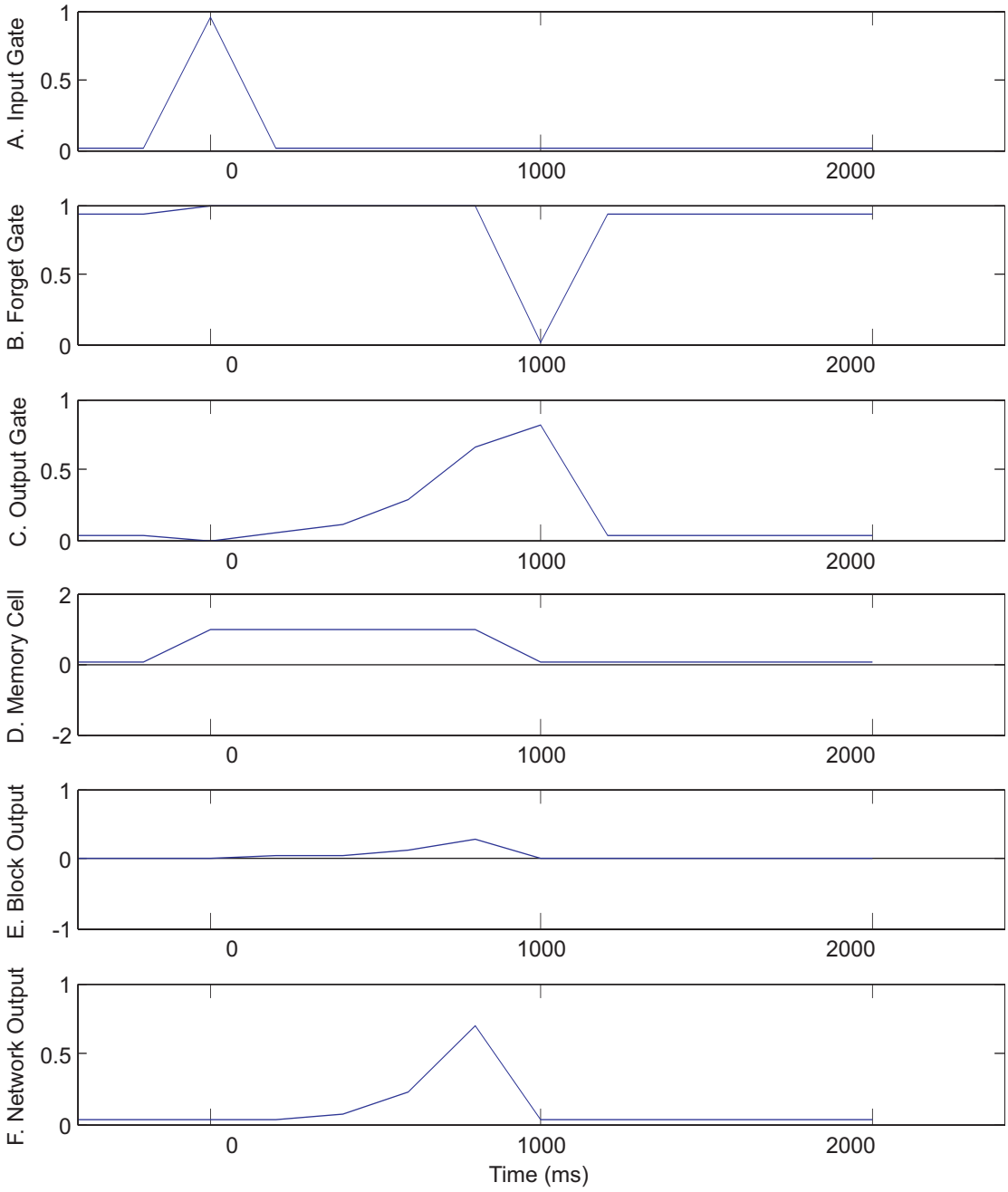
A grid search to optimize α_{LSTM}, β, and λ_{LSTM} (the LSTM eligibility trace factor from previous model) on all three tasks (*Experiment 1*) revealed no significant improvement in learning speed or success rate for the near optimal (α_{LSTM}, β) pair (λ_{LSTM} had not been thoroughly evaluated in (Rivest et al., 2010a)). Therefore, we eliminated the LSTM eligibility trace learning mechanism ($\lambda_{LSTM} = 0$) and selected the best hyper-parameters.

Second, while the preceding version used two memory cells per memory blocks, we found no reason to keep them. The memory cells within the same memory block were usually proportional to each other. Multiple memory cells are more appropriate when something requiring multiple bits needs to be stored in working memory.

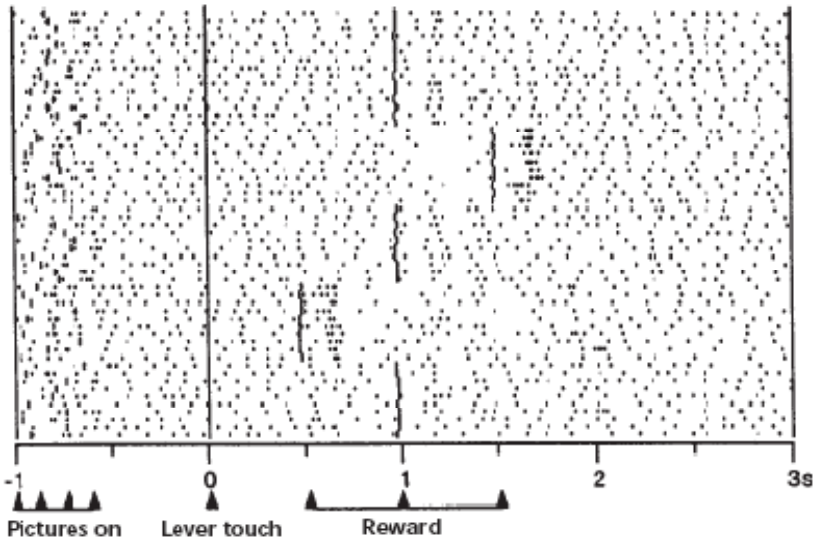
Acknowledgements

This manuscript profited from the comments of James Bergstra. F.R. was supported by doctoral studentships for collaborative projects from the Groupe de recherche sur le système nerveux central (FRSQ). Y.B and J.K. were supported by the CIHR New Emerging Team Grant in Computational Neuroscience, by individual operating grants from the CIHR (JK) and by an infrastructure grant.

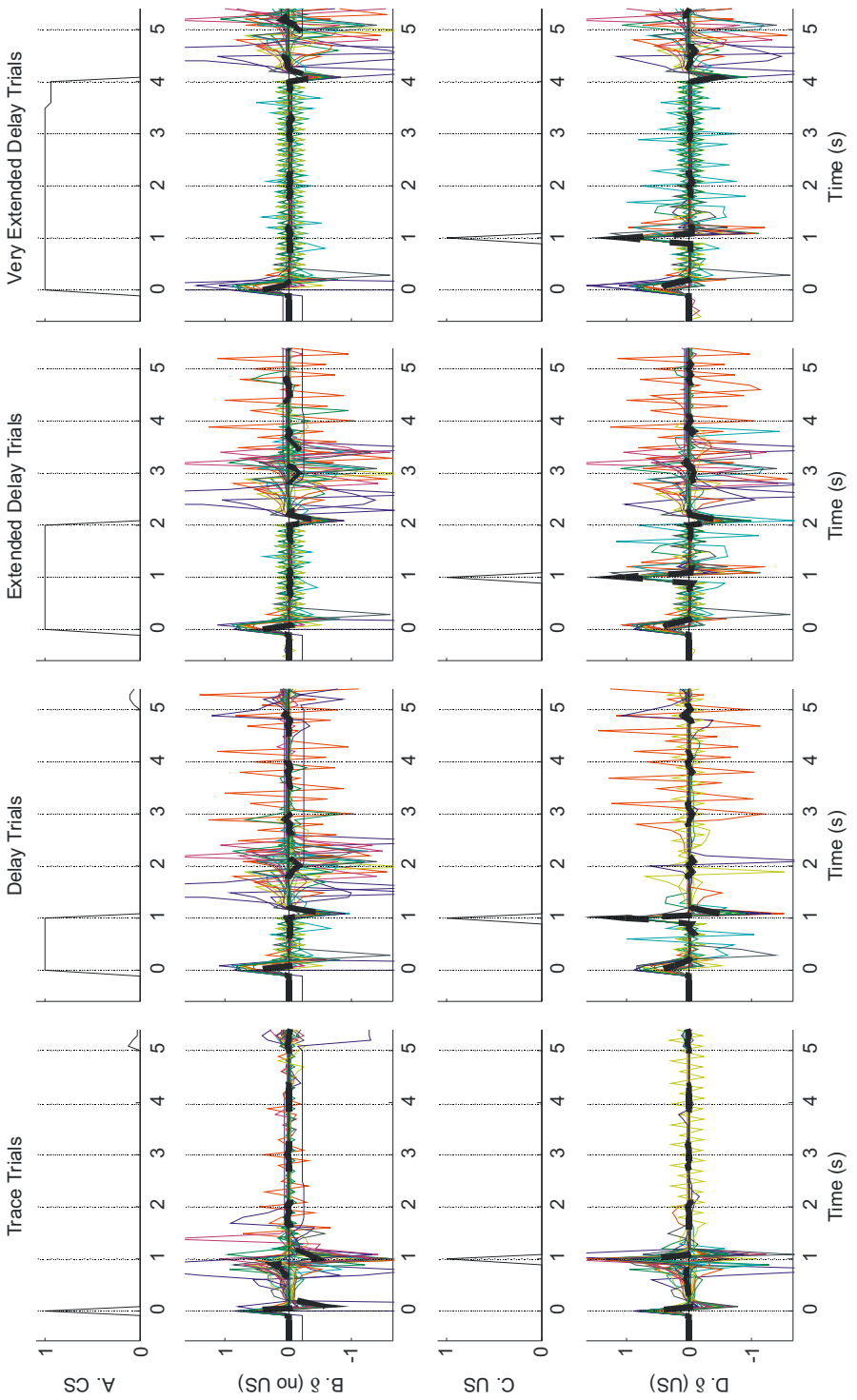
Supplemental Tables and Figures



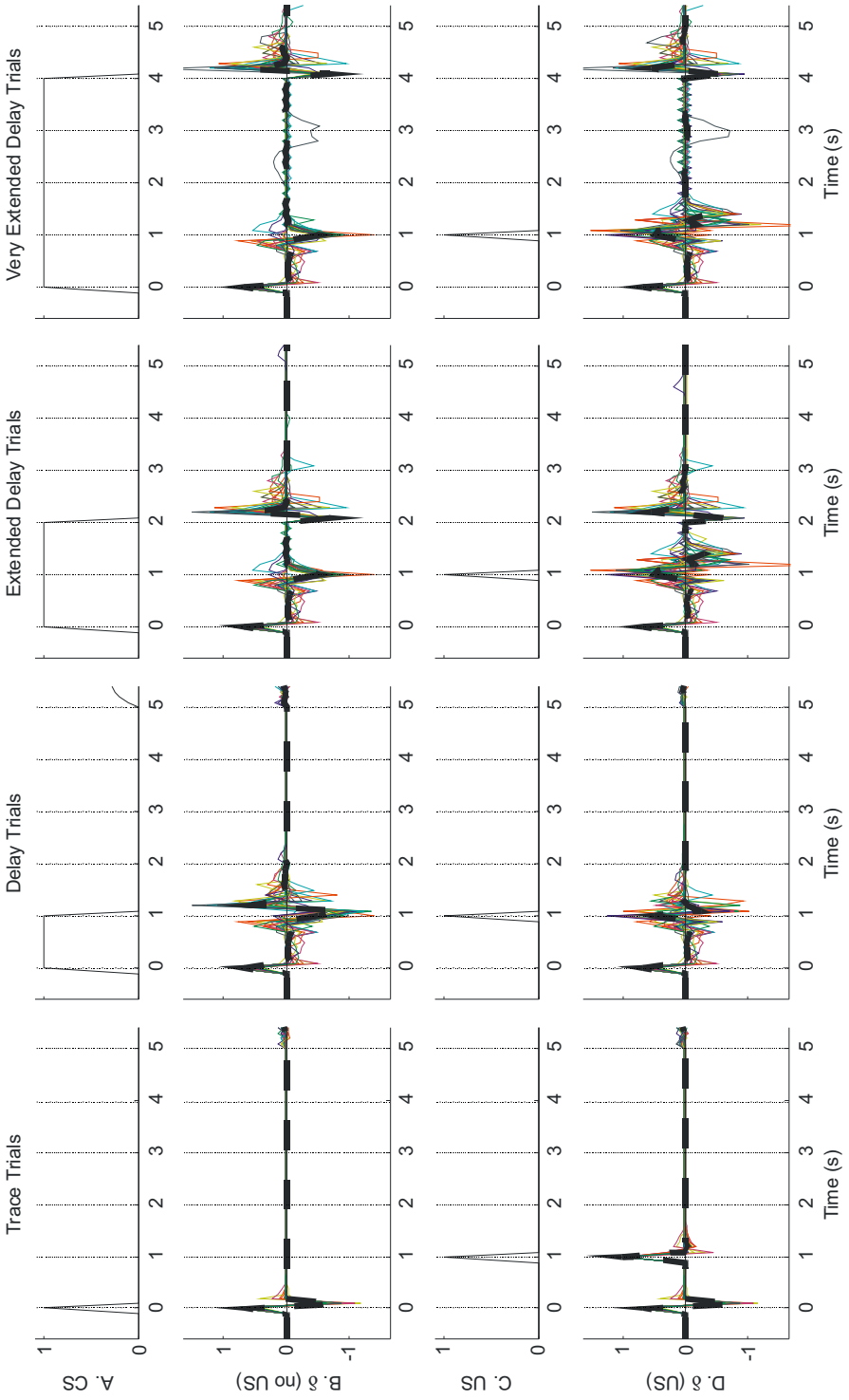
Supplemental Figure 8-17: Typical time representation within an LSTM network under trace conditioning. Data from (Rivest et al., 2010a).



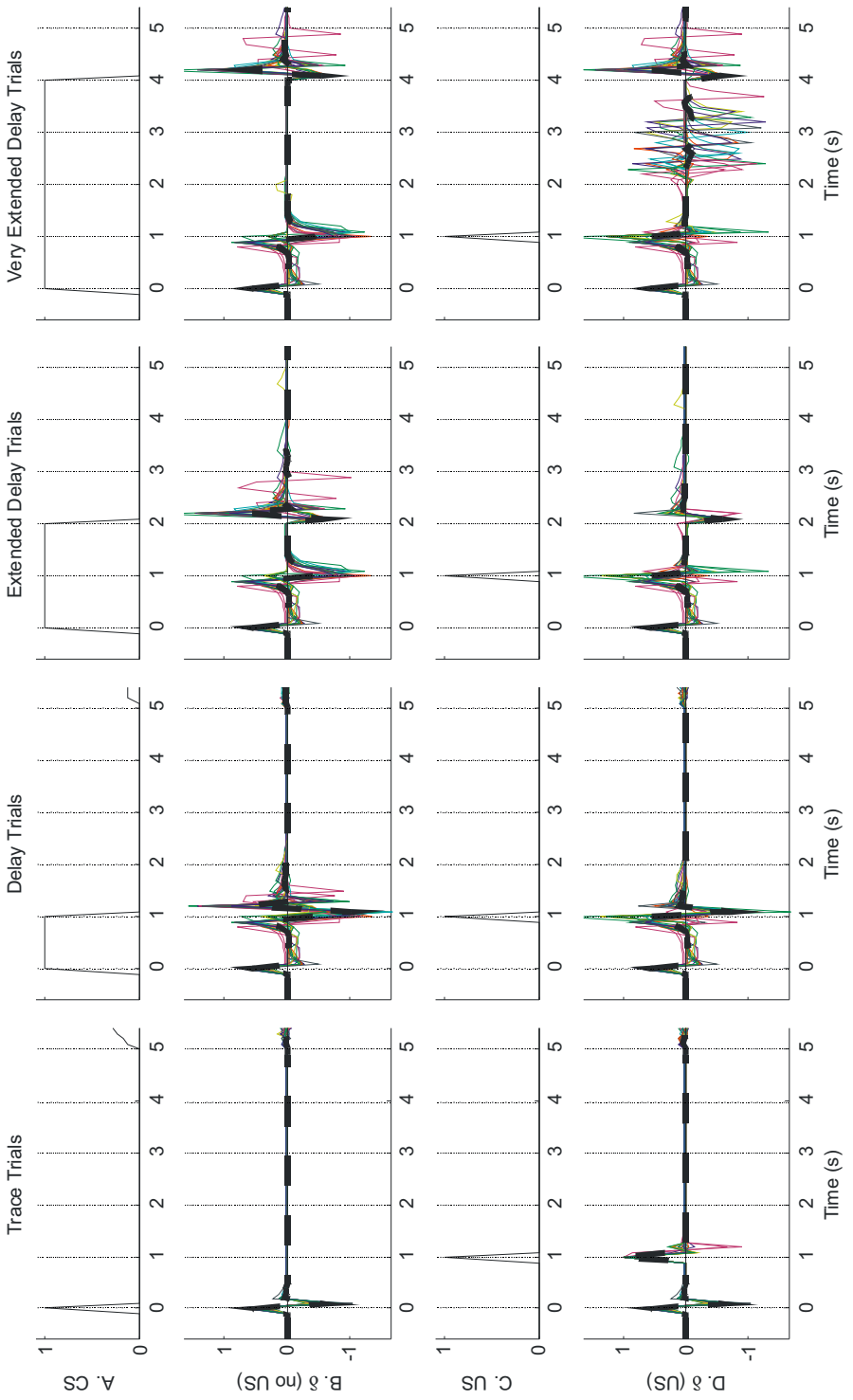
Supplemental Figure 8-18: Response of dopamine neurons on probe trials with different delays. Top, middle and bottom sections show DA activity on normal (1s) trials. The second section shows DA activity on late trials, when the delay is longer than usual (1.5s). The fourth section shows DA activity on early trials, when the delay is shorter than usual (.5s). On late trials, there is a depression at the expected time of reward (1s) and a burst of activity when reward is finally received. On early trials, there is only a burst when reward is unexpectedly received. Reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience (Hollerman & Schultz, 1998, Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat.Neurosci.* 1:304-309), copyright (1998).



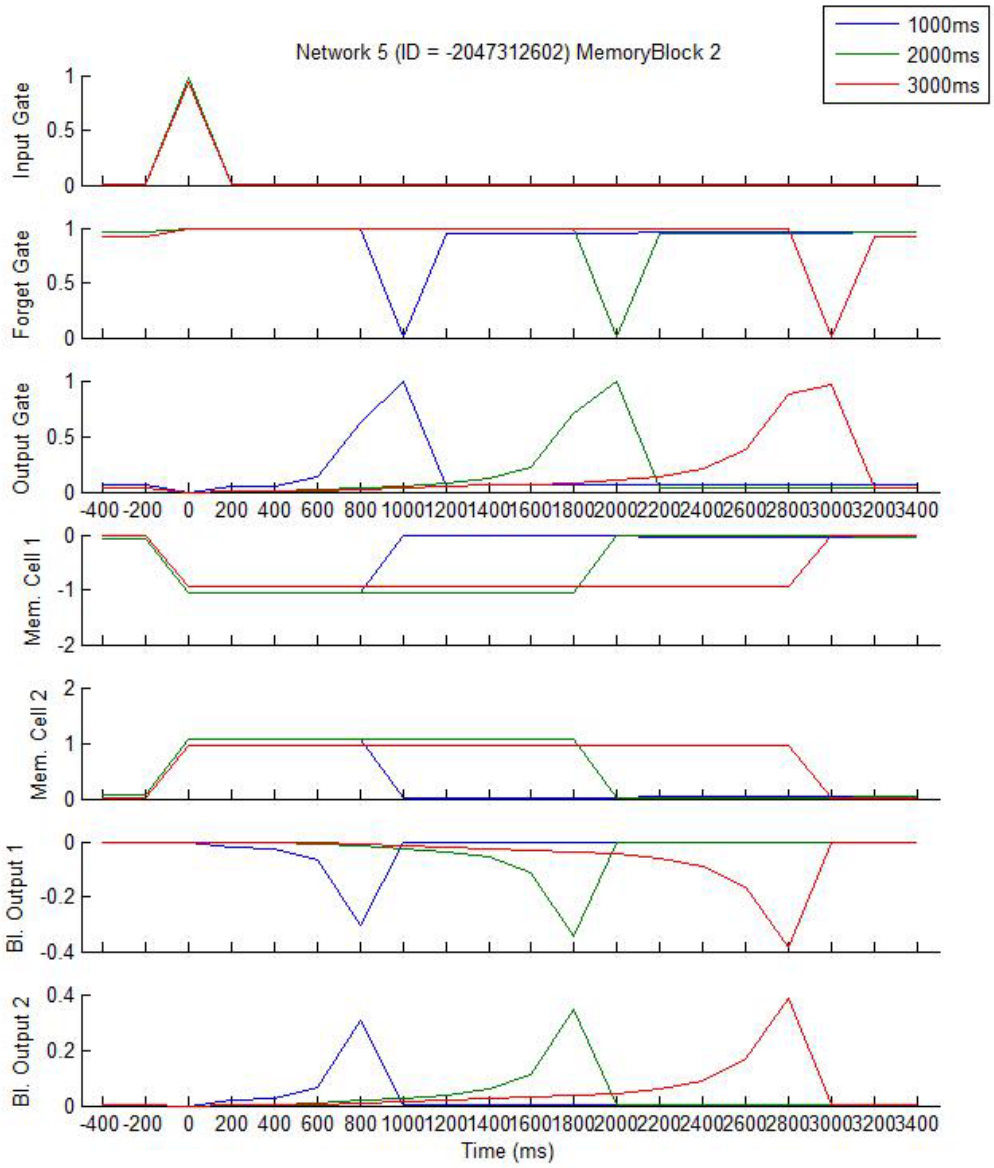
Supplemental Figure 8-19: Mean TD δ for trace networks on probe trials. Each column represents a different type of probe trials. Rows are aligned in each column. The first row is the CS, the second is the δ signal on unrewarded probes, the third is the US, and the forth the δ signal on rewarded probes. Each line represents a different network; the wide dashed line represents the networks population average.



