Neuroscience of decision making : from goal-directed actions to habits

Meropi Topalidou

THESIS

PRESENTED TO

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

by **Meropi Topalidou**

OBTAIN THE TITLE OF

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Neuroscience of decision making: from goal-directed actions to habits

---

Date of defense : 10 October 2016

The jury compose of:

Thomas Boraud ...	Directeur de recherche, IMN,Bordeaux	President of the jury
Anastasia Christakou	Associative Professor, Univ. Reading	Reviewer
Benoît Girard	Directeur de recherche, ISIR,Paris ..	Reviewer
Suzanne Haber ...	Professor, Univ. Rochester	Examiner
Nicolas P. Rougier .	Chargé de recherche, INRIA,Bordeaux	Advisor

– 2016 –

Résumé Les processus de type "action-conséquence" (orienté vers un but) et stimulus-réponse sont deux composants importants du comportement. Le premier évalue le bénéfice d'une action pour choisir la meilleure parmi celles disponibles (sélection d'action) alors que le deuxième est responsable du comportement automatique, suscitant une réponse dès qu'un stimulus connu est présent. De telles habitudes sont généralement associées (et surtout opposées) aux actions orientées vers un but qui nécessitent un processus délibératif pour évaluer la meilleure option à prendre pour atteindre un objectif donné. En utilisant un modèle computationnel, nous avons étudié l'hypothèse classique de la formation et de l'expression des habitudes au niveau des ganglions de la base et nous avons formulé une nouvelle hypothèse quant aux rôles respectifs des ganglions de la base et du cortex. Inspiré par les travaux théoriques et expérimentaux de Leblois et al. (2006) et Guthrie et al. (2013), nous avons conçu un modèle computationnel des ganglions de la base, du thalamus et du cortex qui utilise des boucles distinctes (moteur, cognitif et associatif) ce qui nous a permis de poser l'hypothèse selon laquelle les ganglions de la base ne sont nécessaires que pour l'acquisition d'habitudes alors que l'expression de telles habitudes peut être faite par le cortex seul. En outre, ce modèle a permis de prédire l'existence d'un apprentissage latent dans les ganglions de la base lorsque leurs sorties (GPi) sont inhibées. En utilisant une tâche de bandit manchot à 2 choix, cette hypothèse a été expérimentalement testée et confirmée chez le singe; suggérant au final de rejeter l'idée classique selon laquelle l'automatisme est un trait subcortical.

Title Neuroscience of decision making: from goal-directed actions to habits

Abstract Action-outcome and stimulus-response processes are two important components of behavior. The former evaluates the benefit of an action in order to choose the best action among those available (action selection) while the latter is responsible for automatic behavior, eliciting a response as soon as a known stimulus is present. Such habits are generally associated (and mostly opposed) to goal-directed actions that require a deliberative process to evaluate the best option to take in order to reach a given goal. Using a computational model, we investigated the classic hypothesis of habits formation and expression in the basal ganglia and proposed a new hypothesis concerning the respective role for both the basal ganglia and the cortex. Inspired by previous theoretical and experimental works (Leblois et al., 2006; Guthrie et al., 2013), we designed a computational model of the basal ganglia-thalamus-cortex that uses segregated loops (motor, cognitive and associative) and makes the hypothesis that basal ganglia are only necessary for the acquisition of habits while the expression of such habits can be mediated through the cortex. Furthermore, this model predicts the existence of covert learning within the basal ganglia ganglia when their output is inhibited. Using a two-armed bandit task, this

hypothesis has been experimentally tested and confirmed in monkey. Finally, this works suggest to revise the classical idea that automatism is a subcortical feature.

Keywords habit, goal-directed action, decision-making, Hebbian learning, reinforcement learning, computational neuroscience, cortex, basal ganglia

Mots-clés habitude, action orientée, prise de décision, Hebbian apprentissage, renforcement apprentissage, neuroscience informatique, cortex, ganglion de la base

Laboratoire d'accueil INRIA Bordeaux Sud-Ouest
200, Avenue de la Vieille Tour
33405 Talence Cedex, France

"The only difference between screwing around and science is writing it down."

— Alex Jason

Acknowledgement

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

I would like to gratefully and sincerely thank Dr. Nicolas P. Rougier, my advisor during my PhD, for his guidance, understanding, and most importantly his patience during these three years. He contributed to a rewarding PhD experience by giving me intellectual freedom in my work, supporting my attendance at various conferences, engaging me in new ideas, and demanding a high quality of work in all my endeavors. His mentorship was paramount in providing a well rounded experience consistent my long-term career goals. For everything you've done for me, Dr. Rougier, I thank you.

I would also like to thank Dr. Thomas Boraud for introducing me to the experimental world, and providing me the opportunity to acquire new skills. Furthermore, I would like to thank Dr. Alexandros Elefteriadis for his assistance and guidance in getting my graduate career started.

A big thank you to Charlotte Héricé, my office mate during these years, that provided for some much needed humor and entertainment in what could have otherwise been a somewhat stressful laboratory environment. I would also like to thank all my colleagues from IMN and INRIA for their insightful discussions and friendship.

Finally, and most importantly, I would like to thank my parents, who introduced me to science and more specifically to medicine, and supported me through all my academic studies. I also thanked them for their faith in me and allowing me to be as ambitious as I wanted. And all of my friends, who were honored me with their friendship during these years.

Table of Contents

Neuroscience de la prise de décision : des actions dirigées vers un but aux habitudes

Meropi Topalidou

Résumé étendu

Les habitudes sont une des composantes essentielles du comportement chez les vertébrés supérieurs. Elles peuvent être complexes, de haut niveau et être exécutées rapidement, avec un minimum d'effort et ce, sans mobiliser l'attention, afin notamment de libérer celle-ci pour des fonctions plus importantes comme par exemple la recherche de proies ou l'évitement de prédateurs. Historiquement, Aristote a été le premier à proposer le terme d'habitude en termes de compétences acquises par un individu qui lui sont nécessaires pour améliorer ses performances en vue d'atteindre un objectif donné. Cependant, la recherche en neurosciences sur les habitudes est relativement récente et émane principalement du travail de William James qui a défini les habitudes comme des compétences acquises ne nécessitant pas le contrôle de l'attention. À cette même époque, Edward Thorndike, formulait la «loi d'effet», qui explique en substance que les comportements suivis par des conséquences positives seront répétés alors que des comportements suivis de conséquences négatives seront évités. Plus d'un siècle après, il n'y a toujours pas de consensus quant à la définition des habitudes et celles-ci restent sujettes à controverse tant le terme d'habitude possède de connotations différentes selon les domaines de recherche considérés, où l'on peut préférer les termes de routine, de préférence ou bien de compétences développées. Si la communauté scientifique n'a pas encore convenu des caractéristiques fondamentales d'une action pour la caractériser comme habitude, il existe cependant, un ensemble de caractéristiques qui semble lui faire consensus:

1. Les habitudes sont déclenchées par un stimulus spécifique.
2. Ce sont des actions qui ont été acquises par l'expérience, généralement après un apprentissage approfondi.
3. Les habitudes sont effectuées automatiquement, c'est-à-dire qu'elles sont exécutées rapidement par rapport à des actions dirigées par un but, et inconsciemment, c'est à dire, sans y prêter attention.
4. Enfin, la caractéristique la plus courante utilisée dans les expériences pour définir les habitudes, est leur désengagement par rapport à un but. En d'autres termes, ce type d'actions est exécuté même si leur résultat a été dévalué.

Ce désengagement des habitudes par rapport au but initial a conduit Anthony Dickinson à proposer la division du comportement instrumental en deux types opposés: le comportement dirigé vers un but et le comportement habituel. Suite à cette distinction originelle, de nombreuses études expérimentales ont été développées afin d'identifier et de comprendre les mécanismes cérébraux responsables de la production de ces deux types de comportement. La vision dominante du XX^{ème} siècle se résume par l'idée que les comportements nouveaux nécessitent une attention soutenue et une architecture flexible et dépendraient donc du cortex. A contrario, les comportements automatiques ne

nécessiteraient ni l'une ni l'autre et ne dépendraient donc pas majoritairement du cortex mais plutôt de structures sous-corticales. C'est pourquoi selon cette théorie, un rôle crucial a été attribué aux ganglions de la base (groupe de noyaux sous-corticaux) dans l'acquisition et l'expression des habitudes. Les actions dirigées vers un but pouvant être considérées sous la forme d'actions effectuées afin d'atteindre un but (A-O: *action-outcome*) alors que les habitudes peuvent quant à elles être considérées comme des stimulus déclenchant une réponse automatique (S-R: *stimulus-response*).

Ce travail de thèse porte sur l'étude et la modélisation des mécanismes de prise de décision avec une attention particulière sur les mécanismes relatifs à l'acquisition et à l'expression des habitudes chez le primate. Dans ce cadre précis, nous faisons l'hypothèse que les processus d'acquisition et d'expression des habitudes sont deux processus distincts qui peuvent être expérimentalement dissociés mettant ainsi en lumière les rôles respectifs des ganglions de la base et du cortex. En s'inspirant des travaux théoriques et expérimentaux de *Leblois et al. (2006)* et *Guthrie et al. (2013)*, nous avons conçu un modèle computationnel des ganglions de la base, du thalamus et du cortex qui utilise des boucles distinctes (moteur, cognitif et associatif) ce qui nous a permis de poser l'hypothèse selon laquelle les ganglions de la base ne sont nécessaires que pour l'acquisition d'habitudes alors que l'expression de telles habitudes peut être faite par le cortex seul. Le modèle de *Leblois et al., (2006)* a introduit un mécanisme de sélection d'action, qui dérive de la compétition entre une rétroaction positive par la voie directe et une rétroaction négative par la voie hyper-directe dans la boucle cortex - ganglions de la base - thalamus. Le modèle a été étendu dans *Guthrie et al. (2013)* afin d'explorer l'organisation parallèle des circuits dans le BG. Ce modèle comprend les principaux noyaux des ganglions de la base (saut le Globus Pallidus externe (GPe)) et est organisé le long de trois boucles ségrégées (motrice, associative et cognitive) qui s'étendent sur le cortex, les ganglions de la base et le thalamus. Il intègre une prise de décision à deux niveaux avec une sélection de niveau cognitif (cortex préfrontal latéral, LPFC) basée sur la forme et une sélection de niveau moteur (zone motrice supplémentaire, SMA et cortex moteur primaire, PMC) basée sur la position. Dans ce dernier modèle, le cortex était principalement une structure d'entrée / sortie sous l'influence directe de l'entrée de tâche et de la sortie thalamique résultant des calculs des ganglions de la base. Par conséquent, ce cortex ne pourrait pas prendre une décision, en contradiction avec de nombreuses études. Pour cette raison, nous avons ajouté un mécanisme de compétition latérale au niveau cortical basé sur l'excitation à courte distance et l'inhibition à longue distance. Cette compétition se traduit par la capacité du cortex de prendre une décision, ave cependant une dynamique plus lente par rapport au circuit pasant par les ganglions de la base. Nous avons préservé l'apprentissage modulé par la dopamine via un apprentissage par renforcement (RL) entre le cortex et le striatum et nous avons ajouté un apprentissage de type Hebbien (HL) au niveau cortical qui ne dépend donc pas de la récompense, mais seulement des choix effectués.

Pour tester ce modèle, nous avons utilisé une tâche de type bandit manchot où deux stimuli A et B sont présentés à des positions aléatoires (parmi 4). Le stimulus A est associé à une probabilité de récompense de 0.75 alors que stimulus B est associé à une probabilité de récompense de 0.25. Il est donc évident que le meilleur choix est de choisir A. La difficulté étant cependant que ces probabilités ne sont pas initialement connues du

modèle et requièrent donc d'être approximées par essais-erreurs. Le modèle intact se révèle capable de faire cela assez rapidement, c'est à dire en un peu moins de 60 essais. Après cette première phase, le modèle est lésé au niveau du Globus Pallidus interne (GPi) ce qui entraîne l'impossibilité pour les ganglions de la base d'agir sur le comportement. Or, on constate dans ce cas là que le modèle est malgré tout capable de maintenir sa préférence sur le stimulus A, suggérant qu'un transfert d'apprentissage a eu lieu au sein du modèle. Cela peut-être démontré sur ce modèle lésé en utilisant un nouveau couple de stimuli qui n'a jamais été vu auparavant. Dans ce cas précis, le modèle se révèle incapable de choisir préférentiellement l'un ou l'autre stimulus, démontrant ainsi que les ganglions de la base sont nécessaires pour l'acquisition de la préférence initiale mais pas pour son expression sur le long terme. Autrement si, l'apprentissage par essais-erreurs a été transféré sous forme d'habitudes au niveaux du cortex où la simple apparition d'un stimulus provoque sa sélection préférentielle. Cette même expérience a été confirmée chez le primate au sein de l'institut des maladies neurodégénératives au sein duquel se déroule cette thèse. Via une inhibition reversible (muscimol) au sein du Globus Pallidus interne et après entrainement préalable, les primates se montrent capables de conserver un choix optimum. Une autre prédiction forte de ce modèle est que le protocole peut-être renversé, à savoir que le modèle est initialement lése puis testé en condition intacte. L'hypothèse est que si le modèle va effectuer des choix aléatoires durant la première phase, les valeurs associées aux stimuli respectifs seront apprises et mémorisées au sein des ganglions de la base. Lorsque dans un deuxième temps l'inhibition est levée, ces valeurs apprises vont instantanément guider le comportement du modèle lui conférant ainsi un comportment optimal en terme de récompense. Cela a été confirmé d'une part dans les expériences avec le modèle et d'autres part chez le primate qui ont subi le même protocole sur deux jours. Si les singes ont des réponses aléatoires durant le premier jour, alors qu'ils sont sous l'action du muscimol, dés les premiers essais du deuxème jour, on voit une différence très significative et biaisée vers le stimulus associé à la plus forte probabilité de récompense.

Ce travail de thèse propose de reconsidérer la prise de décision comme étant un processus distribué entre plusieurs structures et en interaction. Nous avons pu montré comment les processus d'acquisition et d'exploitation pouvait être dissocié expérimentalement mettant ainsi en exergue le processus de formation des habitudes au niveau cortical. Par ailleurs, ce travail propose un nouveau cadre théorique et expérimental permettant l'exploration de deux types d'apprentissage en interaction constante; apprentissage Hebbien au niveau cortical, et apprentissage par renforcement au niveau des ganglions de la base. Ainsi, une façon d'étudier la force relative de ces deux types d'apprentissage sur le comportement serait d'effectuer des essais avec choix forcé (un seul stimulus présent) avec un probabilité de récompense donnée. En contrôlant le nombre de fois qu'un stimulus spécifique a été présenté par rapport à la probabilité de récompense associée, nous pourrions mesurer l'influence relative de l'apprentissage de renforcement par rapport à l'apprentissage Hebbien. Ainsi, pour un stimulus A (associé à un probabilité de récompense R_A et une fréquence de présentation F_A) et un stimulus B (associé à un probabilité de récompense R_B et une fréquence de présentation F_B), le modèle prédit que le choix dépendra du ratio entre les probabilités et fréquences respectives.

Enfin, au niveau de notre modèle, nous avons émis l'hypothèse que les habitudes sont stockées au sein du cortex bien que nous n'ayons pas pas de preuves expérimentales

solides quant à cette cette hypothèse. Afin d'évaluer si l'apprentissage cortical Hebbien est effectivement responsable de la mémorisation et de l'expressions des habitudes, il serait nécessaire de pouvoir inactiver sélectivement l'apprentissage associatif dans le cortex dorsolatéral préfrontal ou orbitofrontal et vérifier si les singes seraient à mêm de réussier la tâche précédente sans développer des habitudes.

Au final, il est remarquable que ce modèle renverse l'idée relativement ancienne que l'automatisme est une caractéristique sous-corticale. Le fait que l'association d'entrée / sortie automatique se produise au niveau cortical, contournant un long voyage sous-cortical et donc économisant des ressources cognitives est un argument écologique fort. Si ce modèle est confirmé par d'autres expériences, il ouvre donc de nouvelles questions telles que:

i) Est-ce une spécificité de mammifère?
ii) Une spécificité des primates?
iii) comment ces automatismes sont-ils mis en œuvre chez d'autres vertébrés?

Références

- Piron, C., Kase D., **Topalidou M.**, Goillandeau M., Orignac H., Nguyen T-H., Rougier N.P., and Boraud T. "The globus pallidus pars interna in goal-oriented and routine behaviors: Resolving a long-standing paradox". In: Movement Disorders. 2016.
- **Topalidou M.**, Leblois A., Boraud T., and Rougier N.P. "A Long Journey into Reproducible Computational Neuroscience". In: Frontiers in Computational Neuroscience 9.30.
- **Topalidou M.** and Rougier N.P. "[Re] Interaction between cognintive and motor cortico-basal ganglia loops during decision making: a computational study". In: ReScience 1.1.
- **Topalidou M.**, Kase D., Boraud T., and Rougier N.P. "Dissociation of reinforcement and Hebbian learning induces covert acquisition of values in the basal ganglia". BioRxiv preprint, 2016.

> "The discovery of the habit loop is important because it reveals a basic truth: When a habit emerges, the brain stops fully participating in decision making. It stops working so hard, or diverts focus to other tasks. "
>
> — Charles Duhigg

Introduction

Every morning I take out my dog for a walk. We always follow the same route, because it is the optimal for us. My dog is not on a leash when we are out, so our path should not include big, busy roads. However, when we first came to Bordeaux, I didn't know which route to follow. For this reason, the first days I explored different paths, and by paying attention in the details I evaluated them. When I found the most convenient for us, I started following it for months. During this exploitation phase, we stopped paying attention to the surroundings. That's why one day we had to cross a bridge, as we usually did, to realize that there were constructions at its end, forcing us to modify our route. It took me three days to stop crossing the bridge, and even after, occasionally if I was distracted by my thoughts, I forgot about the constructions and crossed it again. However, my dog needed more than two weeks to express similar behavior. Why were we unable to adapt to the new situation after the first failure? What impelled us to persist in our previous successful route? We were taking this route for more than three months. The reason is that during our exploitation phase we had formed a new habit, which contrary to goal-directed actions, here the exploration of the routes, is difficult to alter.

Historically, Aristotle was the first to propose the term habit to describe acquired skills that are needed by an individual to improve his performances in order to reach a desired goal. However, the research of habits in neuroscience is quite recent, and mostly emanates from the work of William James. He proposed that habits are learned skills that use the optimum amount of fine movements following a cue, and not require conscious attention. At the end of 19th century, Thorndike, one of Jame's student, formulated the "Law of effect", which states: behaviors followed by convivial consequences are likely to be repeated, contrary to behaviors followed by unpleasant consequences are likely to be stopped. Based on this law, Pavlov introduced the "classical conditioning", where an individual learns the association of a response to an antecedent stimuli by being an observer of his environment. On the other hand, Skinner proposed "operant conditioning" (also referred as instrumental) by studying voluntary behavior; now the individual interacts and changes his environment,

not just observe it. In this case, he emits an action as a result of stimulus and depending on if it is rewarded or not (or even punished) then it is reinforced in order to be more probable to be chosen again or not.

More than a century of investigating habits and still today defining them is a controversial subject. The term habit, for instance, has a lot of connotations. It can be a routine, as the story with my dog. Preferences can also be considered as habits, for example one person's favorite drink (tea or coffee) or meal (meat or fish). Another type of habits are developed skills as walking or driving. Aristotle, in his Book II of *Nicomachean Ethics*, suggests that ethics is the result of moral habits. For a example, if someone feels fear or confidence each time in front of a danger that makes him a coward or brave, respectively. Moreover, the research community has not yet agreed on which features of an action are fundamental in order to characterize it as habit. However, the most commonly accepted ones are:

1. Habits are triggered by a specific stimulus.
2. They are actions that have been acquired via experience, usually after extensive training.
3. Habits are performed automatically, id est they are executed fast compared to goal-directed actions, and unconsciously, without paying attention.
4. Finally, the most common feature used in experiments to define habits, is their disengagement from a goal. In other words, this type of actions are executed even if their outcome has been devaluated.

The non-conscious execution of actions is an essential feature of habit for survival in species. This characteristic allows attention to be focused somewhere else and not in a specific action. For instance, when an animal is under attack, it has to flee quickly. If its attention had been focused on how to use its body to run, then it could not capture its predator moves or find a safe place to hide. Another example, related to humans this time, can be the procedure of driving. The driver has to have his full attention to the road in order not to crash or hit somebody. That would be hopeless if he had to think about how to change the gears or push the petals. In general, a well-established habit, depending on the species, is difficult or even impossible to not be expressed and/or be replaced by another behavior. This persistent of habits is useful in a lot of cases, but can also be dangerous for some others. For example, when somebody has health problems and the doctor suggests to follow a healthier diet, but the patient ends up to choose again the unhealthy snacks that used to have. In other words, habits have a strong impact on everyday life of species, either good or bad. Their importance generally in instrumental behavior is

the reason that habits are studied in many different fields, using a variety of species.

The acceptance of habits being independent of goals, led Dickinson to propose the division of instrumental behavior into two opposed types: the goal-directed and habitual. Following this distinction, many experimental studies were developed in order to identify and investigate the mechanisms in the brain responsible for the production of these behaviors. The features that these two types of behavior contain, led to the dominant view of the 20^{th} century: "Novel behaviors require attention and flexible thinking and therefore are dependent on cortex, whereas automatic behaviors require neither of these and so are not mediated primarily by cortex. Instead, it has long been assumed that automatic behaviors are primarily mediated by subcortical structures." Following this theory, a crucial role has been assigned to the basal ganglia (a group of subcortical structures) in habitual learning. The early development of BG both at phylogenetic and ontogenetic level, and the widespread projections from cortex to BG support this assumption. Furthermore, these projections form a mechanism for producing bonds between a sensory input with a motor output that in other words constitutes habits. By investigating the brain areas responsible for instrumental behavior contributed to the proposal that two distinct mechanisms are responsible for expressing this type of behavior: the action-outcome (A-O) which produces goal-directed actions and the stimulus-response (S-R) which expresses habits. It is hypothesized that these systems are implemented in the parallel cortico-basal loops, and either compete for expression or shift from A-O to S-R system. This resulted in the hypothesis that both systems depend on striatum, and as an extension on basal ganglia. However, there are evidence today on the critical role of basal ganglia in goal-directed actions and the initial formation of habits, but not in the expression of habits.

My thesis was dedicated to the investigation of the mechanisms of decision making, with emphasis on the formation of habits in the cortex of primates. I developed a dynamical model of cortical-basal (CBG) loop that incorporates an action selection mechanism and learning, through three segregated loops, that was inspired by previous theoretical and experimental works. This model makes the hypothesis that basal ganglia are only necessary for the acquisition of habits while the expression of such habits can be mediated through the cortex. Furthermore, this model predicts the existence of covert learning within the basal ganglia ganglia when their output is inhibited. Using a two-armed bandit task, this hypothesis has been experimentally tested and confirmed in monkey. We tested it by first inhibiting the GPi and made the monkeys to learn a two armed bandit task using two never seen stimuli. In such condition, monkey's performance is purely random. However, on the second day, we removed the

inhibition and tested the monkey on the same task. Performance were instantly quasi perfect demonstrating the monkeys knew the respective value of the two stimuli even though they were unable to express this knowledge the day before. Finally on the third day, we suppressed again the GPi output. This time, monkey's performances stays at a very good level, demonstrating a transfer has occurred and we hypothesized this could be attributed to the formation of a habit at the cortical level as it is the case in the model. Overall, this work suggests that the classical idea that automatism is a subcortical feature should be revised.

Plan of this thesis

This dissertation is subdivided in four chapters. The first two chapters are reviews of the existing bibliography on: the biology of the structures participating in instrumental learning, and the computational models that implement instrumental behavior. The last two chapters are describing the architecture and the properties of our model, as well as the results of the protocols that have been tested on. The manuscript is organized as follow:

1. An overview of the state-of-the-art of basal ganglia and habits in biological bibliography is provided in the first chapter.

2. The second chapter includes a review of computational models implementing action selection and habits. The choice of these models was based either on their significant contribution in research or the given inspiration to our model.

3. In the third chapter, I introduce our dynamical model of the BG-cortical network, which has been developed to investigate the underlying mechanisms of the acquisition and expression of habits. Before the full description of the model, I recite its history by summarizing the two previous models that it is based on.

4. Finally, I describe the protocols that were used to test either the abilities of our model, or our hypothesis about the elemental mechanisms of instrumental behavior, followed by the analysis of their results. For two of these protocols, we also have conducted experiments on monkeys, and their results are compared to the ones from the model.

"Neuroscience is by far the most exciting branch of science because the brain is the most fascinating object in the universe. Every human brain is different - the brain makes each human unique and defines who he or she is. "

— Stanley B. Prusiner

Chapter 1

Biology Background

Contents

The ability of decision making and learning are essential capacities for the survival of all living organisms. For example, an animal has to learn where to forage for food or how to protect itself from predators in order to survive. The study of animal behavior led to the definition of two learning processes, the operant and classical conditioning, which in turn led to the foundation of behaviorism, a school of psychology that studies the mechanisms of learning and action selection.

At the end of 19^{th} century, Thorndike worked on alearning theory that led to the formulation of the "Law of effect", which states: behaviors followed

by convivial consequences are likely to be repeated, contrary to behaviors followed by unpleasant consequence are likely to be stopped. Based on this law, Pavlov [1927] introduced the "classical conditioning", where an individual is an observer of the relationships among the events in the world. He initially conducted experiments on dogs, where he presented a stimulus (rang a bell) before giving them food. After training he noticed that the dogs produced saliva immediately after the stimulus and before the presentation of the food. Pavlov found that the interval between the conditioned stimulus (CS; sound of the bell) and the appearance of the unconditioned stimulus (UCS; food) affected the strength and the time the dog needed to learn the conditioned response (CR; saliva). In summary, the difference between the classical and operand conditioning is that in the first case, behavior is learned as a response of an antecedent stimuli, whereas in the latter, behaviors are strengthened or weakened by their consequences (*i.e.* reward or punishment). On the other hand, the term "operant conditioning" (also known as "instrumental conditioning") was originated by Skinner [1950], who believed that the observation of the external causes of a behavior is important and not the internal, like thoughts and motivations. In instrumental conditioning, an individual has to obtain knowledge for the actions outcome in an environment through experience, before to acquire this knowledge. Said differently, an individual will emit an action, and, if it provides a reward, then it will be reinforced in order next time to be more likely to be chosen again, although if it provokes punishment, then it will be diminished to be less likely to be chosen. This behavior is equivalent to what is called voluntary behavior. In summary, in classical conditioning the subject learns through observation, on contrary to the operant where he learns through exploring the outcomes of his choices.

These two theories were the starting point of instrumental behavior research on animals, which led to the realization that in order to perform correctly tasks that need this type of behavior, it is necessary the knowledge of: (1) the outcome's value, and (2) the relationship between an action and its outcome are necessary. Yin and Knowlton [2006] highlighted that the manipulation of these variables by the experimenter had altered dramatically the studies of instrumental behavior. A consequence of this realization was the division of instrumental behavior into two types: goal-directed and habitual. The main feature used to differentiate them was their dependency with the expected outcome [Dickinson, 1985; Colwill and Rescorla, 1995; Hammond, 1980; Dickinson and Balleine, 1993]. More precisely, a goal-directed action is driven by the outcome it leads to, whereas a habit is carried out even in the case of outcome devaluation.

Hirsh [1974] was the first to propose a crucial role of basal ganglia (BG) in habitual learning. His argument was based to the early development of BG both in phylogeny and ontogeny, and also to the widespread projections from

cortex to the main input of BG. He proposed that these projections form a mechanism for producing bonds between a sensory input with a motor output that in other words constitutes habits. The crucial role of BG in this type of learning has been supported also from studies in a variety of species and techniques, as well as from computational models [Seger and Spiering, 2011; Frank, 2005; Yin and Knowlton, 2006; Graybiel, 2008; Balleine *et al.*, 2009; Packard, 2009; Ashby and Ennis, 2006; Cohen and Frank, 2009]. Data from studies on rodents and primates revealed the participation of striatum in two distinct systems, and further the existence of parallel functional cortico-striato-thalamic loops responsible for processing different functional types of information (*e.g.* motor or cognitive) [Yin and Knowlton, 2006; Albin *et al.*, 1989; Parent and Hazrati, 1995a].

In this chapter, firstly I summarize the anatomy, as well as the internal and external connectivity of basal ganglia. Also, I refer the existing theories about the functional role of the pathways that are formed inside BG. It has been shown that BG incorporate learning through a dopaminergic signal. For this reason, I describe the role of dopamine in learning, as well as the learning rule that they follow. Finally, I introduce the term habit, and because the definition of habit is a controversial subject, I present different definitions that have been proposed. The description will be kept short, because the focus of this part is the reader to understand the basic and relevant to the model properties (for more details, please refer to the indicated bibliography in the text).

1.1 Anatomy of the basal ganglia

Even though Willis (1667) and Swedenborg (1740) (as cited in Ding and Gold [2013]) talked about the role of striatum (main input of BG) in sensation, their theories were overshadowed by clinical observations in humans of movement disorders. Pathological changes in the basal ganglia have been recognized in diseases as Parkinson and Huntington. This focus of research in motor symptoms led to intensive investigation of the BG role in movement. However, at the end of the previous century, clinical studies indicated the involvement of BG in various cognitive functions [Haber, 2003], such as learning and memory [Hélie *et al.*, 2015]. However, their exact role is still unknown.

Most studies investigating the role of BG during instrumental learning conducted experiments on rodents and primates. For that reason, I will briefly describe the anatomy and connectivity of BG in those species specifically. The description is focused on the structures and their connectivity that are implemented in the model. For more extensive details please refer to the articles of Parent and Hazrati [1995a,b], Haber [2003] and Utter and Basso

[2008], which are excellent for this purpose.

1.1.1 Structures

The basal ganglia are a group of interconnected structures: striatum (Str), subthalamic nucleus (STN), globus pallidus (GP; external [GPE] and internal [GPi]), substantia nigra (SN; pars compacta [SNc] and pars reticulata [SNr]) and ventral tegmental area (VTA). Their location in the brain is shown by sagittal view in Figure 1.1a, and coronal view in Figure 1.1b.

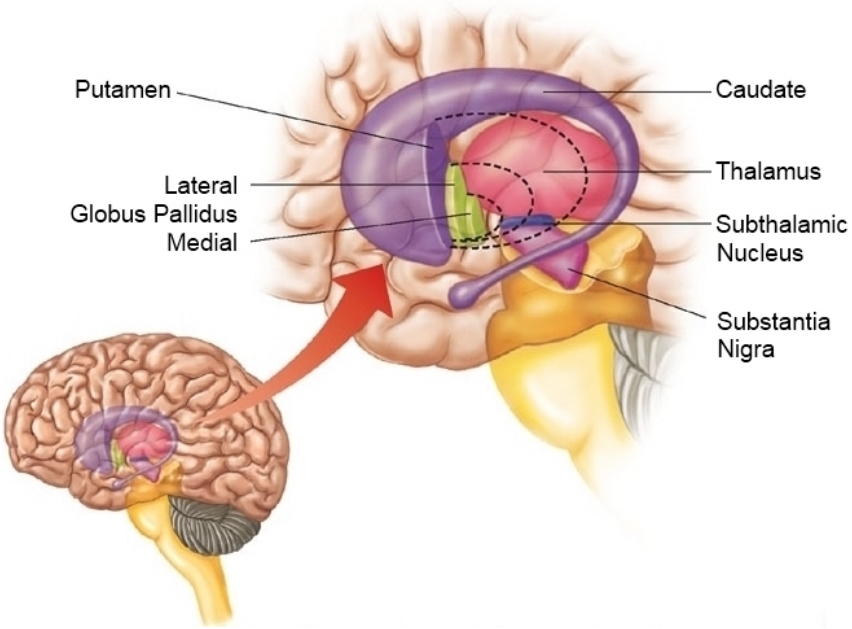
Striatum (Str)

Striatum comprises of the caudate nucleus, the putamen and the nucleus accumbens (in primates [Figure 1.2]; dorsal and ventral compartments in rodents). It is composed principally by medium spiny neurons (MSN) that are projection neurons [Parent and Hazrati, 1995a; Wickens, 1997]. The main neurotransmitter of MSNs is γ-aminobutyric acid (GABA), although a variety of neuroactive peptides are also expressed, such as substance P, enkephalin, dynorphin and neurotensin [Parent and Hazrati, 1995a]. These neurons are separated in two types depending on the dopamine receptors that they contain (D1, D2).

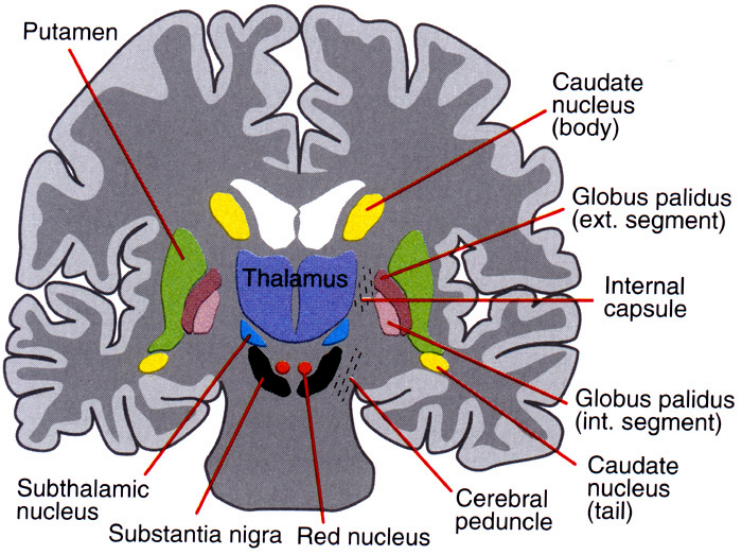
MSNs are usually silent and require concurrent and numerous excitatory cortical input in order to be activated. A variety of interneurons also exist in striatum. One of the types is the GABAergic fast spiking interneurons, which are responsible fro the silence of MSNs. Another type of interneurons are the aspiny cholinergic interneurons that are characterized by spontaneous firing activity, and therefore are referred as tonically active neurons (TANs). Contrary to MSNs, TANs need a relatively small number of extrinsic synaptic input to alter their patterns of activity. They receive excitatory cortical and thalamic input, and dopaminergic input from substantia nigra. In turn, they target primarily MSNs, but also GABAergic interneurons. It has been proposed that TANs participate in reward based learning by modulating the activity of MSNs [Tepper and Bolam, 2008]. The proportion of MSNs versus interneurons is higher in primates (3:1) than the rats (9:1) [Parent and Hazrati, 1995a].