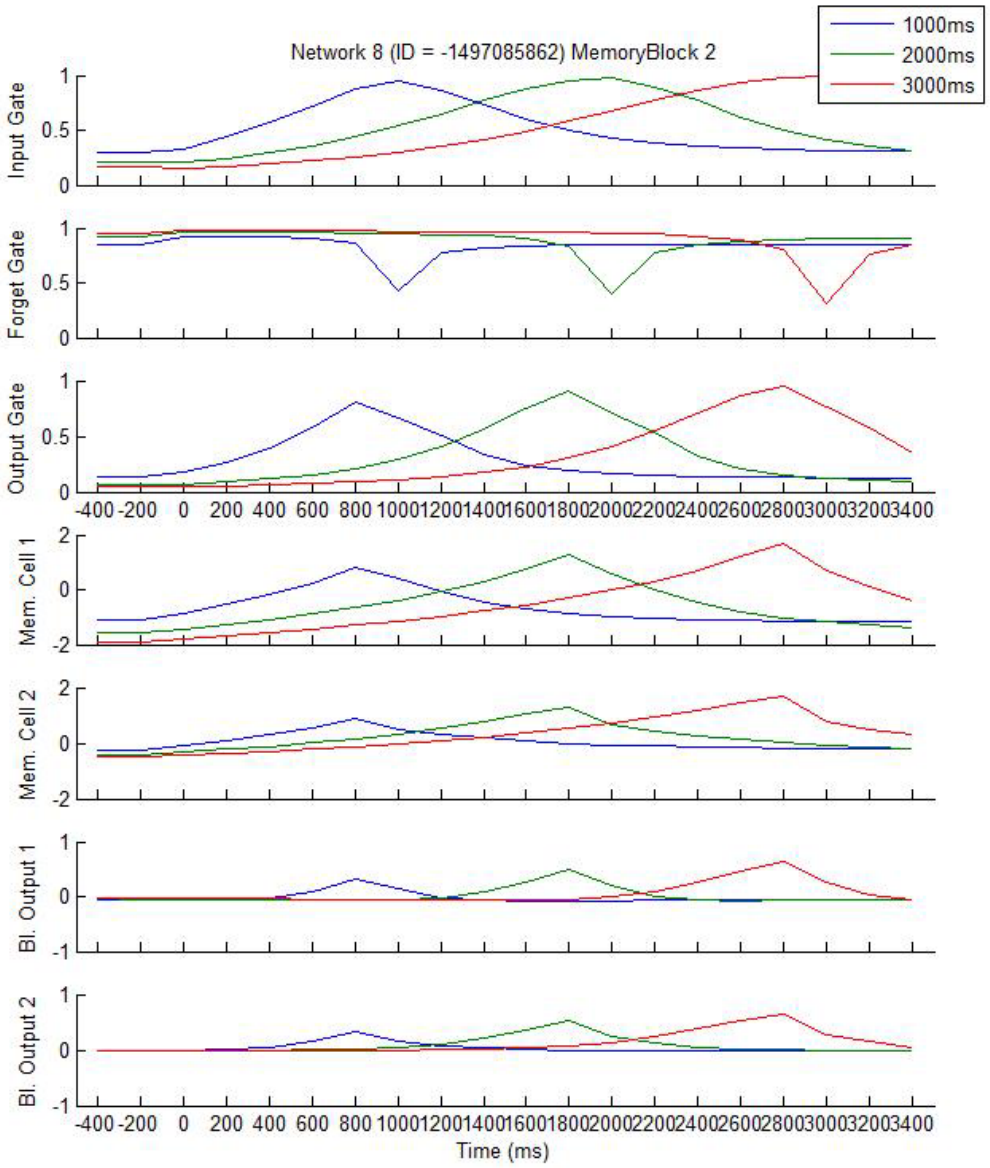
Supplemental Figure 8-20: Mean TD δ for delay networks on probe trials. Same format as in Supplemental Figure 8-19.



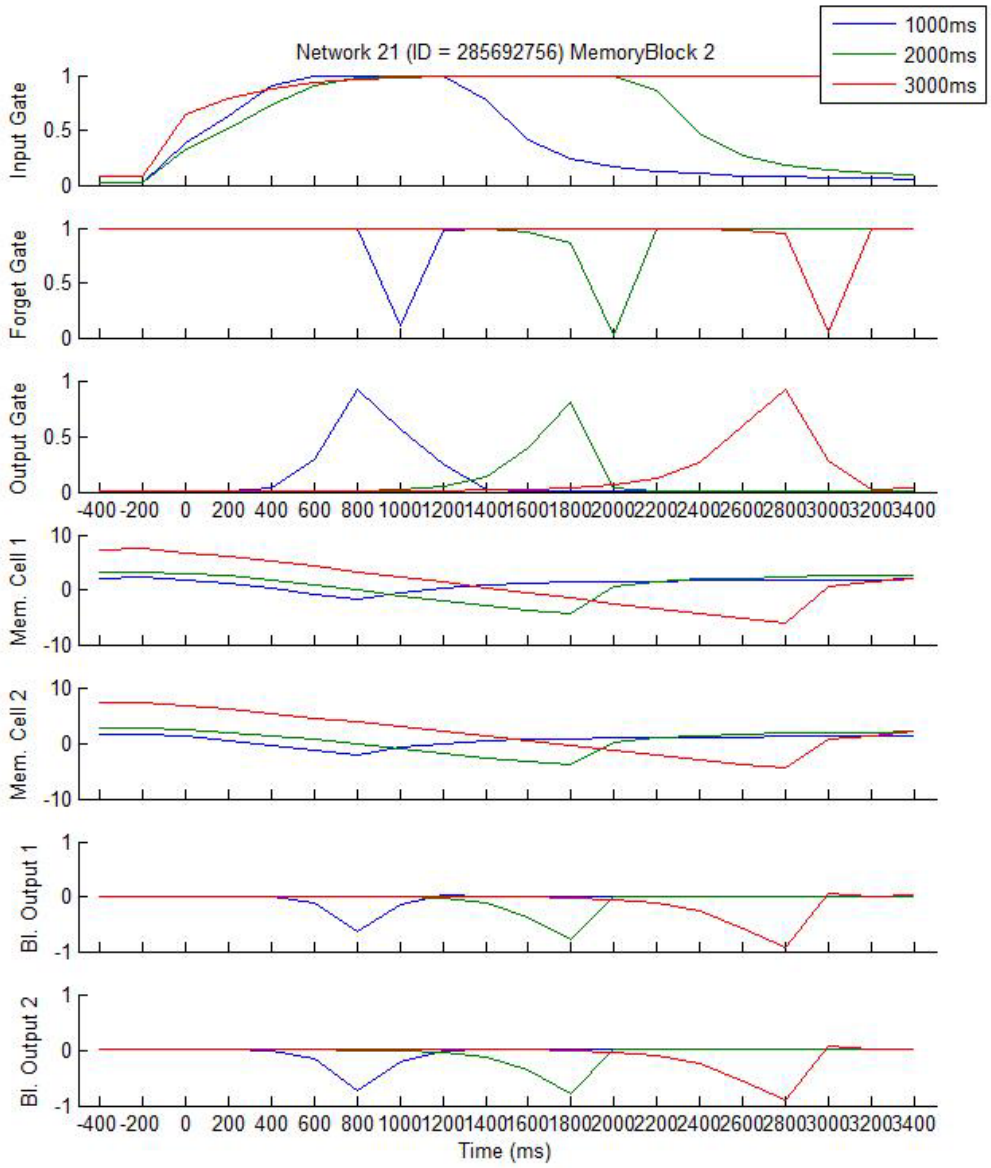
Supplemental Figure 8-21: Mean TD δ for extended delay networks on probe trials. Same format as in Supplemental Figure 8-19.



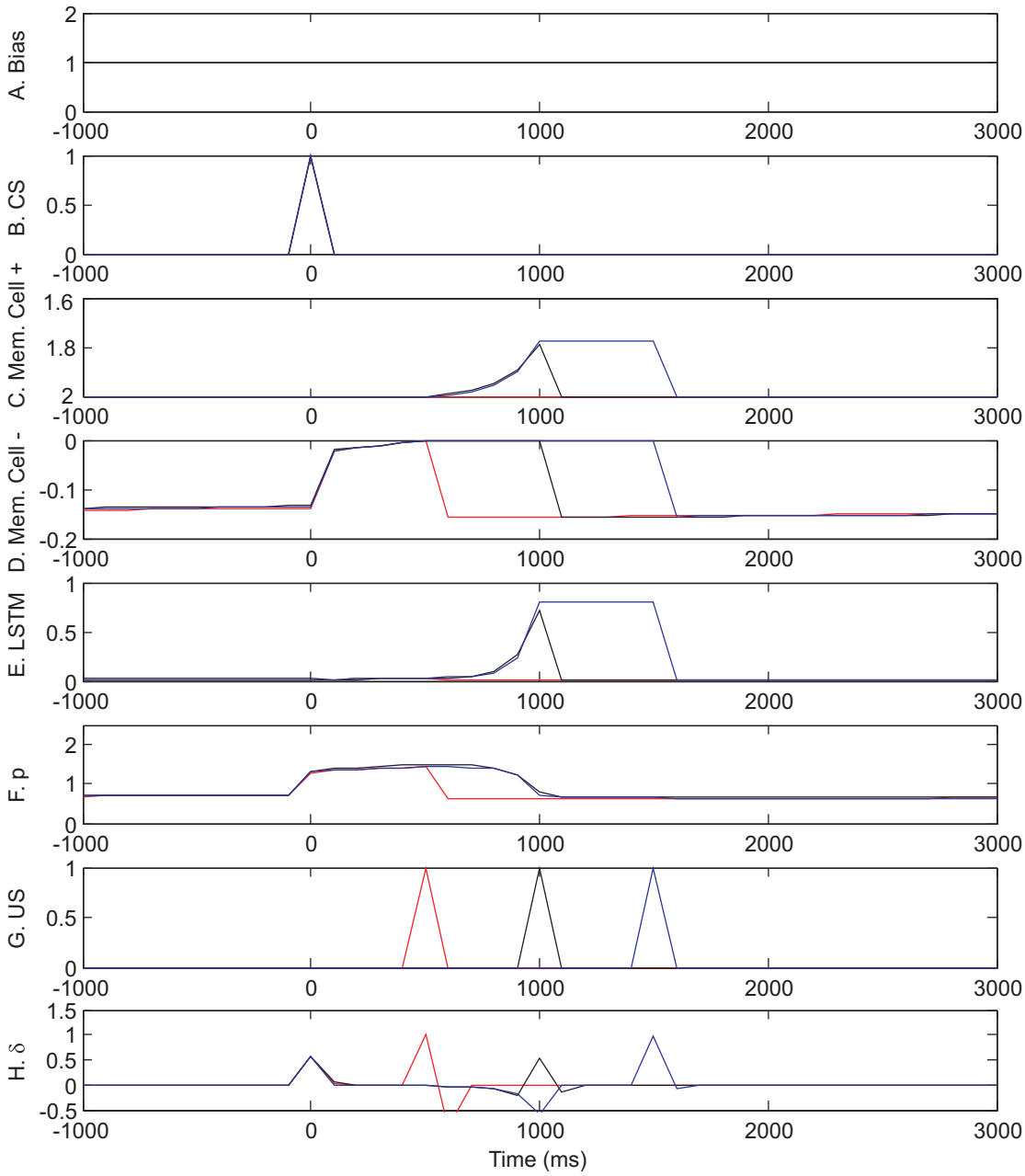
Supplemental Figure 8-22: Example of signals from a memory block of a trace network in *Experiment 4*.



Supplemental Figure 8-23: Example of signals from a memory block of a delay network adapting its input gate signal (group I) to longer time intervals in *Experiment 4*.



Supplemental Figure 8-24: Example of signals from a memory block of a delay network adapting its memory cell range (group II) to longer time intervals in *Experiment 4*.



Supplemental Figure 8-25: Example showing every signal involved in the computation of *p* and *δ* in trace conditioning on early (red), normal (black) and late trials (blue), using a slightly different LSTM to TD connectivity scheme. With this scheme, the contribution of the LSTM representation to the TD expectations and error signal *δ* is much simpler to understand. When CS appears, the TD network uses the memory cell sustained activity (D) to increase its expectation (F); this generates the *δ* (H) CS responses. On early trials (red), the unexpected reward (G) causes the *δ* burst on US. On normal trials (black), the LSTM prediction (E) (or memory cell build-up C) allow a nice cancellation of the expectations (F) and the reward (G), resulting only in a small *δ* (H) burst on US. Finally on late trials (blue), the LSTM predictions (E) cancel the expectations (F) caused by the sustained memory cell (D). This generated the *δ* (H) depression at the expected reward at 1s.

Chapitre 9. Discussion générale

Cette discussion débute par une revue des objectifs de cette thèse, de la contribution des articles ainsi que de leurs limites. La section suivante porte sur les détails des ganglions de la base pour lesquels plus d'information permettrait d'améliorer les modèles et notre compréhension. La troisième section traite des représentations abstraites dans le cortex suivi d'une section sur les représentations temporelles. Finalement, la dernière pose un regard global sur l'apprentissage dans le cerveau et énumère quelques facteurs importants pour guider les recherches futures.

9.1 Modèle développé dans cette thèse

L'objectif de cette thèse était de mieux comprendre le développement de représentations dans un contexte d'apprentissage par renforcement dans le cerveau à l'aide de modèles informatiques. Les ganglions de la base et le système dopaminergique semblent au cœur du phénomène d'apprentissage par renforcement (Chapitre 4) et le cortex est le siège de nombreuses représentations abstraites (section 3.3.1). Par conséquent, cette thèse se limite à l'étude de ces deux systèmes et à leurs interactions potentielles dans l'apprentissage. Deux de ces interactions ont été étudiées : soit l'utilisation de la représentation corticale comme source d'information pour les ganglions de la base (Figure 9-1, flèche rouge) et le rôle possible de la rétroaction mésocorticale dopaminergique (le signal d'erreur supposé d'apprentissage par renforcement) dans l'apprentissage du cortex (longue flèche bleu à gauche). La rétroaction striatopaladothalamocorticale (flèches oranges), jouant possiblement un rôle dans les actions, n'a pas été étudiée.

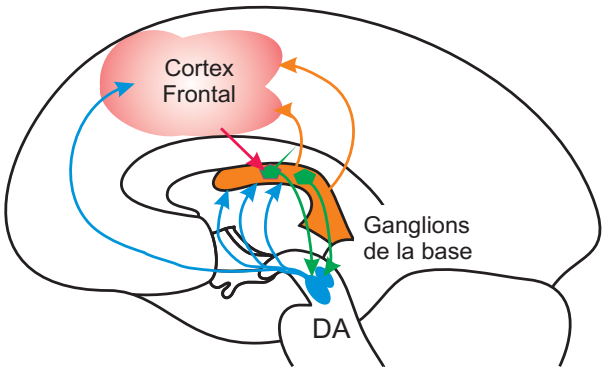


Figure 9-1 : Schéma de connectivité simplifiée entre le cortex et les ganglions de la base.

La première difficulté dans le fait d'apprendre les représentations abstraites de l'environnement en même temps que l'estimation des récompenses à venir, c'est que les algorithmes d'apprentissage par renforcement offrent peu de garanties de convergence dans cette situation (Baird, 1995; Tsitsiklis & VanRoy, 1996; pour de nouvelles approches, voir Mahadevan & Maggioni, 2007; Parr, Painter-Wakefield, Li, & Littman, 2007). Le premier article de cette thèse (Chapitre 6) (Rivest et al., 2005) évalue donc de façon empirique différentes combinaisons d'apprentissage de représentations servant d'entrées supplémentaires à un réseau acteur-critique TD d'apprentissage par renforcement. Dans cette évaluation sommaire (sur une seule tâche), toutes les formes d'apprentissage de représentations ont obtenu de meilleures performances que l'utilisation d'une représentation fixe générée aléatoirement. L'utilisation d'un algorithme non supervisé et complètement indépendant de TD pour apprendre une représentation a généralement diminué les performances en début d'entraînement. Mais l'utilisation du signal d'erreur de TD pour moduler l'apprentissage de représentations a non seulement réglé ce problème, mais a mené à de bien meilleures performances. L'article montre qu'un algorithme d'apprentissage par renforcement peut profiter de l'apprentissage d'une représentation à ses entrées pendant son propre apprentissage dans certaines situations. Il montre aussi que l'utilisation de la rétroaction du signal d'erreur de TD vers la représentation favorise la synchronisation des deux apprentissages parallèles. Les performances de la meilleure combinaison étaient relativement bonnes considérant la capacité limitée du système (nombre de paramètres adaptatifs). L'évaluation de l'apprentissage concurrent de représentations et de la maximisation de récompenses faite dans cet article reste cependant sommaire. L'idée d'utiliser le signal d'erreur de la sortie pour guider l'apprentissage non supervisé d'une représentation intermédiaire est toujours d'actualité (Sussillo & Abbott, 2009).

Pour mieux comprendre l'interaction du cortex et des ganglions de la base, il est utile de trouver la tâche la plus simple possible exigeant d'apprendre une représentation et pour laquelle des données neurophysiologiques sont disponibles. Le conditionnement classique de trace, bien que d'une extrême simplicité, demande une

représentation suffisamment riche pour différentier la période interstimuli de la période interessais. Les données dopaminergiques (Hollerman & Schultz, 1998) sur des délais interstimuli inattendus contiennent suffisamment d'information pour être modélisées et pour inférer en partie le modèle interne de l'animal (Daw et al., 2006), en plus de démontrer que celui-ci doit avoir une représentation du passage du temps pendant la période interstimuli. L'absence d'action dans cette tâche permet d'éliminer l'aspect moteur (ou acteur) du modèle des ganglions de la base, la projection vers le thalamus. Puisque cette projection mène aussi au cortex, on réduit ainsi le nombre d'interactions à l'étude entre le cortex et les ganglions de la base. Cependant, la représentation du temps dans le cerveau reste source de débats (Ivry & Spencer, 2004; Buhusi & Meck, 2005). Mais, certains travaux suggèrent que cette représentation puisse être en partie acquise ailleurs dans le cortex (voir section 3.3.1).

Le second article (Chapitre 7) (Rivest et al., 2010a) vient combler ce vide en proposant un modèle capable d'apprendre une représentation temporelle dans le cortex, source d'information pour l'apprentissage par renforcement dans les ganglions de la base. Le modèle reproduit les effets dopaminergiques classiques déjà simulés par TD (par exemple, Suri & Schultz, 1999; Daw et al., 2006), en plus de pouvoir apprendre la représentation temporelle nécessaire à la reproduction de ces données (Hollerman & Schultz, 1998), une première dans le contexte d'un modèle entièrement naïf. La contribution de cet article est d'autant plus pertinente que les activités de certains neurones du LSTM ressemblent beaucoup à celles d'enregistrements électrophysiologiques de neurones dans le cortex (Niki & Watanabe, 1979; Funahashi et al., 1989; Romo et al., 1999; Lucchetti & Bon, 2001; Brody et al., 2003; Leon & Shadlen, 2003; Reutimann et al., 2004; Lucchetti et al., 2005; Lebedev et al., 2008). Le modèle semble donc favoriser l'idée que les représentations temporelles puissent être apprises dans le cerveau, là où elles sont nécessaires. Le débat sur les représentations temporelles dans le cerveau et leurs structures neurologiques sous-jacentes est toujours vivant au sein de la communauté (Ivry & Schlerf, 2008). L'article montre aussi que si le signal dopaminergique phasique modulait le taux d'apprentissage du cortex de façon additive, alors il permettrait d'accélérer

l'apprentissage d'un modèle de l'environnement par ce dernier. Des travaux récents ont finalement montré que ce signal ne fait pas que corréler avec le signal d'erreur TD, mais qu'il est nécessaire à certaines formes d'apprentissage (Zweifel et al., 2009). D'autres études ont aussi montré un effet de la dopamine sur la plasticité du cortex (Otani et al., 2003). Le modèle reproduit aussi le fait que le conditionnement classique soit plus facile que le conditionnement de trace et suggère que l'historique d'apprentissage (ou le curriculum) puisse influencer le modèle interne final de l'animal (voir aussi, Rivest & Precup, 2003; Krueger & Dayan, 2009; Bengio, Louradour, Collobert, & Weston, 2009).

Le modèle a toutefois de grandes difficultés d'apprentissage et semble beaucoup plus lent que les animaux. Au moins cinq raisons peuvent expliquer cette différence. Premièrement, d'autres représentations temporelles très riches existent déjà chez les animaux suite à leur vécu avant l'expérimentation alors que le modèle commence son entraînement totalement naïf, sans aucune expérience. Deuxièmement, le cortex a surement une bien meilleure capacité d'apprentissage que les quelques unités des réseaux entraînés. Troisièmement, le retour dopaminergique pourrait jouer d'autres rôles plus directs de contrôle sur l'activité corticale tel que le *gating* (Braver & Cohen, 2000; Montague et al., 2004; Rougier et al., 2005). De plus, le retour du signal d'action pallidothalamocortical pourrait aussi jouer un rôle important dans l'apprentissage cortical. Ces deux options n'ont pas été modélisées. Quatrièmement, il est possible que l'hippocampe joue un rôle important comme forme de mémoire épisodique à court terme dans le conditionnement de trace (Clark & Squire, 1998; Beylin et al., 2001), bien que certaines expériences semblent indiquer le contraire (Thibaudeau et al., 2007). Cinquièmement, il se peut fort bien que même si une représentation temporelle de la tâche est éventuellement développée dans le cortex, des mécanismes plus rapides et plus spécifiques au synchronisme, dans le cervelet par exemple (Ivry & Spencer, 2004), soient la source de l'apprentissage très rapide des animaux (Balsam et al., 2002).

Un apport majeur du modèle présenté ici est de permettre d'étudier comment une mémoire de travail générale (modélisée par le LSTM) peut être utilisée pour

représenter le temps dans différentes conditions. On peut ainsi faire des prédictions dans le but de déterminer si le cerveau utilise des fonctionnalités similaires pour représenter le temps. Il permet aussi de prédire l'activité corticale et dopaminergique dans de nouvelles situations en fonction de l'entraînement reçu. Il est donc important de vérifier si le modèle permet de faire des prédictions intéressantes en plus de permettre une explication simple des différents phénomènes naturels qu'il modélise.

Le troisième article (Rivest et al., 2010b) de cette thèse étudie donc la représentation qui se développe dans les LSTM sous trois paradigmes de conditionnement différents. Il étudie aussi comment s'adaptent ces représentations en fonction de la durée interstimuli qui doit être évaluée. Ces expériences ont permis de montrer que le modèle peut être relativement facile à analyser et fournir des explications simples sur les façons dont un réseau de neurones ressemblant à une mémoire de travail peut apprendre à prédire la structure temporelle de différentes tâches. L'article contient, entre autres, des prédictions sur les réponses corticales et dopaminergiques que l'on devrait retrouver chez un animal entrainé dans différentes situations de conditionnement de trace, classique et à délai étendu. Le modèle fait aussi des prédictions intéressantes ayant des antécédents dans la littérature sur les intervalles de temps. Finalement, ces travaux suggèrent trois pistes pour améliorer les capacités d'explication et de prédiction du modèle : améliorer la connectivité du LSTM vers le modèle TD, contrôler le champ de réponse des cellules mémoires du LSTM et trouver une règle d'apprentissage moins rigide permettant des erreurs de précisions temporelles.

Bref, le modèle apporte une contribution à notre compréhension de l'interaction entre le développement possible de représentations abstraites, en particulier temporelles, dans le cortex et l'apprentissage par renforcement par les ganglions de la base et le système dopaminergique.

9.2 Les ganglions de la base et le système dopaminergique

Malgré les avancées majeures reliées aux modèles informatiques de l'apprentissage par renforcement, dont les nombreuses données dopaminergiques reproduites à l'aide de ses modèles (voir section 4.4, Chapitre 7 et Chapitre 8),

plusieurs questions restent à éclaircir au niveau des ganglions de la base et du système dopaminergique qui permettrait le développement de modèles plus complets et plus précis.

Plusieurs détails sur l'implémentation d'un algorithme du style TD par les ganglions de la base restent à éclaircir. Premièrement, seule la moitié des neurones dopaminergiques enregistrés dans le VTA seraient réellement des neurones produisant de la dopamine (Margolis, Lock, Hjelmstad, & Fields, 2006). Les implications que le signal d'erreur de prédiction de récompense ne soit pas codé par les neurones dopaminergiques ou qu'il soit aussi codé par un second groupe de neurones n'ont pas encore été étudiées. Si les neurones dopaminergiques encodent réellement le signal d'erreur de prédiction de récompense, alors il reste à déterminer précisément comment le calcul de la différence temporelle est implémenté par les neurones dopaminergiques et leurs afférences (sections 4.1.6 et 4.4.2.b). La source d'inhibition pourrait provenir du striatum (section 4.4.2.b) autant que de l'habénula (Matsumoto & Hikosaka, 2007). Le rôle des projections corticales vers les neurones dopaminergiques reste lui aussi incertain (sections 4.1.6 et 4.4.2.b). La question suivante est : quel algorithme les ganglions de la base implémentent-ils (section 4.4.2.d)? Certaines expériences visant à élucider cette question favorisent une combinaison acteur-critique (Roesch et al., 2007) alors que d'autres suggèrent plutôt un algorithme comme SARSA ou Q-learning (Morris et al., 2006). Le rôle exact des autres éléments comme le STN (noyau sous-thalamique) dans ces modèles, s'il y en a un, reste lui aussi à déterminer (sections 4.1). Par exemple, les voies de sorties passant par le *globus pallidus* (GP) reçoivent aussi la dopamine et sont probablement aussi sujettes à la plasticité synaptique (sections 4.1). Ces voies de sorties pourraient très bien jouer un rôle important dans l'apprentissage des bonnes actions ou d'autres détails moteurs (section 4.4.2.c). Le modèle développé dans cette thèse s'arrête pour l'instant à une seule unité linéaire (p) représentant un groupe de neurones possiblement striataux ainsi qu'à un signal d'erreur de récompense (δ) corrélant les réponses typiques d'une majorité de neurones dopaminergiques.

Un autre problème intéressant à résoudre est le lien entre drogue, dopamine, et apprentissage. Le modèle TD tel que présenté ici expliquerait bien les problèmes de dépendances liés à la consommation de cocaïne si celle-ci affectait le signal dopaminergique phasique (variation rapide suite à un stimulus) et la plasticité corticostriatale. En supposant que la drogue a l'effet d'un signal d'erreur positif dans TD, alors l'estimé de la récompense associée à la consommation de drogue augmente à chaque fois. De la même façon, les actions menant à sa consommation sont toujours renforcées. Cependant, la cocaïne affecte principalement le niveau tonique de la dopamine (niveau dopaminergique de base sur une plus longue intervalle de temps) et son effet de dépendance passe principalement par la plasticité au niveau du VTA et non du striatum (Kalivas & Alesdatter, 1993). De plus, le niveau tonique de la dopamine ne semble pas être lié à l'apprentissage en général (Zweifel et al., 2009). Les signaux toniques et phasiques pourraient aussi être reliés à la saillance incitative (Berridge & Robinson, 1998; Montague et al., 2004), c'est-à-dire qu'ils pourraient inciter à agir (voir section 4.4.6). Ce rôle possible de la dopamine sur les actes n'a pas non plus été considéré dans cette thèse, en particulier parce que le modèle présenté ici (à l'exception de celui du Chapitre 6) est passif (il n'agit pas). Quoi qu'il en soit, bien qu'il y ait un lien certain entre les signaux dopaminergiques toniques et phasiques, les problèmes de consommation et la théorie de l'apprentissage (TD), la nature exacte de ce lien reste à déterminer.

D'autre part, la dopamine pourrait aussi affecter la perception du temps (Meck, 1996). D'ailleurs, le striatum est souvent présenté comme une structure où des liens entre certains évènements et le passage du temps seraient établis (Brown et al., 1999; Matell & Meck, 2004; Buhusi & Meck, 2005). Cependant, l'approche présentée ici est plutôt que les ganglions de la base intègrent l'information relative au temps en provenance d'autres sources (Ivry & Spencer, 2004), et non que le passage du temps y soit localement mesuré. Le débat sur la perception du temps est toutefois loin d'être terminé (Ivry & Schlerf, 2008). Le cervelet, les ganglions de la base et le cortex semblent tous pouvoir y jouer un rôle important.

Afin d'améliorer notre compréhension des ganglions de la base et du système dopaminergique dans un modèle unique, détaillé, et mieux structuré, il pourrait être avantageux de rassembler les données anatomiques et neurophysiologiques dans une seule et unique base de données. En effet, il est actuellement difficile d'effectuer des travaux de modélisation détaillés à partir des enregistrements rapportés dans la littérature. Les articles bien que rapportant avec précision les éléments les plus importants ne permettent pas toujours de se faire une idée claire de la distribution des réponses neuronales enregistrées. De plus, d'une structure à l'autre (striatum, SNc et VTA, GP, etc.), les enregistrements sont souvent effectués sur des tâches bien différentes, rendant encore plus compliquée l'intégration de ces différents éléments dans un seul et même modèle. En apprentissage machine, il est commun d'avoir des répertoires complets de données sur lesquels les performances des algorithmes peuvent être comparées. De la même façon, il serait certainement profitable de rassembler les distributions de profils de réponses des différents groupes de neurones dans une seule et même base de données pouvant servir à évaluer les différents modèles. Ceci permettrait une évaluation plus rapide des nouveaux modèles, des données qu'ils permettent d'expliquer, et surtout d'isoler rapidement celles qu'ils n'expliquent pas. L'existence d'une quantité suffisante de données pour chaque structure des ganglions de la base sur des tâches suffisamment similaires pourrait même permettre d'inférer des modèles à partir de ces données et des informations neuroanatomiques disponibles. Une analyse automatique des données et des modèles pourrait aussi suggérer des expériences permettant d'acquérir des informations manquantes au développement d'un bon modèle. Développer une telle base de données ne serait cependant pas une tâche simple, mais de telles approches ont de plus en plus d'importance en génomique par exemple (Herrgard et al., 2008; Thorisson, Muilu, & Brookes, 2009). Évidemment, les réponses des neurones d'une même structure comme le striatum peuvent, elles aussi, varier beaucoup en fonction de leur localisation précise (plutôt dorsale ou plutôt ventrale par exemple). Bien que cela puisse laisser croire qu'inférer le modèle à partir des données soit impossible, c'est pourtant ce que l'on fait à plus petite échelle, à chaque fois que l'on tente

d'expliquer le rôle de ces neurones et le fonctionnement de ces structures cérébrales à partir de la littérature et de nouvelles expériences.

9.3 Le cortex : représentation distribuée et hiérarchique

L'idée de représentation distribuée évoque deux éléments importants. Premièrement, qu'aucun neurone seul ne puisse donner une information complète ou précise sur une variable spécifique. Il n'y a pas un neurone spécifique qui répond uniquement au concept « *grand-maman* » et qui peut à lui seul nous indiquer la présence ou l'absence de grand-maman à notre esprit. À l'opposé, si un neurone se résume à une sommation pondérée de ses afférences et une non-linéarité comme une sigmoïde, il n'y aurait pas non plus suffisamment de neurones pour encoder toutes les combinaisons possibles de stimuli permettant de détecter grand-maman et tous les autres objets sur une seule couche (Cybenko, 1989). Une variation d'éclairage, un chandail neuf, et hop, on ne reconnaîtrait plus grand-maman! Mais la représentation de l'information dans le cortex n'est pas non plus sans structure, c'est le deuxième élément.

Les représentations ont aussi une certaine hiérarchie. Par exemple, un neurone de V1 ne couvre qu'une toute petite partie du champ visuel et ne peut donc pas dire grand-chose sur l'objet présent. Sa fenêtre sur le monde n'est pas suffisamment grande. Les couches suivantes intègrent l'information de plusieurs neurones. Elles ont ainsi un plus grand champ visuel et cumulent un plus grand nombre de non-linéarités. C'est ce qui leur permet de détecter des stimuli plus complexes. Alors que des populations de neurones se consacreront plutôt aux textures, d'autres se consacreront aux formes, aux visages ou aux positions des objets qui nous entourent. Certaines représentations seront presque insensibles à la position des objets ou à leur couleur, alors que d'autres seront particulièrement sensibles à cette information, si elle est nécessaire à la tâche. Hiérarchiser les couches représente un avantage computationnel important (Bengio, 2007, 2009). Mais, cette hiérarchie n'est pas purement pyramidale. Certaines informations sont d'abord intégrées pour ensuite être redistribuées à différents systèmes, et ce, même vers le bas (Figure 9-2). Dans un système récurrent comme le cortex où il existe probablement un chemin entre chaque

neurone, il devient parfois difficile de parler de traitement séquentiel de l'information. Même les premières couches du système visuel sont influencées par le traitement des couches supérieures et les attentes de l'individu. Il est aussi difficile d'isoler pourquoi une population de neurones X se spécialise selon une dimension de l'information plutôt qu'une autre. D'une certaine façon, le cortex semble arriver à démêler les facteurs importants des entrées brutes en les recombinant dans de nouvelles représentations distribuées plus utiles et efficaces.

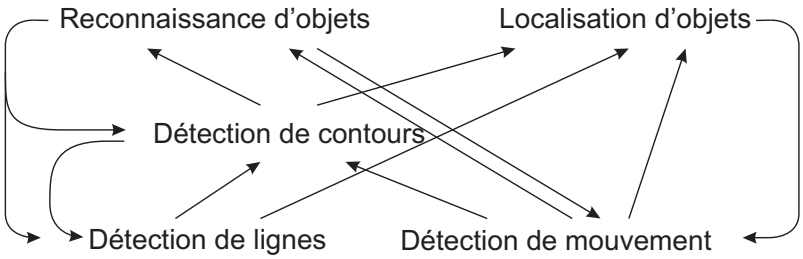


Figure 9-2 : Exemple fictif d'interaction entre différentes représentations.

Comment le cortex ou un quelconque système d'apprentissage pourrait apprendre de telles représentations reste encore à élucider, mais ce domaine de recherche est en pleine évolution (Bengio, 2007, 2009). Il y a bien des modèles de l'apprentissage pour les couches corticales primaires (par exemple, Doi et al., 2003; Hyvarinen et al., 2005) et plusieurs autres régions (par exemple, Rougier et al., 2005; O'Reilly & Frank, 2006), mais c'est bien peu lorsque l'on considère l'ensemble du développement cognitif du cortex et la grande variété de représentations que l'on y retrouve. Bien que le modèle cortical utilisé dans cette thèse (le LSTM) apprend une représentation abstraite de la tâche, il ne fait aucune avancée importante dans la construction de représentations distribuées au-delà de suggérer un rôle de la rétroaction dopaminergique mésocorticale. Le modèle s'avère même plus facile à analyser lorsque le nombre de blocs mémoires est petit, quoique le nombre de blocs mémoires semble avoir peu d'influence sur les résultats dopaminergiques reproduits.

Un autre facteur important dans le succès du cortex à développer des représentations est probablement la chronologie de son développement. Toutes les régions du cerveau n'apprennent pas nécessairement au même rythme au cours d'une vie. Par exemple, les neurones du cortex auditif primaire chez le rat (A1) ont une très

courte période critique pendant laquelle la plasticité est beaucoup plus grande (de Villers-Sidani, Chang, Bao, & Merzenich, 2007). La sensibilité de certains neurones de V1 pourrait être le résultat du développement du système nerveux combiné à un apprentissage prénatal dû à des signaux spontanés en provenance de la rétine (Thivierge & Marcus, 2006), ainsi que d'un raffinement important suivant la naissance (Singer & Tretter, 1976). Chez l'humain, la résolution spatiale et la profondeur de champ du système visuel prennent jusqu'à un an avant d'atteindre leur pleine capacité; un autre phénomène pouvant avoir un effet positif sur l'apprentissage dans le système visuel (Dominguez & Jacobs, 2001). Le langage est une autre faculté pour laquelle il semble y avoir une période critique. Si les stimuli adéquats ne sont pas présents pendant cette période, par exemple si un enfant n'est pas suffisamment exposé à un langage avant l'âge de treize ans, alors il est peu probable qu'il arrive à parler et comprendre convenablement des phrases complètes un jour (Goldin-Meadow, 1978). Selon Piaget (1952), nous ne naîtrions pas avec toutes les représentations dont nous avons besoins, mais nous les développerions avec l'expérience. Le développement cognitif chez les enfants suggère d'ailleurs une évolution des compétences par étapes ou par niveau (Shultz, 2003). Il en est de même pour la perception, l'évaluation et l'utilisation dans nos actions du passage du temps. Bien que des signes de l'évaluation du temps apparaissent très tôt chez le nourrisson, ce n'est que vers l'âge de six ans que l'on semble pouvoir utiliser le plein potentiel de nos différentes représentations temporelles (Droit-Volet, 2000). De la chronologie du développement moléculaire à l'ordre de présentation des concepts (Krueger & Dayan, 2009), il reste beaucoup de chemin à parcourir pour arriver à comprendre, modéliser ou reproduire la grande capacité d'apprentissage du cortex.

9.4 Le développement de représentations temporelles

Au cours de ces travaux sur le développement de représentations dans le cortex, un élément s'est avéré très important à la lumière des modèles et des données existantes : le temps. En effet, aucun modèle des ganglions de la base ne modélisait l'apprentissage de représentations temporelles (section 4.4.3.a) alors que les neurones dopaminergiques (Hollerman & Schultz, 1998; Morris et al., 2004) montrent très bien

que le délai interstimuli est bel et bien appris. Gallistel et Gibbon (2000) suggèrent même que la durée de visibilité des stimuli soit un facteur critique en apprentissage, même dans le simple conditionnement classique.

Bien que l'on ait parfois l'impression de percevoir le temps aussi bien que l'on perçoit une pomme, il n'y a pas d'organe de la transduction du temps de connu pour l'ordre des secondes comme l'œil pour la vision (Buhusi & Meck, 2005). En fait, le temps est constant. Il passe toujours à la même vitesse. Mais les neurones ont un temps de réaction minimum, ce sont des systèmes dynamiques subissant les forces des entrées sensorielles. Par conséquent, le passage du temps peut tout simplement être perçu en fonction de la vitesse relative des stimuli et de leur effet sur les neurones.