Striatum receives dopaminergic input into MSNs from SNc and VTA [Utter and Basso, 2008], and glutamenergic input from multiple nuclei in thalamus [Bar-Gad *et al.*, 2003]. MSNs, also, have a collateral arborization to themselves or adjacent cells. The two types of MSNs project in segregate way to GPi/SNr (D1 receptors) and GPe (D2 receptors).

Extrinsic afferents arise from all cortical areas. Even if there is major convergence at striatal level, anatomical and physiological studies have shown that topography is preserved, *i.e.* functionally distinct cortical areas project in



(a) It can be found as material of a course of the Department of Psychology from the University of Virginia



(b) As in Leisman *et al.* [2013]

Figure 1.1: A schematic illustration of the structures composing basal ganglia: (a) their location within the brain, (b) coronal plane

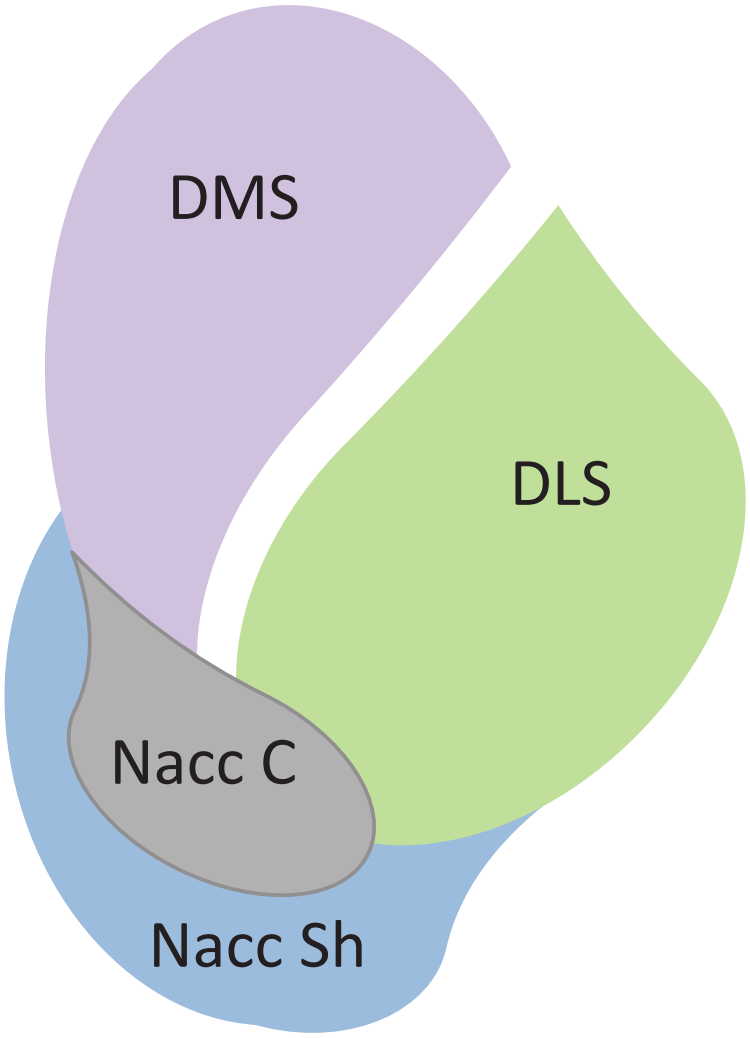


Figure 1.2: As in Liljeholm and O'Doherty [2012]. Schematic representation of striatal compartments and their connectivity with cortex. Afferents arising from different cortical areas project to different sub-regions of the striatum, which project back to the particular cortical areas via the basal-thalamocortical pathway. The circle represents inhibitory connection, whereas the arrow excitatory. DMS, dorsomedial striatum; DLS, dorsolateral striatum; GPi, internal segment of globus pallidus; VP, ventral pallidum; VA, ventral anterior; DM, dorsomedial; VL, ventrolateral; VM, ventromedial; Nacc C, nucleus accumbens core; Nacc Sh, nucleus accumbens shell.

different parts of striatum (Figure 1.2). Based on cortical input, striatum in primates is divided into: (1) *ventral* component, which includes n. accumbens, ventromedial portions of the caudate and putamen, and (2) *dorsal* including the rest caudate and putamen, which can further divided into *associative*, comprising of caudate and the anterior putamen, and *sensorimotor*, including

the posterior putamen [Parent and Hazrati, 1995a]. Homologous division exists also in rodents, the dorsomedial (DMS) and dorsolateral (DLS) striatum, which corresponds to the associative and sensorimotor, respectively [Joel and Weiner, 2000].

Globus Pallidus (GP)

The Globus Pallidus is separated into two nuclei: the external (GPe) and the internal (GPi). In rodents, the latter is referred often as the endopeduncular nucleus.

As mentioned before the two nuclei receive GABAergic input from different types of medium spiny neurons. MSNs with substance P are connected with GPi and with enkephalin are connected with GPe. Although, there are striatal neurons that project to both. Further, anatomical evidence disclosed connectivity directly from GPe to GPi [Parent and Hazrati, 1995b]. In addition, GPe projects to STN, and GPi sends its output to thalamus.

Subthalamic Nucleus (STN)

Traditionally, STN was considered as an intrinsic nucleus. However, anatomical studies revealed direct cortical projections to STN that makes it also an input nucleus. In contrast with cortical input to striatum, STN receives excitatory input form somato-motor areas of frontal lobes [Nambu *et al.*, 2000 b; Mink, 1996].

STN has reciprocal connections to GPe. Its excitatory projections are the only ones in BG circuity. Anatomical studies has shown that STN projects widely and to a variety of GPi neurons [Hazrati and Parent, 1992 a,b].

Substantia Nigra (SN)

Substantia Nigra pars reticulata (SNr) is the other output nuclei of BG. Studies in non-human primates and rodents suggest a role of SNr in movement, but more recent work supports its involvement also in cognitive processes [Utter and Basso, 2008]. Like GPi, it receives inhibitory input from striatum and sends inhibitory input to thalamus.

On the other hand, substantia Nigra pars compacta (SNc) provides dopaminergic input to striatum through its DA cells, and receives back from it. However, the reciprocal connections are neither topographical nor equal in the size of the input. According to Haber [2003]:

> "The ventral striatum receives a limited midbrain input, but projects to a large region. In contrast the dorsolateral striatum receives a wide input, but projects to a limited region".

1.1.2 External connectivity

As Utter and Basso [2008] has emphasized, BG influence many neuronal pathways and information processing systems, because of their major input from the entire cortex and their output through thalamus back to cortex. Several studies have shown that different cortical areas project to explicit regions of striatum, from which they receive back input. Thereby, first Alexander *et al.* [1986] proposed a model composed of a five parallel segregate information processing loops (motor, oculomotor, dorsolateral prefrontal, lateral orbitofrontal). However, a lot of evidence later suggested that not all these five loops are closed, but the existence of cross-talk among the loops. These evidence exhorted several researchers to revise this model, and finally suggest the existence of only three functionally defined loops, the associative, the sensorimotor, and the limbic [Parent and Hazrati, 1995a; Haber, 2003; Seger and Spiering, 2011; Hélie *et al.*, 2015].

Based on this theory and combined with structural evidence, it has been proposed that striatum is divided into three components: the *motor* (dorsolateral striatum in rodents), which includes posterior putamen, the *associative* (dorsomedial striatum in rodents) that consists of all of the caudate and the anterior putamen, and finally nucleus accumbens, ventromedial caudate and putamen comprising the *ventral* component (Figure 1.3) [Hélie *et al.*, 2015; Parent and Hazrati, 1995a; Liljeholm and O'Doherty, 2012; Seger and Spiering, 2011].

Each component is part of one of the functional loops and receives input from particular cortical areas. Sensory and motor cortices send to motor striatum, but associative receives from frontal and parietal association cortices. Finally, amygdala, hippocampus, medial orbitofrontal and anterior cingulate cortices are connected with the ventral striatum. The structural and functional topography is preserved through the BG, to thalamus and back to cortex [Haber, 2003]. However, the integration of information across functional circuits is essential for forming behavioral responses. Haber [2003] suggested that the two intrinsic networks of striato-nigro-striatal and thalamo-cortico-thalamic are responsible for a continuous feedforward mechanism of information flow.

Nowadays, it is widely accepted that the associative loop is involved in goal-directed actions by monitoring recent actions and anticipating their consequences, contrary to the sensorimotor loop, which is related to movements as a response to distinct stimuli, and is independent from reward expectancy, characteristics of habitual behavior [Yin and Knowlton, 2006].

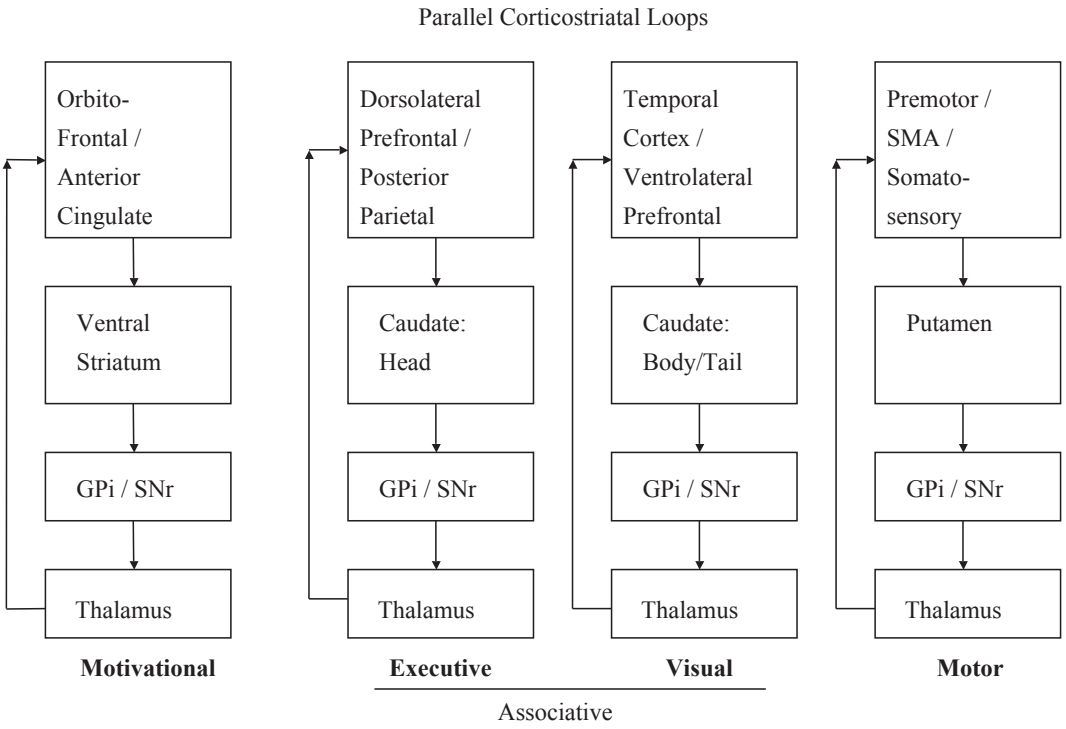


Figure 1.3: As in Seger [2008]. Corticostriatal loops.

1.1.3 Anatomical differences in vertebrates

Invertebrates, despite their small brain size, can perform highly optimized functions for specific behaviors. Although, these functions are part of a genetically pre-programmed fixed repertoire of behaviors [Doya, 1999]. However, decision making is an ability that all vertebrates share, down from Caenorhabditis elegans up to humans. Someone can claim that this is correlated with the brain size, however the brain size varies a lot among these species. The basal ganglia, nonetheless, are present in all species, and they have been associated with voluntary behavior and procedural learning. Figure 1.4 shows that the pallium (analog to cortex in mammals) in lower vertebrates is smaller than the subcortical structures. As a result, the basal-thalamic loop generates most of the behaviors compared to pallium. In birds, on the other hand, the pallium is bigger and is more interconnected with the basal-thalamic loop, such that both parts are used in order distinct behaviors to be generated. Finally, the mammalian cortex is much larger than the subcortical structures, and consequently becomes the main structure that generates behaviors. However, it sends strong input to basal ganglia, and receives back from thalamus. Through this connectivity BG participate in the finally decisions produced by cortex.

These differences in the connectivity among the species are responsible for the variety of abilities that they contain. For this reason, whenever an

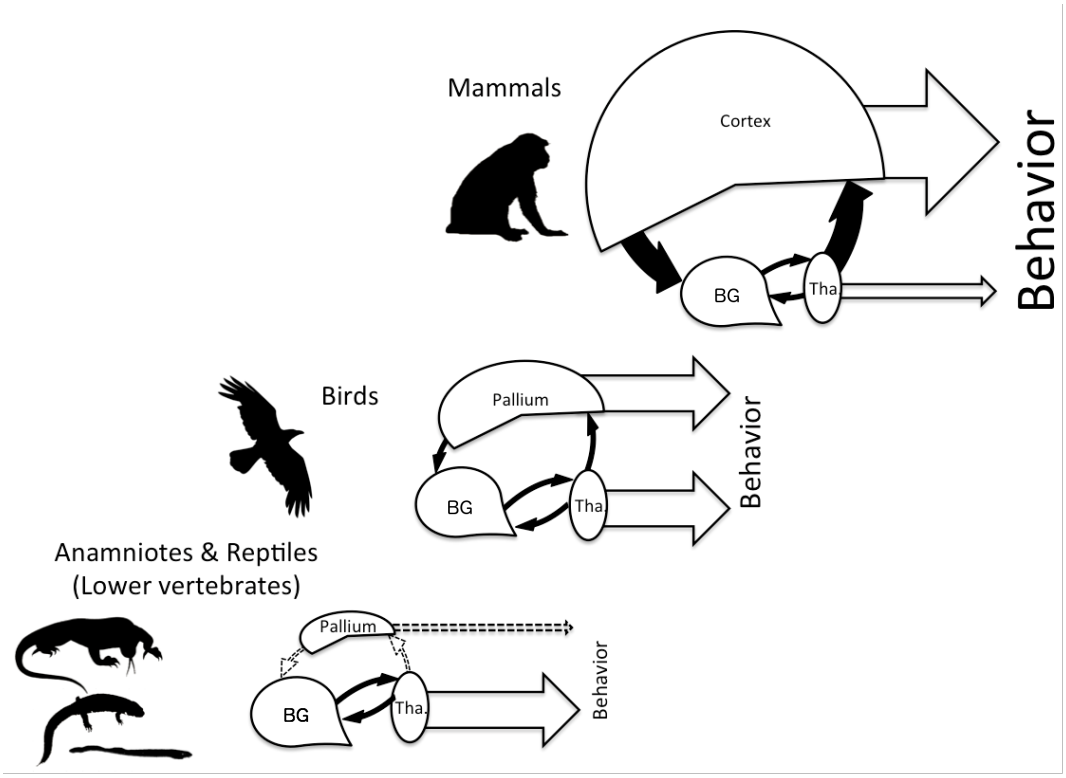


Figure 1.4: As in Boraud [2015]. Schematic illustration of the connectivity of basal ganglia with the rest of the brain, and their contribution in expression of behaviors through the prism of evolution of vertebrates.

experiment is conducted, it must be taken in account the capabilities of the species that is used, and not generalize the results for all species.

1.1.4 Pathology

The association of BG dysfunction with particular diseases, such as Parkinson and Huntington, led to the intensive investigation of their role in movement. Although, different disorders affect different parts of BG. However, from studies of clinical abnormalities combined with the known connectivity among BG and cortex, it has been concluded that BG participate in voluntary motor and cognitive behavior, procedural learning as well as in control of emotions. Some of the most famous diseases associated with BG are: Parkinson, Huntington and obsessive-compulsive disorder.

Parkinson's Disease (PD)

Rigidity, tremor, akinesia (loss of voluntary movements) and bradykineseia (extremely slow movements and reflexes) are some of the motor symptoms encountered in Parkinson's disease. The degeneration of dopamine (DA) neurons

in the SNc is responsible for these symptoms, however still today, the cause of PD is unknown. The loss of DA is hypothesized to generate an imbalance between the activity of the direct and indirect pathways [DeLong, 1990]. In order to conterbalance this loss, patients are prescribed Levodopa (L-Dopa) that is converted to dopamine in the brain. L-Dopa is the most effective medication of PD, however after extended treatment, the efficacy of L-Dopa becomes irregular. It has been observed that lesions on GPi also help the control of the symptoms. Benabid [2003] was the first to observe that high frequency stimulation could improve motor symptoms of PD, so he proposed the use of electrical stimulation to treat them by damaging permanently brain tissue. Because of the inconsistencies associated with drug therapy, Deep Brain Stimulation (DBS) procedure brought a breakthrough in the treatment of the syptoms of PD. DBS involves the implantation of a neurostimulator ("brain pacemaker") that sends electrical stimulus to specific targets in the brain.

Huntington's Disease (HD)

Characteristic symptoms of HD include involuntary spastic movements of the extremities. Degeneration of the GABAergic medium spiny projection neurons in the striatum results in the disruption of proper functioning of the entire BG circuity. Today, there is no cure for HD. However, after the successful use of DBS for PD, the exploration of DBS treatment for HD revealed the amelioration of motor symptoms in a patient with advanced HD by bilateral stimulation of GPi [Moro *et al.*, 2004].

Obsessive-Compulsive Disorder (OCD)

Patients with OCD are plagued by unreasonable thoughts and fears (obsessions), which produce anxiety that lead to repetitive behaviors (compulsions). Human imaging studies of this disorder have shown cortico-striatal dysfunctions, such as an increased or abnormal functional connectivity in a subset of cortico-striatal circuits [Shepherd, 2013]. The dysfunctionality has been located to the anterior part of the caudate nucleus and the ventral striatum [Graybiel, 2008]. Furthermore, Gillan *et al.* [2011] observed that an insensitivity to outcome devaluation and slips of action exists in OCD. Dolan and Dayan [2013] based on evidence by Daw *et al.* [2011] and Maia *et al.* [2008] of abnormalities in components of the model-based system in OCD, concluded that the habitual system overdominate the goal-directed in OCD.

1.2 Functional pathways

1.2.1 D1/D2 Medium spiny neurons (MSN)

Medium spiny neurons contain two types of dopaminergic (DA) receptors, D1 and D2, which are G-protein coupled receptors [Utter and Basso, 2008]. The

bounding of D1 receptors with DA results in depolarization of the neuron, in contrast to D2 receptors, which results in hyperpolarization [Sealfon and Olanow, 2000].

That means that the role of D1 receptors is the enhancement of the cortical-striatal influence, and the role of D2 is the reduction. Evidence indicate that striatal neurons projecting to the two compartments of GP have different receptors depending on their target. The ones with D1 project to GPi, but the ones with D2 project to GPe. This dichotomy inspired Albin *et al.* [1989] to suggest the existence of two pathways: the direct (STR-GPi/SNr), and the indirect (STR-GPe-STN-GPi/SNr). This led to the assumption that the role of the direct pathway is to facilitate movement, whereas the role of the indirect inhibits movements [Gerfen *et al.*, 1990]. Although, new evidence shows that there is co-localization of D1 and D2 receptors on striatal neurons [Aizman *et al.*, 2000; Nadjar *et al.*, 2006]. Through an operant task on mice, Cui *et al.* [2013] observed that both types of MSNs increased their activity during an action, and remained silent when mice were not moving, which contradicts the theory by [Albin *et al.*, 1989].

1.2.2 Direct/indirect/hyperdirect pathways

Albin *et al.* [1989] were the first to propose a model of BG, explaining the functional role of the internal connectivity of BG(Figure 1.5a). This model suggests the existence of two pathways, and it's able to explain different types of behavior in healthy states and in Parkinsonnism. The 'direct' pathway contains STR as the main input structure, which receives input from cortex, and projects to the BG outputs, GPi and SNr. Although the 'indirect' pathway also contains STR as an input structure, but this time its signal reaches the BG outputs through GPe and STN. As discussed previously, STR includes two types of MSNs with different dopamine receptors (D1 and D2), which project in segregate way to GPi/SNr (D1 receptors) and GPe (D2 receptors). Evidence showed that dopamine excites D1 receptors, and enhance cortico-striatal influence, but inhibits D2 receptors leading to reduce of cortico-striatal influence. Based on this differentiation, Albin proposed that the role of the direct pathway is to facilitate movements, contrary to the indirect pathway which is supposed to suppress them. Following this theory, the loss of dopamine that is observed also in Parkinsonism, induces the decrease of direct pathway's activity and increases the indirect's, impending voluntary movement (Figure 1.5b). However, this model is not able to explain other symptom's of Parkinson's disease, such as tremor and rigidity, or evidence for co-existence of D1 and D2 receptors in a single neuron [Bar-Gad *et al.*, 2003].

Few years lated, Mink [1996] proposed a modification of Albin's model, suggesting a center-surround organization between the interaction of the two pathways. In this model, the direct pathway provides an excitatory center

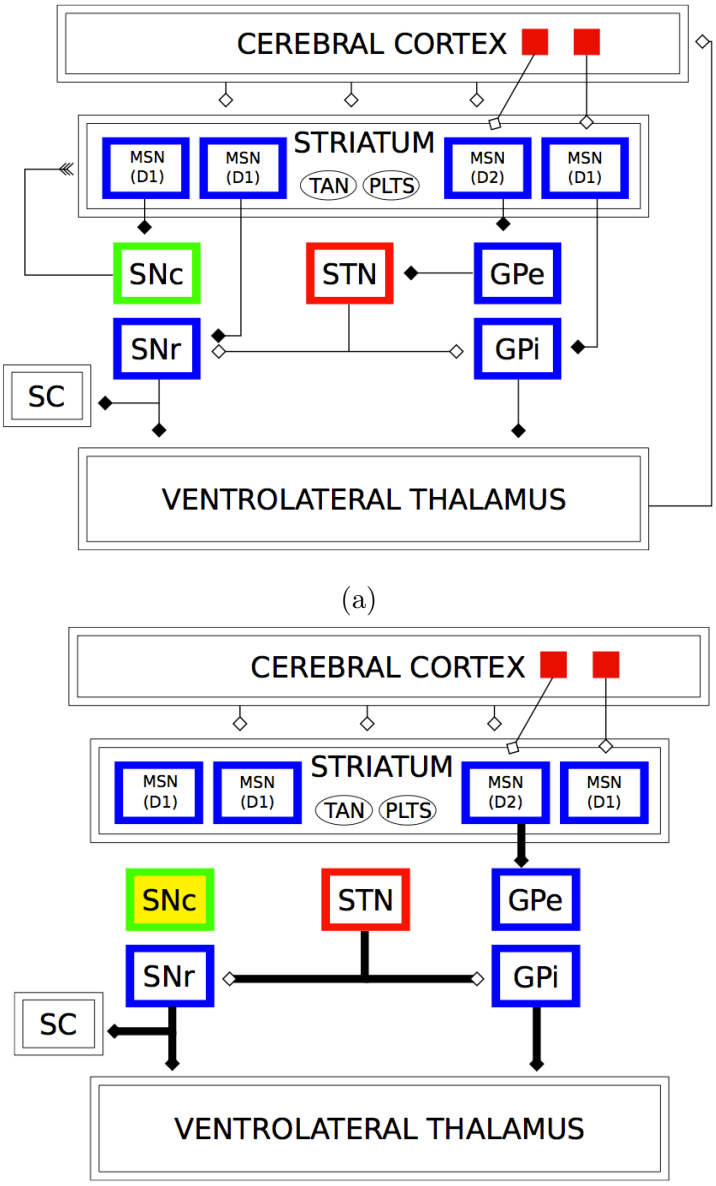


Figure 1.5: BG circuitry in the normal and parkinsonian state. Boxes correspond to projection neurons, ovals to interneurons. The glutamatergic neurons correspond to red boxes and empty endings; the gabaergic neurons correspond to blue boxes and filled endings; the dopaminergic neurons of the SNc correspond to the green box and the arrow tail ending. For the sake of clarity, we characterized the MSN according to their dopaminergic receptor and replaced most of the original notations by the ones presented previously in this manuscript; SC stands for superior colliculus. Impacted areas are shaded in yellow; diminished projections are dashed and grey; augmented projections are wider. Modification from Albin *et al.* [1989] by Liénard [2013], as well as the caption.

by inhibiting GPi and focusing to the desired movement, and the indirect provides inhibitory surroundings by exciting GPi and so inhibiting all the other competing movements. Further, anatomical studies has shown that STN projects widely and to more GPi neurons compared to striatal connections that are topographically and functionally segregated [Hazrati and Parent, 1992 a,b], providing a crosstalk mechanism among different functional areas in addition to the striatal interneurons. These evidence corroborate the existence of a 'center-surround' model in BG.

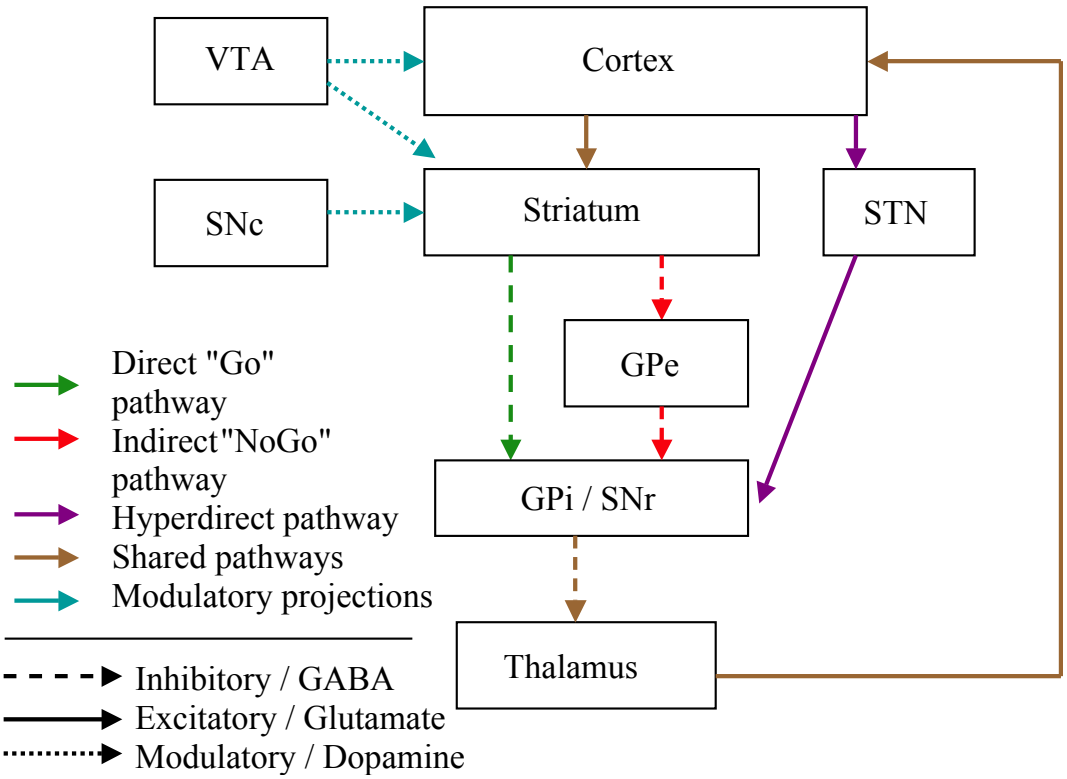


Figure 1.6: As in Seger [2008]. Main pathways through the basal ganglia.

Nambu *et al.* [2002] expanded this model to include the discovery of the fast 'hyper-direct' pathway, where STN receives cortical input (mainly from motor areas) and project to BG outputs [Nambu *et al.*, 2000 b]. As a result, they also revise the functional role of the pathways. They proposed that this third pathway receives cortical input before the initiation of a movement, in order to inhibit all movements prior the facilitation of the execution of only one by the direct pathway. Finally, the indirect pathway signals the end of the movement by increasing GPi activity. All the pathways proposed here are shown in Figure 1.6.

Another equally important loop for the normal functioning of BG is the striato-nigro-striatal pathway. SNc dopaminergic neurons receive and project

back to striatum, participating in learning [Doya, 1999]. It has been shown that the ventral part of striatum projects extensively to SNc but receives limited input. The reverse connectivity exists also between SNc and dorsalateral striatum. This arrangement promotes the flow of information from limbic to motor system [Haber, 2003].

1.3 The role of dopamine

There are two types of dopamine: the phasic and the tonic. The tonic release is a level of dopamine outside the synapse. Although, still today it is unknown what is it for. It may play a role on the balance exploration/exploitation, and it is certainly important for motor functioning, because it is the one which is decreased in Parkinson's disease, and on which L-DOPA and dopaminergic agonist provided to Parkinsonian patients work. On the hand, phasic dopamine is released in a specific level, which increases when unexpected reward is received, or decreases when an expected reward is not given.

In this section, I refer only to the phasic dopamine, because is an important element in instrumental learning.

1.3.1 Reward prediction error

The dopamine neurons are located mostly in substantia nigra pars compacta (SNc), and ventral tegmental area (VTA). These neurons release the neurotransmitter dopamine in the frontal cortex and striatum. The characteristic that distinguishes them from other midbrain neurons is the polyphasic long impulses discharged at low frequencies. These neurons have been correlated with reward characteristics of somatosensory, visual, and auditory stimuli [Schultz, 1998].

Most of the dopamine neurons show phasic activations when animals receive unexpected reward, independently of the nature of the reward; *e.g.* different food objects or liquids (Figure 1.7 *top*; Romo and Schultz [1990]). A small amount of these neurons show phasic activation also to the presentation of primary aversive stimuli [Mirenowicz and Schultz, 1996]. The phasic responses of the dopamine neurons depend on the unpredictability of reward. That is the reason why they are more active during the learning phase, but stop responding to an already conditioned stimuli; *i.e.* the animal knows the existence and the expected timing of the reward (Figure 1.7 *middle*; [Ljungberg *et al.*, 1992; Mirenowicz and Schultz, 1994]). By contrast, when a predicted reward does not occur results in depression of dopamine neurons (Figure 1.7 *bottom*; [Hollerman and Schultz, 1996]). If the reward delivery is delayed for $0.5 to 1.0 s$, at the time it is expected the neurons will be depressed, and activation will occur after the

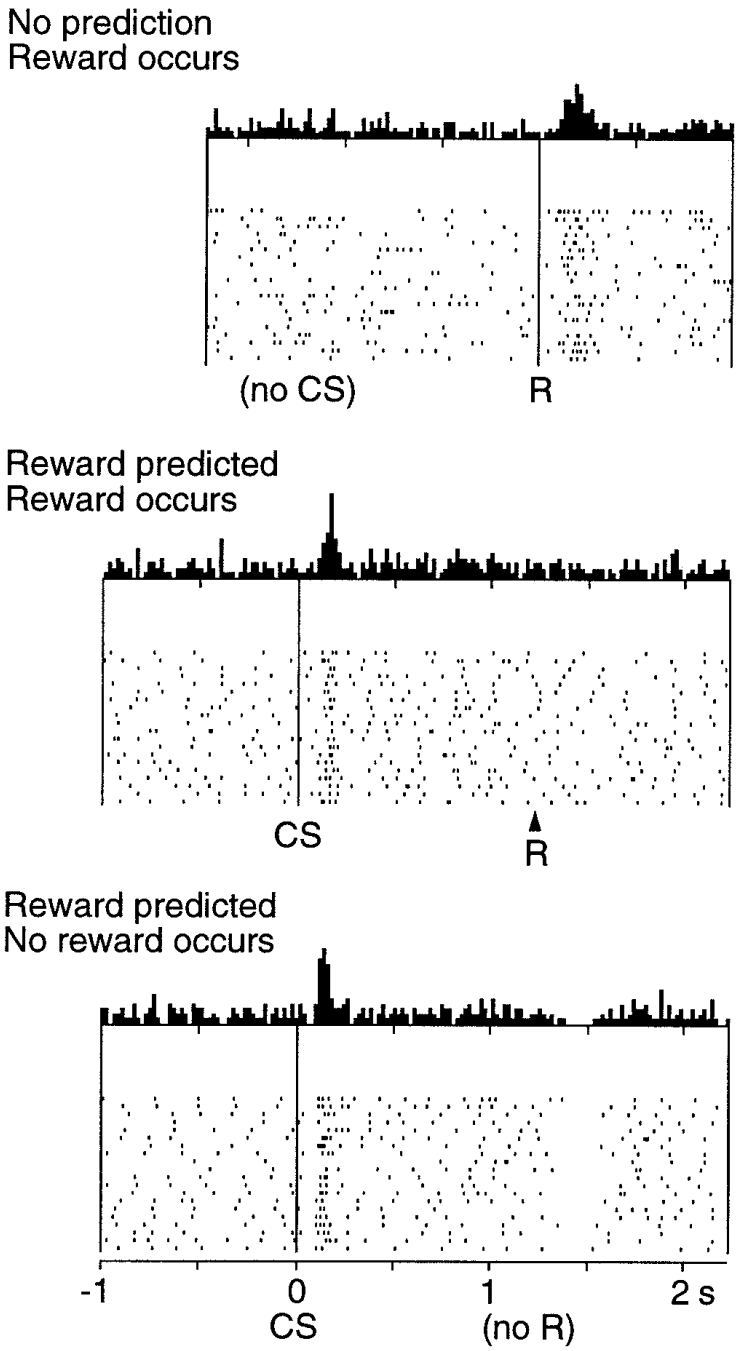


Figure 1.7: As in Schultz *et al.* [1997]. Depending on an error in reward prediction, the dopamine neurons report the predictability of a reward in three cases: *top,* delivery of unpredicted reward, *middle,* delivery of predicted reward to a conditioned stimulus, and *bottom,* lack of an expected reward as a result of a conditioned stimulus.

reward is given. However, earlier deliveries of reward than habitual causes an activation, but not depression at the habitual time.

1.3.2 Reinforcement learning

Computational neuroscience expedited the work on instrumental learning. The pioneering work of Sutton and Barto [1998] that introduces the essential characteristics of reinforcement learning (RL) provided the impetus for the experimental neuroscientists to identify neural signals equivalent to the vital elements of RL models [Daw *et al.*, 2005; Daw and Doya, 2006]. Reinforcement learning is the learning of actions that maximizes the received reward. The learner is ignorant about the outcome of his actions, and must discover them by choosing different actions and observe their consequences. Put it in different words, he has to explore in order to make better actions in the future, but also exploit previously discovered effective in producing reward actions. In summary, the agent in each state has to take an action and wait to obtain or not a reward. After the delivery of the reward, the agent estimates the error between his prediction of being rewarded and the actual reward, and then updates his estimation. As a result, if he gets an unexpected reward then he will increase the value of the action, so next time will be more probable to choose it, but if he does not get an expected reward, he will decrease the value to be less probable to choose it again. In this way, associations between an action and its outcome are established, thereby facilitating the decision making procedure.