Autant le cortex, sous forme de mémoire de travail ou d'accumulation d'activité (sections 3.2.3 et 3.3.1), le striatum (lignes de délai, détecteur de coïncidence) que le cervelet, montrent des signes d'un lien avec l'apprentissage d'intervalles de temps (Buhusi & Meck, 2005; Ivry & Schlerf, 2008). L'arrivée de modèles neurophysiologiques adaptatifs de la perception du temps indépendants des lignes de délai est toute récente (Dragoi et al., 2003; Reutimann et al., 2004; Buonomano, 2005). Deux de ces modèles adaptatifs se proposent pour les intervalles de l'ordre des secondes et sont basés sur l'idée de neurones augmentant leur activité en fonction du temps qui passe (Dragoi et al., 2003; Reutimann et al., 2004). L'idée que les représentations temporelles soient acquises là où elles sont nécessaires n'est toujours qu'une hypothèse, mais on observe des augmentations d'activité corrélées au passage du temps partout dans le cerveau (section 3.2.3) (Ivry & Schlerf, 2008) dont certaines qui s'adaptent clairement au délai (Komura et al., 2001; Leon & Shadlen, 2003; Reutimann et al., 2004). Une contribution importante de cette thèse est de proposer, à l'aide des LSTM, un nouveau modèle dans cette direction (Chapitre 7 et Chapitre 8) (Rivest et al., 2010a; Rivest et al., 2010b). Le modèle montre comment une forme de mémoire de travail et des augmentations d'activités peuvent apprendre à représenter le passage du temps et l'évaluation d'intervalles de temps spécifiques.

Néanmoins, les algorithmes actuels comme TD (sections 2.3.3 et suivantes) et les réseaux de neurones artificiels récurrents (sections 2.1.3 et 2.1.4) sont relativement déficients pour apprendre des propriétés temporelles. TD est totalement dépendant de la représentation des entrées qui lui sont fournies. Il n'a de temporel que le fait qu'il propage vers l'état précédent l'estimation des récompenses à recevoir à partir de l'état présent. Quant aux réseaux de neurones récurrents, leurs architectures limitent mathématiquement leur capacité à faire le lien entre différents éléments éloignés dans le temps (Bengio et al., 1994; Hochreiter & Schmidhuber, 1997). Même les LSTM se sont avérés relativement mauvais s'ils ne pouvaient observer les deux stimuli ayant une relation temporelle à des moments contigus (Chapitre 7) (Rivest et al., 2010a), alors que les animaux semblent apprendre les délais très rapidement (Balsam et al., 2002). Une des raisons possibles de cette difficulté est que ces algorithmes sont entrainés de façon à prédire ce qui doit arriver au temps *t*, et non à prédire *quand* un évènement *x* doit arriver. Par exemple, TD doit apprendre à prédire la somme des récompenses futures (intemporelle) pour l'état courant observé. En prédiction de séries temporelles, le système doit prédire quelles seront les entrées aux temps *t+1* en fonction des entrées aux temps *t, t-1,* .... Bref, il minimise l'erreur de prédiction de la variable *x* au temps *t*, et non l'erreur de prédiction du temps *t* auquel devrait arriver un évènement *x*. Une telle réorientation du problème d'apprentissage le long de l'axe temporel permettrait certainement de développer de meilleurs modèles pour les situations où le temps est une variable importante. Elle permettrait aussi certainement de nouvelles avancées en apprentissage machine, tout particulièrement dans les situations ou les entrées et sorties doivent être traitées en temps continue, comme dans le cas des séries temporelles (sections 2.1.3 et 2.2.5).

9.5 L'apprentissage dans le cerveau : vision globale

L'aspect le plus fondamental du système nerveux central, c'est que tout comme le reste du corps humain, on peut le décomposer en différents systèmes. Toutefois, ceux-ci, de par leurs interactions, forment un tout qui dépasse la somme des parties. Par exemple, le système digestif est important pour la survie, mais il est inutile si on n'est pas capable d'y mettre de la nourriture. Réciproquement, sans le

travail effectué par le système digestif pour transformer la nourriture en énergie nécessaire aux muscles et au système nerveux, il est impossible d'obtenir de la nourriture. Le même principe s'applique pour notre grande capacité d'apprentissage. Il est peu probable que le système nerveux puisse se résumer en une seule et unique formule mathématique répliquée à grande échelle. Certaines composantes du système nerveux sont extrêmement spécialisées, tels que la rétine, la cochlée ou l'hypothalamus, alors que d'autres peuvent jouer des rôles beaucoup plus génériques comme le cortex. Il en est probablement de même pour l'apprentissage. De plus, comme il a été mentionné dans le Chapitre 3, même les régions hautement génériques sont spécialisées et peuvent s'avérer déficientes en isolation.

Bien que le cortex semble le centre de nos capacités cognitives et de nos grandes abstractions, ce dernier a une capacité d'apprentissage limitée sans l'hippocampe. Il est par exemple impossible d'y enregistrer de nouveaux souvenirs (mémoire explicite et épisodique, section 3.1.1) sans ce dernier (section 3.3.3) (Milner et al., 1968). Pourtant, il n'est pas nécessaire pour toutes les formes d'apprentissage. Similairement, les ganglions de la base et la dopamine jouent un rôle important dans plusieurs formes d'apprentissage telles que l'apprentissage procédural (section 3.1.2) et les conditionnements classique et opérant (sections 3.1.4 et 3.1.5). Malgré les modèles d'apprentissage par renforcement bien établis de ces structures, leur rôle exact dans l'apprentissage reste à préciser (Graybiel, 2008; Zweifel et al., 2009).

Les travaux les plus récents suggèrent différents rôles pour ces différents éléments. Par exemple, l'hippocampe pourrait servir entre autres de mémoire tampon cumulant temporairement de nouveaux exemples pour permettre un transfert ultérieur vers le cortex dans une représentation plus distribuée, ou en transformant ces représentations (Takashima et al., 2009). Les ganglions de la base jouent un rôle dans le conditionnement opérant, dans l'automatisation des processus, et dans l'apprentissage de séquence, etc. (Graybiel, 2008). Il n'est cependant pas clair si leur rôle précède celui du cortex, lui succède, ou si cet apprentissage est simplement le résultat d'une coopération permanente entre les deux régions (Laubach, 2005). Le

cortex, les ganglions de la base et l'hippocampe jouent sûrement un rôle crucial dans notre grande capacité d'apprentissage, et plusieurs autres, tels que le cervelet et l'amygdale y jouent aussi sûrement un rôle important. Par exemple, le cortex, les ganglions de la base et le cervelet semblent tous pourvoir jouer un rôle dans l'apprentissage d'intervalles de temps (Buhusi & Meck, 2005; Ivry & Schlerf, 2008). La Figure 9-3 montre quelques-unes de ces structures et de leurs interconnexions.

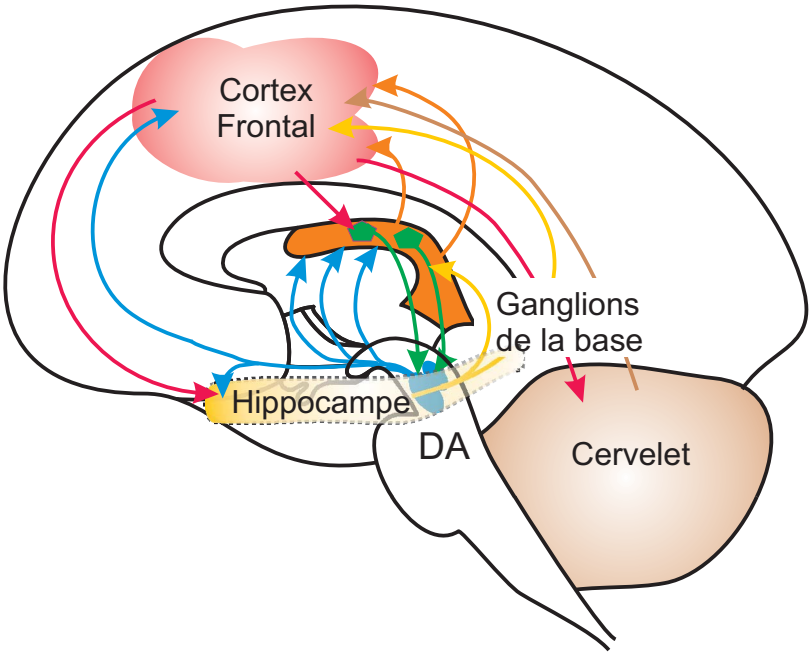


Figure 9-3 : Schéma de connectivité simplifiée entre le cortex frontal, les ganglions de la base, l'hippocampe et le cervelet.

Certaines de ces idées ont aussi été portées en modélisation (Doya, 1999; Doya, 2000; O'Reilly & Munakata, 2000; Khamassi et al., 2006; O'Reilly & Frank, 2006). Par exemple, Doya (1999, 2000) a suggéré que mathématiquement, le cortex fasse de l'apprentissage non-supervisé, le cervelet de l'apprentissage supervisé et les ganglions de la base de l'apprentissage par renforcement (Chapitre 2). L'approche systémique est aussi souvent utilisé dans le domaine du développement cognitif robotique (Huang & Weng, 2002). Certains travaux ont démontré l'effet positif sur l'apprentissage de séparer celui-ci en une partie non supervisée générant des représentations plus générales et une partie plus spécialisée sur la tâche en cours, soit en apprentissage supervisé (Hinton et al., 2006; Bengio et al., 2007; Erhan et al., 2009; Larochelle et al., 2009) ou en apprentissage par renforcement (Foster & Dayan,

2002; Rivest et al., 2005; Khamassi et al., 2006). D'autres ont exploré l'utilité de mixer un système d'apprentissage rapide à un système d'apprentissage plus lent (Ans & Rousset, 2000; Rivest & Precup, 2003; Tieleman & Hinton, 2009). Finalement, l'utilité d'un système de détection de nouveauté pour diriger certains systèmes d'apprentissage a aussi fait son chemin en neuroscience (Bevins et al., 2002; Kakade & Dayan, 2002; Sirois & Mareschal, 2004; Lisman & Grace, 2005) comme en apprentissage machine (Huang & Weng, 2002; Singh, Barto, & Chentanez, 2005). Cette dernière hypothèse pourrait très bien être étudiée à l'aide du modèle développé dans cette thèse en ajoutant à la récompense un signal de nouveauté telle que l'erreur de prédiction du modèle cortical.

Bien que ces systèmes soient plus faciles à étudier individuellement, il est nécessaire d'étudier leurs interactions pour découvrir comment peut en émerger une aussi grande capacité d'apprentissage. Pour l'instant, les modèles les plus avancés pour ce genre d'études semblent ceux dérivés des travaux de O'Reilly (O'Reilly, 1998; O'Reilly & Munakata, 2000). La venue, ces dernières années, d'enregistrements neurophysiologiques dans ces différentes régions du cerveau, sur une même tâche et dès le début de l'entraînement de l'animal, apportera sûrement des informations essentielles à la compréhension de l'apprentissage dans le cerveau et à l'évolution des représentations dans le cortex (Jog et al., 2002; Barnes et al., 2005; Miller & Wilson, 2008). Pour l'instant, le présent modèle se limite aux ganglions de la base, au cortex et à leurs interactions. C'est un bon début.

9.6 Conclusion

Enfin, comprendre, modéliser, ou reproduire la grande capacité d'apprentissage du cerveau demandera bien plus que le seul savoir des neurosciences, des sciences cognitives ou de l'intelligence artificielle. Il faudra intégrer les outils et les indices apportés par toutes ces disciplines pour élucider ce mystère. Dans cet ouvrage, les outils mathématiques issus de l'apprentissage machine ont été utilisés en combinaison aux données neuroanatomiques et neurophysiologiques existantes pour tenter de faire un pas en avant dans notre compréhension des bases computationnelles neurobiologiques de l'apprentissage, en

particulier du développement de représentation dans le contexte de l'apprentissage par renforcement (récompenses). Afin de mieux comprendre l'interaction entre les ganglions de la base et le cortex dans le développement de représentations en apprentissage par renforcement, un nouveau modèle a été développé. Ce modèle contribue à expliquer comment les représentations temporelles pourraient émerger dans une mémoire de travail corticale. Il montre comment l'information de ces représentations pourrait être intégrée par le striatum et expliquer l'activité dopaminergique. Finalement, le modèle suggère une façon par laquelle la dopamine pourrait affecter positivement l'apprentissage dans le cortex en plus de faire de nombreuses prédictions sur l'activité corticale et dopaminergique dans différentes versions du conditionnement classique. Il reste cependant encore beaucoup de chemin à parcourir pour comprendre cet apprentissage dans le cerveau.

Bibliographie

Aguiar, A. & Baillargeon, R. (1998). Eight-and-a-half-month-old infants' reasoning about containment events. *Child Dev., 69,* 636-653.

Aizman, O., Brismar, H., Uhlen, P., Zettergren, E., Levey, A. I., Forssberg, H. et al. (2000). Anatomical and physiological evidence for D1 and D2 dopamine receptor colocalization in neostriatal neurons. *Nat.Neurosci., 3,* 226-230.

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

Albus, J. S. (1971). A Theory of Cerebellar Function. *Math.Biosci., 10,* 25-61.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu.Rev.Neurosci., 9,* 357-381.

Alexander, W. H. (2007). Shifting attention using a temporal difference prediction error and high-dimensional input. *Adaptive Behavior, 15,* 121-133.

Alexander, W. H. & Sporns, O. (2006). Temporal Difference Learning with Learned Attention Shifts. In *Fifth International Conference on Development and Learning*.

Amari, S. (1999). Natural gradient learning for over- and under-complete bases in ICA. *Neural Comput., 11,* 1875-1883.

Ans, B. & Rousset, S. (2000). Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science, 12,* 1-19.

Apicella, P. (2007). Leading tonically active neurons of the striatum from reward detection to context recognition. *Trends Neurosci., 30,* 299-306.

Arbib, M. A. & Dominey, P. F. (1995). Modeling the Roles of Basal Ganglia in Timing and Sequencing Saccadic Eye Movements. In J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 149-162). The MIT Press.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: object permanence in 6- and 8-month-old infants. *Cognition, 23,* 21-41.

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Twelfth International Conference on Machine learning* (pp. 30-37). Morgan Kaufmann: San Francisco, CA.

Bakker, B. (2002). Reinforcement Learning with Long Short-Term Memory. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Neural Information Processing Systems* (pp. 1475-1482). Cambridge, MA: MIT Press.

Bakker, B., Linaker, F., & Schmidhuber, J. (2002). Reinforcement Learning in Partially Observable Mobile Robot Domains Using Unsupervised Event Extraction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*.

Balci, F., Gallistel, C. R., Allen, B. D., Frank, K. M., Gibson, J. M., & Brunner, D. (2009). Acquisition of peak responding: what is learned? *Behav.Processes, 80,* 67-75.

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

Balsam, P. D., Drew, M. R., & Yang, C. (2002). Timing at the start of associative learning. *Learning and Motivation, 33,* 141-155.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature, 437,* 1158-1161.

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

Barto, A. G. & Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behav.Brain Res., 4,* 221-235.

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2002). *Neurosciences : à la découverte du cerveau*. (2e édition, traduction et adaptation française André Nieoullon ed.) Éditions Pradel.

Beiser, D. G. & Houk, J. C. (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *J.Neurophysiol., 79,* 3168-3188.

Bell, A. J. & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput., 7,* 1129-1159.

Bengio, Y. (2007). On the Challenge of Learning Complex Functions. In P.Cisek, T. Drew, & J. F. Kalaska (Eds.), *Computational Neuroscience: Theoritical Insights into Brain Function, 165* (Elsevier.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning, 2,* 1-127.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In B. Scholkopf, J. Platt, & T. Hoffman (Eds.), *Neural Information Processing Systems* (pp. 153-165). Cambridge, MA: MIT Press.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *26th Annual International Conference on Machine Learning* (pp. 41-48). New York, NY: ACM.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions in Neural Networks, 5,* 157-166.

Berns, G. S. & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *J.Cogn Neurosci., 10,* 108-121.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl), 191,* 391-431.

Berridge, K. C. & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res.Brain Res.Rev., 28,* 309-369.

Bertin, M., Schweighofer, N., & Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Netw., 20,* 668-675.

Bevins, R. A., Besheer, J., Palmatier, M. I., Jensen, H. C., Pickett, K. S., & Eurek, S. (2002). Novel-object place conditioning: behavioral and dopaminergic processes in expression of novelty reward. *Behav.Brain Res., 129,* 41-50.

Beylin, A. V., Gandhi, C. C., Wood, G. E., Talk, A. C., Matzel, L. D., & Shors, T. J. (2001). The role of the hippocampus in trace conditioning: temporal discontinuity or task difficulty? *Neurobiol.Learn.Mem., 76,* 447-461.

Blumenfeld, H. (2002). *Neuroanatomy through clinical cases*. Sunderland, MA: Sinauer.

Braver, T. S. & Cohen, J. D. (2000). On the Control of Control: The Role of Dopamine in Regulating Prefrontal Function and Working Memory. In S.Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 713-737). Cambridge, MA: MIT Press.

Breukelaar, J. W. & Dalrymple-Alford, J. C. (1999). Effects of lesions to the cerebellar vermis and hemispheres on timing and counting in rats. *Behav.Neurosci., 113,* 78-90.

Brody, C. D., Hernandez, A., Zainos, A., & Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb.Cortex, 13,* 1196-1207.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J.Neurosci., 19,* 10502-10511.

Buhusi, C. V. & Meck, W. H. (2000). Timing for the absence of a stimulus: the gap paradigm reversed. *J.Exp.Psychol.Anim Behav.Process, 26,* 305-322.

Buhusi, C. V. & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat.Rev.Neurosci., 6,* 755-765.

Buonomano, D. V. (2003). Timing of neural responses in cortical organotypic slices. *Proc.Natl.Acad.Sci.U.S A, 100,* 4897-4902.

Buonomano, D. V. (2005). A learning rule for the emergence of stable dynamics and timing in recurrent networks. *J.Neurophysiol., 94,* 2275-2283.

Carr, C. E. & Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *J.Neurosci., 10,* 3227-3246.

Cerveri, P., Lopomo, N., Pedotti, A., & Ferrigno, G. (2005). Derivation of centers and axes of rotation for wrist and fingers in a hand kinematic model: methods and reliability results. *Ann.Biomed.Eng, 33,* 402-412.

Church, R. M. (2003). A Concise Introduction to Scalar Timing Theory. In W.H.Meck (Ed.), *Functional and Neural Mechanisms of Interval Timing* (pp. 3-22). Boca Raton: CRC Press.

Church, R. M., Meck, W. H., & Gibbon, J. (1994). Application of scalar timing theory to individual trials. *J.Exp.Psychol.Anim Behav.Process, 20,* 135-155.

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

Comon, P. (1994). Independent Component Analysis, A New Concept. *Signal Processing, 36,* 287-314.

Conde, H. (1992). Organization and physiology of the substantia nigra. *Exp.Brain Res., 88,* 233-248.

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

Cromwell, H. C. & Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J.Neurophysiol., 89,* 2823-2838.

Cybenko, G. (1988). *Continuous Valued Neural Networks with Two Hidden Layers Are Sufficient.* Technical Report, Department of Computer Science, Tufts University, Medford, MA.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems, 2,* 303-314.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Neural Information Processing Systems* (pp. 83-90). Cambridge, MA: MIT Press.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Comput., 18,* 1637-1677.

Daw, N. D. & Doya, K. (2006). The computational neurobiology of learning and reward. *Curr.Opin.Neurobiol.*.

Daw, N. D. & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Comput., 14,* 2567-2583.

Dayan, P. (1992). The Convergence of Td(Lambda) for General Lambda. *Machine Learning, 8,* 341-362.

de Villers-Sidani, E., Chang, E. F., Bao, S., & Merzenich, M. M. (2007). Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. *J.Neurosci., 27,* 180-189.

DeCasper, A. J. & Fifer, W. P. (1980). Of human bonding: newborns prefer their mothers' voices. *Science, 208,* 1174-1176.

Diamantaras, K. I. & Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications*. Toronto: Wiley.

Dickinson, A. & Mackintosh, N. J. (1978). Classical conditioning in animals. *Annu.Rev.Psychol., 29,* 587-612.

Doi, E., Inui, T., Lee, T. W., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Comput., 15,* 397-417.

Dominey, P. F. & Arbib, M. A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb.Cortex, 2,* 153-175.

Dominguez, M. & Jacobs, R. A. (2001). Visual Development and the Acquisition of Binocular Disparity Sensitivities. In C. E. Brodley & A. P. Danyluk (Eds.), *Eighteenth International Conference on Machine Learning* (pp. 114-121). San Francisco, CA: Morgan Kaufmann.

Dormont, J. F., Conde, H., & Farin, D. (1998). The role of the pedunculopontine tegmental nucleus in relation to conditioned motor performance in the cat. I.

Context-dependent and reinforcement-related single unit activity. *Exp.Brain Res., 121,* 401-410.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw., 12,* 961-974.

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

Doya, K., Kimura, H., & Miyamura, A. (2001). Motor control: Neural Models and System Theory. *International Journal of Applied Mathematics and Computer Science, 11,* 77-104.