Doya [1999] observed that the cortico-striatal synapses follow the reinforcement learning module to evaluate a given state and produce an action based on this evaluation. The learning among the neurons emerges from the strength of the synapses. Furthermore, a large literature provides evidence that dopamine fires above baseline after unexpected reward, and below when expected reward has not been delivered [Montague *et al.*, 1996; Wickens, 1993; Hollerman and Schultz, 1998; Bar-Gad *et al.*, 2003]. So, this dopaminergic signal is competent for the role of reward-mediating training signal [Schultz, 2002; Ashby *et al.*, 2007; Seger and Spiering, 2011]. In order for a synapse to be strengthen, it requires strong presynaptic and postsynaptic activation, and release of dopamine [Calabresi *et al.*, 1996; Arbuthnott *et al.*, 2000; Wickens, 1990, 1993; Ashby *et al.*, 2007]. The cortical neurons send glutamate input to the striatal neurons, which they receive it through N-methyl-D-aspartate (NMDA) receptors. NMDA receptors have a high threshold for activation, which is an important factor for discriminating the actual input from noise, and consequently crucial in long-term potentiation (LTP). In this way, it is ensured that only the synapses driven by the cortical cells responding to the stimulus will be strengthen. Cortico-striatal synapses receive dopaminergic input originated by VTA and SNc. In the case of no postsynaptic activation or dopamine re-

lease, long-term depression (LTD) will occur; i.e. the synapses will be weaken [Arbuthnott *et al.*, 2000; Calabresi *et al.*, 1996]. Another necessary feature of dopamenergic signal to striatum that makes it ideal as learning signal, is that DA is released in the relevant synapses quickly, and also is cleared quickly from the synapses. If it is not, either the activation of the neurons will not be enough and LTD will occur, or in the next trial inappropriate synapses will be strengthen [Hélie *et al.*, 2015].

In contrast, cortical DA levels change slowly. The single delivery of reward increases DA levels in PFC over the baseline for several minutes [Seamans and Robbins, 2009; Feenstra and Botterblom, 1996]. That means that all the active synapses in this period will be strengthen regardless of whether it is associated with the appropriate behavior or not. Thus, Doya [1999] identified cortical learning as Hebbian learning. In this case, LTP occurs at synapses when pre and post synaptic activity is strongly correlated and LTD when is weakly correlated [Ashby *et al.*, 2007; Hélie *et al.*, 2015].

1.4 Habits

Historically, Aristotle was the first to propose the term habit, in order to describe different types of acquired skills of an individual needed to improve his performances leading to a desired goal [Bernacer and J.I., 2014]. The first type is the theoretical, and regards the habits of basic associations that acquired by comprehension ("knowing that x is so") and not repetition, which can be used after for understanding new concepts and propositions. An example is the perception of mathematics, in which someone has to understand the theorems or the concepts to become a mathematician, and not only repeat operational routines. A behavioral habit is the learned best option for the agent in a situation ("knowing how to behave"). The last type, the technical habits, are acquired motor skills needed to achieve an external goal.

The genealogical map of the concept of habit by Barandiaran and Di Paolo [2014] (Figure 1.8) shows the history, starting from Ancient Greece to the late 1980s, and th richness of this notion. They authors identify two major trends, consisting of seven schools of thought that include 77 thinkers. However, the research of habit in neuroscience is quite recent, and mostly emanates from the work of William James. He proposed that habits are learned skills, which use the optimum amount of fine movements (need the least effort and are the most precise) as response to a cue, and do not require conscious attention. As Bernacer and J.I. [2014] describes, Jame's proposal was based on associationism trend, which follows the idea that:

> "... habits are based on the plasticity of matter, and they subserve adaptive purposes. Moreover, a habit can be chunked into smaller pieces that are automatically assembled: this is the main feature

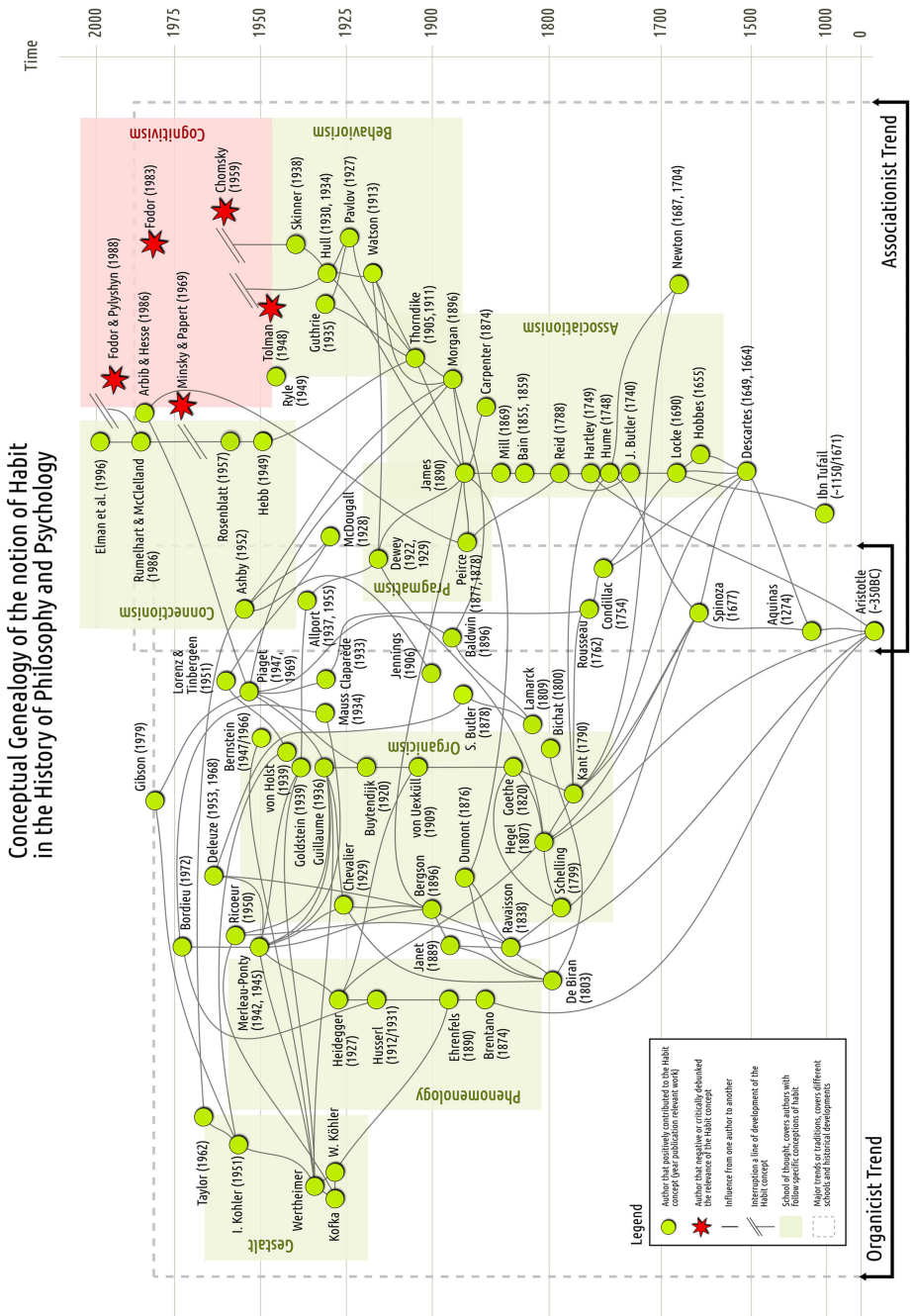


Figure 1.8: As in Barandiaran and Di Paolo [2014]. Conceptual genealogy of the notion of habit. Positive contributors to the notion of habit are indicated by green circles, whereas breaks in the development or significance of the habit concept by red stars.

> of associationism, and the start point of the Pavlovian stimulus-response pairing."

Thorndnike, Pavlov and Skinner followed, and, based on their studies of animal behaviors, establish a new trend, the behaviorism, which supports that behavior can be described and explained through external observations (from the environment) without making ultimate reference to internal psychological processes (the mind, internal state). Soon after the advent of computational and information theory, the notion of habit was replaced by "mental representation" [Tolman, 1948; Chomsky, 1959; Fodor, 1983]. This led to the use of artificial neural networks in computational neuroscience. These networks composed by a number of units and weights that measure the strength of the connections among these units, which are analogous to neurons and their synapses with other neurons.

1.4.1 Definition of Habits

The definition of habit varies among research domains, such as cognitive psychology, cognitive neuropsychology, and animal learning.

In cognitive psychology, the term habit learning was not used at first, but from the late 1960s through the 1980s, several concepts were embodied with this theory [Seger and Spiering, 2011]. Graf and Schacter [1985] defined *implicit memory* as the previous experience used to ease performance on a task without conscious or intentional recollection of those experiences. In implicit memory tasks, priming improve in accuracy and/or processing time after repeated stimuli. Three rules for implicit learning were delineated by Seger [1994][p. 164]:

1. "the knowledge gained in implicit learning is not fully accessible to consciousness, in that subjects cannot provide a full .. verbal account of what they have learned",

2. "information [learned] .. is more complex than a single simple association or frequency count", and

3. "implicit learning does not involve processes of conscious hypothesis testing but is an incidental consequence of the type and amount of cognitive processing performed on the stimuli".

Further, Seger and Spiering [2011] identified as influential to the notion of habit the concept of automaticity that was developed by Shiffrin and Schneider [1977], who characterized a process to be automatic based on the following criteria:

1. short-term memory capacity limitations do not constrain automatic processes, in which attention is not required

2. after the initiation of an automatic process, it is completed without the subjects' intention, as a result of a too quick performance

3. extensive training is required in order a shift from controlled to automatic processing to occur

4. the modification of an automatic process is difficult, once it has been acquired.

The investigation of learned behaviors, such as habits, led Cohen and Squire [1980] to define *procedural learning* in cognitive neuropsychology as "operations governed by rules or procedures". However, this term was insufficient to describe all types of learning and memory, and therefore the "non-declarative" term was defined later by Squire and Zola-Morgan [1988][p. 171] as: " a heterogeneous collection of abilities: motor skills, perceptual skills, and cognitive skills (these abilities and perhaps others are examples of procedural memory); as well as simple classical conditioning, adaptation level effects, priming, and other instances where experience alters performance independently of providing a basis for the conscious recollection of past events". The authors in Squire and Zola-Morgan [1988, 1991] developed a figure in order to illustrate the differenttypes and subtypes of declarative and non-declarative memory (Figure 1.9).

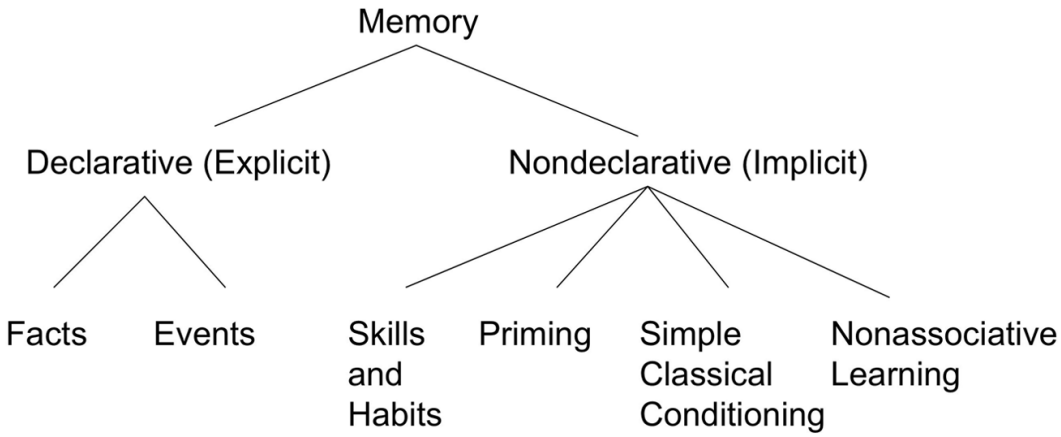


Figure 1.9: The division of long-term memory proposed by Squire and Zola-Morgan [1991]. It has been redrawn by Seger and Spiering [2011].

Parallel, through experimentation on animals, Dickinson [1985] suggested the division of instrumental behavior into two opposed types: the goal-directed and the habit. The diversity of these two types stems from the dependency of the animal behavior on the expected outcome. Later on, a wide definition of habit learning was given by Graybiel [2008][p. 361]:

1. Habits (mannerisms, customs, rituals) are largely learned; in current terminology, they are acquired via experience-dependent plasticity.
2. Habitual behaviors occur repeatedly over the course of days or years, and they can become remarkably fixed.
3. Fully acquired habits are performed almost automatically, virtually non-consciously, allowing attention to be focused elsewhere.
4. Habits tend to involve an ordered, structured action sequence that is prone to being elicited by a particular context or stimulus.
5. And finally, habits can comprise cognitive expressions of routine (habits of thought) as well as motor expressions of routine.

"These characteristics suggest that habits are sequential, repetitive, motor, or cognitive behaviors elicited by external or internal triggers that, once released, can go to completion without constant conscious oversight."

1.4.2 Acquisition vs expression

Identifying the characteristics of a behavior is the guide for investigating which brain areas participate into the acquisition and expression of the different types of behavior. Using this approach, Sherrington [1906] proposed a theory, which is summarized in Ashby *et al.* [2007][p. 632]:

"Novel behaviors require attention and flexible thinking and therefore are dependent on cortex, whereas automatic behaviors require neither of these and so are not mediated primarily by cortex. Instead, it has long been assumed that automatic behaviors are primarily mediated by subcortical structures."

Along with this view, Lashley [1950] supported the existence of a shift of memory traces from cortex to subcortical structures after extensive training. Another supporter of this theory, Fuster [2001] endorsed it with the argument that prefrontal cortex does not participate in automatic overlearned sequences of behaviors. This view assumes that two distinct systems regulate instrumental behavior: the action-outcome (A-O) which produces goal-directed actions and the stimulus-response (S-R) which expresses habits [Yin and Knowlton, 2006; Daw *et al.*, 2005; Niv *et al.*, 2006; Dayan and Berridge, 2014]. It has been hypothesized, for quite long time now, that these systems are parallel and either compete for expression or a shift from A-O to S-R system exists [Yin and Knowlton, 2006].

Nowadays, there is evidence showing that cortex acquires the habits, but BG are needed during training. Piron *et al.* [2016] showed that after the inactivation of GPi (main output of BG) primates are incapable to learn new contingencies, contrary to well-learned ones that are expressed.

In this work, we argue that there is no shift from the goal-directed system to habits. Instead, our proposal is that both habits and goal-directed actions are learned simultaneously, but in a different pace and based on different learning rules. In the beginning of habit acquisition, the goal-directed system leads the final decision, whereas after the establishment of habits, the habit system dominates in the decision procedure. To explain better, how this is possible, I describe the chronicle of the acquisition of habits. Simultaneously, when an action starts to be associated with an outcome in basal ganglia, also an association between a stimulus and a particular action starts to exist at cortical level. However, basal ganglia learn through the feedback they receive from the environment (reward), so they learn fast. Contrary, cortical learning starts at the first received reward, but continues for a lot subsequent actions, regardless of their outcome. In this way, cortex cannot learn the best action as a response to a stimulus alone. That's why it needs basal ganglia as its trainer. After they associate the actions with their outcome, basal ganglia influence cortical decision by leading it to the best choice, and consequently cortex learns only the optimal actions. In other words, cortical learning is based on the statistics provided by the basal ganglia.

"Even if a scientific model, like a car, has only a few years to run before it is discarded, it serves its purpose for getting from one place to another. "

— David L. Wingate

"A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant. "

— Manfred Eigen

Chapter 2

Computational Background

Contents

The complexity of basal ganglia circuity revealed the necessity to develop models for investigating the functional role of their internal and external connectivity, as well as the different mechanisms that are implemented inside their architecture. First Albin *et al.* [1989] used the box-and-arrow representations, schematic representations of neurons (boxes) and axonal projections (arrow), to provide an unified view of BG. However, these representations can be confusing to interpret, and are not suitable for examining the interactions among subsystems or the evolution of the dynamics during learning. On the other hand, biologically constrained computational models provide a useful framework to explain the results from different studies, and even compare findings among different experiments, species and level of analysis [Schroll and Hamker, 2013; Cohen and Frank, 2009]. Furthermore, they are able to produce predictions that derive from the model properties and not directly from the assumptions. Another advantage of computational models is that they can test different hypothesis faster and easier than conducting an experiment, providing preliminary results, which can reveal any deficiencies of these assumptions. The development of the computational neuroscience field expedited the research on basal ganglia functioning, providing evidence that experimental research benefits from these kind of models. For example, the pioneering work of Sutton and Barto [1998] that introduces the actor-critic model and the focused selection model by Mink [1996] gave a new perspective on the role of the nuclei of BG and the pathways that they comprise, as well as on the role of dopamine during learning.

Our model has been developed to investigate action selection, and particularly the formation of habits. For this reason, in this chapter, I introduce models which investigate one of the two systems. Some of the models are biological inspired, meaning that they are developed to explore how the anatomy, connectivity or other biophysical properties participate in these mechanisms. However, other models investigate the principles that underlie these behaviors, without taking in account how exactly they emerge from specific brain areas. Because it would be impossible to review all the existed models, the choice of the presented models was based on either their significant contribution in research of these mechanisms or the given inspiration to our model. Table 2.1 summarizes which models will be reviewed here and the mechanisms they explore. For further information, someone can refer to Liénard [2013].

References	*Action Selection*	*Learning*	*Habit Formation*
Gurney *et al.* [2001a,b]	✓		
Daw *et al.* [2005]			✓
Leblois *et al.* [2006]	✓		
Girard *et al.* [2008]	✓		
Ashby *et al.* [2007]	✓	✓	✓
Guthrie *et al.* [2013]	✓	✓	
Dezfouli and Balleine [2013]			✓
Baldassarre *et al.* [2013]	✓	✓	✓

Table 2.1: Overview of the computational models reviewed and the mechanisms they explore

2.1 Reinforcement learning

Twenty years ago, the study of learning methods by which agents can improve their performances by interacting with the environment was a hot area (and still is) of research in artificial intelligence. Inspired by the animal research, Barto [1995] proposed a new learning paradigm called *reinforcement learning (RL)*. RL systems differ from their predecessors (supervised learning) in such way that the goal of RL systems is to maximize the frequency of reinforcing events through time by adjusting their behavior, contrary to the latter system which learn appropriate behaviors by a set of examples of correct input/output behavior.

Theories derived from animal research, as the classical and operant conditioning, were the core of reinforcement learning. However, the research of learning in neuroscience was expedited by the development of the reinforcement learning theory, as a result of the effort to relate these architectures with structures and functions of certain brain regions.

2.1.1 Prediction error learning theory

The theory described in this section explains how a stimulus is paired with a reward. When the presence or the absence of a stimulus ends up to the same reward, then the stimulus induces no surprises, resulting in a correct prediction from the object. Hence, no learning occurs about this stimulus. In other words, an object will learn only when a prediction error exists. The procedure of learning a stimulus-reward pair is outlined below.

A reward prediction error (PE) is defined by:

$$PE(t) = \lambda(t) - P(t) \tag{2.1}$$

where λ is the received reward in trial t and P the prediction of the reward by the object. So PE is the difference between the real and expected reward. Using this PE, the object updates the value of the expected reward of a stimulus, in order to minimize the error and become more accurate in its predictions. This update follows the equation:

$$P(t+1) = P(t) + \alpha * PE(t) \tag{2.2}$$

where α is the learning rate. Said differently, the calculated prediction error weighted by the learning rate is added to the current prediction . The diagram in Figure 2.1A displays how the predictions are updated based on the errors [Schultz, 2015]. The learning rate is usually < 1 in order the error to decline gradually, following the typical asymptotic learning curve (Figure 2.1B).

From the equations above, it is concluded that predictions worse than the received reward generate positive prediction errors and lead to learning of a behavior, although the extinction of a behavior derives from negative prediction errors, when the reward is less than expected. Finally, there is no learning and the prediction stays unchangeable when the error becomes zero. Thus, as Schultz [2015] draws up:

> This formalism views learning intuitively as a change in behavior that occurs when encountering something new or different than predicted, whereas behavior stays the same when everything occurs according to "plan".

2.1.2 Temporal difference

The prediction error method implies that the value of an action is updated based only on the last outcome. Sutton and Barto [1998] suggested that this update also depends on their reward history, and based on this assumption they proposed the *temporal difference learning theory*. Put it differently, the object predicts the reward by taking in account a prior event, so the learning

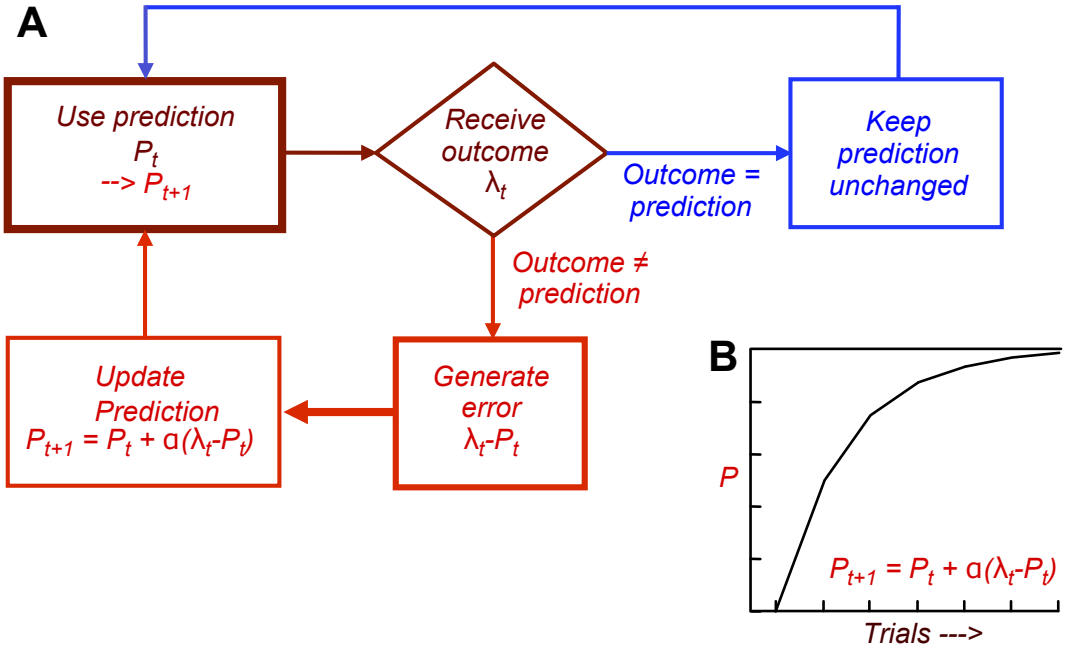


Figure 2.1: A: Feedback circuit diagram that illustrates how learning occurs through the update of a prediction by errors. B: Gradually declining prediction errors generate typical learning curve. Taken from Schultz [2015].

derives directly from experience without a priori model of the environment [Sutton and Barto, 1998; Bar-Gad *et al.*, 2003].

After an action has occurred, its value is contrasted with the observed reward plus the updated prediction for the future rewards. This comparison result is called *temporal difference error (TD)* and its formula is:

$$TD(t) = [r(t) + \gamma P(t)] - P(t-1) \tag{2.3}$$

where γ is a discount factor. This TD error is used to update the value of the action as in PE method.

$$P(t+1) = P(t) + \alpha * TD(t) \tag{2.4}$$

Sutton and Barto [1998] TD learning theory has been associated with the phasic changes of the midbrain dopaminergic neurons in their firing rate [Schultz *et al.*, 1997].

2.1.3 Actor/critic

Influenced by the basic idea of Throndike's "Law of effect", Sutton and Barto proposed the actor-critic architecture to describe the reinforcement learning rules underlying this idea; if an action followed by a desired (absence) outcome,

then the action is reinforced (weaken), in order to be more (less) probable to be produced again. In this architecture, there are two different memory systems, the actor which is responsible for selecting the actions, and the critic which evaluates the actions made by the actor.

The critic estimates the action value in consequence of a reinforcement signal, that follows the temporal difference learning rule, and the actor based on the same signal learns to produce or not an action as a result of a stimulus. The actor learning rule is based on the idea that if the comparison of the expected with the observed outcome of an action differs, then learning occurs, otherwise no. Now, if the expected consequence is better (worse) than the observed, then the tendency of the action, responsible for this outcome, to be repeated in the future will be weakened (strengthened). When the TD error is positive (negative), the tendency of choosing the action is strengthen (weaken), because it means that the outcome is better (worse) than expected.

Suppose a collection of possible actions is represented by a collection of units, and the actor chooses the more active one. Let a_t to the activity of a unit at time t analogous to action a, which is computed by:

$$a_t = \sum_{i=1}^{n} w_t^i \, x_t^i \tag{2.5}$$

where w_t^i are the weights of the activation and x_t^i the input activity at time t. If an action is executed at time t, then the following learning rule is applied on the corresponding unit:

$$w_t^i = w_{t-1}^i + \alpha \, TD(t-1) \, x_{t-1}^i \tag{2.6}$$

where α is the learning rate and $TD(t)$ the TD error as described in 2.3.

Figure 2.2 illustrates a neural network implementing the actor-critic architecture. The same learning rule and modulator signal is used by both units, actor and critic, in order to update their synaptic weights. However, the learning rule is applied only to the winning actor unit, after competing with the others. Lateral inhibition is a suitable implementation of this competition. Also, the actor unit mechanisms must remember past connection of pre- and postsynaptic activity, in contrast to the critic unit which needs only past presynaptic activity.

2.1.4 Model free / model based

Experimental and theoretical evidence support that actions are guided by two distinct systems. These mechanisms are trained with different methods, to learn and make predictions about an action's outcome; *i.e.* if it will receive reward or punishment. With the *model-based* method, the system learns to

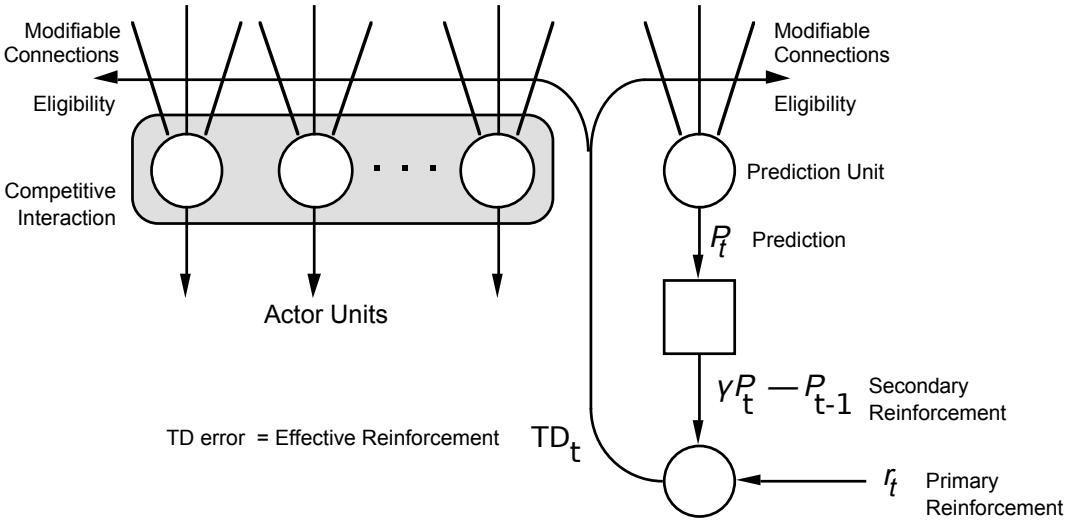


Figure 2.2: As in Barto [1995]. Network Implementation of the Actor-Critic Architecture. The same learning rule, as well as the same TD signal are used in both units, actor and prediction. A competition among the actor units results to a winner that determines the action, which will be executed, followed by the application of the learning rule only to the winner.

make predictions of an action outcome based on representations of the environment, expectations and prospective calculations [Dayan and Berridge, 2014]. Contrary, *model-free* method uses retrospective experience to train the system to estimate the value of an action. In other words, model-based models are capable to predict an outcome of an action by seeing in the future, and model-free models by seeing in the past and choosing an appropriate action from the previously experienced ones (*i.e.* they take advantage of the collected knowledge of what an action results).

Computational analyses of instrumental and Pavlovian learning implicate this distinction between model-free and model-based forms of learning and computation [Dayan and Berridge, 2014]. Goal-directed actions have been correlated with model-based strategies, which employs an internal model of the external world [Daw *et al.*, 2005; Doya, 1999]. On the other hand, model-free strategies, as habits, are cached information of the action outcomes that were received in past interactions with the environment. In the Figure 2.3 by Dayan and Berridge [2014], the authors summarize these two computational approaches through their applications in instrumental versus Pavlovian forms of reward learning.

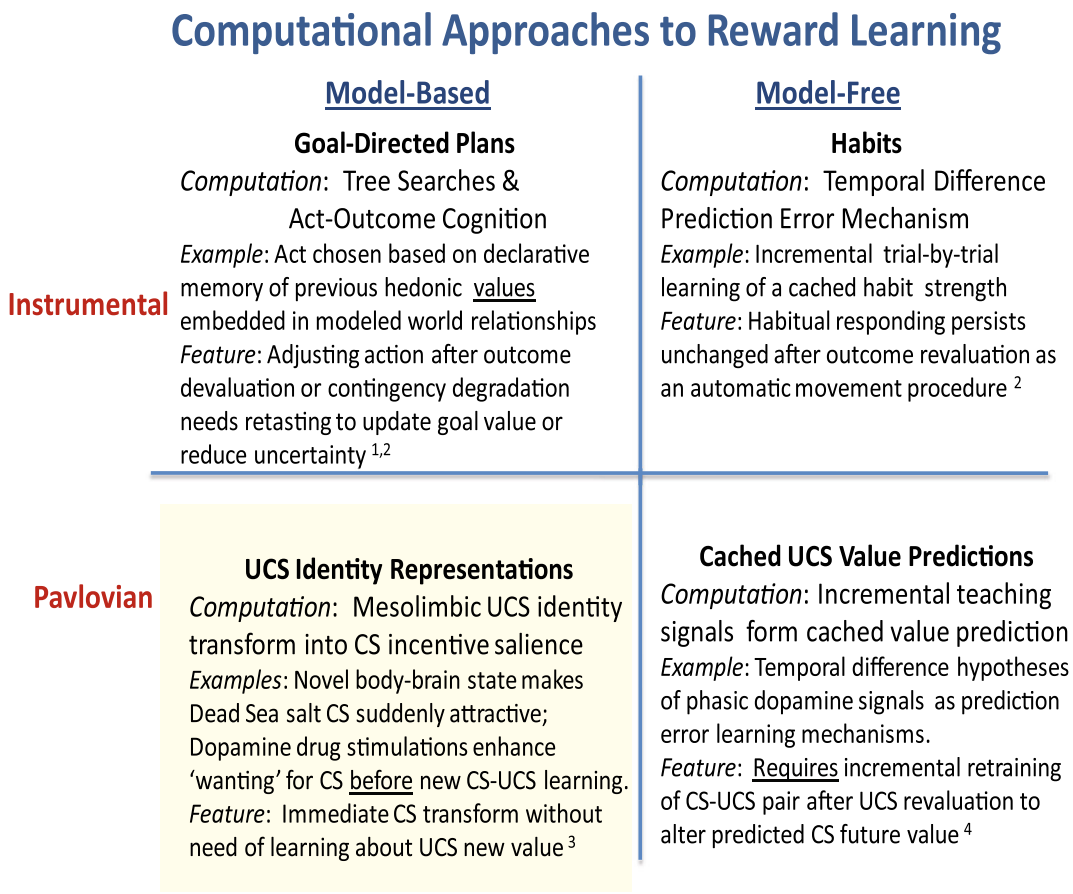


Figure 2.3: As in Dayan and Berridge [2014]. A comparison of model-based and model-free approaches to reward learning. Applications of both approaches in instrumental and Pavlovian conditions are also proposed. A brief description of computations, an example of behavioral or neural demonstration, and a feature by which it can be recognized are presented in each cell. Citations: [1] Dickinson and Balleine [2010]; [2] Daw *et al.* [2005]; [3] Robinson and Berridge [2013]; [4] Schultz *et al.* [1997]

2.2 Models of decision making

2.2.1 Gurney *et al.* [2001a,b]

Gurney *et al.* [2001a,b] presented a model of basal ganglia to explore the intrinsic processes responsible for action selection. Through the revision of these networks and the anatomy of basal ganglia (BG), they proposed a novel functional architecture of BG, which is dissociated into two pathways: the 'selection' pathway, which is constituted by a feedforward off-center on-surround network, and makes the selection per se, and the 'control' pathway, which regulates the selected action (by the 'selection' pathway) to ensure a plain

performance.

Previously, Redgrave *et al.* [1999] argued that centralized switching mechanisms are capable to approximate action selection, in terms of connectivity productiveness. Additional to this work, Prescott *et al.* [1999] showed that their hypothesis of selection as a major function of basal ganglia can be justified by the known anatomy and physiology. In that paper, the authors developed a biologically inspired model of BG, referred to as GPR2, to provide answers to the following questions: How 'selection' can be quantitatively articulated and implemented as basal ganglia intrinsic model? How basal ganglia anatomy can be explained as a specialized set of neural mechanisms for selection?

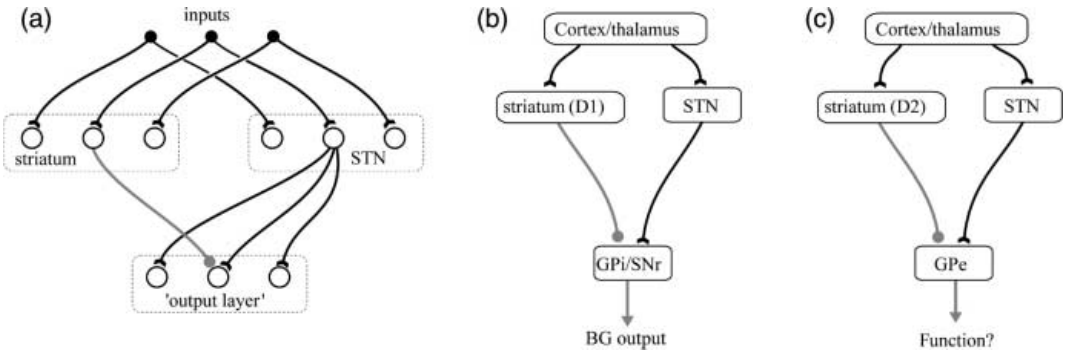


Figure 2.4: As in Gurney *et al.* [2001a]. The new functional architecture divided in the component parts. (a) Illustration of the network with two separated input nuclei, one for excitation and one for inhibition, but a unique output. (b) An instantiation of the 'selection' pathway. (b) An instantiation of the 'control' pathway.

Answering to the first question, they proposed that actions are presented to BG as *saliences* (i.e. the overall activity level of neural representation corresponding to an action), and they are mediated to the outputs by the release of inhibition. Thus, the process of 'action selection' is redefined as 'signal selection', in which large salience input to striatum and STN (inputs of BG) inhibits the outputs of BG (GPi and SNr) resulting in low signal outputs. Selection among a variety of actions is derived by the competition of neural populations (called here 'channels') based on their saliences, in a winner-take-all manner. The channels with the lowest salience in the outputs of BG will trigger an action. Therefore, a selection mechanism can be implemented by an off-center on-surround neural network as shown in Figure 2.4 (a).

Based on the connectivity of BG, and the previously presented approach of selection process, the authors proposed two co-existing modules in the architecture of BG, the selective and the control. The selective module contains an off-center part with the projection from D1 to GPi/SNr, and an on-surround part with the distributed STN to GPi/SNr projection of the hyperdirect path-

way (Figure 2.4 (b)). This module contains the selection mechanism per se, since GPi/SNr are the BG outputs; therefrom the name selective. Contrary, the control module's function is to regulate the properties of the main selection mechanism by increasing the striatal inhibition to the outputs structures. This is accomplished directly by the signals sent from GPe to GPi/SNr, and indirectly to STN (Figure 2.4 (c)). The control module includes the focused projections from striatal populations containing D2 receptors and diffused projections from STN to GPe. The overall architecture of the model is shown in Figure 2.5

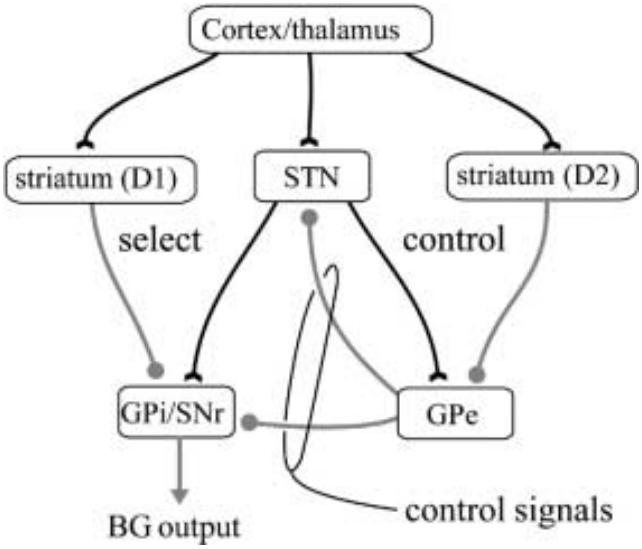


Figure 2.5: As in Gurney *et al.* [2001a]. The functional architecture of the full model demonstrating the combination of the selection and control pathways.

Figure 2.6 shows the ability of the model to choose the channel with the highest salience. The figure displays the salience of the input to three channels and their output signal level through time. Initially there is no salience on all channels, until time $t = 1$ when the channel 1 receives a salience of 0.4, resulting in a decrease of its output activity and an increase of the other channels. If the output of channel 1 is less than a given threshold it is hypothesized that the corresponding action is performed. At $t = 2$, the salience of channel 2 increases to 0.6, which is sufficient to entirely inhibit its output while forcing an increase to the output of channel 1 above the threshold, as well as a further increase to channel 3. Thus, the selection of channel 2 causes the interruption of the prior selection of channel 1, if it had previously occurred. Then, from $t = 3$ to $t = 4$ the two first channels receive the same salience, which prompts a inter-channel competition that leads to a higher common output level, and therefore none of them is selected.

Compared to the model of [Albin *et al.*, 1989], the authors highlighted two main differences. First, the direct pathway of Albin contains the whole

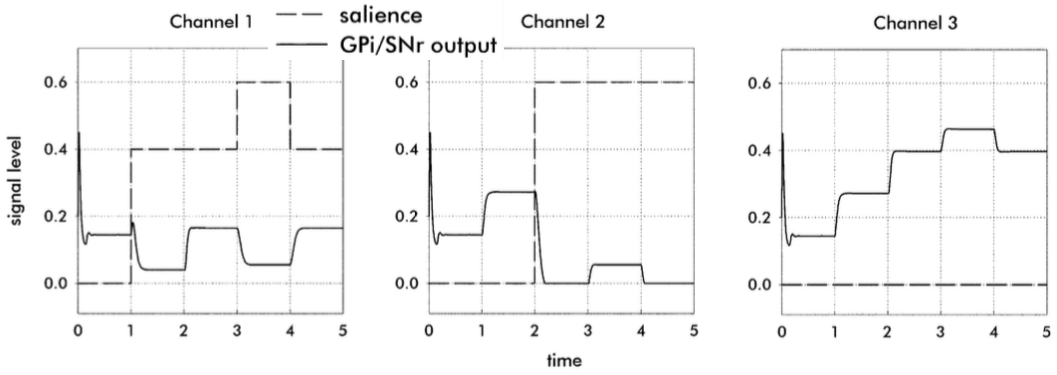


Figure 2.6: As in Gurney *et al.* [2001b]. Simulation results for three channels in a 6-channel model. Only channels 1 and 2 have non-zero salience; channel 3 is shown as representative of the other, non-active channels. For each channel, the solid line indicates the GPi/SNr output and the dashed line the input salience. Time is measured in arbitrary units. a Results for an intact model with synaptic weights on STN e erents having 0.9 times the value of those from striatum.

striatum, although the selective module comprises of only the striatal sub-populations of D1 receptors, and furthermore incorporates STN as an input nucleus. Secondly, the control module does not contain the GPi/SNr nuclei, but contains projections from STN to GPE, compared to the indirect pathway. Also, in GPR2 model there is the distinction of striatal sub-population (D2 receptors) that participate in the control module. Lastly, the two pathways have different functional roles between the two models: selection and control in GPR2, facilitation and inhibition of actions in Albin.

2.2.2 Girard *et al.* [2008]

Inspired by the GPR model, Girard *et al.* [2008] explored the functional perspective of the BG anatomy in action selection through a computational model of the basal ganglia (hence name CBG) as shown in Figure 2.7. The GPR model was extended to contain usually neglected neural projections, such as the GPe projections to STN, GPi/SNr and Str. Also, except the D1 and D2 striatal neurons, it also includes GABAergic fast-spiking interneurons (FS).

The architecture of CBG model includes a closed cortico-baso-thalamo-cortical loop which emanates by and ends up to frontal cortex (FC). However, basal ganglia receive also cortical input from sensory cortex (SC) that projects to all types of striatal neurons (D1, D2, FS). SC projects also to the excitatory cortico-thalamic loop through its projection to FC. The thalamic reticular nucleus (TRN) is included as a global regulatory inhibitor of thalamus. Fur-

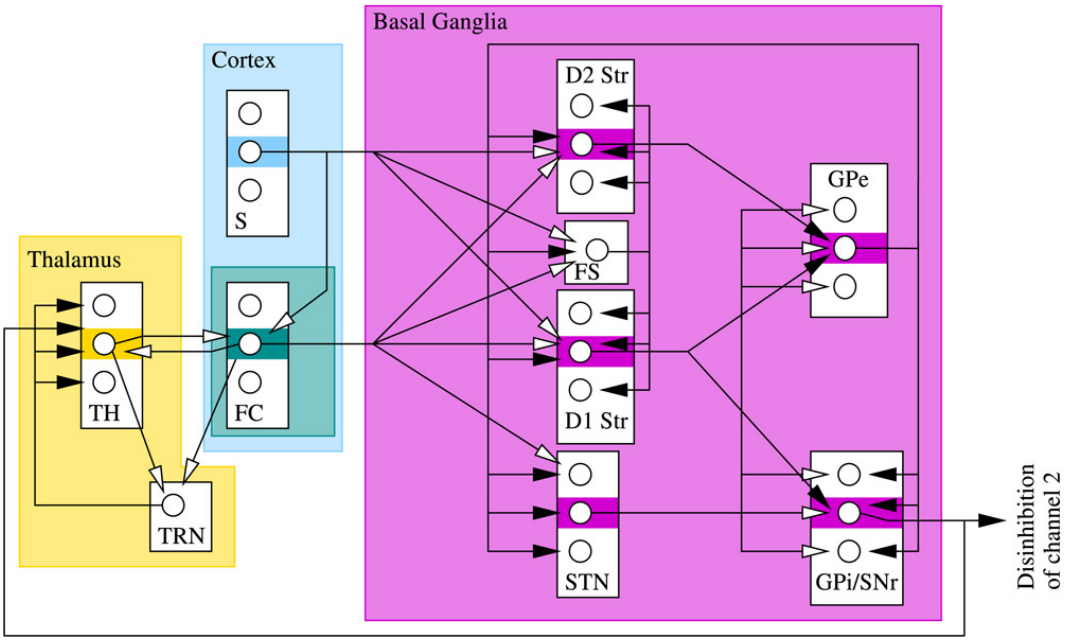


Figure 2.7: As in Girard *et al.* [2008] Diagram of the architecture of the basal ganglia model (CBG). Each box presents a nucleus, and the circles inside a neuron. Each nucleus contains three neurons (referred in the passage as channels), which compete for selection. The colored shading represents the second channel. All the connections related to this channel are illustrated, and they are identical for the other channels. Excitatory projections are shown with white arrowheads, and inhibitory with black arrowheads.

thermore, GPe modulates the activity of FS, which in turn exert feedforward inhibition on the striatal neurons.

GPe projects also to all the other basal nuclei, striatum, GPi/SNr, and STN [Staines *et al.*, 1981; Bevan *et al.*, 1998; Kita *et al.*, 1999]. STN and GPi/SNr neurons receive diffused input from GPe, influencing large sets of STN and GPi/SNr neurons, while it projects in a focused manner to striatum. In this way, the focus feedback from GPe to striatum makes the selection of a channel sharpened by promoting the channel with the highest salience in D1 and D2. Also, the amplitude of the unselected channels inhibition in GPi/SNr is limited by the interaction between the global inhibition of GPe and excitation of STN on GPi/SNr. As a result, the thalamo-cortical loop receives inhibitory projections from BG that consequently causes a selective amplification of the winning channels while limits the amplification of the unselected channels. This architecture provides to frontal cortex a mechanism for selective amplification of the winning channel, through the thalamo-cortical loop, and an inhibitory mechanism of the unselected channels, through the subcortical circuits of BG, ensuring no interference with unwanted motor commands.

They tested the model with a similar task as in Gurney *et al.* [2001a,b] (in Gurney protocol the saliences are changed every $1s$, here they change every $2s$), and compared their results with a more recent version of GPR model [Prescott *et al.*, 2006]. As shown in Figure 2.8, the model expresses the same behaviors as in GPR model, except at $t = 6 \rightarrow 8$, when the first two channels receive the same salience (emphasized with asterisk in the bottom row of the Figure 2.8). The CBG model selects both channels (both behaviors are hypothesized to be executed), contrary to GPR model that selects the second one; this comes as a result of the previous step where the channel 2 was selected and remains selected in the current step whereas channel 1 is fully inhibited (higher inhibition level than the inhibition at rest).

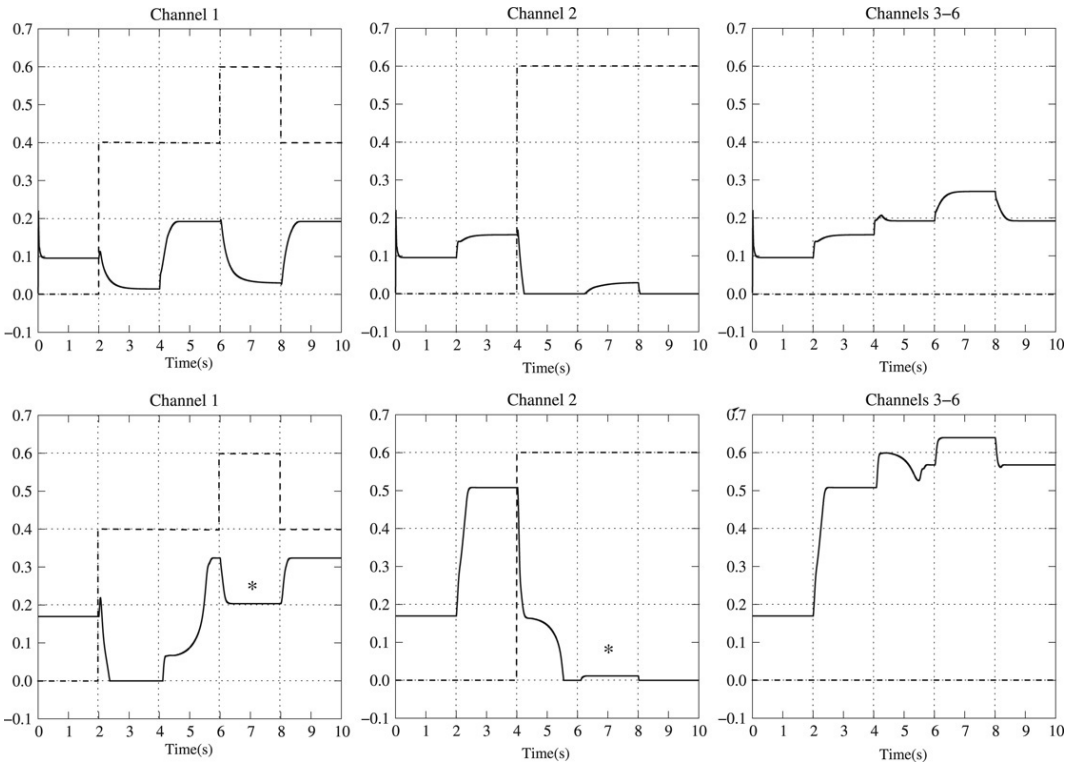


Figure 2.8: As in Girard *et al.* [2008]. At the top are presented the variation of the GPi/SNr inhibitory output of CBG during the Gurney *et al.* [2001b] test, and at the bottom the result of the output of GPR. Dashed and solid lines represent the input salience of the channel and the output of the channel respectively. The asterisk denotes the different behaviors that the two models express during the fourth step (6 s $<$ t $<$ 8 s). CBG selects both channels 1 and 2, whereas GPR only channel 2.

2.2.3 Leblois *et al.* [2006]

In Leblois *et al.* [2006], the authors proposed a model for the function and dysfunction of the motor part of basal ganglia. Albin *et al.* [1989] and DeLong [1990] introduced one of the first models of motor symptoms of Parkinson's disease (PD) and Huntington's disease (HD), which relies on a segregation between the BG's direct and indirect pathways. Later, Mink and Thach [1993] have suggested that a center surround inhibition of pallidal activity originated by the focused striatal inhibition combined with the diffused excitation from STN provides the basis of action selection. Based on this idea, Leblois *et al.* [2006] assign a primary role to the hyperdirect pathway. They demonstrated that action selection in cortico-basal ganglia loop arise from the competition between the positive feedback of the direct pathway and the negative of the hyperdirect, resulting to the emergence of symmetry-breaking.

Our model is an extension of this model, and will be described more thoroughly in chapter 3.1.

2.2.4 Guthrie *et al.* [2013]

Guthrie *et al.* [2013] extended the model by Leblois *et al.* [2006] in order to explore the parallel organization of circuits in BG [Alexander *et al.*, 1991] through a two-armed bandit task. In this task, two cues associated with a reward probability are presented in two positions, and the model has to choose one of them. After the choice has been made, a reward is delivered or not according to the reward probability of the chosen cue. The model contains the intrinsic connectivity presented in Leblois *et al.* [2006] with the addition of two action selection modules: the cognitive for the selection of the cue, and the motor for the position. The two modules are considered parallel, because of their distinct cortical input. Action selection is based on symmetry breaking that emanates from the addition of internal noise in the structures, which further allows the system to explore possible actions. Furthermore,the model incorporates learning between the cues and their outcome. A simulated reward signal evokes cortico-striatal learning inside the cognitive module, resulting the selection of an action based on the learned cues and not the noise as before learning.

This model constitutes the previous generation for our own model, and is described extensively in 4.1.2.

2.3 Models of habit formation

2.3.1 Daw *et al.* [2005]

The traditional theory of behavioral control supports the existence of two systems emanated by: the dorsolateral striatum and its dopaminergic afferents, from which habitual control derives, and the prefrontal cortex that executes goal-directed actions. However, such a distinction of behavioral control raise two questions as identified by Daw *et al.* [2005]: "why should the brain use multiple action controllers, and how should action choice be determined when they disagree?". The authors answered these questions by suggesting a Bayesian principle for the arbitration between the two controllers based on uncertainty. They assume that the most accurate system is unfolded in order a decision to be made.

The main idea behind this model is that the goal-directed and habit controllers compete each other for expression. The controller with the strongest salience wins this competition and its action is finally executed (Figure 2.9).

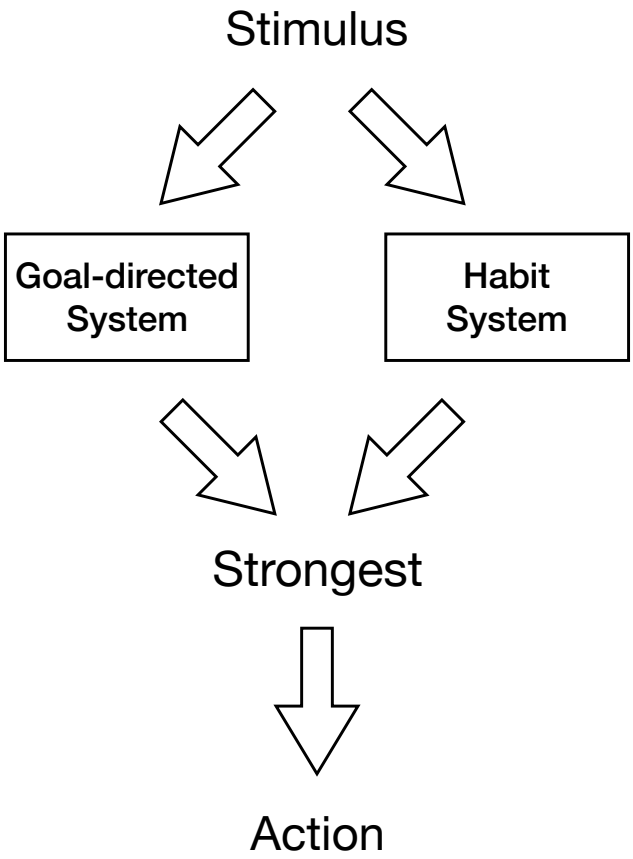


Figure 2.9: Illustration of the proposal by Daw *et al.* [2005] about how a goal-directed action or a habit is chosen to be executed.

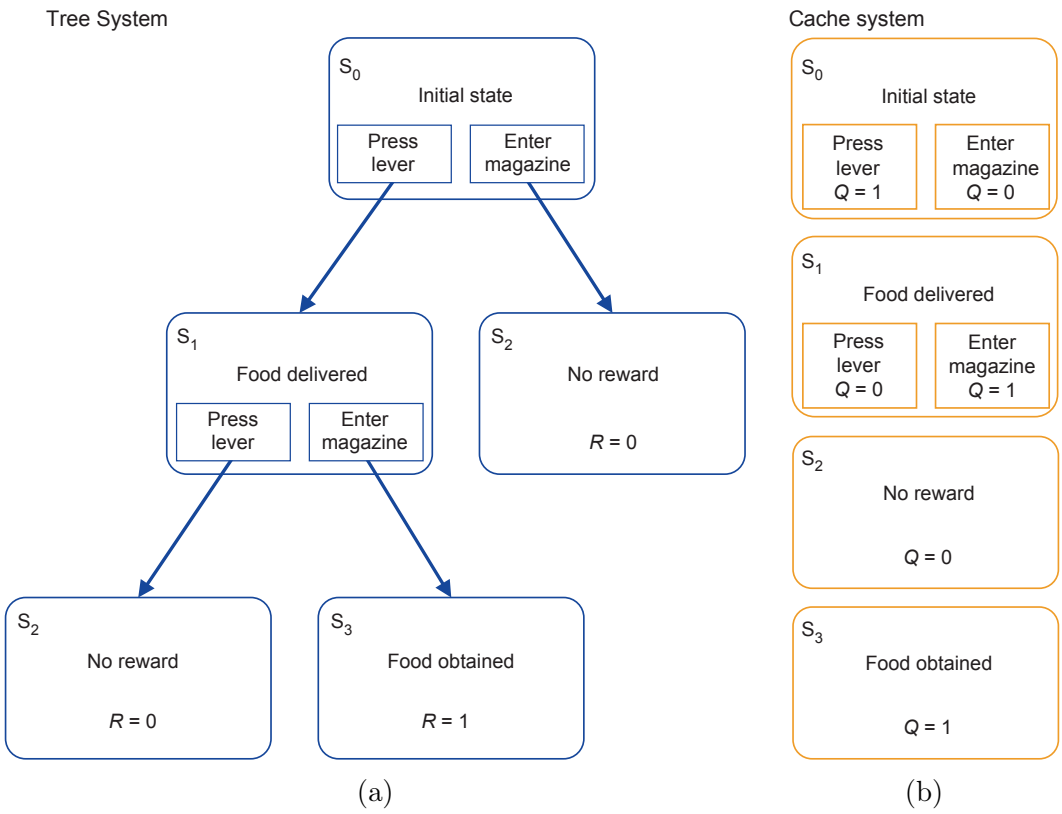


Figure 2.10: As in Daw *et al.* [2005]. Representations of an instrumental conditioning task used by the (a) model-based and (b) model-free methods. Reward is represented by $R = 1, 0$; 1 if reward was attained, 0 if it wasn't. Q is the expected future value for each action in each state in a model-free controller.

They proposed that both systems are based on reinforcement learning framework, although they comprise of methods that make approximations by a different manner in particular circumstances. Model-free approaches are associated with the habitual control, whereas model-based with goal-directed control. The basis of the first method is the association of an action with a summary of its future value (referred as 'cached' value). Working with this kind of values results in the separation of the values from the outcomes, which make this method inflexible to re-evaluation of the outcomes. This is a characteristic of habitual control. By contrast, the model-based method is constructed by chaining predictions in a sequence about the consequences of an action. This method triggers quicker reactions when outcomes are changed than the model-free, however it needs more time to select an action, because of the need for exploration of different branches of future situations.

The two controller models were simulated on the tasks illustrated on Figure

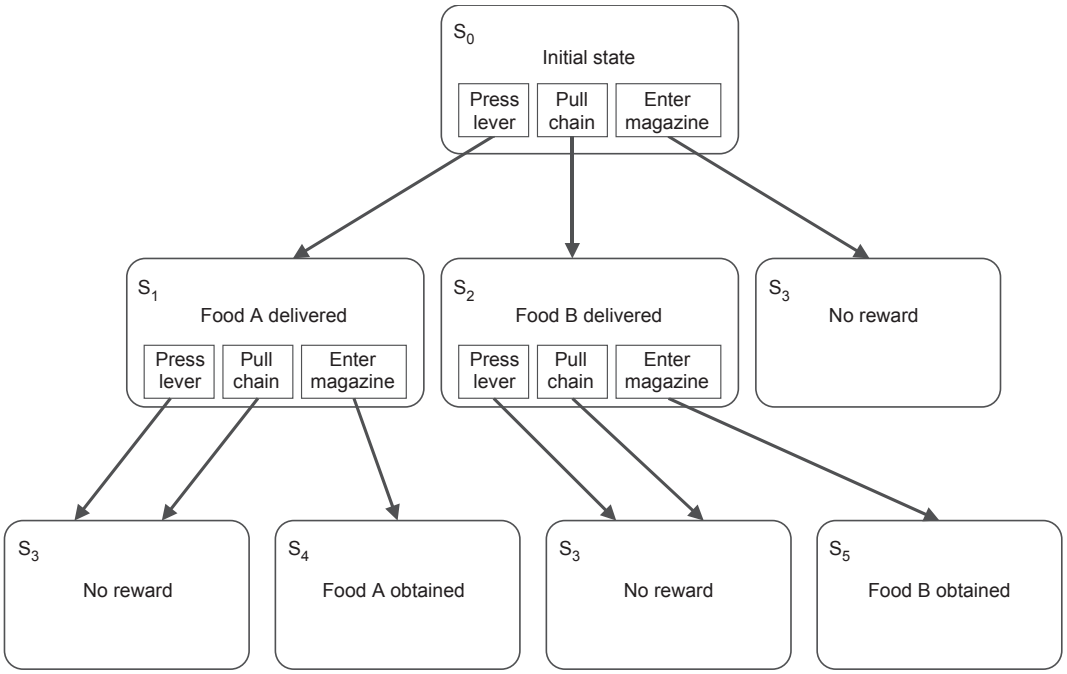


Figure 2.11: As in Daw *et al.* [2005]. Tree representation of a two-actions instrumental conditioning task, where two rewards can be obtained.

2.10 and 2.11. The subject is given different choices in two states in order to achieve a reward, although only one sequence results to reward. As shown in Figure 2.10b, the cache system (model-free) evaluates the value of an action independently of any outcome information. On the contrary, the tree system keeps track of the expected consequences for a sequence of actions (Figure 2.10a). The action values of each system along with the uncertainty of the system about those values were studied as functions of the amount of training and the position of the action in the behavioral sequence relative to the reward.

Figure 2.12 and 2.13 show that in both systems prior ignorance was gradually replaced by certainty of the expected value. The tree system was more confident earlier in training compared to the cache, because any part of experience is used to influence the estimation of action values at all states. However, the nature of the learning in the cache system is to delay this transmission.

In the first task, for the action proximal to reward (magazine entry), the enhancement of data efficiency allowed the tree system to be asymptotically more certain (Figure 2.12b), contrary to the distal action (lever press) where the system is less confident, because of the requirement of an extra iteration resulting to the addition of noise (Figure 2.12a). By contrast, the recall of values make the cache system to stay intact by noise (there are no computations) resulting to the same confidence in both states. Different results were found on the second task, where a second choice with a different rewarded outcome

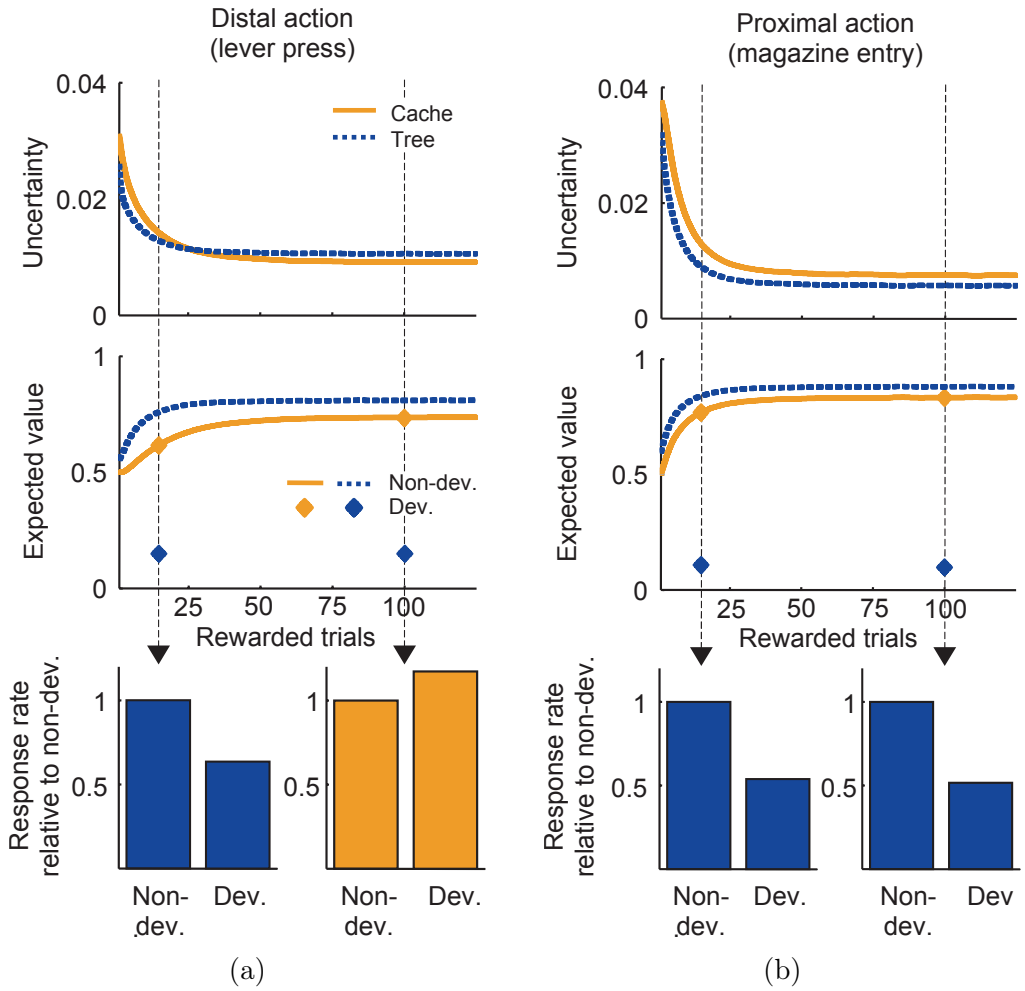


Figure 2.12: As in Daw *et al.* [2005]. Simulation of the two controllers in the task of Figure 2.10 for the first (a) an the second (b) action. The topmost figures show the uncertainty of the value estimation in the two systems, and the middle ones, their value estimations. The diamond indicate the estimation of the value after reward devaluation at various training stages. The bar plots illustrate the probability of choosing an action before and after devaluation of their contingencies, normalized to the non-devalued level. Which system controlled the action is denoted by the bar color.

exists (Figure 2.13). Here, the experience of the agent was distributed among more states and actions, resulting to fewer relevant data about any particular action value. This ensued the preservation of uncertainty in high levels in the cache system even after extensive training. In both tasks, the sensitivity to reward devaluation was kept whenever the tree system dominated, whereas the domination of the cache system the insensitivity in devaluation. Overall these results support the authors' hypothesis that brain selects among the controllers

based on expected accuracy in order to choose an action.

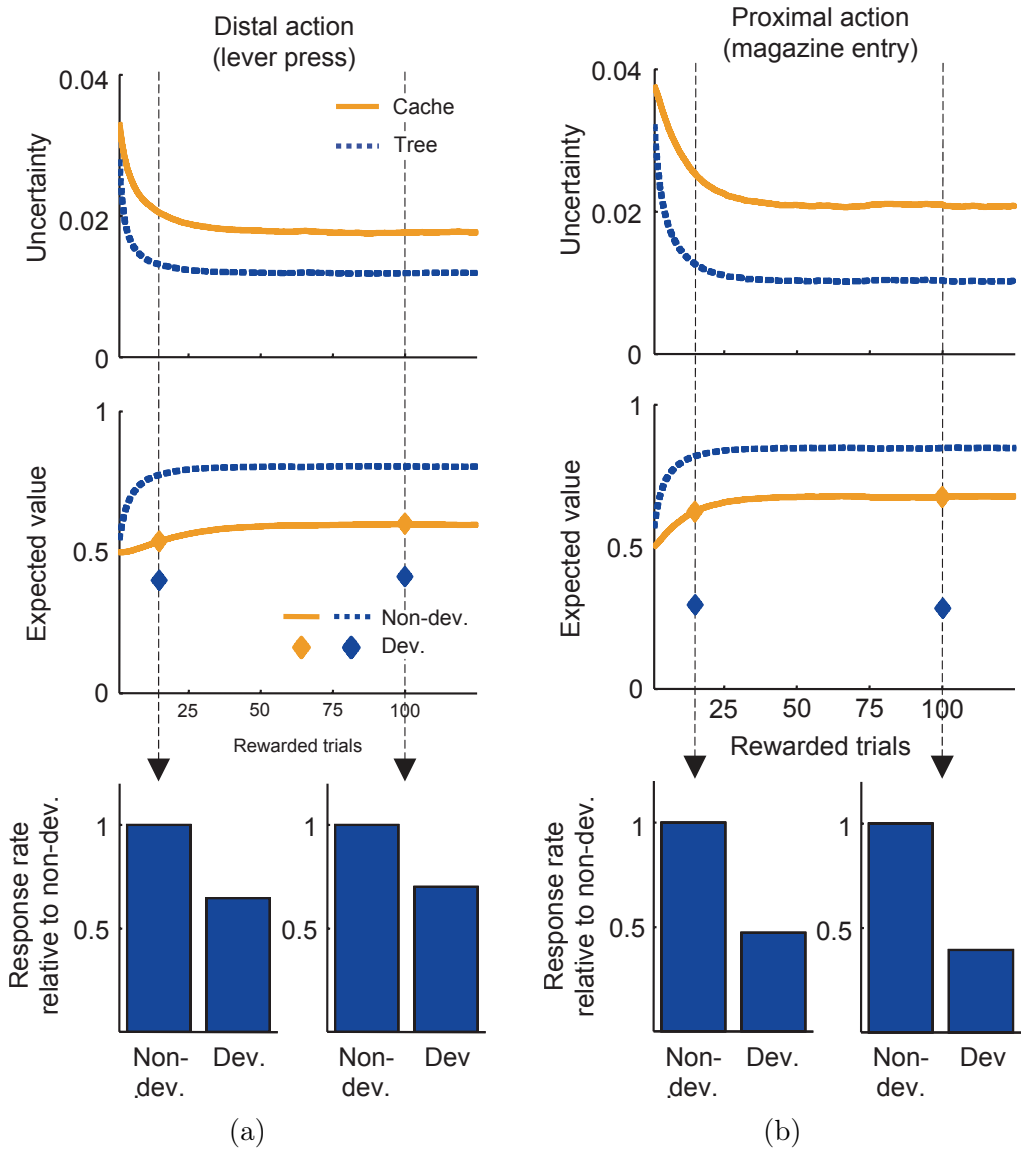
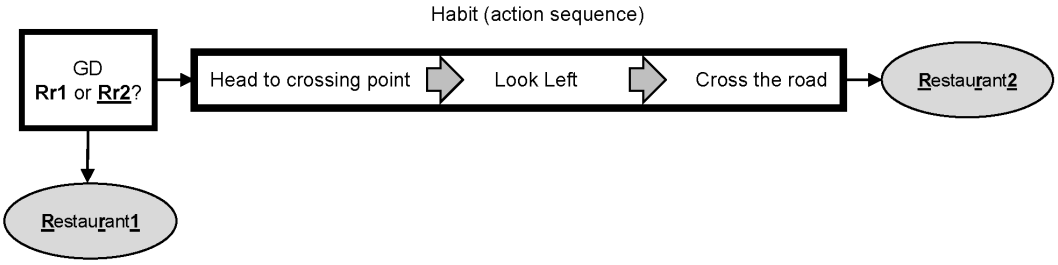


Figure 2.13: As in Daw *et al.* [2005]. Simulation of the two controllers in the task of Figure 2.11.