Dragoi, V., Staddon, J. E., Palmer, R. G., & Buhusi, C. V. (2003). Interval timing as an emergent learning property. *Psychol.Rev., 110,* 126-144.

Droit-Volet, S. (2000). L'estimation du temps : perspective développementale. *L'année psychologique, 100,* 443-464.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification*. (2nd ed.) New York: NY: Wiley-Interscience.

Durstewitz, D. (2004). Neural representation of interval time. *Neuroreport, 15,* 745-749.

Duva, M. A., Tomkins, E. M., Moranda, L. M., Kaplan, R., Sukhaseum, A., & Stanley, B. G. (2005). Origins of lateral hypothalamic afferents associated with N-methyl-d-aspartic acid-elicited eating studied using reverse microdialysis of NMDA and Fluorogold. *Neurosci.Res., 52,* 95-106.

Eck, D. & Schmidhuber, A. (2002). Learning the long-term structure of the blues. *Artificial Neural Networks - Icann 2002, 2415,* 284-289.

Egelman, D. M., Person, C., & Montague, P. R. (1998). A computational role for dopamine delivery in human decision-making. *J.Cogn Neurosci., 10,* 623-630.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14,* 179-211.

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P. (2009). The Difficulty of Training Deep Architectures and the effect of Unsupervised Pre-Training. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*.

Fadel, J. & Deutch, A. Y. (2002). Anatomical substrates of orexin-dopamine interactions: lateral hypothalamic projections to the ventral tegmental area. *Neuroscience, 111,* 379-387.

Fiorillo, C. D., Newsome, W. T., & Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nat.Neurosci., 11,* 966-973.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299,* 1898-1902.

Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput., 19,* 1468-1502.

Foster, D. & Dayan, P. (2002). Structure in the space of value functions. *Machine Learning, 49,* 325-346.

Foster, D. J., Morris, R. G., & Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus, 10,* 1-16.

Fromkin, V., Rodman, R., Hultin, N., & Logan, H. (1997). *An Introduction to Language*. (First Candian Edition ed.) Toronto: Harcourt Brace.

Fukuda, M., Ono, T., Nakamura, K., & Tamura, R. (1990). Dopamine and ACh involvement in plastic learning by hypothalamic neurons in rats. *Brain Res.Bull., 25,* 109-114.

Fukuda, M., Ono, T., Nishino, H., & Nakamura, K. (1986). Neuronal responses in monkey lateral hypothalamus during operant feeding behavior. *Brain Res.Bull., 17,* 879-883.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J.Neurophysiol., 61,* 331-349.

Gallistel, C. R. (2003). Time has come. *Neuron, 38,* 149-150.

Gallistel, C. R. & Gibbon, J. (2000). Time, rate, and conditioning. *Psychol.Rev., 107,* 289-344.

Geisler, S., Derst, C., Veh, R. W., & Zahm, D. S. (2007). Glutamatergic afferents of the ventral tegmental area in the rat. *J.Neurosci., 27,* 5730-5743.

Gerfen, C. R. (1984). The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature, 311,* 461-464.

Gerfen, C. R. (1985). The neostriatal mosaic. I. Compartmental organization of projections from the striatum to the substantia nigra in the rat. *J.Comp Neurol., 236,* 454-476.

Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annu.Rev.Neurosci., 15,* 285-320.

Gerfen, C. R., Engber, T. M., Mahan, L. C., Susel, Z., Chase, T. N., Monsma, F. J., Jr. et al. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science, 250,* 1429-1432.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Comput., 12,* 2451-2471.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research, 3,* 115-143.

Gibbon, J. (1977). Scalar expectancy theory and Weber's Law in animal timing. *Psychol.Rev., 84,* 279-325.

Gillies, A. & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Mov Disord., 15,* 762-770.

Gluck, M. A. & Myers, C. E. (2001). *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus and Learning*. MA: MIT Press.

Goldin-Meadow, S. (1978). A Study in Human Capacities. *Science, 200,* 649-651.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu.Rev.Neurosci., 31,* 359-387.

Graybiel, A. M., Canales, J. J., & Capper-Loup, C. (2000). Levodopa-induced dyskinesias and dopamine-dependent stereotypies: a new hypothesis. *Trends Neurosci., 23,* S71-S77.

Groenewegen, H. J., Berendse, H. W., & Haber, S. N. (1993). Organization of the output of the ventral striatopallidal system in the rat: ventral pallidal efferents. *Neuroscience, 57,* 113-142.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol.Cybern., 84,* 401-410.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biol.Cybern., 84,* 411-423.

Gurney, K. N., Humphries, M., Wood, R., Prescott, T. J., & Redgrave, P. (2004). Testing computational hypotheses of brain systems function: a case study with the basal ganglia. *Network., 15,* 263-290.

Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *J.Chem.Neuroanat., 26,* 317-330.

Hahnloser, R. H., Kozhevnikov, A. A., & Fee, M. S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature, 419,* 65-70.

Harrington, D. L., Lee, R. R., Boyd, L. A., Rapcsak, S. Z., & Knight, R. T. (2004). Does the representation of time depend on the cerebellum? Effect of cerebellar stroke. *Brain, 127,* 561-574.

Hassani, O. K., Cromwell, H. C., & Schultz, W. (2001). Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *J.Neurophysiol., 85,* 2477-2489.

Hawkins, J. & Blakeslee, S. (2004). *On Intelligence*. New York: Times Books.

Hebb, D. O. (1949). *The organization of behaviors*. New York: Wiley.

Hebb, M. O. & Robertson, H. A. (2000). Identification of a subpopulation of substantia nigra pars compacta gamma-aminobutyric acid neurons that is regulated by basal ganglia activity. *J.Comp Neurol., 416,* 30-44.

Hernandez, G., Hamdani, S., Rajabi, H., Conover, K., Stewart, J., Arvanitogiannis, A. et al. (2006). Prolonged rewarding stimulation of the rat medial forebrain bundle: neurochemical and behavioral consequences. *Behav.Neurosci., 120,* 888-904.

Herrgard, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M. et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat.Biotechnol., 26,* 1155-1160.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput., 18,* 1527-1554.

Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313,* 504-507.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput., 9,* 1735-1780.

Hollerman, J. R. & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat.Neurosci., 1,* 304-309.

Hopson, J. W. (2003). General Learning Models: Timing without a Clock. In W.H.Meck (Ed.), *Functional and Neural Mechanisms of Interval Timing* (pp. 23-60). Boca Raton: CRC Press.

Horgan, P. & Cummins, F. (2006). Modeling Dopamine Activity by Reinforcement Learning Methods: Implications from Two Recent Models. *Artif.Intell.Rev., 26,* 49-62.

Hornick, K., Stinchcombe, M., & White, H. (1989). Feedforward Networks are Universal Approximators. *Neural Networks, 2,* 359-366.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience, 96,* 651-656.

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 J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 249-270). The MIT Press.

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

Huang, X. & Weng, J. (2002). Novelty and reinforcement learning in the value system of developmental robots. In *Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*.

Hyland, B. I., Reynolds, J. N., Hay, J., Perk, C. G., & Miller, R. (2002). Firing modes of midbrain dopamine cells in the freely moving rat. *Neuroscience, 114,* 475-492.

Hyvarinen, A., Gutmann, M., & Hoyer, P. O. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC.Neurosci., 6,* 12.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis*. Toronto: Wiley.

Ikemoto, S. & Panksepp, J. (1996). Dissociations between appetitive and consummatory responses by pharmacological manipulations of reward-relevant brain regions. *Behav.Neurosci., 110,* 331-345.

Ito, M. (2000). Mechanisms of motor learning in the cerebellum. *Brain Res., 886,* 237-245.

Ivry, R. B. & Keele, S. W. (1989). Timing Functions of The Cerebellum. *Journal of Cognitive Neuroscience, 1,* 136-152.

Ivry, R. B. & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends Cogn Sci., 12,* 273-280.

Ivry, R. B. & Spencer, R. M. (2004). The neural representation of time. *Curr.Opin.Neurobiol., 14,* 225-232.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb.Cortex, 17,* 2443-2452.

Jaakkola, T., Jordan, M. I., & Singh, S. P. (1994). On the Convergence of Stochastic Iterative Dynamic-Programming Algorithms. *Neural Computation, 6,* 1185-1201.

Jimenez-Castellanos, J. & Graybiel, A. M. (1989). Compartmental origins of striatal efferent projections in the cat. *Neuroscience, 32,* 297-321.

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

Joel, D. & Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience, 96,* 451-474.

Jog, M. S., Connolly, C. I., Kubota, Y., Iyengar, D. R., Garrido, L., Harlan, R. et al. (2002). Tetrode technology: advances in implantable hardware, neuroimaging, and data analysis techniques. *J.Neurosci.Methods, 117,* 141-152.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, A. M. (1999). Building neural representations of habits. *Science, 286,* 1745-1749.

Kaiser, D. H. (2008). The proportion of fixed interval trials to probe trials affects acquisition of the peak procedure fixed interval timing task. *Behav.Processes, 77,* 100-108.

Kaiser, D. H. (2009). Fewer peak trials per session facilitate acquisition of peak responding despite elimination of response rate differences. *Behav.Processes, 80,* 12-19.

Kakade, S. & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw., 15,* 549-559.

Kalivas, P. W. & Alesdatter, J. E. (1993). Involvement of N-methyl-D-aspartate receptor stimulation in the ventral tegmental area and amygdala in behavioral sensitization to cocaine. *J.Pharmacol.Exp.Ther., 267,* 486-495.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science*. (4th ed.) McGraw-Hill.

Karmarkar, U. R. & Buonomano, D. V. (2007). Timing in the absence of clocks: encoding time in neural network states. *Neuron, 53,* 427-438.

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

Khamassi, M., Lacheze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor-Critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior, 13,* 131-148.

Khamassi, M., Martinet, L. E., & Guillot, A. (2006). Combining self-organizing maps with mixtures of experts: Application to an actor-critic model of

reinforcement learning in the basal ganglia. *From Animals to Animats 9, Proceedings, 4095,* 394-405.

Khamassi, M., Mulder, A. B., Tabuchi, E., Douchamps, V., & Wiener, S. I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *Eur.J.Neurosci., 28,* 1849-1866.

Killeen, P. R. & Taylor, T. J. (2000). How the propagation of error through stochastic counters affects time discrimination and other psychophysical judgments. *Psychol.Rev., 107,* 430-459.

Kirkpatrick-Steger, K., Miller, S. S., Betti, C. A., & Wasserman, E. A. (1996). Cyclic responding by pigeons on the peak timing procedure. *J.Exp.Psychol.Anim Behav.Process, 22,* 447-460.

Kobayashi, S. & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *J.Neurosci., 28,* 7837-7846.

Kobayashi, Y. & Okada, K. (2007). Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. *Ann.N.Y Acad.Sci., 1104,* 310-323.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science, 297,* 846-848.

Kolodziejski, C., Porr, B., & Worgotter, F. (2009). On the Asymptotic Equivalence Between Differential Hebbian and Temporal Difference Learning. *Neural Comput., 21,* 1173-1202.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature, 412,* 546-549.

Krueger, K. A. & Dayan, P. (2009). Flexible shaping: how learning in small steps helps. *Cognition, 110,* 380-394.

Lang, C. E. & Bastian, A. J. (1999). Cerebellar subjects show impaired adaptation of anticipatory EMG during catching. *J.Neurophysiol., 82,* 2108-2119.

Lang, K. J., Waibel, A. H., & Hinton, G. E. (1990). A Time-Delay Neural Network Architecture for Isolated Word Recognition. *Neural Networks, 3,* 23-43.

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research, 10,* 1-40.

Lau, B. & Glimcher, P. W. (2007). Action and outcome encoding in the primate caudate nucleus. *J.Neurosci., 27,* 14502-14514.

Laubach, M. (2005). Who's on first? What's on second? The time course of learning in corticostriatal systems. *Trends Neurosci., 28,* 508-511.

Lebedev, M. A., O'Doherty, J. E., & Nicolelis, M. A. (2008). Decoding of temporal intervals from cortical ensemble activity. *J.Neurophysiol., 99,* 166-186.

LeCun, Y. (1985). Une procédure d'apprentissage pour réseau à seuil assymétrique (A Learning Scheme for Assymmetric Threshold Network). In (pp. 599-604). Paris, France.

Lejeune, H. & Wearden, J. H. (2006). Scalar properties in animal timing: conformity and violations. *Q.J.Exp.Psychol.(Colchester.), 59,* 1875-1908.

Leon, M. I. & Shadlen, M. N. (2003). Representation of time by neurons in the posterior parietal cortex of the macaque. *Neuron, 38,* 317-327.

Levesque, M. & Parent, A. (2005). The striatofugal fiber system in primates: a reevaluation of its organization based on single-axon tracing studies. *Proc.Natl.Acad.Sci.U.S.A, 102,* 11888-11893.

Lewis, P. A. (2002). Finding the timer. *Trends Cogn Sci., 6,* 195-196.

Linsker, R. (1988). Self-Organization in A Perceptual Network. *Computer, 21,* 105-117.

Lisman, J. E. & Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron, 46,* 703-713.

Littman, M., Sutton, R. S., & Singh, S. (2002). Predictive representations of state. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Neural Information Processing Systems* (pp. 1551-1561). Cambridge, MA: MIT Press.

Livesey, A. C., Wall, M. B., & Smith, A. T. (2007). Time perception: manipulation of task difficulty dissociates clock functions from other cognitive demands. *Neuropsychologia, 45,* 321-331.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J.Neurophysiol., 67,* 145-163.

Lucchetti, C. & Bon, L. (2001). Time-modulated neuronal activity in the premotor cortex of macaque monkeys. *Exp.Brain Res., 141,* 254-260.

Lucchetti, C., Ulrici, A., & Bon, L. (2005). Dorsal premotor areas of nonhuman primate: functional flexibility in time domain. *Eur.J.Appl.Physiol, 95,* 121-130.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput., 20,* 3034-3054.

Ludvig, E. A., Sutton, R. S., Verbeek, E., & Kehoe, E. J. (2009). A Computational Model of Hippocampal Function in Trace Conditioning. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Neural Information Processing Systems* (pp. 993-1000).

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychol.Rev., 104,* 241-265.

Mahadevan, S. & Maggioni, M. (2007). Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. *Journal of Machine Learning Research, 8,* 2169-2231.

Margolis, E. B., Lock, H., Hjelmstad, G. O., & Fields, H. L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *J.Physiol, 577,* 907-924.

Marr, D. (1969). A theory of cerebellar cortex. *J.Physiol, 202,* 437-470.

Matell, M. S. & Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Brain Res.Cogn Brain Res., 21,* 139-170.

Matsumoto, M. & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature, 447,* 1111-1115.

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends Neurosci., 26,* 423-428.

Meck, W. H. (2003). *Functional and Neural Mechanisms of Interval Timing*. Boca Raton: CRC Press.

Meck, W. H. (1996). Neuropharmacology of timing and time perception. *Brain Res.Cogn Brain Res., 3,* 227-242.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & miel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29,* 143-178.

Miller, E. K. & Wilson, M. A. (2008). All my circuits: using multiple electrodes to understand functioning neural networks. *Neuron, 60,* 483-488.

Milner, B., Corkin, S., & Teuber, H. L. (1968). Further Analysis of the Hippocapal Amnesic Syndrome: 14-Year Follow-up Study of H.M. *Neuropsychologia, 6,* 215-234.

Mirenowicz, J. & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J.Neurophysiol., 72,* 1024-1027.

Mirenowicz, J. & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature, 379,* 449-451.

Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., & Sejnowski, T. J. (1993). Using Aperiodic Reinforcement for Directed Self-Organization During Development. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Neural Information Processing Systems* (pp. 969-976). San Mateo, CA: Morgan Kaufmann.

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

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature, 431,* 760-767.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron, 43,* 133-143.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat.Neurosci., 9,* 1057-1063.

Mountcastle, V. B. (1978). An Organizing Principle for Cerebral Function: The Unit Module and the Distributed System. In G.M.Edelman & V. B. Mountcastle (Eds.), *The Mindful Brain: Cortical Organization and the Group-Selectivity Theory of Higher Brain Function* (pp. 7-50). Cambridge, MA: MIT Press.

Mountcastle, V. B. (2003). Introduction. Computation in cortical columns. *Cereb.Cortex, 13,* 2-4.

Moustafa, A. A. & Maida, A. S. (2007). Using TD Learning to Simulate Working Memory Performance in a Model of the Prefrontal Cortex and Basal Ganglia. *Cognitive Systems Research, 8,* 262-281.

Nakahara, H., Amari, S. S., & Hikosaka, O. (2002). Self-organization in the basal ganglia with modulation of reinforcement signals. *Neural Comput., 14,* 819-844.

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J.Cogn Neurosci., 13,* 626-647.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41,* 269-280.

Nakamura, K. & Ono, T. (1986). Lateral hypothalamus neuron involvement in integration of natural and artificial rewards and cue signals. *J.Neurophysiol., 55,* 163-181.

Niki, H. & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Res., 171,* 213-224.

Niv, Y. & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends Cogn Sci., 12,* 265-272.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38,* 329-337.

O'Reilly, R. C. (1998). Six Principles for Biologically Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences, 2,* 455-462.

O'Reilly, R. C. & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput., 18,* 283-328.

O'Reilly, R. C. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C. & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus, 10,* 389-397.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J.Math.Biol., 15,* 267-273.

Otani, S., Daniel, H., Roisin, M. P., & Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cereb.Cortex, 13,* 1251-1256.

Padoa-Schioppa, C. & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441,* 223-226.

Padoa-Schioppa, C. & Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat.Neurosci., 11,* 95-102.

Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J.Neurosci., 25,* 6235-6242.

Parent, A. & Hazrati, L. N. (1995a). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res.Brain Res.Rev., 20,* 91-127.

Parent, A. & Hazrati, L. N. (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,* 128-154.

Parr, R., Painter-Wakefield, C., Li, L., & Littman, M. (2007). Analyzing feature generation for value-function approximation. In *24th international conference on Machine learning* (pp. 737-744). New York, NY: ACM.

Pavlov, J. P. (1960). *Conditionned reflexes: an investigation of the physiological activity of the cerebral cortex*. New York: Dover.

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychol.Rev., 101,* 587-607.

Pearce, J. M. & Bouton, M. E. (2001). Theories of associative learning in animals. *Annu.Rev.Psychol., 52,* 111-139.

Piaget, J. (1952). *The Origins of Intelligence in Children*. New York, NY: International University Press.

Potjans, W., Morrison, A., & Diesmann, M. (2009). A Spiking Neural Network Model of an Actor-Critic Learning Agent. *Neural Comput., 21,* 301-339.

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat.Neurosci., 2,* 79-87.

Rao, R. P. & Sejnowski, T. J. (2001). Predictive learning of temporal sequences in recurrent neocortical circuits. *Novartis.Found.Symp., 239,* 208-229.

Redgrave, P. & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat.Rev.Neurosci., 7,* 967-975.

Redhead, E. S. & Pearce, J. M. (1995). Stimulus salience and negative patterning. *Q.J.Exp.Psychol.B, 48,* 67-83.

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science, 306,* 1944-1947.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol.Rev., 114,* 784-805.

Rescorla, R. A. & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H.Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Reutimann, J., Yakovlev, V., Fusi, S., & Senn, W. (2004). Climbing neuronal activity as an event-based cortical representation of time. *J.Neurosci., 24,* 3295-3303.

Reynolds, J. N. & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw., 15,* 507-521.

Rivest, F., Bengio, Y., & Kalaska, J. F. (2005). Brain Inspired Reinforcement Learning. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Neural Information Processing Systems* (pp. 1129-1136). Cambridge, MA: The MIT Press.

Rivest, F., Kalaska, J. F., & Bengio, Y. (2010a). Alternative Time Representation in Dopamine Models. *Journal of Computational Neuroscience, 28,* 107-130.

Rivest, F., Kalaska, J. F., & Bengio, Y. (2010b). *Conditioning and Time Representation in the Long Short-term Memory Networks* (in preparation).

Rivest, F. & Precup, D. (2003). Combining TD-learning with Cascade-correlation Networks. In *Twentieth International Conference on Machine Learning* (pp. 632-639). AAAI Press.

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annu.Rev.Neurosci., 27,* 169-192.

Roberts, P. D., Santiago, R. A., & Lafferriere, G. (2008). An implementation of reinforcement learning based on spike timing dependent plasticity. *Biol.Cybern., 99,* 517-523.

Roberts, S. (1981). Isolation of an internal clock. *J.Exp.Psychol.Anim Behav.Process, 7,* 242-268.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat.Neurosci., 10,* 1615-1624.

Romo, R., Brody, C. D., Hernandez, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature, 399,* 470-473.

Rosenbaum, D. A. (1991). *Human Motor Control*. San Diego: Academic Press.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proc.Natl.Acad.Sci.U.S.A, 102,* 7338-7343.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature, 323,* 533-536.

Russell, S. J. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

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

Sanabria, F. & Killeen, P. R. (2007). Temporal generalization accounts for response resurgence in the peak procedure. *Behav.Processes, 74,* 126-141.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J.Neurophysiol., 80,* 1-27.

Schultz, W. (2007). Behavioral dopamine signals. *Trends Neurosci., 30,* 203-210.

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. *J.Neurosci., 13,* 900-913.

Schultz, W., Apicella, P., Romo, R., & Scarnati, E. (1995). Context-dependent Activity in Primate Striatum Reflecting Past and Future Behavioral Events. In J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 11-27). The MIT Press.

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

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593-1599.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related Signals Carried by Dopamine Neurons. In J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 233-248). Cambridge, MA: The MIT Press.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cereb.Cortex, 10,* 272-284.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2003). Changes in behavior-related neuronal activity in the striatum during learning. *Trends Neurosci., 26,* 321-328.

Schweighofer, N., Doya, K., & Lay, F. (2001). Unsupervised learning of granule cell sparse codes enhances cerebellar adaptive control. *Neuroscience, 103,* 35-50.

Sejnowski, T. J. & Rosenberg, C. R. (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems, 1,* 145-168.

Semba, K. & Fibiger, H. C. (1992). Afferent connections of the laterodorsal and the pedunculopontine tegmental nuclei in the rat: a retro- and antero-grade transport and immunohistochemical study. *J.Comp Neurol., 323,* 387-410.

Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.

Singer, W. & Tretter, F. (1976). Unusually large receptive fields in cats with restricted visual experience. *Exp.Brain Res., 26,* 171-184.

Singh, S., Barto, A. G., & Chentanez, N. (2005). Intrinsically Motivated Reinforcement Learning. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Neural Information Processing Systems* (pp. 1281-1288). Cambridge, MA: MIT Press.

Sirois, S. & Mareschal, D. (2004). An interacting systems model of infant habituation. *J.Cogn Neurosci., 16,* 1352-1362.

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century-Crofts.

Strick, P. L., Dum, R. P., & Picard, N. (1995). Macro-organization of the Circuits Connecting the Basal Ganglia with the Cortical Motor Areas. In J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 117-130). Cambridge, MA: The MIT Press.

Suri, R. E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Exp.Brain Res., 140,* 234-240.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Netw., 15,* 523-533.

Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience, 103,* 65-85.

Suri, R. E. & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp.Brain Res., 121,* 350-354.

Suri, R. E. & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience, 91,* 871-890.

Suri, R. E. & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comput., 13,* 841-862.

Sussillo, D. & Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron, 63,* 544-557.

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

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychol.Rev., 88,* 135-170.

Sutton, R. S. & Barto, A. G. (1990). Time-Derivative Models of Pavlovian Reinforcement. In M.Gabriel & J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (pp. 497-538). Cambridge: MIT PRess.

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

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function aproximation. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Neural Information Processing Systems* (pp. 1057-1063). Cambridge, MA: MIT Press.

Sutton, R. S. & Tanner, B. (2005). Temporal-Difference Networks. In L. K. Saul, C. Weiss, & L. Bottou (Eds.), *Neural Information Processing Systems* (pp. 1377-1384). Cambridge, MA: MIT Press.

Takashima, A., Nieuwenhuis, I. L., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernandez, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *J.Neurosci., 29,* 10087-10093.

Teyler, T. J. & DiScenna, P. (1984). Long-term potentiation as a candidate mnemonic device. *Brain Res., 319,* 15-28.

Thibaudeau, G., Potvin, O., Allen, K., Dore, F. Y., & Goulet, S. (2007). Dorsal, ventral, and complete excitotoxic lesions of the hippocampus in rats failed to impair appetitive trace conditioning. *Behav.Brain Res., 185,* 9-20.

Thivierge, J.-P. & Marcus, G. F. (2006). Computational Developmental Neuroscience: Exploring the Interactions Between Genetics and Neural Activity. In IEEE Press (Ed.), *International Joint Conference on Neural Networks, 2006.* (pp. 4630-4637).

Thorisson, G. A., Muilu, J., & Brookes, A. J. (2009). Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat.Rev.Genet., 10,* 9-18.

Tieleman, T. & Hinton, G. (2009). Using Fast Weights to Improve Persistent Contrastive Divergence. In L. Bottou & M. Littman (Eds.), *26th International Conference on Machine Learning* (pp. 1033-1040). Omnipress.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J.Neurosci., 23,* 10402-10410.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307,* 1642-1645.

Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Neural Information Processing Systems* (pp. 1689-1696). Cambridge, MA: The MIT Press.

Tsitsiklis, J. N. (1994). Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning, 16,* 185-202.

Tsitsiklis, J. N. & VanRoy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning, 22,* 59-94.

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C. et al. (2001). I know what you are doing. a neurophysiological study. *Neuron, 31,* 155-165.

Vasta, R., Haith, M. M., & Miller, S. A. (1999). *Child Psychology: A Modern Science*. (3rd ed.) New York: John Wiley & Sons.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In A. McCallum & S. Roweis (Eds.), *25th International Conference on Machine learning* (pp. 1096-1103). Omnipress.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci., 27,* 468-474.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412,* 43-48.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-Learning. *Machine Learning, 8,* 279-292.

Whiteson, S. & Stone, P. (2003). Concurrent layered learning. In *International Joint Conference on Autonomous Agents & Multi-agent Systems*.

Wickens, J. & Kotter, R. (1995). Cellular Models of Reinforcement. In J.C.Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 187-214). Cambridge, MA: The MIT Press.

Williams, R. J. & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y.Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, Architectures, and Applications* (pp. 433-486). Hillsdale, NJ.: Erlbaum.

Worgotter, F. & Porr, B. (2005). Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Comput., 17,* 245-319.

Yelnik, J. (2002). Dysfonctionnement des noyaux gris centraux. *Rev.Neurol.(Paris), 158,* 33-41.

Zipser, D. & Torres, E. (2007). Computing movement geometry: a step in sensory-motor transformations. *Prog.Brain Res., 165,* 411-424.

Zweifel, L. S., Parker, J. G., Lobb, C. J., Rainwater, A., Wall, V. Z., Fadok, J. P. et al. (2009). Disruption of NMDAR-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. *Proc.Natl.Acad.Sci.U.S A, 106,* 7281-7288.

Annexe I. Dérivée, gradient et optimisation

La plupart des modèles d'apprentissage décrits dans cette thèse ont recours à des méthodes d'optimisation. Optimiser les paramètres d'une fonction consiste à trouver les valeurs de ces paramètres pour lesquelles la fonction est à son minimum (ou maximum). Cette annexe explique comment optimiser les paramètres d'une fonction en utilisant la dérivée, c'est-à-dire la pente (dans l'espace des paramètres) de la fonction. Ces principes de base sont essentiels à la bonne compréhension du Chapitre 2.

I.1 Optimisation unidimensionnelle et dérivée première

Soit une fonction *f()* qui calcule l'erreur d'un système donné en fonction d'un seul paramètre *w*. Si tout le reste est constant (les données d'entrées $\boldsymbol{x_p}$ du système par exemple), la fonction peut alors s'écrire sous la forme *f(w)*. Un objectif fréquent est de trouver la valeur de *w* pour laquelle l'erreur du système (la fonction *f()*) est minimale. Dans plusieurs des situations décrites dans cette thèse, la valeur de *f()* ne peut être calculée que pour une seule valeur de *w* à la fois et le résultat doit être utilisé tout de suite. Par exemple, si calculer *f()* est très couteux, il est préférable de choisir les valeurs de *w* avec parcimonie et de façon intelligente. Ça peut aussi être une contrainte contextuelle, comme lorsque le système est biologique.

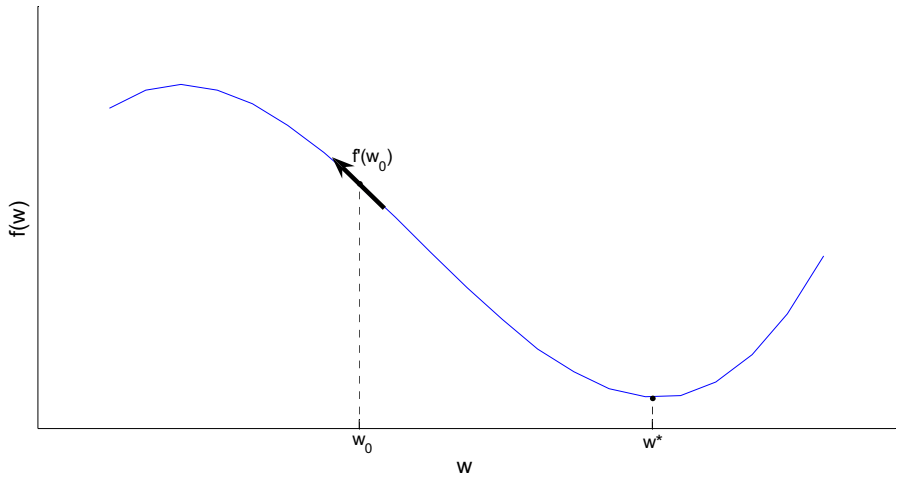


Figure I-1 : Courbe des valeurs d'une fonction *f* (Équation I-3) en fonction du paramètre *w* et dérivée de *f()* au point $w = w_0$.

Soit un point de départ $w = w_0$. Une méthode simple pour trouver le paramètre w minimisant f dans un voisinage de w_0 est de calculer la dérivée de $f()$ par rapport à w pour la valeur de w courante (w_0). La dérivée $df()/dw$ sera ici noté plus simplement par $f'()$. $f'(w_0)$ est donc la pente de $f()$ au point w_0 (flèche noire sur la Figure I-1). Pour trouver le point w^* où $f()$ est minimale, il suffit de calculer une nouvelle valeur pour w, dénotée ici par w_1, en faisant un petit pas dans la direction opposée à la dérivée :

$$w_1 = w_0 - \alpha f'(w_0)$$ **Équation I-1**

où α est une constante (généralement $0 < \alpha < 1$) appelée facteur d'apprentissage (ou *learning rate*, en anglais). En répétant plusieurs fois cette étape (appelé une itération),

$$w_{k+1} = w_k - \alpha f'(w_k)$$ **Équation I-2**

w_k se rapproche peu à peu de la valeur optimale w^* comme le montre la Figure I-2.

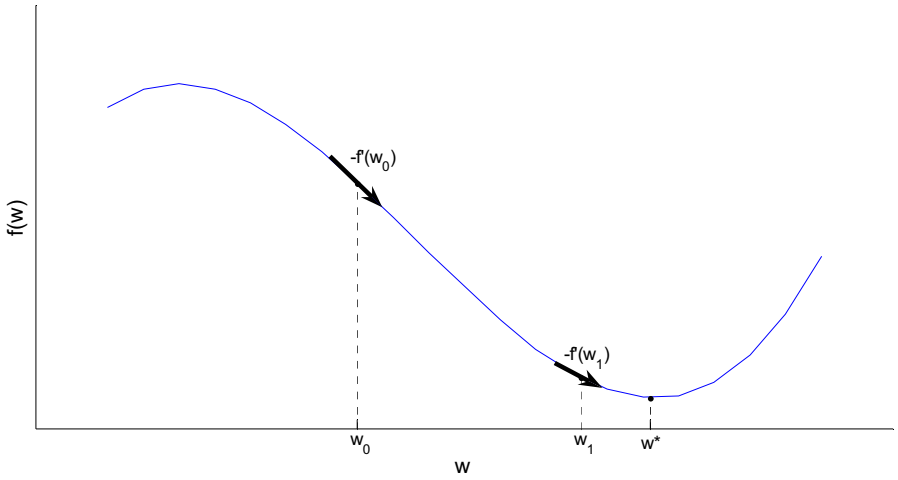


Figure I-2 : Courbe des valeurs d'une fonction f (Équation I-3) en fonction du paramètre w et dérivées négatives de $f()$ aux points $w = w_0$ et $w = w_1$ (après une itération).

Par exemple soit la fonction $f()$ (Figure I-1 et Figure I-2) à optimiser avec sa dérivée $f'()$ suivante :

$$f(w) = w^3 - 2w^2 + 5$$ **Équation I-3**

$$f'(w) = 3w^2 - 4w$$ **Équation I-4**

Si le point de départ est $w_0 = 0.5$, alors la dérivée est $f'(0.5) = -1.25$. L'application de la règle de l'Équation I-1 (avec $\alpha = 0.5$ par exemple) donne $w_1 = w_0 - \alpha f'(w_0)$

= 0.5 - 0.5 x -1.25 = 1.125 (Figure I-2). $w_2 = w_1 - \alpha f'(w_1)$ = 1.1250 - 0.5 x -0.7031 = 1.4766 est ensuite calculé de la même façon (Équation I-2). Les résultats pour les six premières itérations figurent dans le Tableau I-I ci-dessous.

Tableau I-I : Tableau des valeurs de w_k, $f'()$, et $f()$ pour les six premières itérations.

k=	w_k=	$f'(w_k)$	$f(w_k)$
0	0.5000	-1.2500	4.6250
1	1.1250	-0.7031	3.8926
2	1.4766	0.6345	3.8588
3	1.1593	-0.6052	3.8701
4	1.4619	0.5639	3.8500
5	1.1799	-0.5430	3.8583
6	1.4514	0.5142	3.8444

D'une façon générale $f(w_k)$ diminue et rapidement, w_k oscille autour de l'optimum local w^* tout en réduisant sa distance à celui-ci. Lorsque la valeur de $f()$ ne varie presque plus, le processus peut être arrêter. Il est important de noter que ce processus conduit à un minimum local dans le voisinage de w_0, mais n'assure en rien de trouver le minimum global de la fonction.

I.2 Optimisation multidimensionnelle et gradient

Le même principe s'applique lorsque la fonction $f()$ dépend de plusieurs paramètres comme le vecteur $\boldsymbol{w} = [w_1, \ldots, w_N]$, plutôt que d'un seul paramètre w (à partir d'ici, le temps ne sera plus indicé pour les paramètres et les équations récursives pour les paramètres, comme l'Équation I-2, seront écrites sous la forme de l'Équation I-5 ou l'Équation I-6). Dans ce cas, il faut calculer la pente par rapport à chaque dimension (chaque w_i). Le vecteur de ces dérivées partielles, dénotées individuellement par $\partial f()/\partial w_i$, s'appelle le gradient. Dans ce cas, il faut donc se diriger dans la direction opposée au gradient (Figure I-3). Pour se faire, il suffit d'utiliser la dérivée partielle de chaque dimension et d'en changer le signe. Pour chacun des paramètres w_i, l'Équation I-2 devient donc l'Équation I-5. Plus généralement, la règle de mise à jour des paramètres $\boldsymbol{w}$ à chaque itération sera notée par l'Équation I-6. La

nouvelle valeur des paramètres du côté gauche sera déterminée par le résultat de l'équation à droite.

$$w_i \leftarrow w_i - \alpha \frac{\partial f(\boldsymbol{w})}{\partial w_i}$$ **Équation I-5**

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} f(\boldsymbol{w})$$ **Équation I-6**

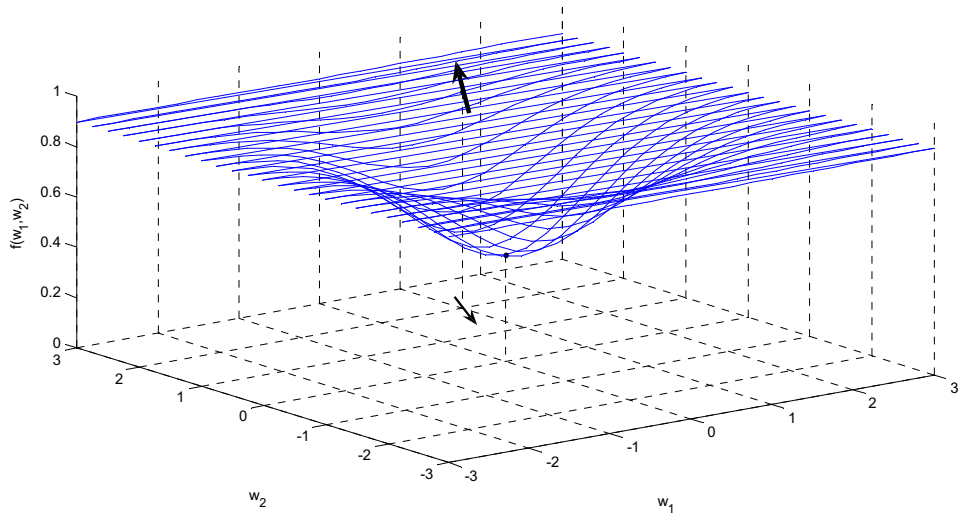


Figure I-3 : Valeur de la de *f* en fonction d'un vecteur de paramètres *w* à deux dimensions. La flèche supérieure indique la direction du gradient, la petite flèche projetée sur le plan *(w_1,w_2)* indique la direction de la correction à apporter pour se rapprocher du minimum de la fonction *f*.

Par exemple, soit une nouvelle fonction *f()* (Figure I-3) avec deux paramètres à optimiser, w_1 et w_2, et ses dérivées partielles suivantes :

$$f(w_1, w_2) = -\frac{3}{2\pi} e^{-\frac{1}{2}(w_1^2 + w_2^2)} + .9$$ **Équation I-7**

$$\frac{\partial f(w_1,w_2)}{\partial w_i} = \frac{3}{2\pi} e^{-\frac{1}{2}(w_1^2 + w_2^2)} w_i$$ **Équation I-8**

Les résultats des premières itérations de l'Équation I-6 (combinée à l'Équation I-8) sont compilés dans le Tableau I-II suivant (α = 0.99) :

Tableau I-II : Tableau des valeurs pour chaque itération.

k=	$w_{1,k}$=	$w_{2,k}$=	$\partial f(w_{1,k},w_{2,k})/\partial w_1$	$\partial f(w_{1,k},w_{2,k})/\partial w_2$	$f(w_{1,k},w_{2,k})$
0	0.8000	1.5000	0.0400	0.0750	0.7874
1	0.7604	1.4257	0.0437	0.0820	0.7706
2	0.7171	1.3445	0.0477	0.0894	0.7505
3	0.6699	1.2561	0.0516	0.0968	0.7267

4	0.6188	1.1603	0.0553	0.1037	0.6989
5	0.5640	1.0576	0.0584	0.1094	0.6672
…	…	…	…	…	…
20	.0294	0.0551	0.0062	0.0117	0.4235

Le chemin suivi par le processus d'optimisation de *f()* dans l'espace des paramètres est montré à la Figure I-4. Celui-ci ressemble beaucoup à une boule déposée sur une surface ayant un relief. La boule se déplace tranquillement vers le creux le plus proche, où elle peut osciller (la colonne w_k du Tableau I-I est un bon exemple d'oscillation) un peu avant de s'arrêter.

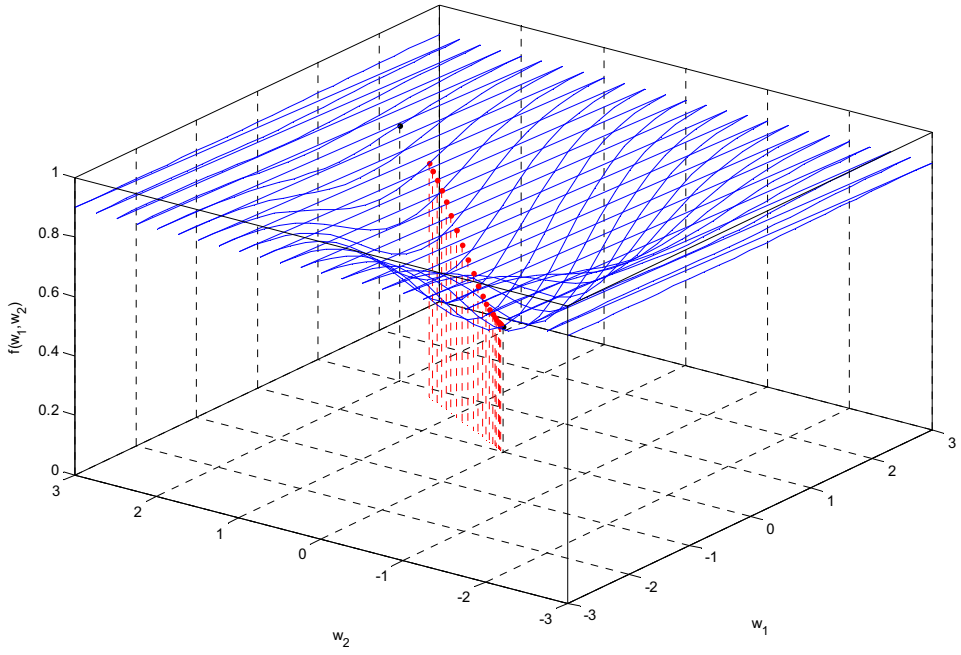


Figure I-4 : Valeur de la de *f* en fonction d'un vecteur de paramètres *w* à deux dimensions sur une vingtaine d'itérations. Chaque itération est marquée en rouge.