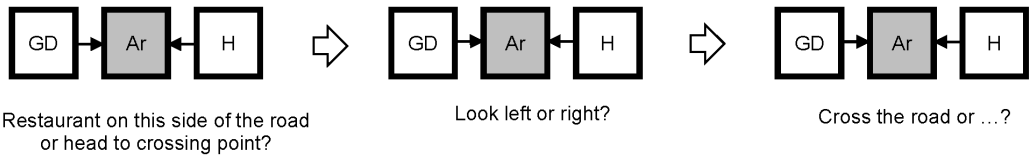
The existence of two separate controllers is based also on the idea that habitual control derives from dorsolateral striatum and further from the loop that it forms with cortex, and the goal-directed from prefrontal cortex and thus the dorsomedial striatal loop. This parallelism endorses the suggestion of the authors that the competition between model-free and model-based control must be considered between the two loops and not between cortex and striatum as previously believed. Although, the role of dopamine projection to dorsomedial striatal neurons cannot be explained by this hypothesis.

2.3.2 Dezfouli and Balleine [2013]

Following the general idea of two forms of action control to govern instrumental conditioning (goal-directed actions and habits) and experimental results from a two-stage task on humans, Dezfouli and Balleine [2013] suggested a new model for understanding the interaction between goal-directed and habitual action control. This model proposes that a goal-directed mechanism selects if a goal-directed action or a habitual sequence of actions will be executed to reach a goal; *i.e.* the model uses a hierarchical manner for the interaction between the two processes (Figure 2.14a). Furthermore, it proposes that when a habitual sequence is selected, the action in the second stage depends on the choice of the first stage. This contradicts the flat architecture proposed by Daw *et al.* [2011], who argued that the selection of an action in the second stage is independent from the first, and that an external arbitration mechanism coordinates the interactions between the two systems (Figure 2.14b).



(a)



(b)

Figure 2.14: As in Dezfouli and Balleine [2013]. Illustration of the (A) hierarchical and (B) flat organization in an example. Goal-directed system (GD); Restaurant on this side of the road (Rr1) and on the other side (Rr2); Arbitration mechanism(Ar).

Overall this model suggests the existence of a goal-directed system, which (explicitly) selects the type of the action (goal-directed or habit) that will be executed (Figure 2.15). In this hierarchical model, there is no need for both systems to make their own choice and then compete for expression.

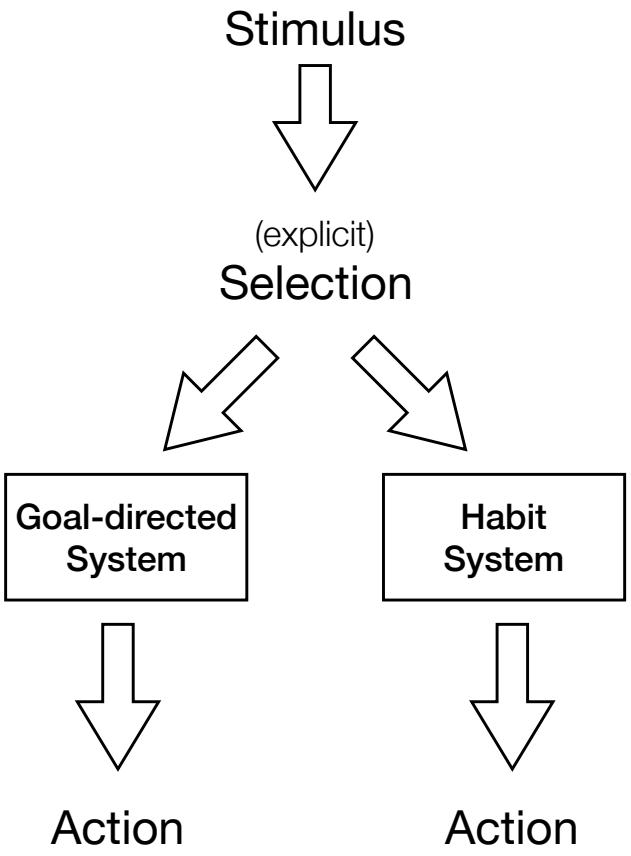


Figure 2.15: Illustration of the proposal by Dezfouli and Balleine [2013] about how a goal-directed action or a habit is chosen to be executed.

The structure of the task used to test this theory is illustrated in Figure 2.16. In the first stage (black screen), two actions are available that lead to one out of two slot machines by a fixed probability. For example, the action A1 provides access to slot machine S1 70% of the times, whereas S2 is reached only 30% of the times. Choices in the second stage are rewarded with a high (0.7) or low (0.2) probability. The association of the reward probability with each slot machine changes randomly with a small probability (1/7).

The authors compared the ability of the flat and the hierarchical architectures to explain experimental data from the same task. First, they queried whether the decisions in this task are goal-directed, habitual or a mixture of both. To answer this question, they analyzed the data to identify the tendency of the subjects staying on the same action in the first stage based on the received feedback of the previous trial. The results indicated that a previous reward trial increased the possibility of repeating an action regardless whether the reached slot machine was the common or the rare result of the action (Figure 2.17). Although this increase was higher in the case of a common

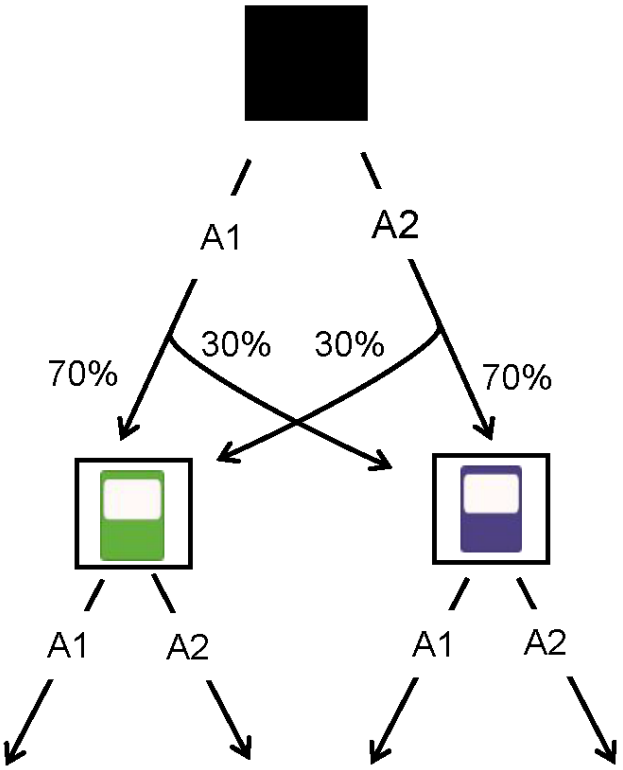


Figure 2.16: As in Dezfouli and Balleine [2013]. Structure of the task. The key presses leads to one of the slot machines 70% of the time, and to the other slot machine 30% of the time. There is a high (0.7) or a low probability (0.2) a choice at the second stage to be reinforced, which can change randomly with a small probability (1/7).

transition, revealing the knowledge about the task structure by the subjects. Therefore, they expressed both goal-directed and habitual action, but with a bias of staying on the same action independently from previous trial reward and transition type. As shown in Figure 2.17, both the flat and the hierarchical model captures the subjects' behavior.

In a second analysis, the authors investigated the interaction of goal-directed actions an habit sequences in stage 2. Their hypothesis is that a habitual response is implied by the repetition of a first stage action which was rewarded in the previous trial, and it is expected to choose the same action also in the second stage, even when different slot machine is accessed. On the contrary, if another first action is chosen, it means that it is independent from the previous sequence of action, and so the repetition of the second stage action with a different slot machine in the next trial is not expected. Figure 2.18 shows the probability of selecting the same second stage action as a function of if a same first stage was chosen after a rewarded trial. The data pinpointed that subjects had the drift to stay on the second stage action when habitual actions

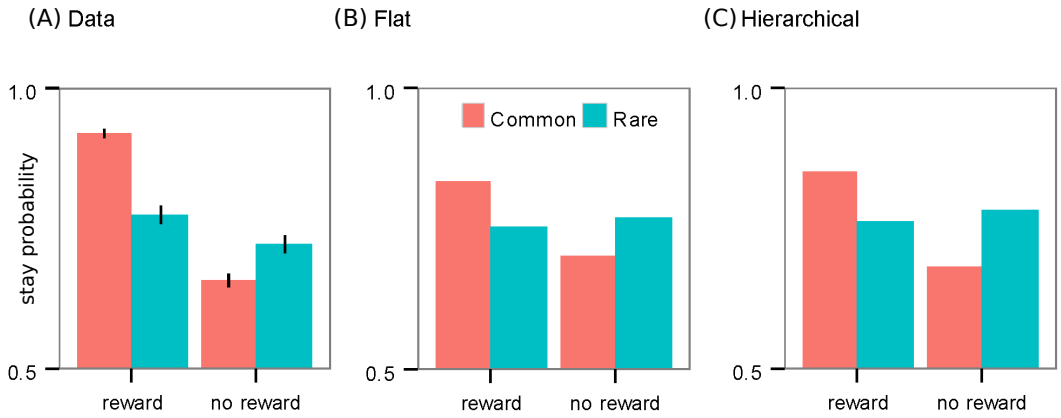


Figure 2.17: As in Dezfouli and Balleine [2013]. (A) Data from the experiment: The probability of staying on the first stage action was higher when the previous trial was rewarded, and even higher when it contained a common transition. Thus, the subjects expressed both habitual (1^{st} case) and goal-directed (2^{nd} case) action control. (B) & (C) show the responses of the flat and hierarchical architecture, respectively, in the same task. Both architectures are able to model the pattern of the data from the subjects.

were executed. The hierarchical model was able to capture the interaction of the first stage action and the reward, whereas the flat structure wasn't.

Figure 2.18: As in Dezfouli and Balleine [2013]. The stay probability on the second stage action when the slot machine differs from the one in the previous trial. (A) The probability is higher to stay on the same second stage action when the subjects stay on the same first stage action after being rewarded in the previous trial. (B) The flat architecture expresses different behavior than the observed data from the subjects. (C) The hierarchical model captures the observed pattern in actual stay probabilities.

Also, the reaction times were faster when a subject executed a habitual

response contrary to goal-directed. When a previously rewarded first stage action was repeated, the reaction time was increased when the second stage action was not compared to the repeated ones (Figure 2.19). The same effect was observed after no rewarded trials, excluding the fact that the increase was due to the cost of switching to another second stage action. The results of the hierarchical model implies that the second stage is inversely related to action sequence value.

Overall these results indicate that when the reward of the previous trial is followed by the same first stage action and decreased reaction time, then most probably the subject performs an action sequence. This results to the expectation that a second stage action will be repeated regardless of the accessed slot machine.

Figure 2.19: **As in Dezfouli and Balleine [2013]** The probability of staying on the second stage action when the same (A) or different (B) first stage action is taken, as a function of whether the previous trial is rewarded, and whether the second stage state is the same or different from the previous trial.

The hierarchical model proposed by Dezfouli and Balleine [2013] assumes that during the execution of an action sequence, its performance reveals insensitivity to the received feedback. That contrasts previous work of hierarchical RL theory, which considers that the state of the environment induces the action selection. Finally, the authors propose that action sequences, similarly to single actions, are prompted by goal-directed behavior.

2.3.3 Ashby *et al.* [2007]

Ashby *et al.* [2007] supported the idea that the goal-directed and habit system compete for expression, as in Daw *et al.* [2005], with the difference that this time the fastest system is expressed (and not the one with the strongest salience). Another difference is that in this model the habit system, except the stimulus, receives also input from the goal-directed system(Figure 2.20). The authors hypothesized that at the begin of training the goal-directed system is faster, but progressively it also trains the habit system. The latter one, once it has learned the correct action for a specific stimulus, becomes faster, and so leads the decision.

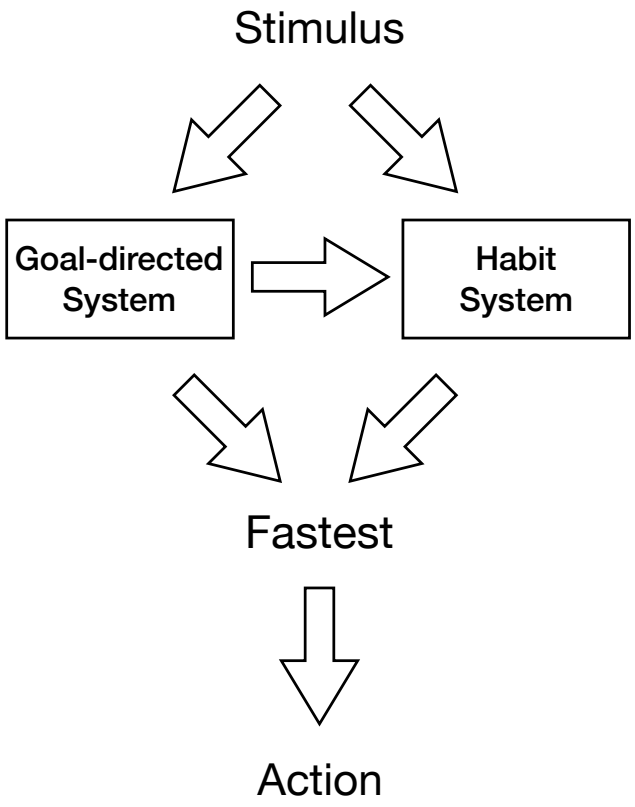


Figure 2.20: Illustration of the proposal by Ashby *et al.* [2007] about how a goal-directed action or a habit is chosen to be executed.

Ashby *et al.* [2007] designed a model called SPEED (Subcortical Pathways Enable Expertise Development) to study how automatic perceptual categorization is acquired in procedural learning tasks. The model comprises of two neural pathways (thus forming two systems) responsible for action selection, which both originate from sensory association cortex and end up to premotor area: a faster cortical path, and a slower subcortical path through basal

ganglia. Both systems incorporate learning, but based on different learning rules. The first follows the 2-factor learning rule (Hebbian learning [HL]), and the second the 3-factor (reinforcement learning [RL]). Their main hypothesis is that before learning, the BG learn fast through RL which ensues that the subcortical system dominates the decision. However, after cortical learning has been established, a shift of control to the cortical system takes place; a characteristic of automaticity. This is compatible with our own hypothesis.

The model includes the sensory association cortex (SC), premotor area (PrM), striatum (Str), globus pallidus internal segment (GPi) and thalamus (Th). It assumes that the SC influences the activity of neurons in premotor area and consequently the execution of a single action through two routes: indirectly via the direct pathway of basal ganglia, and by direct projections in cortical level (Figure 2.21). The role of the indirect route is to facilitate the

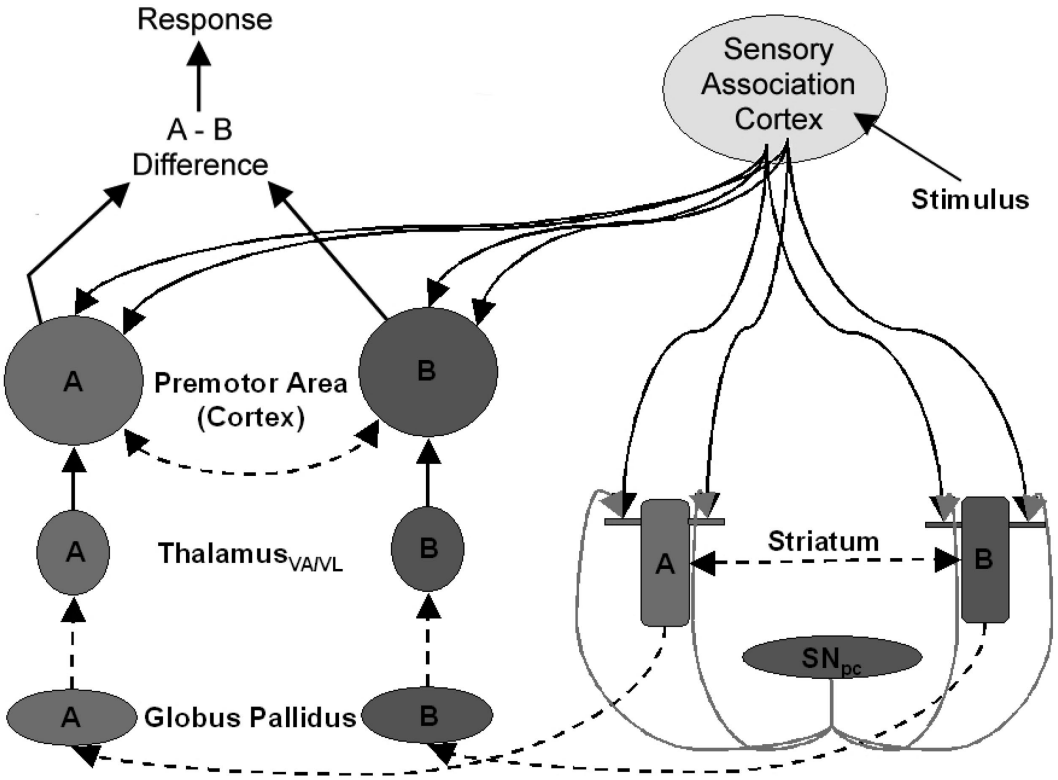


Figure 2.21: As in Ashby *et al.* [2007]. Schematic illustration of SPEED in the case of two contrasting categories, A and B. Excitatory projections are denoted by solid black lines, inhibitory by dashed lines, and dopaminergic by solid gray lines.

development of more permanent cortico-cortical connectivity. Whereas the strong cortico-cortical projections result to faster decisions, as a consequence of the involvement of only one synapse, and the avoidance of the long subcor-

tical path consisting of at least four synapses.

Ashby *et al.* [2007] tested SPEED in different protocols inspired by experiments on monkeys and rats. One of them was equivalent to a series of single-unit recording studies by Romo, Merchant, and their colleagues [Merchant *et al.*, 1997; Romo *et al.*, 1995, 1997]. In these experiments, monkeys learned to push one button in response to five low-speed vibrations, and another one for five high-speed. To simulate this experiment in SPEED, the SC contained one-dimensional array of 100 sensory cortical cells, and because of the existence of two contracting categories all the other structures contained only two units. All sensory cells projected on both striatal units. Each stimulus maximized the activation of one sensory cell and activated less the nearby cells. A response was considered to be initiated when the activation of the two premotor units exceeded a threshold. Finally, the cortico-cortical and cortico-striatal connections were altered between trials.

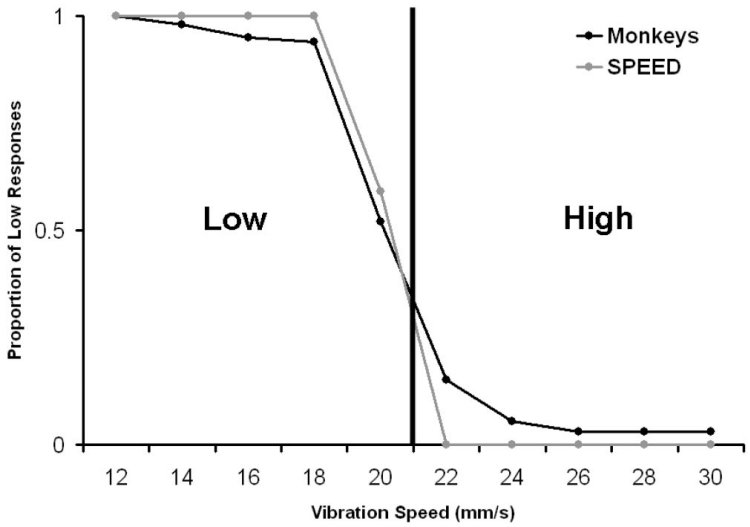


Figure 2.22: As in Ashby *et al.* [2007]. Proportion of "low" responses given to each of 10 stimuli by monkeys (black line) [Merchant *et al.*, 1997, p. 1151] and by SPEED (gray line).

Figure 2.22 demonstrates that both the monkeys and the SPEED learned the two categories by showing the proportion of low responses for each of the 10 stimuli given by them. The boundary of the category is indicated by the solid vertical line. Figure 2.23a illustrates the response of low and high speed striatal and premotor cells of monkeys in both types of vibration. Note that each type of cells in both structures becomes active only when the corresponding vibration is given as an input. SPEED mimics these results as shown in Figure 2.23b.

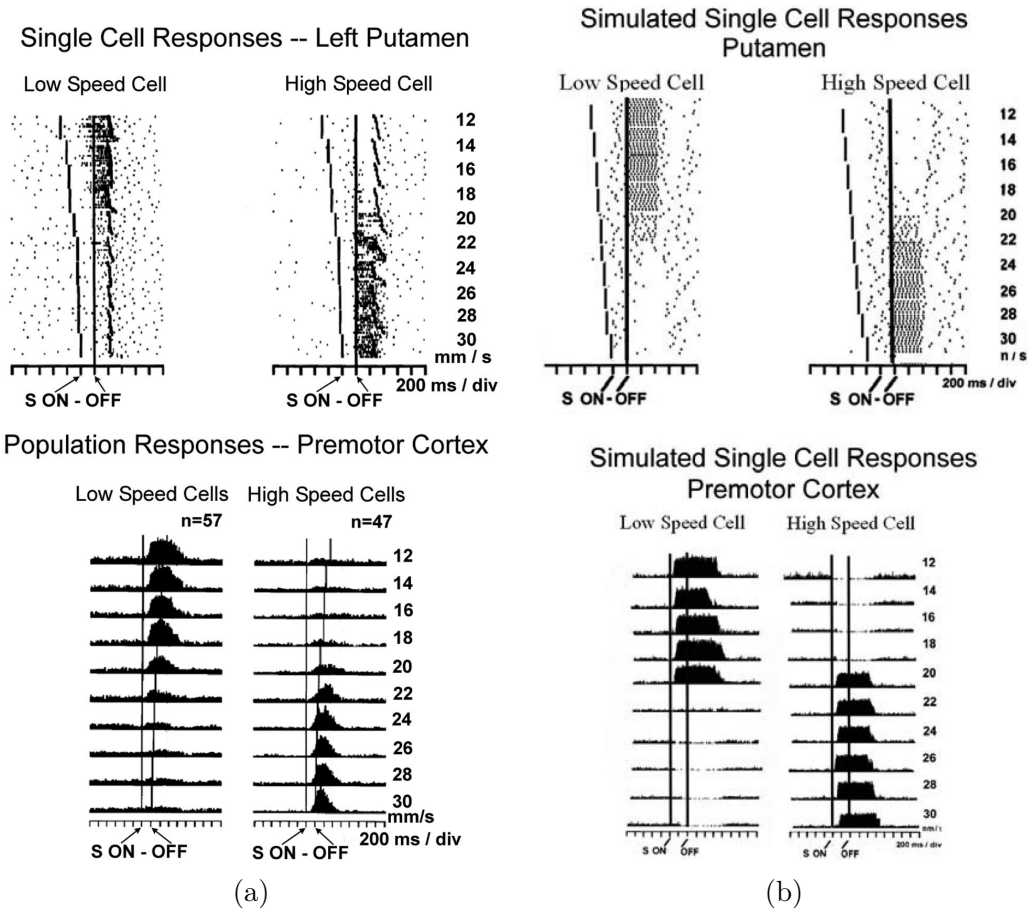
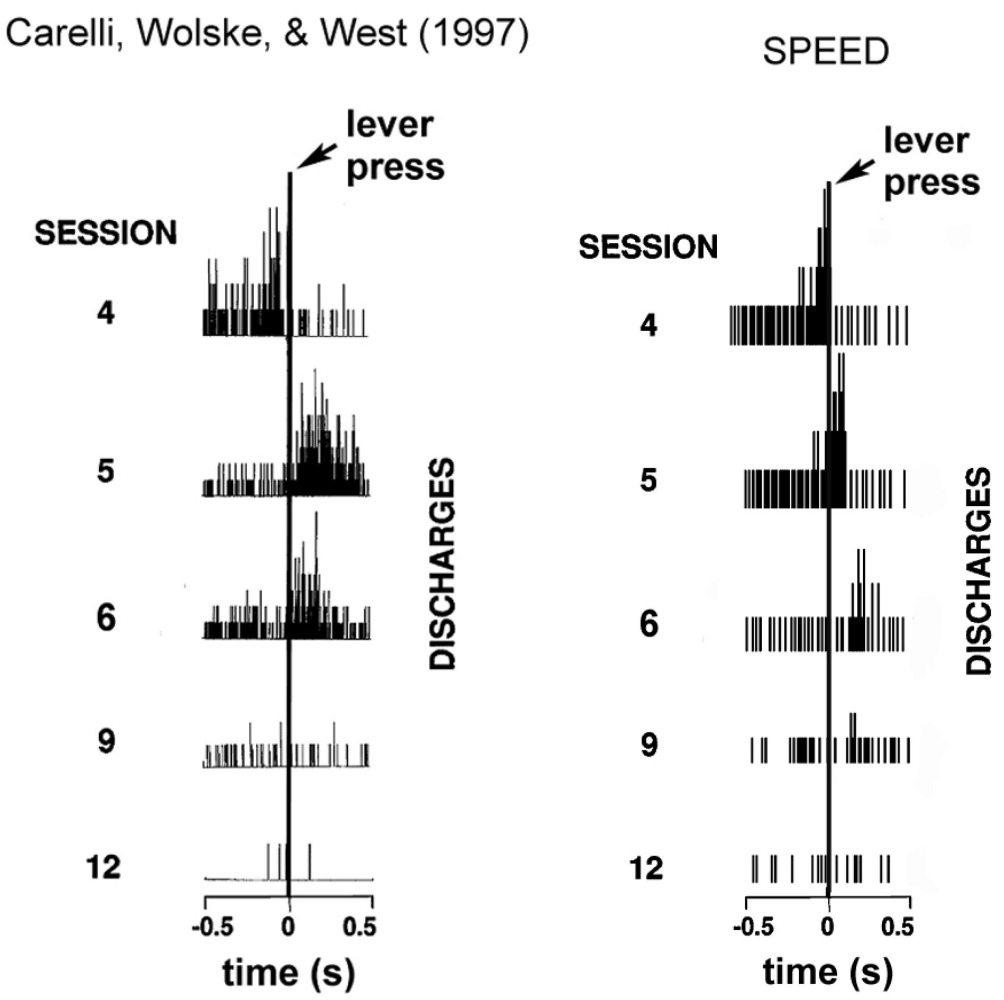


Figure 2.23: As in Ashby *et al.* [2007]. (a) Single-cell responses from left putamen [Merchant *et al.*, 1997] and population responses from premotor cortex [Romo *et al.*, 1997] to each of 10 stimuli in a tactile category-learning experiment. (b) Single-cell responses from the striatum and premotor cortex of SPEED in the tactile category-learning experiment as on the monkeys.

In another experiment Carelli *et al.* [1997] trained rats to press a lever as a response to a tone, and they recorded striatal single units. As shown in Figure 2.24a at session 4 the striatal units burst before the lever is pressed. Notice that at session 5 and 6, the same units still burst, but this time after the response. However, in later sessions, no activity is elicited by the tone or the lever press. The authors presumed that the last behavior comes as a result of the establishment of automaticity. In order to test SPEED on the equivalent protocol, they modified the architecture of SPEED. Only one sensory cortical unit exists, which is activated by the presence of the stimulus. Similarly, all brain regions include only one unit, because of the unique possibility of response. Figure 2.24b shows that SPEED expresses analogous behavior with the rats.

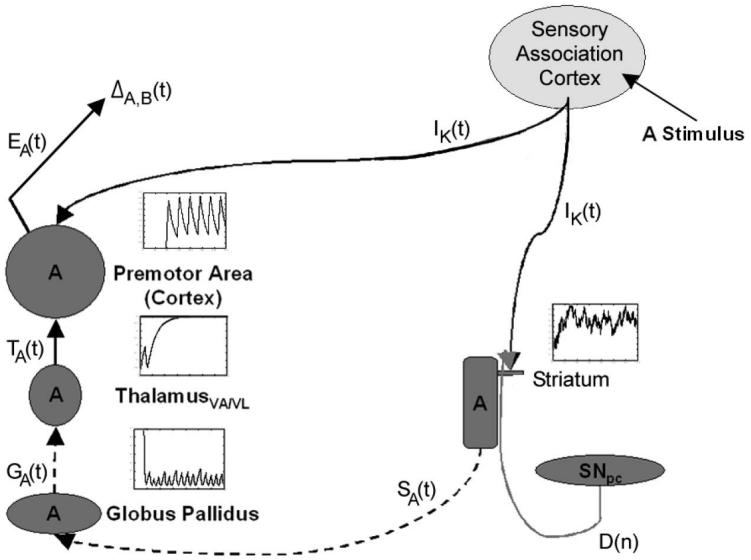


(a) (b)

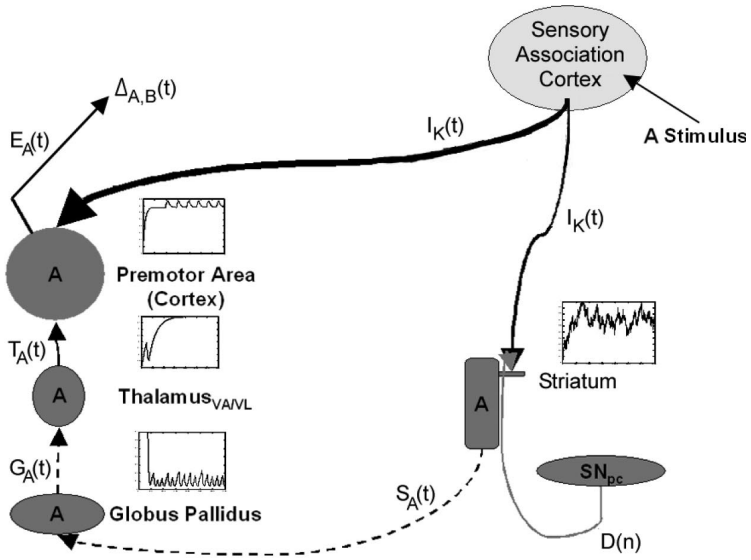
Figure 2.24: As in Ashby *et al.* [2007] (a)Striatal single-cell responses of a rat in the instrumental learning task of [Carelli *et al.*, 1997] (b) Striatal responses from SPEED

Further, Figure 2.25 presents the architecture of SPEED for this experiment, as well as the activity of all brain regions units before and after learning. At the beginning of training, striatum chooses first and then leads the premotor area to perform the appropriate action (Figure 2.25a). Contrary, premotor decision is much faster after training (Figure 2.25b). Furthermore, both the monkeys and SPEED gradually decrease their reaction time over training.

From these data derives that: (a) differences in reaction time between the early stages of training (slow) and the late stages (fast) exist, (b) information-integration category learning is initially mediated within the striatum, and (c) automaticity is acquired through repetition in cortico-cortical connectivity.



(a)



(b)

Figure 2.25: As in Ashby *et al.* [2007]. Schematic illustration of SPEED when a stimulus from Category A is presented. Illustration of the simulated solutions of the differential equations of each brain area in a trial (a) early and (b) later in learning.

2.3.4 Baldassarre *et al.* [2013]

Baldassarre *et al.* [2013] developed a system-level bio-constrained computational model to demonstrate how actions that have been previously learned by an 'intrinsic motivations' (IM) system can be used to achieve goals that emerge from an 'extrinsic motivations' (EM) system. The model explores these two systems through three segregated cortico-basal-thalamic loops (similar to Guthrie *et al.* [2013]): an arm loop, which selects the arm actions, an oculomotor loop that selects the eye gaze, and finally a goal loop, which selects the goals to pursue (Figure 2.26). Also, they implemented the intrinsic basal connectivity of the direct and hyperdirect pathways (as in Leblois *et al.* [2006]) combined with the off-center on-surround network for selection through diffused STN and focused striatal projections to GPi as introduced in Gurney *et al.* [2001a,b].

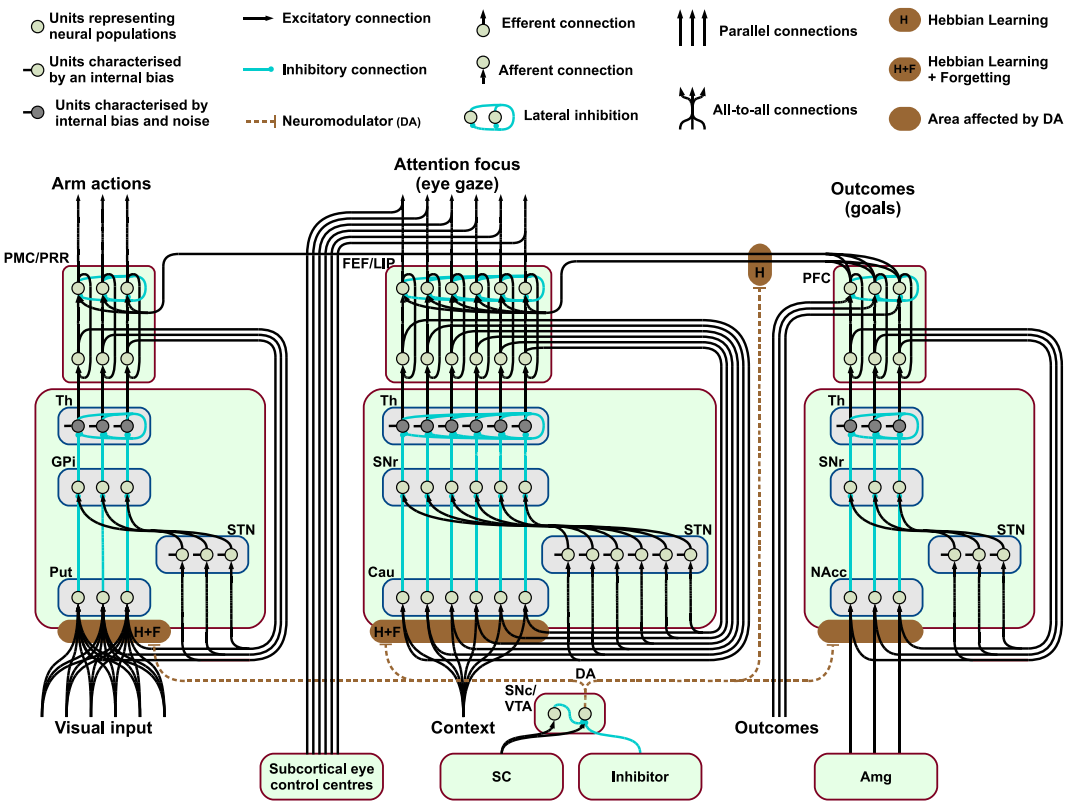


Figure 2.26: As in Baldassarre *et al.* [2013]. Detailed architecture of the model. The small circles indicate neural units of a populations, and the boxes are rate-coded neural populations. The abstractly implemented components are represented with boxes that include text.