Annexe II. Alternative Time Representation in Dopamine Model, Supplemental Material

II.1 Supplemental Pseudocode

```
Model()
//Activity variables (reset to 0 between each simulation block)
in_td,t        //Activity of cortico-striatal projections(toTD)(vector)
r_td,t         //Hedonic reward signal (to TD)
δ_td,t         //TD error signal (dopaminergic neuron)
target_lstm,t  //Target to predict in the cortex (LSTM)
in_lstm,t      //Activity projections to cortex (to LSTM)(vector)
//Code
t=0
{while environment block running
      //Read the environment
      Read x_cs,t x_us,t
      //Process and update TD
      in_td,t = [x_cs,t c_1,1,t-1 c_1,2,t-1 c_2,1,t-1 c_2,2,t-1 y_us,t-1]
      r_td,t = x_us,t
      δ_td,t = processTD(in_td,t, r_td,t)
      //Update the LSTM weights
      target_lstm,t = x_us,t
      if t≠0 then trainLSTM(target_lstm,t, δ_td,t)
      //Process the LSTM activity
      in_lstm,t = [x_cs,t x_us,t]
      [c_1,1,t c_1,2,t c_2,1,t c_2,2,t y_us,t] = bound(processLSTM(in_lstm,t),[0 1])
      //Next time step
      t = t+1
}
```

Pseudocode II-1: Model with mesocortical projection main routine.

```
Model()
//Activity variables (reset to 0 between each simulation block)
in_td,t        //Activity of cortico-striatal projections(toTD)(vector)
r_td,t         //Hedonic reward signal (to TD)
δ_td,t         //TD error signal (dopaminergic neuron)
target_lstm,t  //Target to predict in the cortex (LSTM)
in_lstm,t      //Activity projections to cortex (to LSTM)(vector)
//Code
t=0
{while environment block running
      //Read the environment
      Read x_cs,t x_us,t
      //Process and update TD
      in_td,t = [x_cs,t c_1,1,t-1 c_1,2,t-1 c_2,1,t-1 c_2,2,t-1 y_us,t-1]
      r_td,t = x_us,t
      δ_td,t = processTD(in_td,t, r_td,t)
      //Update the LSTM weights
      target_lstm,t = x_us,t
      if t≠0 then trainLSTM(target_lstm,t, 0)
      //Process the LSTM activity
      in_lstm,t = [x_cs,t x_us,t]
```

```
      [c1,1,t c1,2,t c2,1,t c2,2,t yus,t] = bound(processLSTM(in1stm,t),[0,1])
      //Next time step
      t = t+1
}
```

Pseudocode II-2: Basic model main routine.

```
y = bound(x, [a b])
{
      y = min(max(a,x),a)
}
```

Pseudocode II-3: bound: Trim the value of x within boundaries [a b].

```
[c1,1,t c1,2,t c2,1,t c2,2,t yus,t] = processLSTM(int)
//Arguments
int           //Pre-synaptic extra-cortical afferances activity(vector)
//Returns
Ci,j,t        //Memory cell (i,j) activity
Yus,t         //Network output (US prediction)
//Constants
λ = .8        //Eligibility trace discounting factor
//Activity variables (reset to 0 between each simulation block)
xt            //Pre-synaptic afferances activity
ut            //Afferances eligibility traces (vector)
zin,j,t       //Input gate j dendritic activity
zφ,j,t        //Forget gate j dendritic activity
zout,j,t      //Output gate j dendritic activity
yin,j,t       //Input gate j axonal activity
yφ,j,t        //Forget gate j axonal ctivity
yout,j,t      //Output gate j axonal activity
zc,i,j,t      //Memory cell (i,j) dendritic activity
ci,j,t        //Memory cell (i,j) axonal activity
ui,j,t        //Memory cell (i,j) eligibility traces
//Synaptic weights (initialized only once with Uniform([-.1.1])
Win,j         //Input gate dendritic synapses (vector)
Wφ,j          //Forget gate dendritic synapses (vector)
Wout,j        //Output gate dendritic synapses (vector)
Wc,i,j        //Memory cell (i,j) dendritic synapses (vector)
//Functions
fin ≡ sig[0,1] //Input gate activation function
fφ ≡ sig[0,1]  //Forget gate activation function
fout ≡ sig[0,1]//Output gate activation function
g ≡ sig[-1,1]  //Memory cell input side activation function
h ≡ sig[-1,1]  //Memory cell output side activation function
fus ≡ sig[0,1] //Network output (US predictor) activation function
//Code
{
      //Create afferance vector and e-traces for memory blocks
      xt = [1 int y1,1,t-1 y1,2,t-1 y2,1,t-1 y2,2,t-1]
      ut = bound(etrace(λ, ut-1, xt), [-1 1])
      //Loop over memory blocks
      for j = 1 to 2
            //Process input gate
            zin,j,t = Win,j*[xt c1,j,t-1 c2,j,t-1]
            yin,j,t = fin,j(zin,j,t)
```

```
            //Process forget gate
            z_φj,t = W_φ,j*[x_t c_1,j,t-1 c_2,j,t-1]
            y_φ,j,t = f_φ,j(z_φ,j,t)
            //Loop over memory cells
            for i = 1 to 2
                  z_c,i,j,t = W_c,i,j*x_t
                  c_i,j,t = y_in,j,t*g(z_c,i,j,t) + y_φ,j,t*c_i,j,t-1
                  u_i,j,t = bound(etrace(λ, u_i,j,t-1, c_i,j,t), [-1 1])
            next i
            //Process output gate
            z_out,j,t = W_out,j*[x_t c_1,j,t c_2,j,t]
            y_out,j,t = f_out,j(z_out,j,t)
            //Process memory block outputs
            for i = 1 to 2
                  y_i,j,t = y_out,j,t*h(c_i,j,t)
            next i
      next j
      //Process network output
      z_us = [1 y_1,1,t y_1,2,t y_2,1,t y_2,2,t]
      y_us = f_us(W_us*z_us)
}
```

Pseudocode II-4: LSTM activity processing pseudocode. $sig_{[a,b]}$ means sigmoidal activation function with range [a b]. There is also some gradient computations made in this procedure that are used in the weight update procedure, but that are not detailed here. See (Hochreiter & Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2002) for details.

```
u_t = etrace(λ, u_t-1, x_t)
{
      //If same sign, e-trace as usual
      if u_t-1*x_t > 0 then u_t = λ*u_t-1 + x_t
      //If sign changed, reset traces
      else u_t = x_t
}
```

Pseudocode II-5: etrace: Eligibility trace for signed variables resets the trace when the sign of the variable changes.

```
trainLSTM(target_t, δ_td,t)
//Arguments
target_t      //Target value to be predicted
δ_td,t        //Mesocortical projections
//Constants
α_0 = .5      //Basic learning rate
β = .5        //DA factor
//Local variables
a_t           //Effective learning rate
//Code
{
      //Compute effective learning rate
      α_t = α_0 + β|δ_td,t|
      //Compute weight updates
      //Whenever some input x_t should be used, use trace u_t instead.
      ...}
```

Pseudocode II-6: LSTM synaptic weight update pseudocode. See (Hochreiter & Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2002) for details.

```
δ_t = processTD(in_t, r_t)
//Arguments
in_t          //Pre-synaptic afferances activity (vector)
r_t           //Hedonic reward signal
//Returns
δ_t           //Dopaminergic neuron
//Constants
α = .1        //Learning rate
γ = .98       //Discounting factor
λ = .9        //Eligibility trace discounting factor
//Activity variables (reset to 0 between each simulation block)
p_t           //Predictive neuron
u_t           //Afferances eligibility traces (vector)
//Synaptic weights (initialized only once)
W_p = .1      //Predictive neuron's dendritic synapses (vector)
//Code
{
      //Prediction neuron
      p_t = W_p * in_t
      //Dopaminergic neuron
      if t = 0 then δ_t = 0  //On first time step, there is no update
      else δ_t = r_t + γ*p_t - p_t-1
      //Eligibility traces for prediction neurons
      u_t = bound(λ*u_t-1 + in_t-1, [-1 1])
      W_p = W_p + α*δ_t*u_t
}
```

Pseudocode II-7: TD(λ) pseudocode.

Annexe III. Affiches

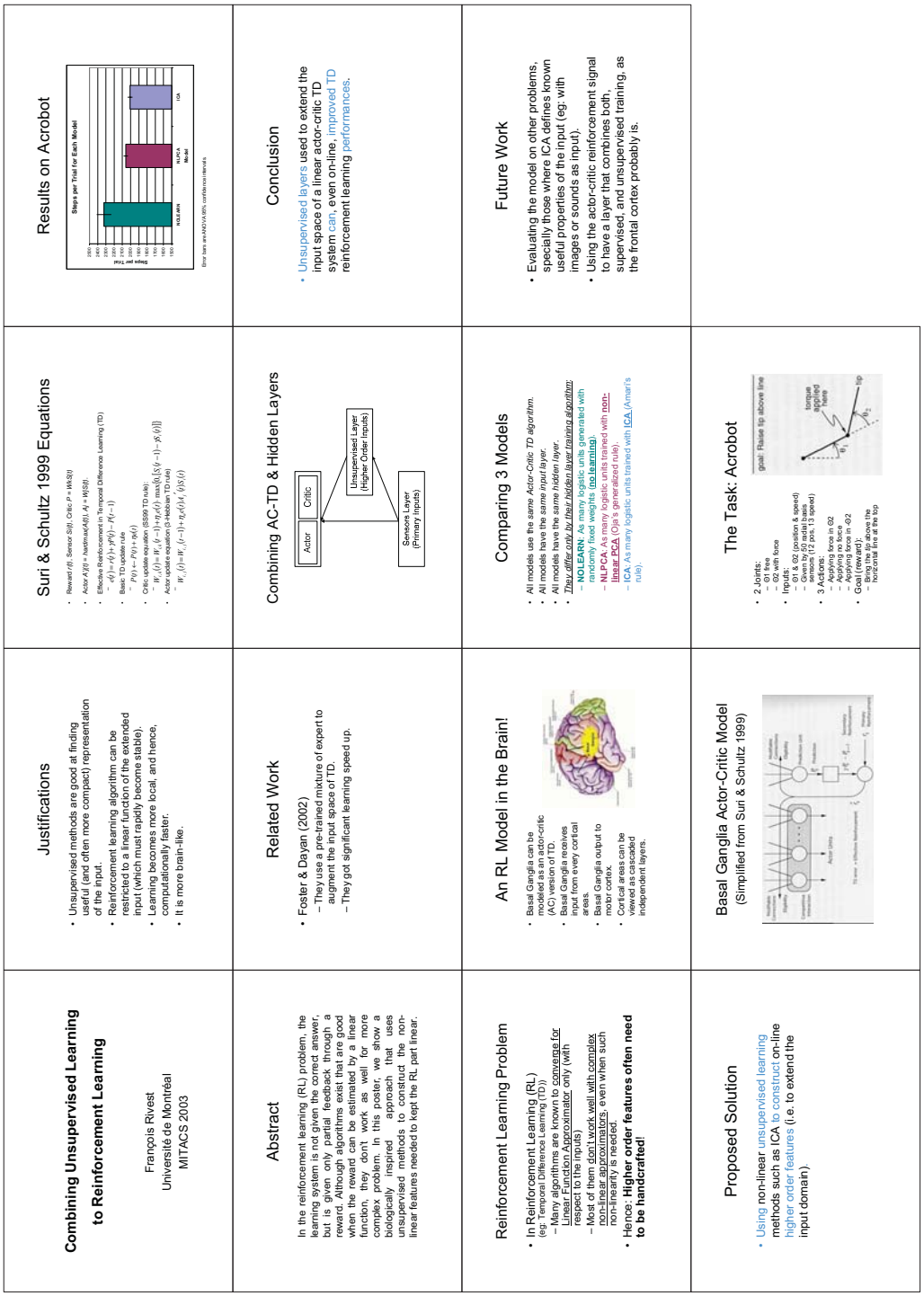


Figure III-1 : Rivest, F. (2003) Combiner l'apprentissage non-supervisé à l'apprentissage par renforcement/Combining Unsupervised Learning to Reinforcement Learning. MITACS Quebec Interchange, Mathematics of Information Technology and Complex Systems.

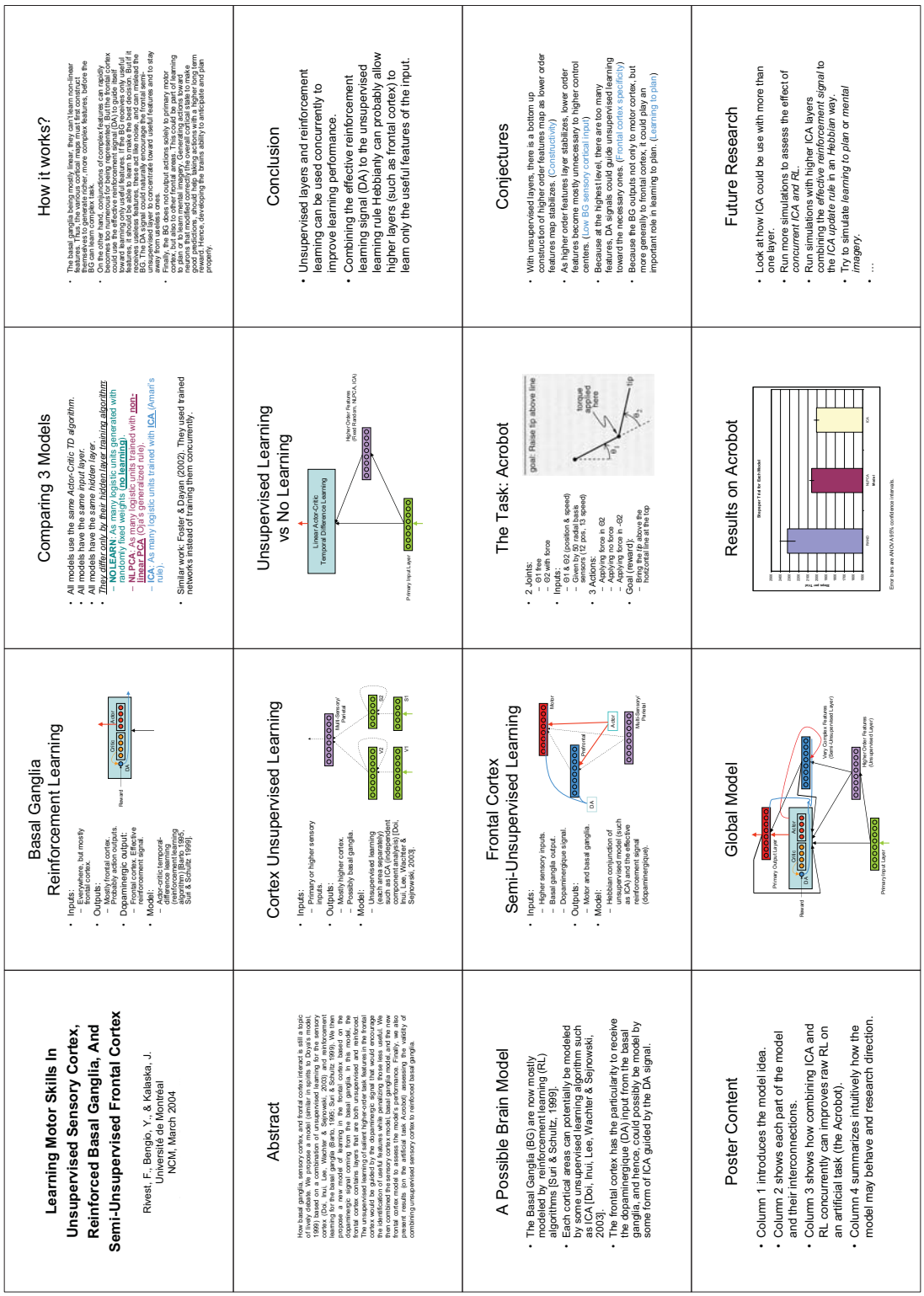


Figure III-2 : Rivest, F., Bengio, Y., & Kalaska, J.F. (2004) Learning Motor Skills In Unsupervised Sensory Cortex, Reinforced Basal Ganglia, And Semi-Unsupervised Frontal Cortex, Motor Learning & Plasticity Satellite Meeting, Neural Control of Movement Conference.

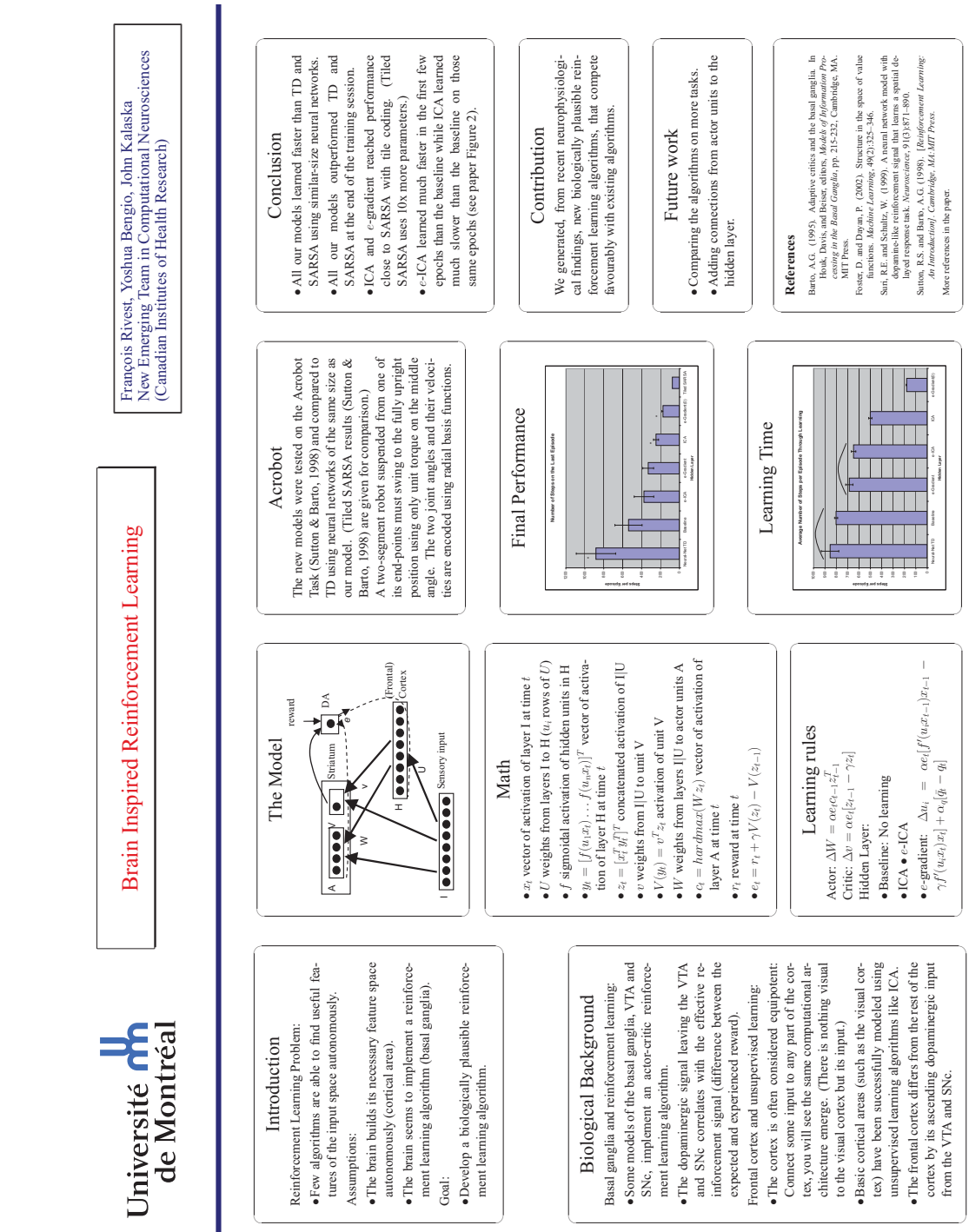


Figure III-3 : Rivest, F., Bengio, Y, & Kalaska, J.F. (2005) Brain Inspired Reinforcement Learning. ***Neural Information Processing Systems, NIPS 2004.***