The experiment includes two phases: a training and a testing. In the training phase, the system is let to randomly choose one eye gaze and one arm action. The set-up comprises of six possible gazes, one for each of the three

buttons and the three boxes available, and three arm actions, one 'reach to and press the looked-at object' action and two 'dummy' that are not useful for the task. When the eye gaze is on certain button and the arm reaches and presses it, then a box is opened. The goal of the task is to learn which button opens which box. This learning occurs at cortical and striatal level. An example of a trial during the training phase is shown in Figure 2.27. After the system discovers a combination of an arm action and a eye gaze that opens a box, it repeats this combination until the two actions (arm and eye gaze) are learned, and only then the focus is moved on others that are still unexplored. When

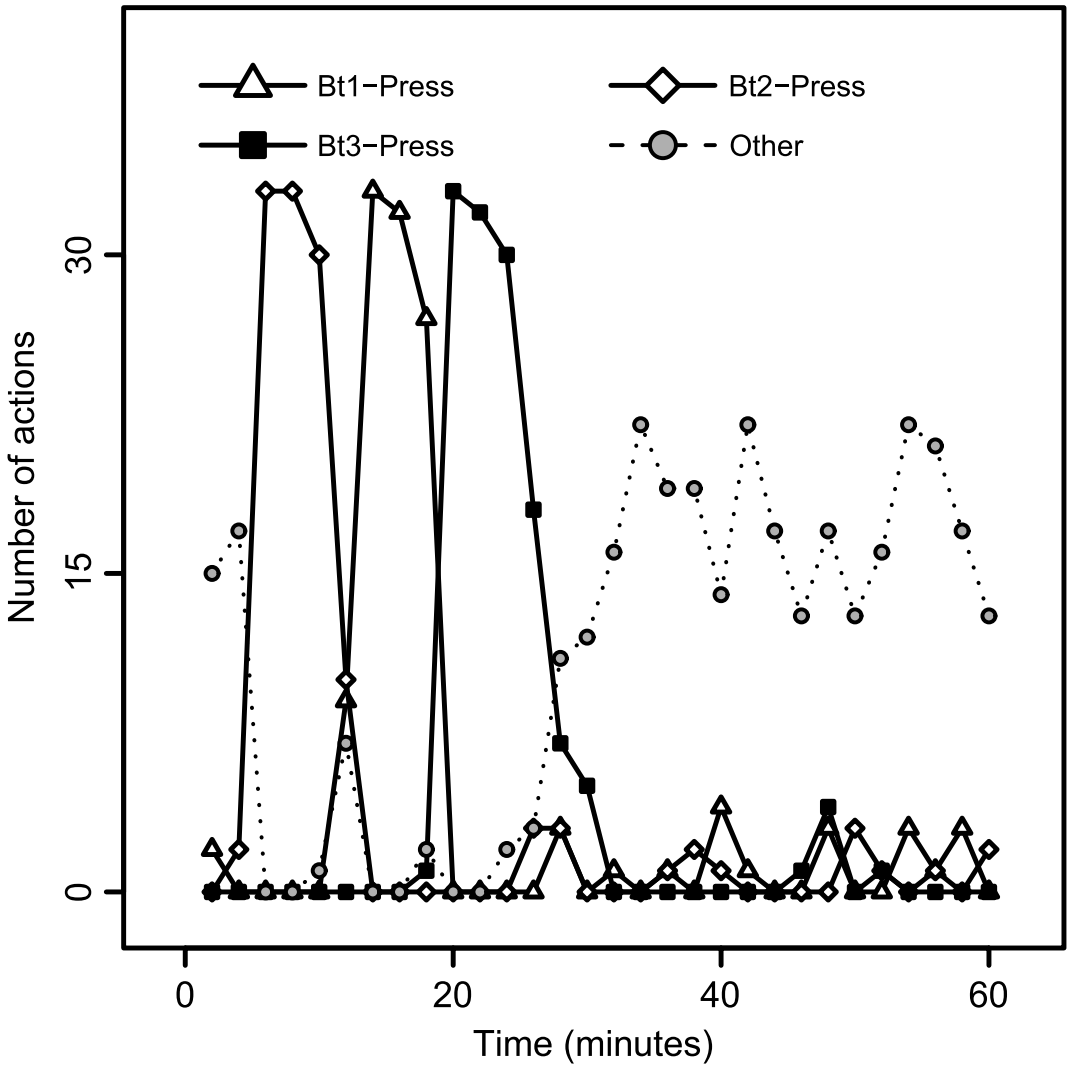


Figure 2.27: As in Baldassarre *et al.* [2013]. Behavioral example of the intact model during the training phase. The y-axis shows the number of executions of the action that opens finally one of the three boxes (Bt1-Press, Bt2-Press, Bt3-Press), and also the number of executions of all the other available actions considered together (Other).

a box is opened, phasic DA is released and reaches the striatal connections of the arm and oculomotor loops, resulting in strengthening the connections between the seen object and the performed action, and between the view of the button with the eye saccade, respectively. This plasticity is based on a 3-factor learning rule [Reynolds and Wickens, 2002], which increases the weight of the connection between two neurons when DA is released and the pre and post synaptic neurons are more active than their thresholds. However, an inhibitor component also exists, in order to suppress DA expression so a decay of the weight to be ensued. They added the inhibitor to ensure that after the system explores repeatedly, and consequently learns the desired association, it will be able to unlearn it in striatal level providing the system the ability to explore new contingencies (Figure 2.28 a, b). The cortical learning follows a Hebbian learning rule and is applied on the cortical projections from goal loop to the arm and oculomotor loops. In this way, an association between the outcome of the performed action and the action itself is formed. The authors implented a different from classical Hebbian learning rule, which involves dopamine. Here, the learning occurs also as a result of DA release, but its absence does not evoke unlearning of the associations as in striatal case (Figure 2.28 c, d). After the acquisition of the associations during the intrinsically motivated learning phase, the ability of the model to engaged them for reaching a rewarded goal is tested. During the testing phase, a goal (open a specific box) is activated externally in the goal loop, and if the system has learned properly, then the equivalent actions will be executed by the other two loops.

They authors explored the behaviors of the intact model, as well as of four lesioned models with inactive: (a) the putamen, (b) the caudate, (c) the putamen & the caudated, and finally (d) the inhibitor. The results of the five models are shown in Figure 2.29. During the learning phase, a test procedure is performed in which the three goals are sequentially activated for 2 *minutes* each. This can give an evaluation of the model performances in different timing of the learning. At the beginning of the learning, the intact model explores its options and acquires the desired ones, resulting to efficient use of them after the whole training. When putamen or/and caudate are lesioned, the arm and oculomotor loops are still able to choose action, but only in a random manner. This results in slow learning. The lesion of putamen allows the system to explore more (than in the intact model) the 'dummy' actions for a specific gaze, and consequently to not learn efficiently the good action. On the other hand, the caudate lesion lets the system to interact with all buttons to a certain extent, even with no attentional focus, resulting to better performances from putamen lesion, but also equivalent learning for all buttons. As expected the lesion of both parts of striatum slows further the learning as a consequence of performing wrong actions on buttons and interacting with all the boxes. Finally, the inhibitor lesion eliminates the ability of the system to explore other buttons.

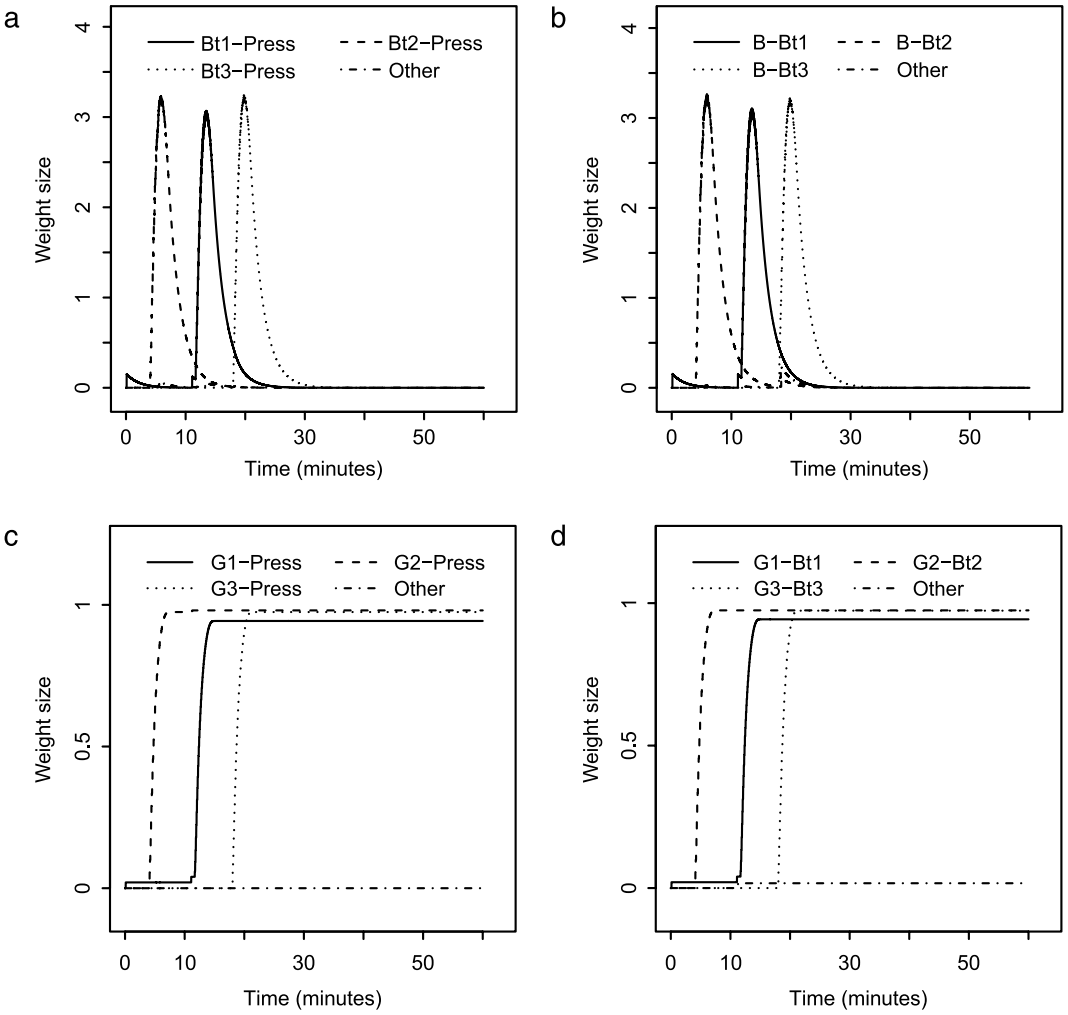


Figure 2.28: As in Baldassarre *et al.* [2013]. Example of evolution of the trained connection weights of the model during the learning phase. (a) Cortico-striatal (Put) weights, (b) Cortico-striatal (Put) weights (Cau), (c) Cortico-cortical weights from the PFC to the PMC/PRR , and (d) Cortico-cortical connection weights from the PFC to the FEF/LIP.

In summary, the proposed model by Baldassarre *et al.* [2013] demonstrates how actions are acquired by intrinsic motivation (*i.e.* randomly perform actions and observe their outcomes) can be later used to achieve goals emanated by extrinsic motivation; *e.g.* reach a reward, through a repetition bias developed in the learning phase.

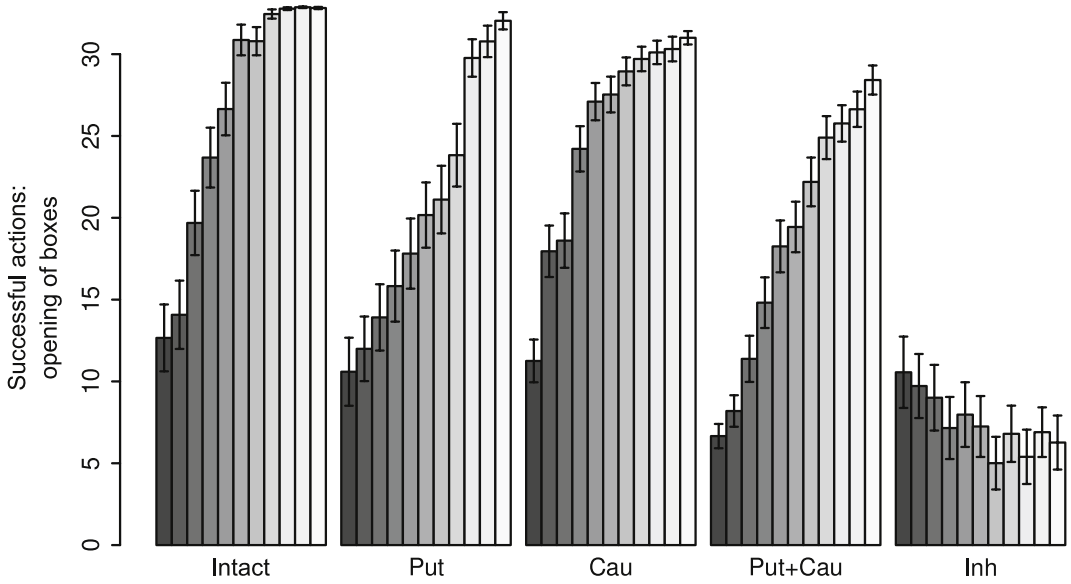


Figure 2.29: As in Baldassarre *et al.* [2013]. Performance of the non-lesioned model ('Intact') and four lesioned versions of the putamen (Put), the caudate (Cau), both the putamen and the caudate (Put-Cau), and the inhibitor (Inh). For more details refer to the text, or the original article.

"Essentially, all models are wrong,
but some are useful."
— George E. P. Box

Chapter 3

A computational model

Contents

In this chapter, I introduce a dynamical model of the BG – cortical network that we developed, which is able to solve a two-armed bandit task. The model is based on previously published models by Leblois *et al.* [2006] and Guthrie *et al.* [2013]. The first model introduces an action selection mechanism originated by the competition between a positive feedback, provided by the direct pathway, and a negative feedback emanated from the hyperdirect pathway. Guthrie *et al.* [2013] have further extended the model to explore the parallel organization of circuits between cortex and basal ganglia [Alexander *et al.*, 1986; Albin *et al.*, 1989; Parent and Hazrati, 1995a] by using segregated modules: one for the selection between two presented cues and the other for

the selection between two possible movement directions corresponding to the positions of the cues. However, to solve the task properly, it is necessary for the model to choose the cue shape and select the corresponding movement. For this reason a third module exists to provide solution to the binding problem, as well to the need for cross-talking between these two modules. Additionally, learning occurs between cortex and striatum using a simple reinforcement learning rule where the values of the different cues are updated a each decision has occurred.

We further refined the model such as to have a competition mechanism within each cortical group. Using short-range excitation and long-range inhibition, this competition ensures that a unique cognitive and motor decision eventually emerges. Hebbian learning is also added at the cortical level, and is enforced after a move has been executed, independently from the actual reward.

Through the sections below, I recount the history of our model by summarizing the two previous models that it is based on [Leblois *et al.*, 2006; Guthrie *et al.*, 2013], and describing the modifications and additions that lead to the final architecture. Trying to implement the model by Guthrie *et al.* [2013], we confronted some difficulties such as erroneous and ambiguous information, which I describe and explain how we managed to overstep them.

3.1 First generation [Leblois *et al.*, 2006]

In Leblois *et al.* [2006], the authors proposed a dynamical model of the motor part of basal ganglia (BG), providing an explanation of how action selection is generated from BG. Contrary to Albin *et al.* [1989] and DeLong [1990] model of basal ganglia circuitry that relies on a segregation between the direct and the indirect pathways of BG, Leblois *et al.* [2006] assigned a fundamental role to the hyperdirect pathway. They showed that BG functions and dysfunctions arise from the competition of direct and hyperdirect pathways. The aim of this model was to investigate the physiology and pathophysiology of BG, however I focus only to the action selection mechanism, which we use in our model.

The model consists of five populations: cortex (CTX), striatum (STR), subthalamic nucleus (STN), internal segment of globus pallidus (GPi) and thalamus (Th) [Figure 3.1]. Based on anatomical and electrophysiological studies supporting that there is a topographic organization in the direct pathway, the model contains two parallel circuits that are hypothesized to control two distinct motor programs [Alexander *et al.*, 1986; Nakano, 2000]. However, an interaction between the two circuits exists at the level of STN to GPi connectivity, expressing divergence at this level [Parent and Hazrati, 1995b]. Each

population contains two neurons of rate model [Hopfield, 1984; Wilson and Cowan, 1972; Shriki *et al.*, 2003], which are engaged in distinct circuits.

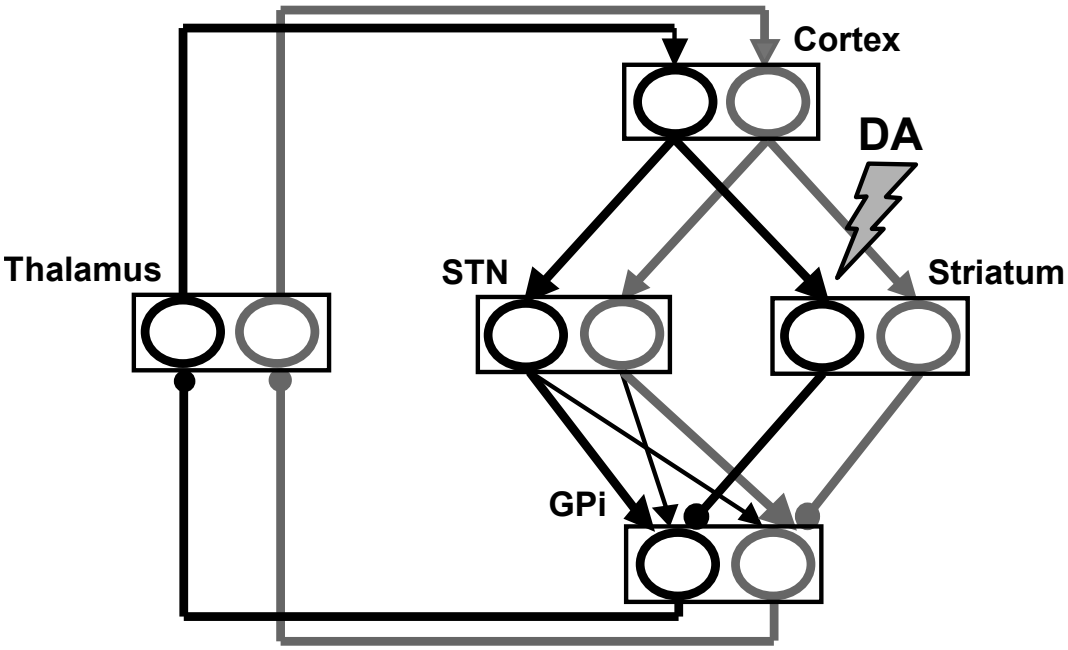


Figure 3.1: As in Leblois *et al.* [2006]. Architecture of the model. Each structure contains two populations that implement two circuits. The excitatory connections are indicated by arrows, and the inhibitory by dots. The network does not include the substantia nigra pars compacta.

Figure 3.1 displays the two circuits that contain two feedback loops: one global positive close feedback loop (direct loop; $CTX \rightarrow STR \rightarrow GPi \rightarrow Th \rightarrow CTX$), and one global negative close feedback loop (hyperdirect loop; $CTX \rightarrow STN \rightarrow GPi \rightarrow Th \rightarrow CTX$). The positive or negative characterization of the loops emanates from the effect of each loop on thalamus. For example in the direct loop, thalamus is disinhibited through one excitatory and two inhibitory set of projections. On the other hand, the hyperdirect loop inhibits thalamus through two excitatory and one inhibitory set of projections.

The relation between the products of synaptic strength of these two loops affects the behavior of the model ($G_+ = G_{CtxTh}G_{ThGPi}G_{GPiStr}G_{StrCtx}$, $G_- = G_{CtxTh}G_{ThGPi}G_{GPiSTN}G_{STNCtx}$, where G_{ba} denotes the strength of the interaction between two neurons in the same circuit), as shown in Figure 3.2. In the oscillations state, the external input that is sent to cortical neurons provoke oscillations in all populations, with the same phase in the two circuits (Figure 3.3a). By contrast, a small asymmetric disturbance leads to symmetric instability in the symmetry breaking state, which is originated by a strong feedback that is sent to thalamus from both loops, direct and hyperdirect. If

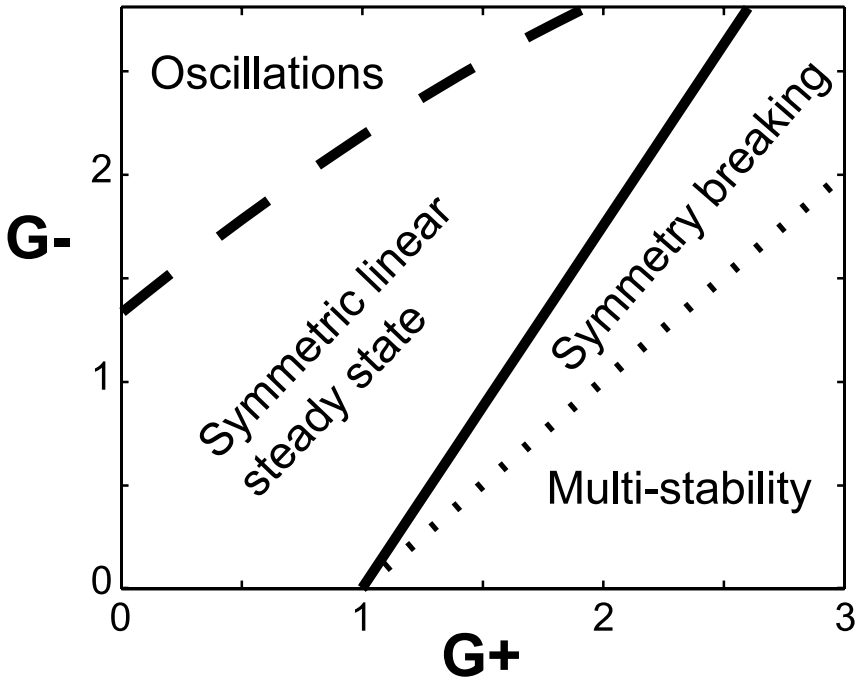


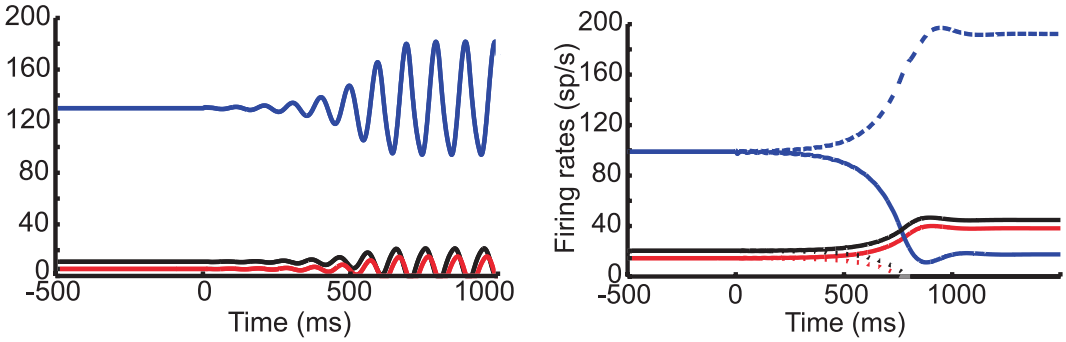
Figure 3.2: As in Leblois *et al.* [2006]. Various dynamical regimens of the reduced model as a function of G_+ and G_- are presented in the phase diagram.

one of the two cortical populations increases its activity, then the direct loop amplifies it. On the contrary, strong cortical input in STN leads to increase of activity in GPi for both populations. Although, this increase tends to reduce the activity of the weaker in cortical level, resulting in less inhibition from GPi to thalamus to the more active population. So in the end, only one cortical population stays active. Figure 3.3 demonstrates the evolution of activity of the two populations in GPi, thalamus and cortex, before and after external input is sent to cortex in the symmetry breaking state.

Other models than Leblois *et al.* [2006], assumed that a selection is made at cortical level, and this information is processed by BG in a feedfoward manner. However, in this model BG have an active role in action selection by inducing cortical activity after symmetry breaking has been achieved.

3.2 Second generation [Guthrie *et al.*, 2013]

Guthrie *et al.* [2013] extended the model by Leblois *et al.* [2006] in order to explore the parallel organization of circuits in BG [Alexander *et al.*, 1991] through a two-armed bandit task. Generally, in this type of tasks two cues are presented to the subjects. Each of the cues is associated with a hidden reward probability, and the subjects are able to learn these probabilities by exploring



(a) (b)

Figure 3.3: As in Leblois *et al.* [2006]. The network from $t = -500ms$ remains stable until $t = 0$, when a brief external current is sent in cortical population of circuit 1 resulting in (a) oscillatory instability for $G_+ = 0.18$ and $G_- = 2.85$, and (b) symmetry breaking for $G_+ = 2.47$ and $G_- = 2.85$. The activities in circuit 1 (respectively circuit 2) are donated be solid (resp. dashed) lines. The colors of the lines indicate the different structures: red for cortex, black for thalamus, blue for GPi.

the outcome of the cues. In their paradigm, Guthrie *et al.* [2013] used the task as described in Pasquereau *et al.* [2007], where two shapes are presented in two distinct positions on a screen in front of the monkeys.

To be compatible with the nature of the task, the model implemented two segregated modules: one devoted to shape selection (cognitive), and one to position selection (motor). However, to solve the task successfully the model should be able to exchange information between the modules, so that when a shape is chosen, the corresponding position will also be chosen. Therefore, a third open module (associative) exists. This module receives the info about which shape is placed at which position, solving the binding problem, and further integrates the needed mechanism for transmission of information between the segregated loops. Furthermore, noise has been added to the the system to produce symmetry breaking, when two identical inputs are given. In addition, the model incorporates cortico-striatal learning inside the cognitive module by a simulated reward signal, showing that after training, the action selection is generated based on the learned cues and not the noise as before learning.

3.2.1 Architecture

The general architectural diagram of the model is illustrated in Figure 3.4. Following the architecture of Leblois *et al.* [2006], the two modules, cognitive and motor, comprise of two segregated BG-cortical closed loops: the direct (Ctx-Str-GPi-Th-Ctx) and hyperdirect (Ctx-STN-GPi-Th-Ctx respectively).

The authors consider that the two modules are parallel, with inputs from

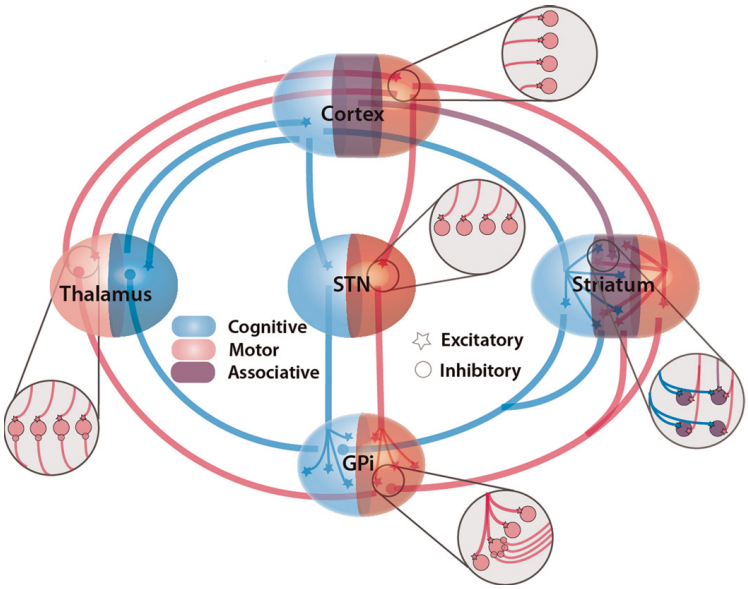


Figure 3.4: As in Guthrie *et al.* [2013]. Architecture of the model of the basal ganglia. The two levels of decision making are represented by the two cortico-basal ganglia: the cognitive loop in blue, and a motor loop in red.

distinct areas of cortex. The third module is an open loop, which contains only cortical and striatal populations. One of its roles is to provide cross-talking between the parallel modules. This is achieved by the projection from cortical to associative striatal populations, and in turn they project to GPi populations of both of the modules. Another role of associative module is to solve the binding problem generated by the distinct cortical input that the two parallel modules receive. For instance, cognitive cortex receives information of which pair of shapes is presented and motor cortex which positions contain a shape. However, there is no information about which position each shape occupies. This obstacle is overstepped by providing this information through an external input to the associative cortex.

Each structure consists of two segregated groups of ensembles (one ensemble represents a population of neurons), the cognitive and the motor, except cortex and striatum, which contain also an associative group. Individual groups are components of the corresponding modules. Also, each group comprises of four ensembles, which within cortical area represent a possible choice. In associative groups, sixteen ensembles exist in order all the possible combinations of cues in corresponding positions to be included. The connectivity among the populations are shown in Figure 3.5. All the projections in cognitive and motor modules and the associative cortico-striatal ones are somatotopic, except of divergence from STN to GPi, where one STN ensemble projects to all ensembles of GPi in the same module. Finally, divergence occurs from cortical motor and cognitive groups to the associative striatum, followed by

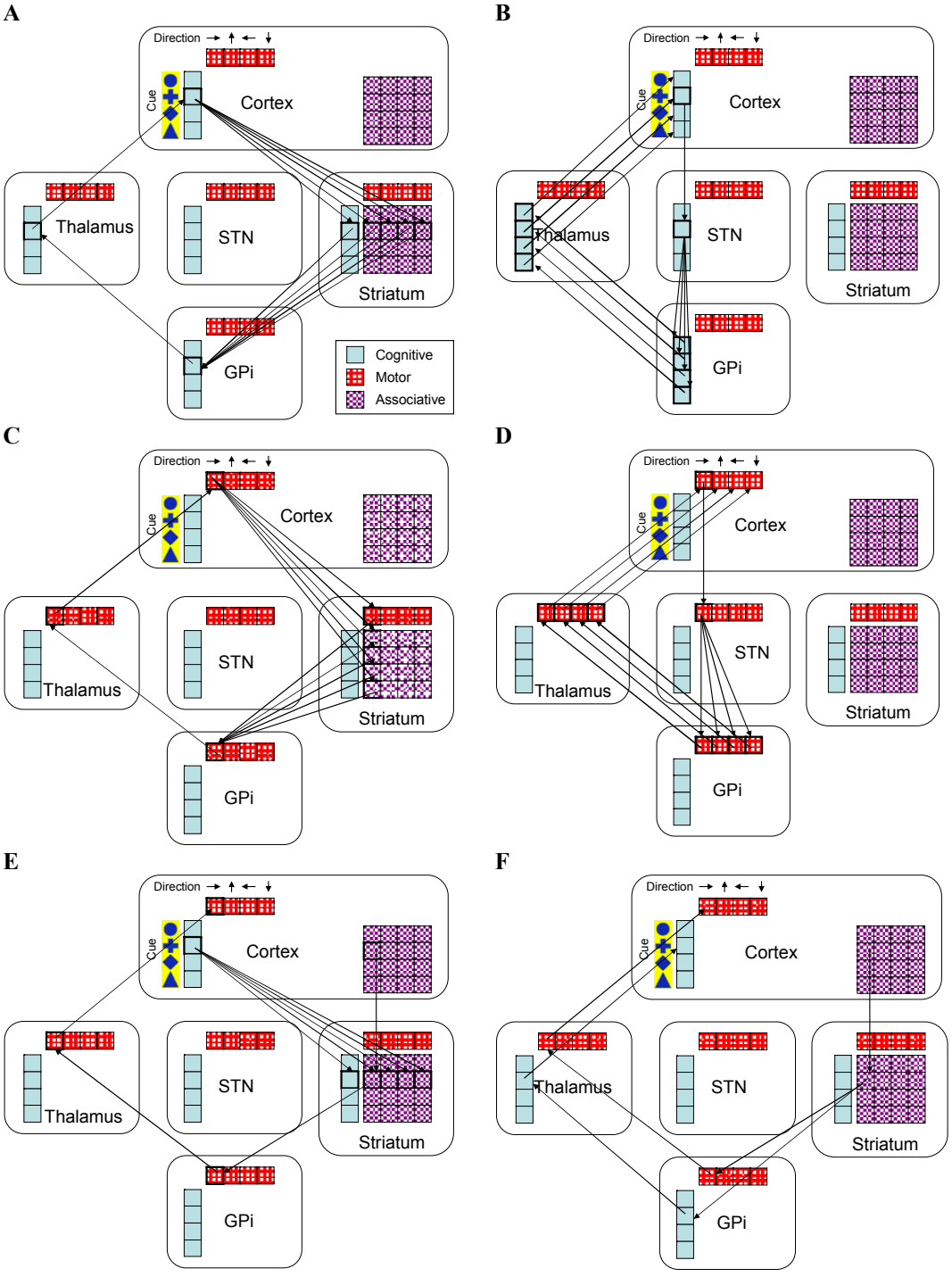


Figure 3.5: As in Guthrie *et al.* [2013]. Connectivity of the model. For more details please refer to the original paper.

reconvergence from associative STR to motor and cognitive GPi.

3.2.2 Neuron model

Neural activity in all ensembles is described by the same equation, except a variation in parameters among populations (Table A.1). Each ensemble is hypothesized to simulate the mean activity of a population of neurons, which is the reason for using a leaky integrator as neuron model [Leblois *et al.*, 2006; Hopfield, 1984]:

$$\tau \frac{dm}{dt} = -m + I_S + I_{ext} - T \tag{3.1}$$

where τ is a decay time constant of synaptic input, m is the output of the ensemble, I_s is the synaptic input to the neuron, I_{ext} is the external input representing the visual salience of the cue and presented only to the cortical structues, and T is the threshold of the ensemble. In Leblois *et al.* [2006], the authors applied a bias signal to one of the ensembles to break symmetry and thus action selection. However, Guthrie *et al.* [2013] added Gaussian noise to the activity of each ensemble at each time step in order to generate symmetry breaking.

The total input to a neuron B is fully described from:

$$I_S^B = \sum_A G_B^A \times m_A \tag{3.2}$$

where A is the presynaptic neuron, B the postsynaptic, G_B^A the gain of the synaptic connectivity from A to B respectively.

Many studies showed that striatal neurons follow a sigmoidal function, because of their particularity of being silent without coordinated input [Nisenbaum and Wilson, 1995; Sandstorm and Rebec, 2003; Wilson and Groves, 1981]. This is implemented by the Boltzmann equation:

$$m_{out} = V_{min} + (V_{max} - V_{min})/(1 + e^{((V_h - m_{in})/V_C)}) \tag{3.3}$$

where m_{in} is the input to the transfer function (the activation level of the cortical inputs in this case) and V_{out} is the output, V_{min} is the minimum activation, V_{max} the maximum activation, V_h the half-activation, and V_c the slope.

3.2.3 Learning

Striatal learning is implemented among cognitive cortico-striatal synapses. The cognitive module was chosen over the motor, because in this particular task the received reward is a result of the chosen cue and not the chosen position. So, the desired association of the reward with the cue should be learned in the cognitive module. It has been show, that learning at striatal level follows reinforcement learning rules. That has been based on the role of

dopamine that is sent to striatal neurons, which is to provide an error prediction signal and strengths the synapses. For this reason, at the end of each trial, reinforcement learning rules are applied as follows:

$$\Delta w_{A \to B} = \alpha_\alpha \times PE \times m_B \tag{3.4}$$

where $\Delta w_{A \to B}$ is the change in the weight of the corticostriatal synapse from cortical population A to striatal population B, PE is the prediction error, the amount by which the actual reward delivered differs from the expected reward, m_B is the activation of the striatal ensemble, and α_α is the global actor learning rate. The generation of long-term potentiation (LTP) and long-term depression (LTD) in striatal neurons has been found to be asymmetric [Pawlak and Kerr, 2008]. Therefore, the actor-learning rate in the model is $\alpha_\alpha = 0.002$ for LTP and $\alpha_\alpha = 0.001$ for LTD.

The PE is calculated by using a simple critic-learning algorithm.

$$PE = R + V_i \tag{3.5}$$

where i is the number of the cue chosen, and V_i is the value of cue i. Then, the value of the chosen cue is updated by using the PE.

$$V_i \leftarrow V_i + PE \cdot \alpha_c \tag{3.6}$$

where α_c is the critic learning rate, set to 0.05.

The weights are bounded to absolute maximum 0.75 and absolute minimum 0.25.

3.2.4 Results

Figure 3.6 shows the evolution of cortical activity during a trial of a two-armed bandit task. The network is let to stabilize for $500ms$, when the external input is applied to cortex, representing the display of two cues (out of four) in two distinct positions (out of four). Each cue is associated with a unique reward probability ($c_1 = 1.00$, $c_2 = 0.66$, $c_3 = 0.33$, $c_4 = 0.00$). The two ensembles that don't receive input, are immediately inhibited, contrary to the other two which increase their activity and start to compete each other. A decision is considered to be made, when one motor cortical ensemble is $40sp/s$ more active than the others. When the network is naive, the motor and cognitive selection is random. Also, the motor decision can be preceded by the cognitive in some occasions, as shown in the inset of Figure 3.6, because their inputs have the same value.

In the beginning of a learning simulation, containing 120 trials, the network has a chance level probability (0.5) to choose the optimal cue (the cue with a higher probability of being rewarded). However, as Figure 3.7 shows, the network gradually learns the optimum choices among the different pairs, and

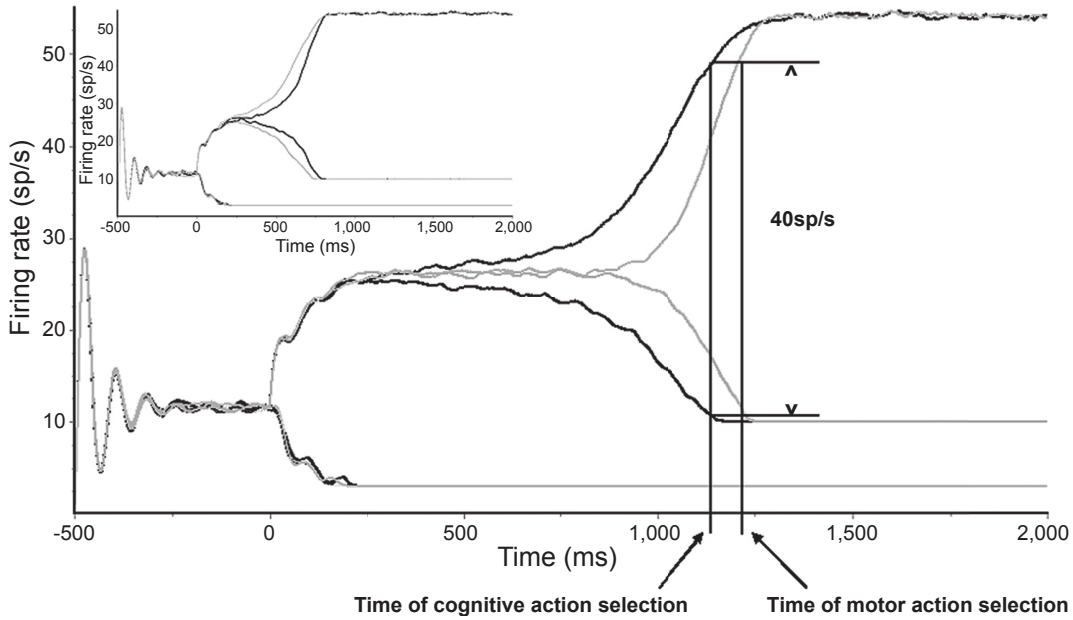


Figure 3.6: As in Guthrie *et al.* [2013]. Cortical activation of the motor (grey line) and cognitive (black line) populations during a single trial. *Inset:* an example trial where the motor action selection precedes and leads the cognitive action selection.

at the end of the simulation, there is an increase of probability for the optimal choice (0.95±0.01). The analysis of the evolution of the cortico-striatal weights (Figure 3.8) revealed that the weight associated with the best cue (highest reward probability) increases to 96% of the absolute maximum by the end of the simulation. On the other hand, the weight of the worst cue decreases only to 44%. These results reveal that the system learns by actions.

3.3 A long journey into reproducible science

[Topalidou *et al.*, 2015; Topalidou and Rougier, 2015]

In the long journey of understanding the brain, computational neuroscience is a powerful ally. The development of models, even simple ones, is an invaluable tool, providing the opportunity of exploring this or that structure and propose new hypothesis concerning the overall brain organization. However, the uniqueness of this tool is based on the possibility of the existing models to be produced, and further extended. In this manner, every extension of a model gets us nearer to an incremental computational knowledge of the brain. Unfortunately, when we tried to reproduce the model by Guthrie *et al.* [2013],

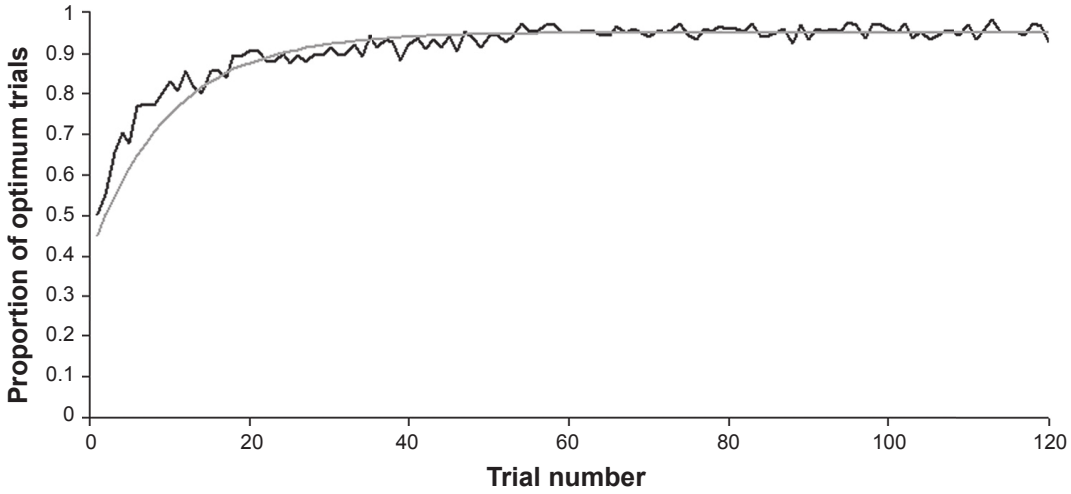


Figure 3.7: As in Guthrie *et al.* [2013]. Learning curve (black) and an approximation of this curve (light gray; $y = 0.5 + 0.5 \times 1 - Exp[-(t-1)/13.7]$) for the choice of the optimum cue.

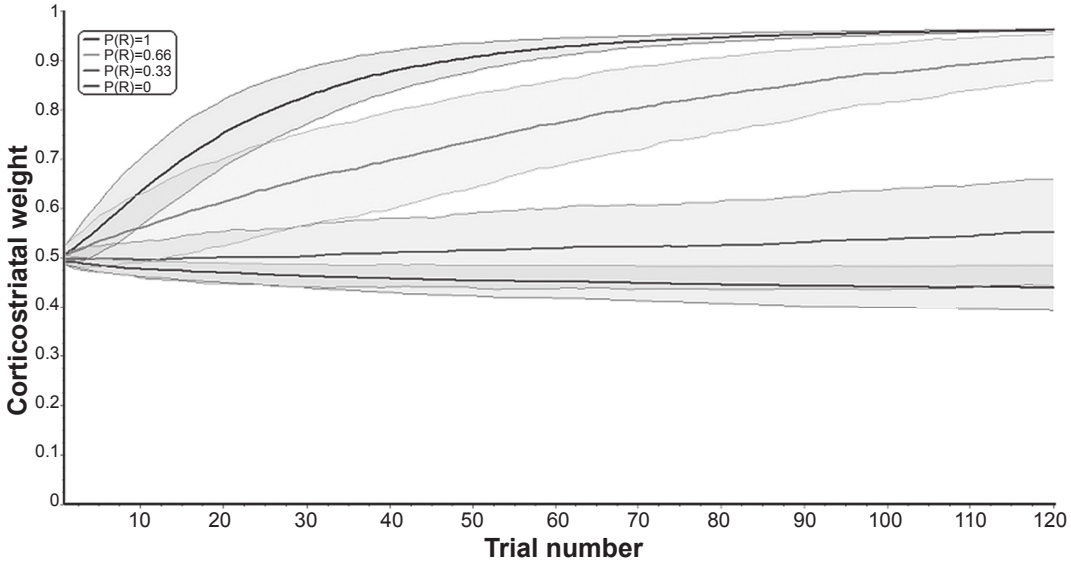


Figure 3.8: As in Guthrie *et al.* [2013]. Evolution of normalized cognitive cortico-striatal synaptic weights during a session for each of the four cues.

we confronted a common problem in the field, the inability of reproducing a model from the bibliography. The information provided by the article was not sufficient to allow the direct reproduction of the model, either the source code that the authors provided us (hundred of files and 6000 lines of Delphi [Object Pascal]). Overall, the model is described quite precisely, with a lot of important information, but still, some information are ambiguous or erroneous and some others are just missing.

After some minor corrections and modifications of the original description of the model, we were able at last to reproduce the original results, confirming the correctness of the original implementation of the model [Topalidou and Rougier, 2015]. We did not reproduce all analyses of the original article but the main results which are illustrated on figures 4 & 5 in the original article Guthrie *et al.* [2013] (3.6 & 3.7 in section 3.2).

We first reproduce the activity in the cortical populations during a single trial, prior to learning. Noise has a great influence on the overall dynamic and

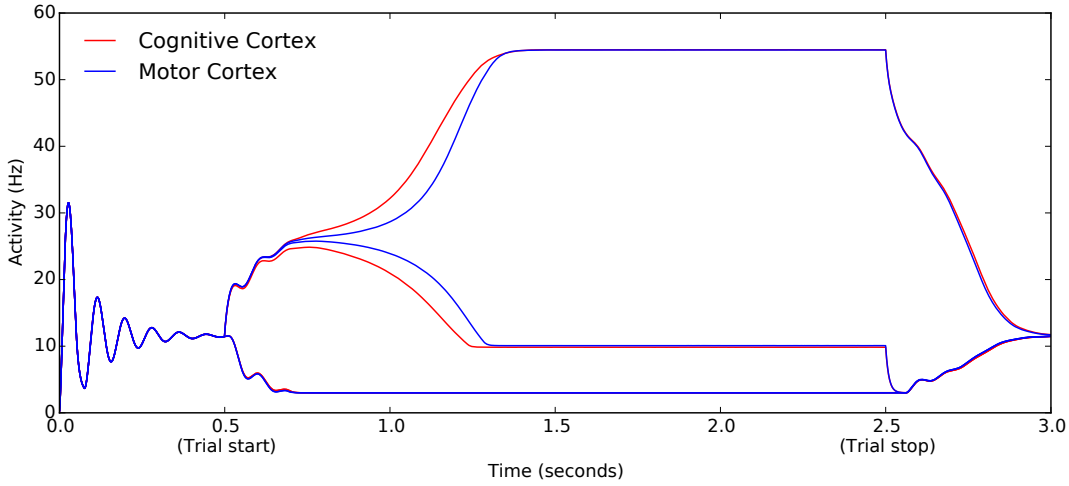


Figure 3.9: **Activity in the cortical population during a single trial of action selection.** This is the reproduction of figure 3.6 in Guthrie *et al.* [2013].

it is not possible to exactly reproduce figure 4 in the original article without precise information on the underlying random generator(seed). Consequently, we can only report a qualitatively equivalent figure where the most critical feature is the existence of bifurcation in cognitive and motor activities after stimulus onset. Since no learning has occurred yet, it is also possible to have the motor decision occurring before the cognitive decision. Figure 3.9 shows an example of the cortical dynamics with an oscillatory regime between time t=0 and time t=500ms that is a characteristic of the model. Finally, we tested the learning capacity of the model (Figure 3.10) by reproducing the same procedure as in the original article (250 experiments, 120 trials). We also established the tabular description of the model as proposed in Nordlie *et al.* [2009] [Table A.1], which allow anyone to rewrite the model using a different language, tools or software.

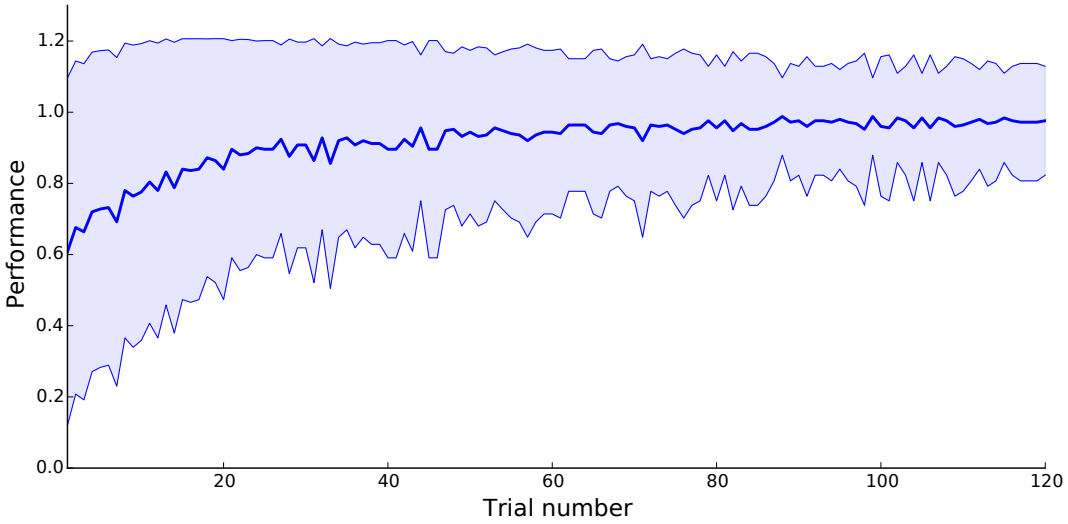


Figure 3.10: **Learning time course over 120 trials, averaged over 250 simulations.**The blue filled area indicates the standard deviation of the mean performance. This is the reproduction of figure 3.7.

3.4 Third generation [Topalidou *et al.*, 2016]

The distinct somatatopic, functional, cortical projections to explicit regions of striatum[Alexander *et al.*, 1986; Deniau *et al.*, 1996; Nakano, 2000] inspired first Alexander *et al.* [1986] to propose a model of five parallel, segregate loops that process information, which was revised later by several researchers suggesting the existence of only three functionally defined loops (sensorimotor, associative, limbic) [Parent and Hazrati, 1995a; Haber, 2003; Seger and Spiering, 2011; Hélie *et al.*, 2015]. Sensorimotor loop has been associated with habitual behavior, contrary to the associative which is believed to participate in goal-directed actions. Several theoretical and experimental studies have shown that dorsomedial striatum (DMS in rodents; associative in primates) is responsible for goal-directed actions, and dorsolateral striatum (DLS in rodents; sensorimotor in primates) for storing the habits [Daw *et al.*, 2005; Yin *et al.*, 2005; Yin and Knowlton, 2006]. Nowadays, new evidence bring to light the ability of cortex to express habits without the feedback from BG[Hélie *et al.*, 2015; Piron *et al.*, 2016]. For example in Piron *et al.* [2016], we demonstrated the ability of the monkeys to make choices and execute actions when the main output of BG (GPi) was inactive.

The model presented in the previous section (3.2) illustrates how the functionally parallel closed CBG loops participate in action selection. However, if GPi is inactive, as in monkey experiments, it is no more able anymore to choose among options, much less expressing habits. Thus, we extended the model to explore the mechanisms of the acquisition and expression of habits.

In this section, I will present the final architecture and properties of our

model by emphasizing the modifications and additions that we have conducted in the model by Guthrie *et al.* [2013].

3.4.1 Architecture

The architecture of the model is equivalent to the one described in the section 3.2 and is illustrated in Figure 3.11. It contains five structures: cortex (Ctx), striatum (Str), subthalamic nucleus (STN), internal segment of globus pallidus (GPi), and thalamus (Th). Also, it implements the direct (Ctx-Str-GPi-Th-Ctx) and hyperdirect (Ctx-STN-GPi-Th-Ctx) pathway of BG, and three functional modules, cognitive, associative and motor (for more details refer to Section 3.2).

The model has been refined such as to have a competition mechanism within each cortical group. Using short-range excitation and long-range inhibitions, this competition ensures that a unique cognitive and motor decision eventually emerges, even if these decisions might be unrelated at this stage. In other words, this mechanism provides the ability of action selection to the model when the connectivity between GPi and Th is interrupted (equivalent to the inactivation of GPi in monkey experiments [Piron *et al.*, 2016]). However, this addition does not provide solution to the binding problem (where each cue is positioned). To overcome this obstacle, we added cross-connectivity among the three cortical groups, from cognitive and motor to associative groups. In this case, the associative part of cortex is employed as mediator between the cognitive and motor part. Furthermore, we Hebbian learning (LTP) has been added at the cortical level between the cognitive and the associative cortical group. This learning is enforced once per trial, at the time a move is made and independently of the actual reward. Because of these additions, the dynamic of the model have been changed, and so we modified the parameters of the Guthrie model (Table A.2).

3.4.2 Neuron model

Each ensemble is hypothesized to simulate the mean activity of a population of neurons. The neuron model that we used is a leaky integrator as in Leblois *et al.* [2006] and Guthrie *et al.* [2013]. The membrane potential $V(t)$ and firing rate $U(t)$ are described by the following equations:

$$\tau \frac{dV}{dt} = -V + I_S + I_{ext} - h \tag{3.7a}$$

$$U = max(V, 0) \tag{3.7b}$$

where τ is a decay time constant of synaptic input, V is the activity of the neuron, I_s is the synaptic input to the neuron, I_{ext} is the external input received only by the cortical structures (for the other structures $I_{ext} = 0$), and h is the

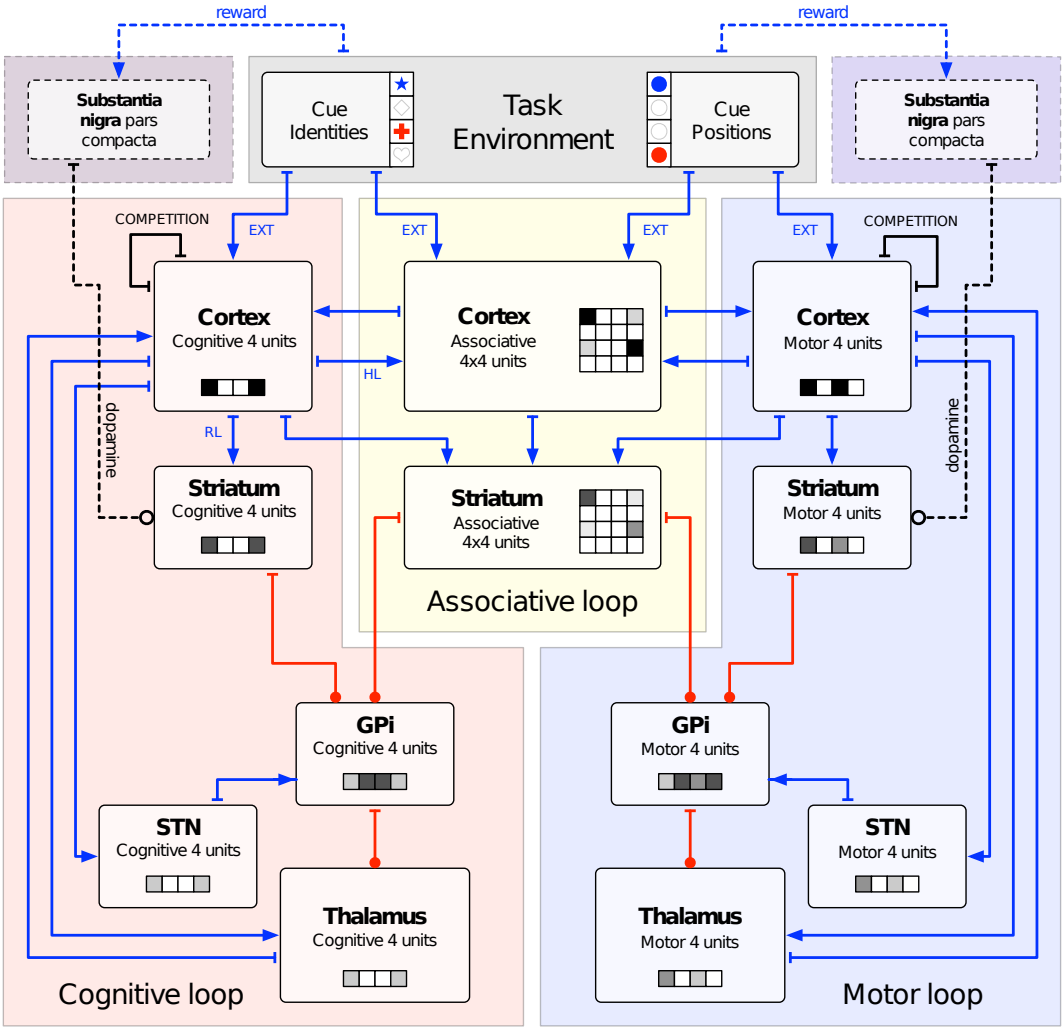


Figure 3.11: **Model Architecture:** The model contains 5 structures: cortex (Ctx), striatum (Str), subthalamic nucleus (STN), internal segment of globus pallidus (GPi), and thalamus (Th). Two of the three known cortico-basal-thalamic pathways are also implemented: direct (Ctx-Str-GPi-Th-Ctx), and hyperdirect (Ctx-STN-GPi-Th-Ctx). In order the model to be able to solve the two-armed bandit task, three loops are added, the cognitive (responsible for cognitive selection), the motor (responsible for action selection), and finally the associative (necessary for binding reasons, to keep which shape is where; more details in the text [3.2, 3.4.1]. All structures include 2 populations for the cognitive and motor loop, except cortex and striatum which have one more for the associative loop. Finally, owing to the nature of the task, each population comprise of 4 neurons.

threshold of the neuron. Notice that firing rates cannot be negative, so with equation 3.7b we ensure that the minimum obtained value is zero. The cortical

cognitive ensembles receive information about the presented cues, the motor about their positions, and finally the combination (where is what) is given as input to the associative ensembles, as shown in Figure 3.12. Also, symmetry breaking is generated by Gaussian noise to the activity of each ensemble at each time step.

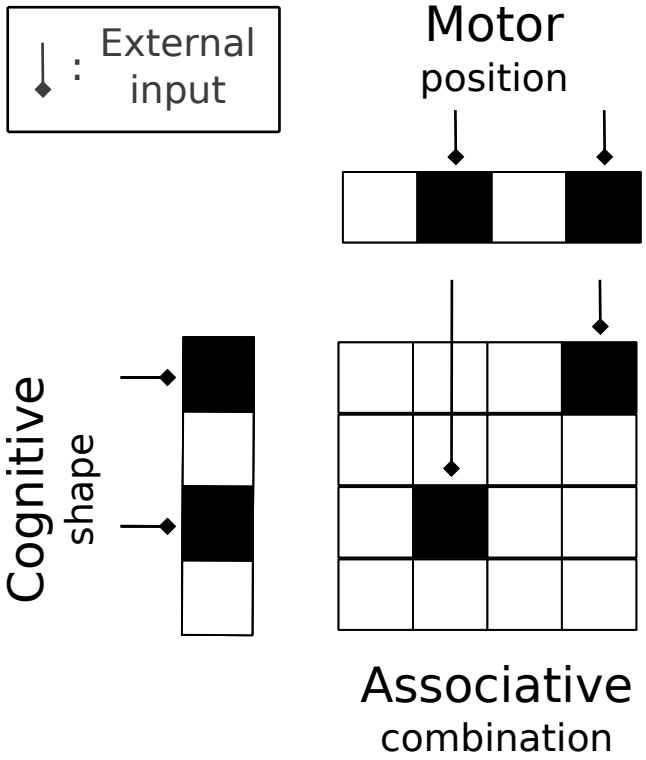


Figure 3.12: Representation of the external input received by the cortical groups.

The total input to a neuron B is fully described from:

$$I_S^B = \sum_A G_B^A \times W_B^A \times U_A \tag{3.8}$$

where A is the presynaptic neuron, B the postsynaptic, G_B^A and W_B^A the gain of the synaptic connectivity and weights from A to B respectively.

Striatal neurons silence without coordinate input is implemented by the Boltzmann equation:

$$U = V_{min} + (V_{max} - V_{min})/(1 + e^{((V_h - V)/V_C)}) \tag{3.9}$$

where V is the input to the transfer function (the activation level of the cortical inputs in this case) and U is the output, V_{min} is the minimum activation, V_{max} the maximum activation, V_h the half-activation, and V_c the slope.

3.4.3 Learning

Striatal

Striatal learning is implemented among cognitive cortico-striatal synapses, in order to associate the cues with the reward probability. At the end of each trial, a reinforcement learning rule is applied as follows:

$$\Delta W_{A\to B} = \alpha_\alpha \times PE \times U_B \times U_A \tag{3.10}$$

where $\Delta W_{A\to B}$ is the change in the weight of the cortico-striatal synapse from cortical population A to striatal population B, PE is the prediction error, the amount by which the actual delivered reward differs from the expected reward, U_B is the activation of the striatal ensemble, and α_α is the global actor learning rate. Generation of LTP and LTD in striatal MSNs has been found to be asymmetric [Pawlak and Kerr, 2008]. Therefore, the actor-learning rate in the model is $\alpha_\alpha = 0.01$ for LTP and $\alpha_\alpha = 0.008$ for LTD.

The PE is calculated using a simple critic-learning algorithm.

$$PE = R + v_i \tag{3.11}$$

where i is the number of the cue chosen, and v_i is the value of cue i. Then, the value of the chosen cue is updated using the PE.

$$v_i \leftarrow v_i + PE \cdot \alpha_c \tag{3.12}$$

where α_c is the critic learning rate, set to 0.002.

The authors in Guthrie *et al.* [2013] are referred to bounded weights to absolute maximum 0.75 and absolute minimum 0.25. However, the bounding algorithm is not described in the article. We hypothesized that it is based on the estimation of the weight gradient along the sigmoid. So, we used an Oja-like rule for our model:

$$S = (W_{A\to B} - W_{min})(W_{max} - W_{A\to B}) \tag{3.13}$$

Finally the updated weights are given by the following equation:

$$W_{A\to B} = W_{A\to B} + \Delta W_{A\to B} \cdot S \tag{3.14}$$

The model as described above is capable of associating the different reward probabilities of the cue (exploration phase), and once it has learned, to choose always the cue with the higher probability (exploitation phase). These results imply that BG are able to learn the respective value of A (low reward probability) and B (high reward probability), and also learn to select B to get the higher probability to be rewarded.

Hebbian

In Piron *et al.* [2016], we showed that monkeys are capable to make optimal choice even without the feedback from BG, and concluded that habits are stored outside of BG, probably in cortex (not a lot of evidence exists today to support or reject this hypothesis). To provide this ability to the model, we implemented learning at cortical level, and we particularly chose Hebbian learning based on evidence from different studies [Doya, 1999; Hélie *et al.*, 2015]. Cortico-cortico synapses from cognitive to associative populations are updated after a decision has been made, according to the following rule:

$$\Delta W_{Cog \to Ass} = \alpha \times U_{Cog} \tag{3.15a}$$

$$W_{Cog \to Ass} = W_{Cog \to Ass} + \Delta W_{Cog \to Ass} \cdot S \tag{3.15b}$$

where $\Delta W_{Cog \to Ass}$ is the change in the weight of the cortico-cortical synapse from the cognitive to the associative cortical group, U_{Cog} is the activation of the cognitive ensemble, and α_α is the global actor learning rate, set to 0.0005.

The choice of the learning rate was based on the realization of a rapid basal ganglia learning compared to the cortical learning [Paspathy and Miller, 2005].

All the weights are initialized to the absolute values displayed in Table except the ones that are altered in learning (striatal & cortical). The lateral ones are initialized to 0.5 (SD: 0.005) at the beginning of each session.

3.5 Conclusions

Our model is an extion of previously introduced models by Leblois *et al.* [2006] and Guthrie *et al.* [2013], in order to investigate the acquisition and expression of habits. In the model of Guthrie *et al.* [2013], action selection at the cortical level occurs because of the interaction between the direct and hyperdirect pathway. However, it has been shown that monkeys are able to choose among options when the main output of BG is inactive [Piron *et al.*, 2016]. In order the model to obtain this capability, we added a competition mechanism, short-range excitation and long-range inhibitions, within each cortical group. This competition ensures that eventually a unique cognitive and motor decision emerges. Hebbian learning has also been included at the cortical level, and is enforced once per trial, after a move has been executed, independently from the actual reward. We assume that habits are stored outside of basal ganglia, and more precisely at cortical level, even though we still do not have enough evidence to exclude other areas (*e.g.* cerebellum) [Hélie *et al.*, 2015; Piron *et al.*, 2016].

In the next chapter, I will introduce the protocols that the model has been tested on. Then, I will report the results of the model on these protocols, and

compare them with the monkey results whenever they exist. I will close the chapter by our interpretation of these results, and the comparison of our model with the ones introduced in the previous chapter.

"When it comes to exploring the mind in the framework of cognitive neuroscience, the maximal yield of data comes from integrating what a person experiences - the first person - with what the measurements show - the third person."

— Daniel Goleman

Chapter 4

Experimental and computational results

Contents

Our model has been developed based on anatomical and physiological data of primates, and its aim is to explore the decision mechanisms underlying goal-directed actions and habits. Therefore, we test the model on equivalent tasks,

such as n-armed bandit paradigms which used in experimental psychology and neuroeconomics.

In total, we used four protocols of two-armed bandit task. The first one is the same as described in Guthrie *et al.* [2013] (adaptation of the protocol used on monkeys in Pasquereau *et al.* [2007]). With this protocol, the model ability for action selection and learning hidden reward probabilities is tested. Furthermore, the experiments on monkeys described by Piron *et al.* [2016] provide evidence for the crucial role of basal ganglia during initial stages of learning, alongside cortex ability to make optimal choices alone, without feedback from the BG after learning. Thus, we implemented the same protocol to investigate the role of BG during the acquisition and expression of habits through the model. These results combined with the properties of the model led us to a new hypothesis about how cortex is able to acquire (store) habits and express them. We developed a new protocol to test our theory, and then conducted experiments on monkeys that confirmed it.

The characterization of an action as a habit in many studies is provided by the inability of the subjects to adapt their behavior according to new internal or external conditions (food satiation or outcome devaluation) [Yin and Knowlton, 2006; Liljeholm and O'Doherty, 2012]. We finally developed the last protocol in order to observe how the model will handle these kind of changes (devaluation of the outcome), and explore its capacity.

4.1 Protocols

4.1.1 Overall structure of the task

In a n-armed bandit task, the subject must choose repetitively among different options. Each choice results to an outcome that is unknown to the subject before the start of a session. The subjects generally assess the outcomes during exploratory trial-and-error phase, and then choose preferentially, but not always exclusively, the choice associated with the best outcome, in an exploitation phase. This type of task allows the testing of deliberative decision-making process built on the accumulation of evidence (learning).

Generally, in this kind of tasks, each cue (S_i) is associated with three values: the probability ($P_i \in [0, 1]$), the quality ($Q_i \in (0, 1]$), and the amount ($R_i \in (0, 1]$) of reward. If the probability of a cue is 1, it means that each time the cue is chosen it will be rewarded, and when it is 0 then no reward will ever be received. Analogically, the amount of reward depends on the value R_i, with 1 being the most reward that can be achieved by the subject. The quality value characterizes how desirable the reward is to the subject. For example, cappucin monkeys like cucumber, but grapes are much better food for them, so the cucumber will be correlated with a smaller quality value than the grapes. The motivation of the subject to choose a particular cue or the

amount of learning during one trial varies depending on the different values of reward. For instance, if the value of the quality or amount of reward associated with one cue is high, then theoretically the subject will learn faster this cue compare to others with more medium values. But if its probability is low, then maybe he will prefer to choose one cue with less quality or amount, but more probable to receive some reward.