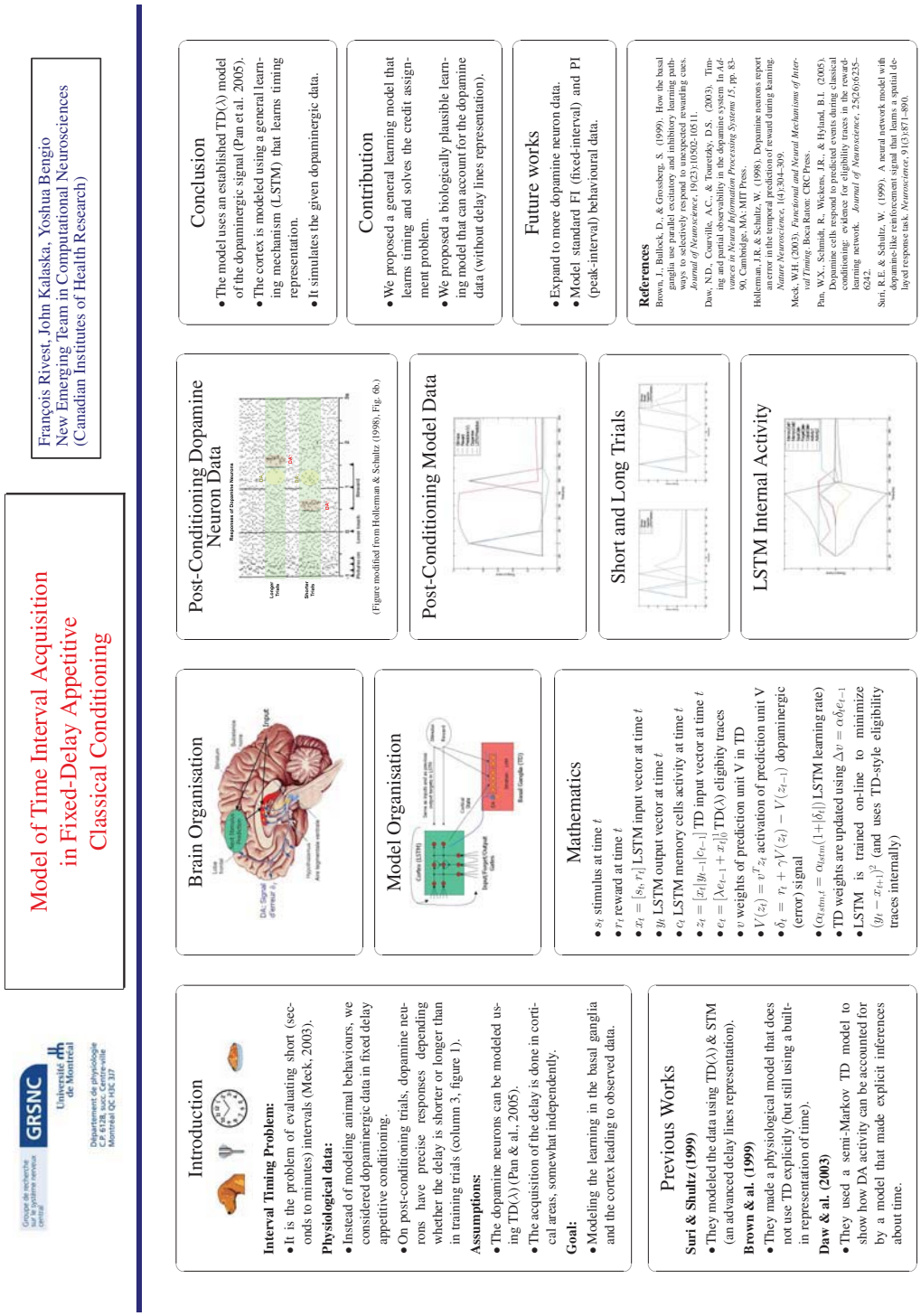


Figure III-4 : Rivest, F., Kalaska, J.F., & Bengio, Y. (2006) Model of Time Interval Acquisition in Fixed-Delay Appetitive Classical Conditioning. ***XVIIIe symposium international, Computational Neuroscience Computationnelle*****, Groupe de recherche sur le système nerveux central.**

Modèle neuroinformatique du signal dopaminergique et de l'intervalle de temps en conditionnement à délai fixe

François Rivest, John Kalaska, Yoshua Bengio

Groupe de Recherche sur le système nerveux central
Université de Montréal

Introduction

Les neurones dopaminergiques et l'intelligence artificielle:

- L'activité des neurones dopaminergiques semble être corrélée au signal d'erreur d'un d'algorithme d'apprentissage machine nommé TD dans différentes situations de conditionnement classique et instrumental.

Limite:

- Cependant, les modèles de TD reproduisant l'activité de ces neurones dans des tâches à délais fixes ne construisent pas de représentation raisonnable du temps ou de l'environnement comme le cerveau.

Problème:

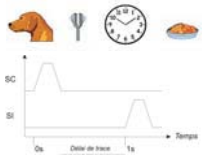
- Même dans la littérature sur la perception du temps, il n'y a pas vraiment de modèle neuro-biologique de la façon dont le cerveau représente le temps.

Mais:

- En conditionnement à délai fixe, le cerveau acquiert le délai (voir figure en bas).

Tâche

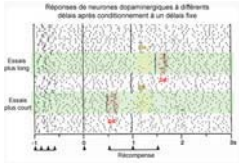
- Conditionnement de trace à délai fixe.



Hypothèse

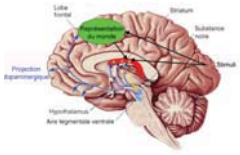
- C'est un mécanisme d'apprentissage général dans le cortex, qui apprend son environnement, et acquiert le délai.
- C'est cette représentation de l'état de l'environnement qui fournit l'information nécessaire à TD pour se comporter de cette façon.

Réponse des neurones dopaminergiques

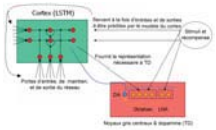


(Figure modifiée de Hollerman & Schultz (1998), Fig. 6b.)

Organisation du cerveau

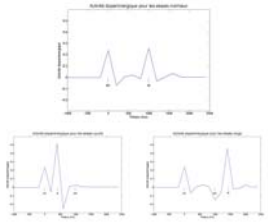


Organisation du modèle

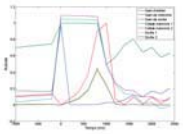


- Le cortex apprend de façon non-supervisée, à prédire son environnement. On modélise le cortex frontal à l'aide d'un réseau de neurones appelé *réseau de longue mémoire à court terme* (LSTM) qui apprend à prédire le stimulus suivant.
- Ce réseau développe ainsi une représentation de l'état de l'environnement. C'est cette représentation qui sert de base à l'algorithme TD pour apprendre à prédire les récompenses à venir.
- Pour faire cet apprentissage, TD doit calculer un signal d'erreur de prédiction de récompense. C'est ce signal qui est corrélé au signal des neurones dopaminergiques.

Résultats



Neurones du LSTM



Conclusion

- Un réseau de neurones récurrent (LSTM) réussit très bien à apprendre le délai dans la tâche (hypothèse d'Hopson, dans Meck 2003).
- La représentation développée par le réseau est suffisante pour expliquer le comportement du signal d'erreur dans TD et donc des neurones dopaminergiques.

Contribution

- Ce modèle du signal dopaminergique acquiert complètement la tâche (sans ligne de délai).
- Ce modèle biologique du temps résout seul le problème de découvrir quand commencer et arrêter.
- Ce modèle réagit différemment aux conditionnements de trace et de délai.
- Ce modèle du temps n'est pas basé sur une horloge interne.
- L'étude de la représentation du LSTM semble correspondre à ce que l'on retrouve parfois dans le cortex.
- Permet l'étude des rôles possibles de la dopamine dans le cortex frontal.
- Permet l'étude de l'interaction de l'apprentissage du cortex et des noyaux gris centraux.

References

Hollerman, J.R. & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4):304–309.

Suri, R.E. & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871–890.

Meck, W.H. (2003). *Functional and Neural Mechanisms of Interval Timing*. Boca Raton: CRC Press.

Daw, N.D., Courville, A.C., & Touretzky, D.S. (2006). Representation and Timing in Theories of the Dopamine *Neural Computation*, 18:1637–1677.

Contact

Financement

Un projet de recherche de l'Équipe en voie de formation en neuroscience computationnelle des Instituts de Recherche en Santé du Canada.

Groupe de recherche sur le système nerveux central
GRSNC
Université de Montréal
Département de physiologie
C.P. 6128, succ. Centre-ville
Montréal QC H3C 3J7

Figure III-5 : Rivest, F., Kalaska, J.F., & Bengio, Y. (2007) Modèle neuroinformatique du signal dopaminergique et de l'intervalle de temps en conditionnement à délai fixe. ***L'approche transdisciplinaire des sciences cognitives, ACFAS 2007.***

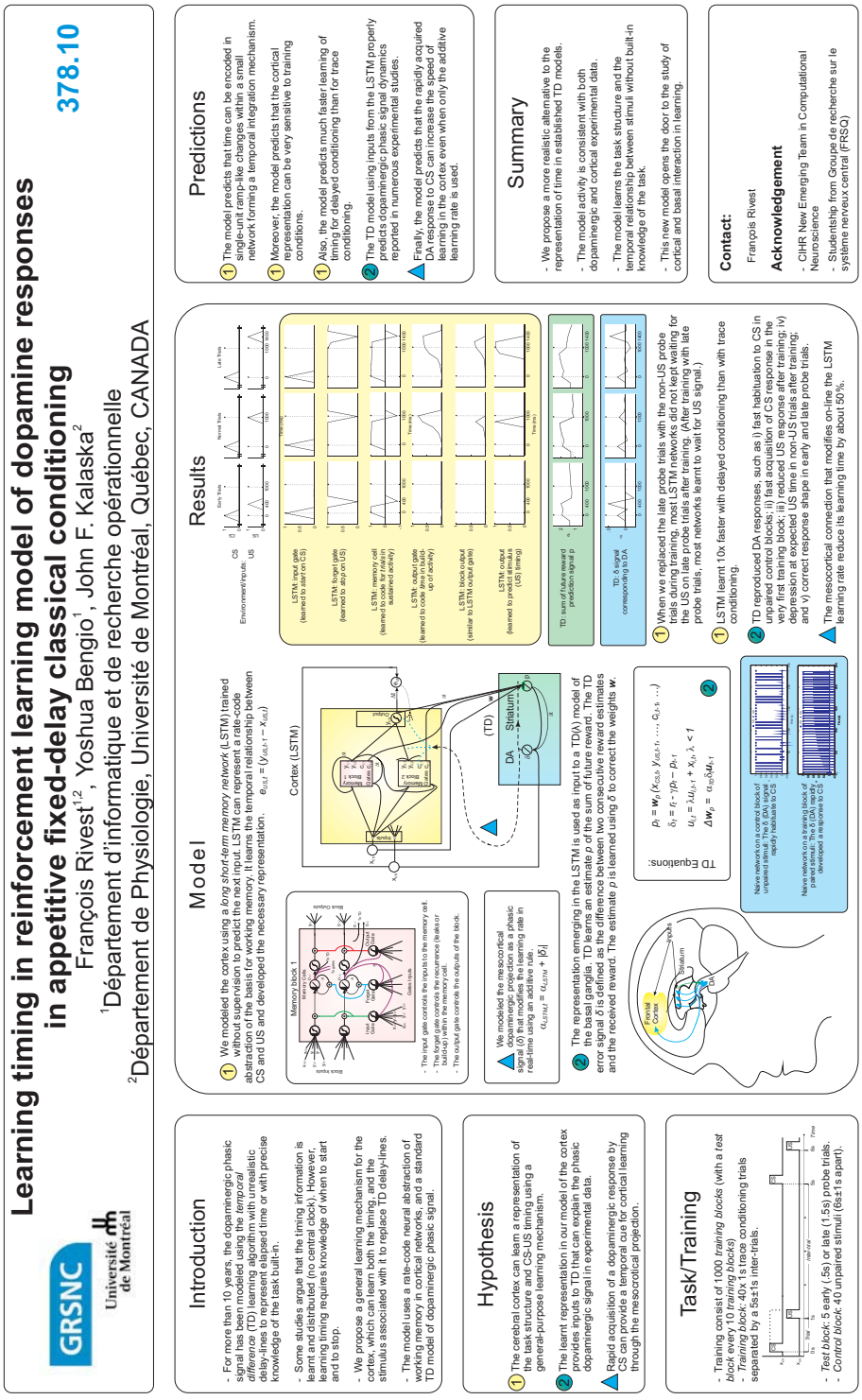


Figure III-6 : Rivest, F., Bengio, Y., & Kalaska, J.F. (2008) Learning timing in reinforcement learning model of dopamine responses in appetitive fixed-delay classical conditioning. ***Society for Neuroscience Abstracts, SFN 2008.***

Annexe IV. Publications

Rivest, F., Bengio, Y., & Kalaska, J.F. (2010) Alternative Time Representation in Dopamine Models. *Journal of Computational Neuroscience* **28**(1):107-130.

Dandurand, F., Shultz, T.R., & **Rivest, F.** (2007) Complex problem solving with reinforcement learning. In *Proceeding of the 6th IEEE International Conference on Development and Learning (ICDL-2007)*, pp. 157-162. IEEE.

Shultz, T.R., **Rivest, F.**, Egri, L., Thivierge, J.-P., & Dandurand, F. (2007) Could Knowledge-based Neural Learning Be Useful in Developmental Robotics? The Case of KBCC. *International Journal of Humanoid Robotics* (Special Issue on Autonomous Mental Development) **4**(2):245-279.

Thivierge, J.-P., **Rivest, F.**, & Monchi, O. (2007) Spiking Neurons, Dopamine, and Plasticity: Timing Is Everything, But Concentration Also Matters. *Synapse* **61**:375-390.

Shultz, T.R., **Rivest, F.**, Egri, L., & Thivierge, J.P. (2006) Knowledge-based learning with KBCC. *Proceedings of the Fifth International Conference on Development and Learning ICDL 2006.* Department of Psychological and Brain Sciences, Indiana University, Bloomington.

Rivest, F., & Shultz, T.R. (2005) Learning with Both Adequate Computational Power and Biological Realism. *Proceedings of the 2005 Canadian Artificial Intelligence Conference: Workshop on Correlation Learning*, pp. 15-23. University of Victoria, Victoria, BC.

Rivest, F., Bengio, Y, & Kalaska, J.F. (2005) Brain Inspired Reinforcement Learning. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pp. 1129-1136. MIT Press, Cambridge, MA.

Bellemare, M.G., Precup, D., & **Rivest, F.** (2004) Reinforcement Learning Using Cascade-Correlation Neural Networks. *Technical Report RL-3.04.* School of Computer Science, McGill University.

Rivest, F., & Shultz, T.R. (2004) Compositionality in a Knowledge-based Constructive Learner. Papers from the *2004 AAAI Symposium*, Technical Report FS-04-03, pp. 54-58. AAAI Press: Menlo Park, CA.

Rivest, F., & Precup, D. (2003). Combining TD-learning with Cascade-correlation Networks. *Proceedings of the Twentieth International Conference on Machine Learning,* pp. 632-639. AAAI Press.

Thivierge, J.-P., **Rivest, F.**, & Shultz, T.R. (2003). A Dual-phase Technique for Pruning Constructive Networks. *Proceedings of the IEEE International Joint Conference on Neural Networks 2003*, pp. 559-564.

Shultz, T. R., & **Rivest, F.** (2003). Knowledge-based cascade-correlation: Varying the size and shape of relevant prior knowledge. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics,* pp. 631-638. Tokyo: Springer-Verlag.

Rivest, F. (2002) *Knowledge-Transfer in Neural Network: Knowledge-Based Cascade-Correlation.* M.Sc. Thesis, School of Computer Science, McGill University.

Rivest, F. & Shultz, T.R. (2002) Application of Knowledge-based Cascade-correlation to Vowel Recognition, *IEEE International Joint Conference on Neural Network 2002*, pp. 53-58. IEEE Society Press.

Shultz, T.R. & **Rivest, F.** (2001) Knowledge-based Cascade-correlation: Using Knowledge to Speed Learning, *Connection Science* **13**:1-30.

Shultz, T.R. & **Rivest, F.** (2000) Knowledge-based Cascade-correlation: An Algorithm for Using Knowledge to Speed Learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 871-878. San Francisco, CA: Morgan Kaufmann.

Shultz, T.R. & **Rivest, F.** (2000) Knowledge-based Cascade-correlation, *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Network 2000*, pp. V641-V646. Los Alamitos, CA: IEEE Society Press.

Curriculum Vitae

Formation

- Candidat au PhD en Informatique à l'Université de Montréal. (2002-2009 prévue)
- *Computational Neuroscience Summer Course*, Center for Neural Dynamics, University of Ottawa. (2007)
- Maîtrise en Science informatique mise au tableau d'honneur du doyen de l'Université McGill. (2000-2002)
- Bac en Science avec spécialisation double en mathématique et Science informatique et mineure en Science cognitive de l'Université McGill. (1996-2000)
- *Second Annual Undergraduate Summer Workshop in Cognitive Science*, Institute for Research in Cognitive Science, University of Pennsylvania. (1999)

Bourses

- Bourse pour projet collaboratif du Groupe de recherche sur le système nerveux central (GRSNC). (2007-2009)
- Bourse étudiante de l'équipe en voie de formation en neuroscience informatique (NET) des Instituts de recherche en santé du Canada (IRSC). (2002-2007)
- Bourse de recherche en milieu pratique du Fonds pour la Formation de Chercheur et l'Aide à la Recherche (FCAR-MRST) au Centre de Recherche Informatique de Montréal (CRIM). (2001-2002)

Expériences de travail

- Professionnel de recherche au *Département de génie électrique, École de Technologie Supérieure*, Université du Québec. (2001-2002)
- Étudiant boursier au *Centre de recherche informatique de Montréal* (CRIM). (2001-2002)
- Consultant pour la firme *Les Services Topo-Info* dans le développement d'un système de dessin automatisé de données topographiques. (Hiver 2000)
- Aide-enseignant à *l'École de sciences informatiques*, Université McGill. (1998-2000)

- Assistant de recherche au *Laboratory for Naturel and Simulated Cognition,* Université McGill. (1997-2000)

Formations données et présentations invitées sélectionnées

- *Computationnal Neuroscience of Reinforcement Learning*, cours PSYCH 532 *Cognitive Science*, Department of Psychology, McGill University. (2007-2009)
- *Apprentissage par renforcement*, cours IFT3395/6390 *Fondements de l'apprentissage machine*, Département d'informatique et de recherche opérationnelle, Université de Montréal. (2008)
- Atelier Matlab (5 séances de 2 heures) pour les membres (étudiants, techniciens ou professeurs) du Groupe de recherche sur le système nerveux central (GRSNC), Université de Montréal. (2007)
- *Real Neurons for Machine Learning,* Séminaires UdeM-McGill-MITACS. (2007)
- *L'apprentissage dans le cerveau,* Séminaires UdeM-McGill-MITACS. (2006)

Autres activités universitaires

- Évaluateur pour un article soumis à *PLOS One.* Évaluateur (décliné) pour un article soumis à *Transactions on Autonomous Mental Development.* (2009)
- Étudiant bénévole pour l'*International Conference on Machine Learning* et le *Multidisciplinary Symposium on Reinforcement Leanring* à Montréal. (2009)
- Fondateur et administrateur de la liste de diffusion *Neuroscience Computationnelle au Québec* (neuralcomp@listes.umontreal.ca). (2006-2009)
- Cofondateurs des *Rencontres bimensuels Math-Neuro* (qui ont encore lieu régulièrement) avec Paul Cisek à l'Université de Montréal. (2006-2007)
- Coorganisateur des *Séminaires en neuroscience informatique* avec Paul Cisek à l'Université de Montréal. (2006-2007)
- Évaluateur d'articles soumis au *Inductive Transfer Workshop*, lors de la conférence *Advances in Neural Information Processing Systems.* (2005)