In our protocols, we used exclusively a two-armed bandit task meaning that only two options are given in each trial. Also in our case, the amount of reward is always $R_i = 1$, as well its quality ($Q_i = 1$). The probability of reward although varies among the protocols.

Implementation

As described in the previous chapter (3.4.1), the cortical cognitive part of our model comprises of four ensembles, representing equal number of possible cues, the motor part of possible positions, and the associative of sixteen possible combinations of cues and positions. In each trial, external input of 7Hz is sent to two ensembles of each cortical group (Figure 4.1), representing the display of two shapes in two specific positions. For convenient reasons, the external

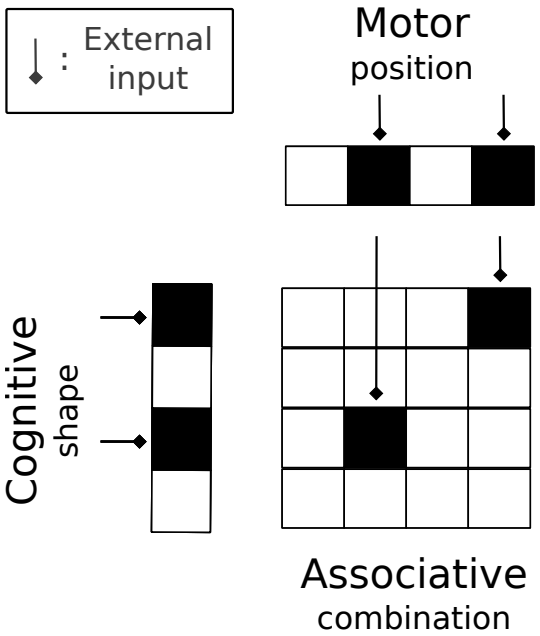


Figure 4.1: The cognitive and motor cortex receive external input to two random ensembles, representing which shapes are given to the network and which positions are occupied. The associative cortex contains the information of which shape is presented in which position.

inputs to cortical cognitive and motor groups will be referred as shape and

position respectively. The choice of two presented shapes and positions are made randomly in each simulated trial, and then their combination is derived.

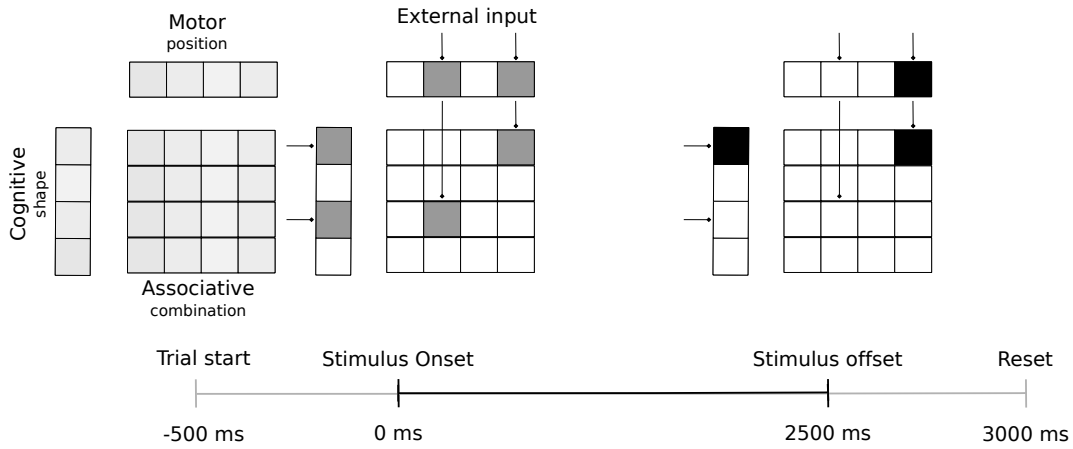


Figure 4.2: Timeline: The model is left for $500ms$ to stabilize before the cortical populations receive the external input at $t = 0ms$. It has $2.5s$ to make a choice, and then the activities of all the populations in the network are reset.

The timeline of the task is shown in Figure 4.2. Each trial lasts 3 sec of simulated time. The system is let to settle into its stable point for $500ms$. At this moment, the 3 cortical groups receive an external input. A movement is considered to be performed when one of the motor neurons is 40Hz more active than the others, and the trial ends. A trial is considered successful if a movement has occurred within 2.5 sec subsequently to the reception of the external input. In a successful trial, it is ensured that a motor decision has been made. However, there is the chance that the cognitive decision would occur before or after the motor, or not at all. In the monkey experiments, we observe only the push of a button as a result of a decision, but we cannot be sure if the respective shape was chosen first or not. For that reason, in the model as chosen shape is considered the one in the chosen position. After the end of a successful trial, the reward is delivered or not, according to the reward probability of the chosen shape, and learning occurs in cortical and striatal level. To ensure independence among sessions and trials, the model is initialized before the start of each session (weights initialization), and the activity of all ensembles before each trial.

A choice is defined as optimal when the shape at the chosen position is associated with the highest reward probability. Furthermore, the reaction time is defined as the latency between the reception of the input at cortical level and the execution of a simulated movement (a choice in motor level). Finally, the beginning of the session is defined as the first 25 trials, and the end of the session as the last 25 trials.

The performance of the model in a protocol is evaluated by the mean success rate; in other words, the mean of optimal choices for each trial in all

sessions is computed, and then the mean of all trials outlines the performance of the model in a protocol.

Task Set

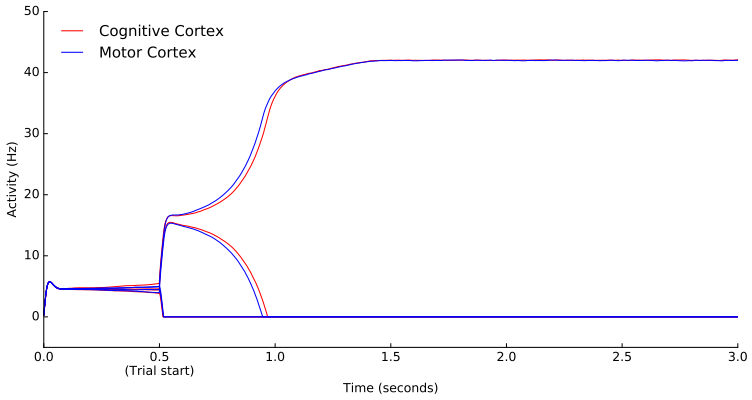
For each trial, two out of the four cognitive and motor cortical ensembles receive external input representing the display of shapes in specific positions. Two associative cortical ensembles also receive external input providing the information of where each shape is placed. Each protocol determines how many shapes are presented during a session, and which reward probability are associated with. If the association between the individual shape and reward probability changes or not during each session and among sessions, it depends on the protocol. The choice of the positions are random among the four possibilities in all protocols. However, the presented pair of shapes can be random , but can also be the same during a session. The protocol defines the number of trials in a session, although each protocol contains 25 sessions.

Example

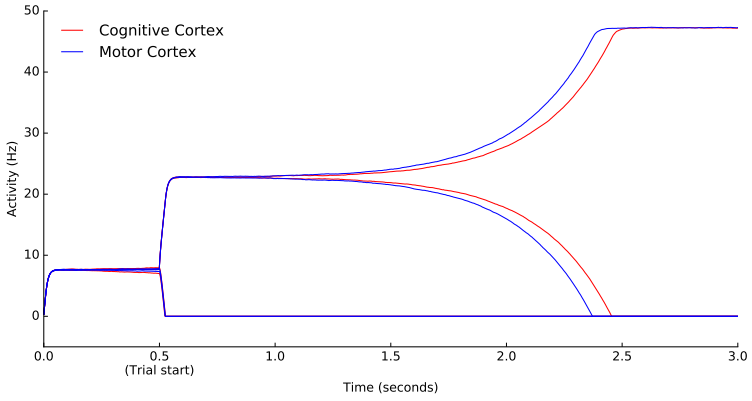
An example of cognitive and motor cortical activity during a trial in the beginning of a session is shown in figure 4.3a (red and blue line respectively). The system is let to stabilize for $500ms$, and then external input is sent to the cortical groups. The two ensembles that receive the input inhibit the other two due to lateral connectivity, and start to compete with each other. Due to the noise, the symmetry breaks and the activation of one ensemble is amplified, when simultaneously inhibits the other one.

Figure 4.3b is equivalent to 4.3a, but when the BG output (GPi) is inactive. It shows that even in this situation, the model is able to make a decision due to the lateral cortical connectivity and connectivity among cortical structures absence. The motor decision precedes the cognitive one in he latter figure. Knowing the task, someone could say that this is a deficiency of the model. However, at the beginning of a session, there is no evidence of which component of the task is important, the shape or the position. The choices based on each of them are equally optimal.

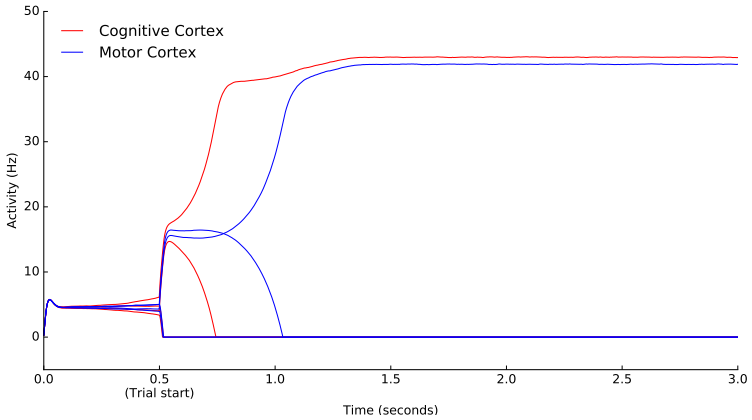
In Figure 4.3c, we notice that the cognitive choice has been made faster than the motor, providing time to cognitive cortex to interfere in the motor. Furthermore, we observe that in motor cortex one ensemble starts to win but few milliseconds later the other one becomes more active and win at the end. This change occurs as a result of a faster cognitive choice that influences the motor through the associative group. Finally, Figure 4.4 demonstrates the evolution of activity for all structures during a single trial.



(a)



(b)



(c)

Figure 4.3: The activity of the cognitive and motor cortical populations is presented at the beginning of learning with (a) active and (b) inactive the connections between GPi and thalamus, and after learning has occurred with active connections.

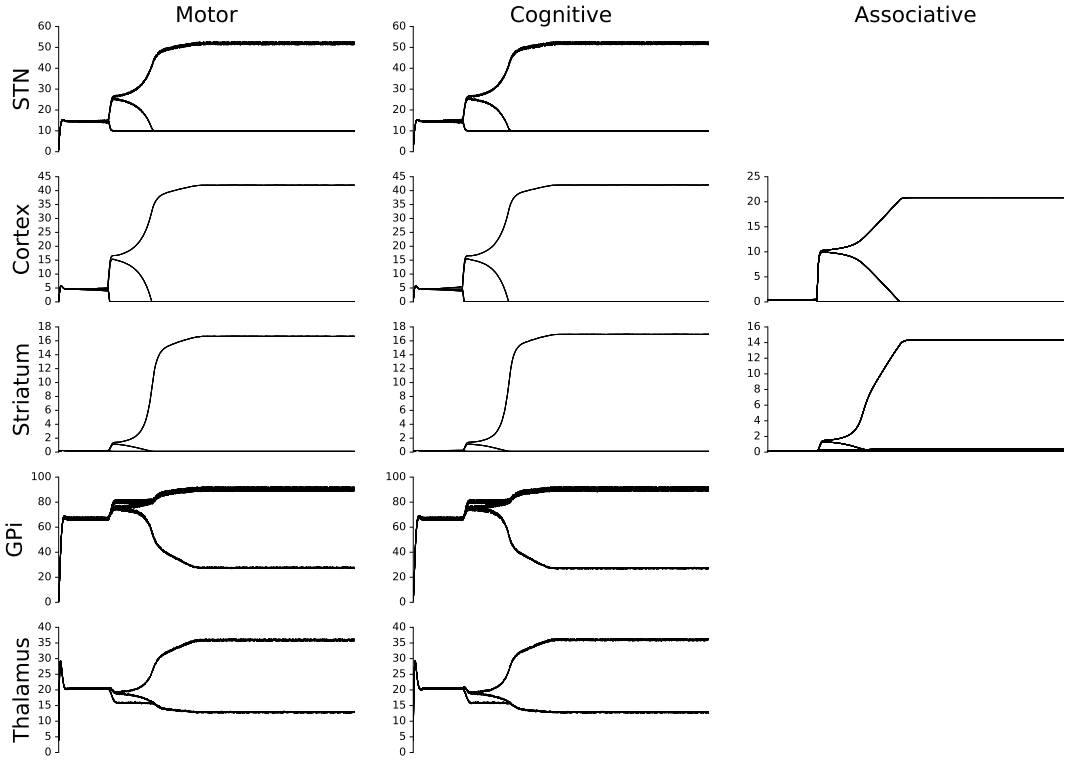


Figure 4.4: Activity of all populations at the beginning of learning.

4.1.2 Protocol A: Control

The purpose of this protocol is to test the ability of the model to learn the hidden reward probabilities in a two-armed bandit task. We hypothesize that the model is initially naive, choosing randomly, but after training it is able to learn the values of new contingencies.

This protocol follows the task set as described in section 4.1.1, and is shown in Figure 4.5. Four shapes are presented in a session (180 trials in total; two shapes in each trial), and each is associated with a unique reward probability: $n_1 = 1.00, n_2 = 0.66, n_3 = 0.33, n_4 = 0.00$.

During one session and among all sessions, the association between the individual shape and its reward probability does not change.

4.1.3 Protocol B: Formation of habits

To address the contradiction between experimental data showing that the basal ganglia are involved in goal-oriented and routine behaviors, and clinical observations which have reported no sever impairment in goal-directed or automatic movement after lesion or disruption by deep brain stimulation of the globus pallidus interna, Piron *et al.* [2016] designed an experimental paradigm based on a two-armed bandit task that combines pre-learned choice behavior, delib-

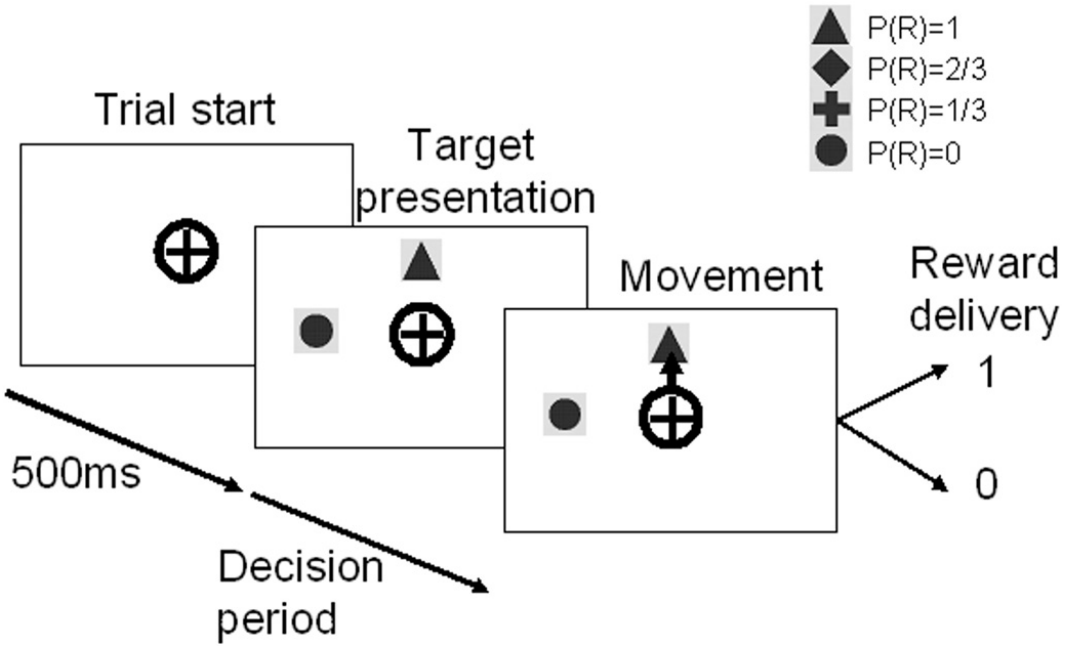


Figure 4.5: As in Guthrie *et al.* [2013]. Schematic representation of the time line of the task. During a session, four cues associated with reward probabilities [P(R)] are used, and only two out of the four are chosen randomly for each trial.

erative decision making, and procedural learning. The experiment was carried out on nonhuman primates with pharmacological inactivation of the GPi.

I will first present the monkey set up, task and protocol and then the equivalent protocol for the model.

Monkey set up

Two female macaque monkeys (Macaca mulata weighing 4.9 and 5.6 kg, respectively) took part in this experiment. The setup consists of four buttons placed on a board at different locations (0°, 90°, 180°, and 270°), and a further button in a central position, which detects contact with a monkey's hand (Figure 4.6). The primates are placed in front of a screen at a distance of 50 cm, seated in chairs. A cursor appears on the screen to one of the 5 possible display positions, when the corresponding button is pressed.

Monkey Task

A trial is initiated, when the monkeys keep their hands on the central button, resulting in the appearance of the cursor at the central position of the screen. Two shapes are presented in two out of the four positions, randomly determined

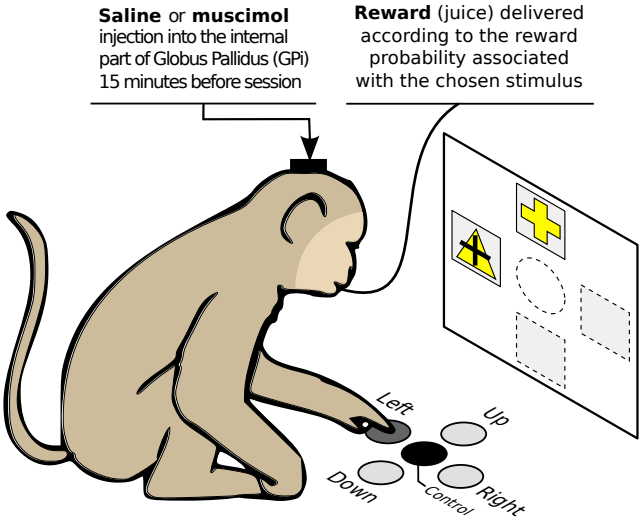


Figure 4.6: Monkey set up

for each trial, after a random delay $(0.5 - 1.5s)$. Once the cues are shown, the monkeys must press one of the periferl button corresponding to one of the shapes in a random duration time window (0.5-1.5 s). They have to maintain this position for 0.5 s to 1.5 s, in order to be rewarded (0.3 ml of water). If the chosen button corresponds to one of the presentes shapes, only then reward is delivered according to the reward probability of the chosen target. The disappearance of the cursor from the chosen shape indicates the finish of the trial. The next one begins after an inter-trial interval between 0.5 s and 1.5 s. The task is summarized in the Figure 4.7.

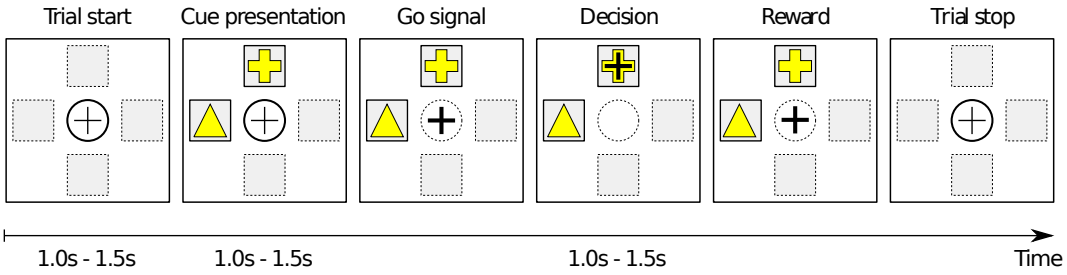


Figure 4.7: Schematic representation of the behavioral task

Bilateral inactivation of GPi

Microinjections were delivered bilaterally 15 minutes before a session. We assume that our injection encompassed a significant proportion of the GPi including motor and associative areas. For both animals injections of the GABA agonist muscimol hydrobromide (Sigma) or saline (NaCl 9‰) were randomly assigned each day. Muscimol was delivered at a concentration of

$1\mu g/\mu l$ (dissolved in a NaCl vehicle). The effect of the muscimol injection is presented in Figure 4.8. For more details, please refer to Appendix.

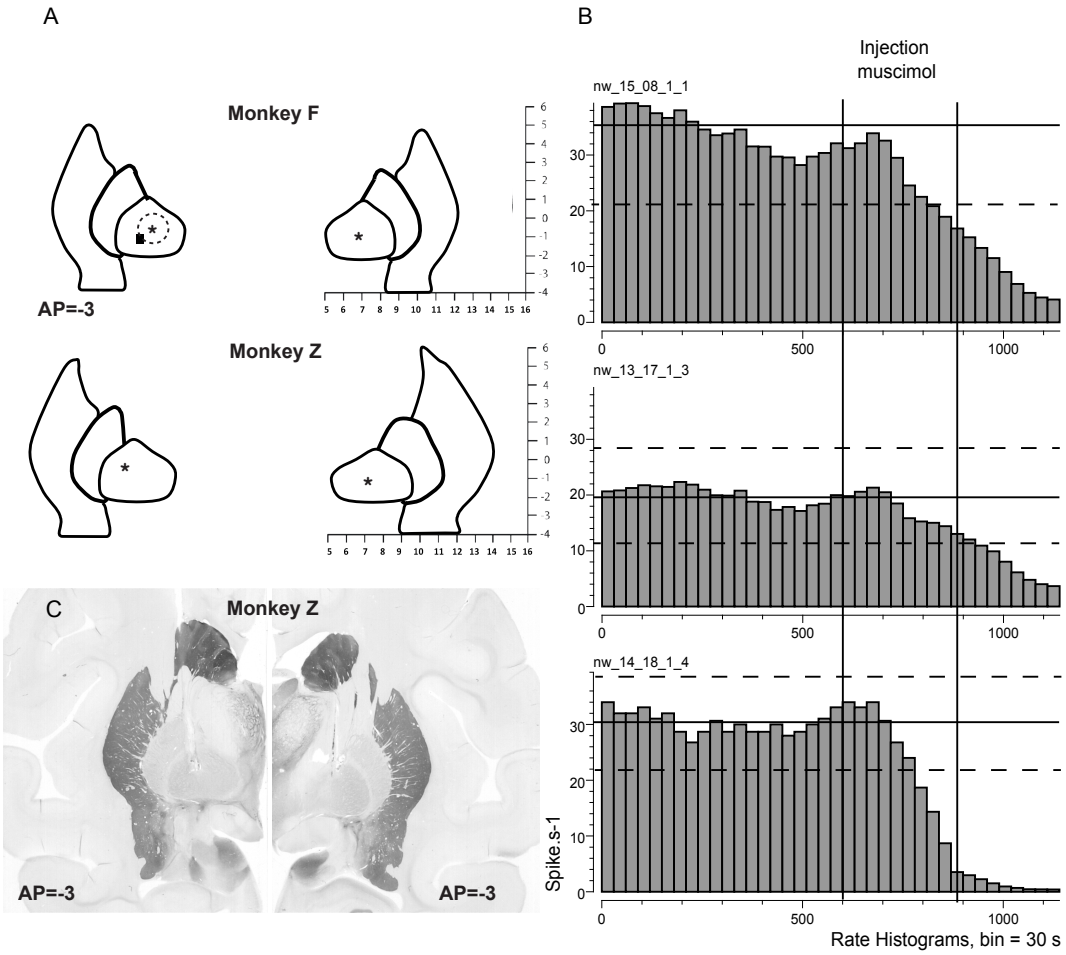


Figure 4.8: As in Piron *et al.* [2016]. A) Theoretical injection sites for both monkeys in the right and left hemispheres. B) Time histogram of the firing rate of three neurons recorded before and after muscimol injection in GPi. The two vertical lines represent the beginning and the end of the injection. The three neurons recorded were sensitive to the muscimol injection. The timescale is in ms. C) Histological display of the position of the cannulae in monkey Z. For more details please refer to the original paper.

Monkey Protocol

The two experimental conditions were alternated in blocks of 10 trials: the Habitual Condition (HC) and the Novelty Condition (NC). In the HC always the same pair of shapes was presented (HC_1 and HC_2), on which the animals had been previously trained (during 8 months for monkey Z and 12 months

for monkey F). Each shape was associated with a fixed reward probability ($P_{HC_1} = 0.75$ and $P_{HC_2} = 0.25$) [Figure 4.9]. In the NC, each session contains a unique pair of shapes that was never presented before (NC_1 and NC_2) with fixed probabilities of reward of 0.75 and 0.25 respectively ($P_{NC_1} = 0.75$ and $P_{NC_2} = 0.25$) [Figure 4.9].

20 sessions (10 for each monkey) with saline injections (Saline) and 20 (10 for each monkey also) with muscimol injections (Muscimol) were performed. The proportion of trials in which the animals chose the optimal target (i.e. HC_1 or NC_1 respectively) was defined as the success rate, normalized by the number of trials in which a choice was made. When a trial was interrupted before a choice had been made and validated, it was counted as an error trial.

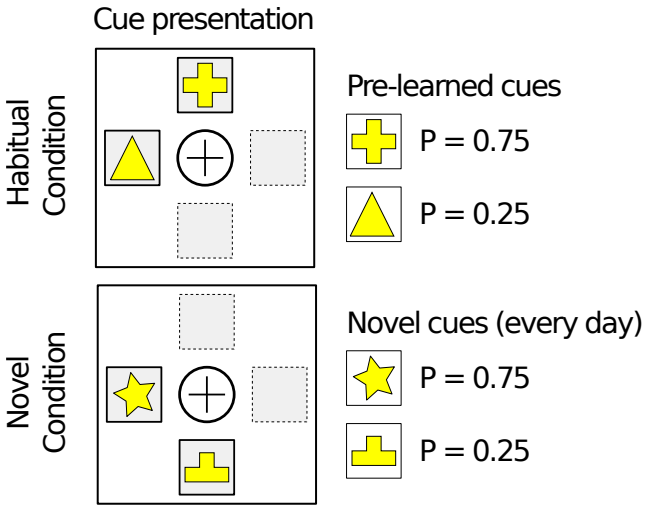


Figure 4.9: As in Piron *et al.* [2016]. The monkeys have been trained on the cues (HC_1, $P = 0.75$ and H_2C, $P = 0.25$), which are used in the habitual condition (top). In the novelty condition (bottom), the cues (NC_1 and NC_2) have the same reward probabilities ($P = 0.75$ and $P = 0.25$), respectively), but the pairs are changed (new shape and colors) in each session.

Model Protocol

The protocol is divided in three sections, which I will refer to them as training, habitual, and novel. The training section simulates the monkeys' pre-training to a particular pair of shapes before they are tested on them in the Habitual condition. In training and habitual sections, only the two first cognitive cortical ensembles receive input, with 75% and 25% probability to be rewarded ($n_1 = 0.75$, $n_2 = 0.25$). The two last ensembles receive input during the novel section, with reward probabilities 75% and 25% ($n_3 = 0.75$, $n_4 = 0.25$). With this separation, we ensure that a different pair of shapes is presented during this part.

At the first part, the model is trained for 360 trials, and then it is tested on the same inputs for 120 trials in the habitual part. Lastly, inputs to the other pair of neurons are presented to the model for 120 trials during the novel part. The cortical groups receive input as described in 4.1.1. The model is tested in the two conditions (Habitual and Novelty) separately, contrary to the experiments where two blocks of ten trials of each condition alternate to ensure that the monkeys do not use only their working memory.

This protocol includes also two states: one with active connectivity between GPi and Thalamus, and one with inactive; simulating the effect of saline and muscimol (*i.e.* inactivation of GPi) injections. Even if it is inaccurate, these two conditions will be referred as GPi On and GPi Off for simplicity reasons. We separate these two states in two experiments, where we conduct the experiment as described above with GPi On during all parts of the protocol with GPi On. In the second state, the connectivity is active during the training part, but inactive during the two conditions. During and among sessions the association between the individual ensemble and its reward probability remains the same.

4.1.4 Protocol C: Storage of habits

The aim of this protocol is to investigate where and when the acquisition and expression of habits occur, and the role of BG in these procedures.

Model

In this protocol, only the two first ensembles of cognitive cortex receive external input, associated with the reward probabilities of 75% and 25% respectively ($n_1 = 0.75$ and $n_2 = 0.25$). We also included two states, GPi On and Off, as described in section 4.1.3.

The protocol is divided in three phases with alternation of the two states. First the model is tested with GPi Off (Day 1), followed by GPi On (Day 2), and lastly with GPi Off again (Day 3). Each session includes the three phases consisting of 120 trials each. The learning weights are initialized at the start of each session to ensure the presentation of new contingencies in the first part.

Our hypothesis is that during the whole first part the model will choose randomly between new inputs, because of the absence of feedback from BG, like in Protocol B. However, it will be able to perform over the chance level (0.5) at the beginning of the second part and improve until its end. That is because we expect striatum to have already learned the values of each input, so it can influence and lead cortical choices to optimum options. Finally, at the beginning of the third part the performances are also predicted to be better than chance level, as a consequence of habit acquisition in cortical level during the second part.

Monkeys

Following the experimental rules at Piron *et al.* [2016], on Day 1 muscimol is injected, before the monkeys to be tested on an unknown pair of shapes. On Day 2, after saline injection, the same pair is presented for testing. Unfortunately, the monkey experiments do not include the third part of the model's protocol. However, we conducted another experiment that can help us make a hypothesis for the last part. The monkeys after saline injection on Day 1 were tested on novel pair of shapes, and on Day 2, muscimol was injected before testing them on the same pair of Day 1. Each day comprised of 60 trials in a total of 5 experiments.

4.1.5 Protocol D: Characterizing habits

One characteristic of habits widely accepted is the insensitivity to devaluation; i.e. reduction of an action outcome. This protocol has been designed to test the ability of the model to follow the changes of reward probabilities.

In this protocol, only the two first cortical cognitive ensembles receive input during a session. At the beginning, the 1^{st} has 75% probability to be rewarded ($n_1 = 0.75$) and the 2^{nd} 25% ($n_2 = 0.25$) until trial t. From the trial t+1 and until the end of the session, the probabilities are reversed ($n_1 = 0.25$, $n_2 = 0.75$). We conducted different experiments with a variety of t, from 50 to 950 per 50 trials. This variety of reverse time is convenient for the demonstrating the model behavior after short or extensive training. Finally, we assume that the model explores when its performances are over 0.0 and less than 0.90 and exploits when they are more than 0.9.

4.2 Computational results

4.2.1 Protocol A: Control

Figure 4.10 shows the mean success rate for all the trials of one session over all sessions. At the first trials, the model makes choices randomly (58.6%±10.0%), showing its ignorance to the hidden reward probabilities associated with the distinct shapes. Then, we observe an exploration phase for about 100 trials, when the model starts to learn the worth of each input. For the rest of the session, the learning continues but with a slower progress, reaching finally $93 \pm 0.06\%$ of success ($95 \pm 10\%$ for the last 30 trials and 250 sessions in Guthrie *et al.* [2013]). More trials would improve the performances, but in this task would never be perfect. The reason is that the input with no reward is almost never chosen, so striatum cannot learn that it is a bad choice. We can see that from figure 4.11, where the value of the input and the cortico-striatal weights don't decrease, resulting to an insignificant difference with the second

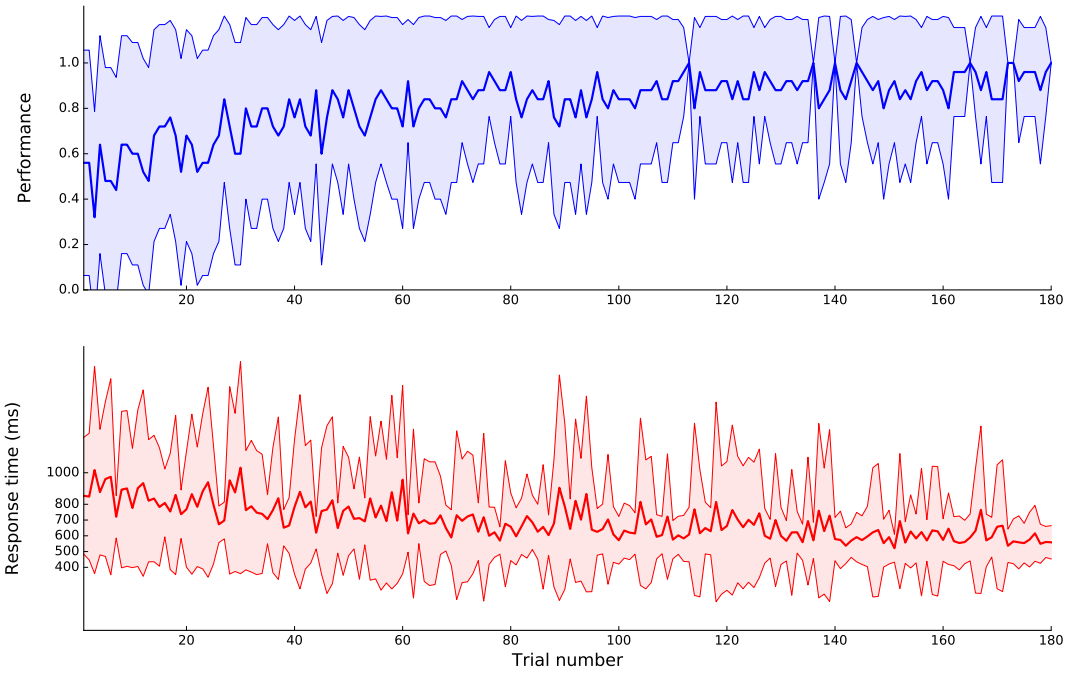


Figure 4.10: Performances (*top*) and reaction time (*bottom*) of the model over 180 trials, averaged over 25 simulations. The filled area indicates the standard deviation.

worst one. Thus, when these two inputs compete, even if the one with 33% chance of reward will be more likely to be chosen, it's not absolute certain such as with the other combinations, because of the noise in the system. Finally, it is worth mentioning that as learning improves the time for the model to choose an action decreases from $851.74 \pm 76.16ms$ to $596.01 \pm 50.21ms$ (lower part of panel in Figure 4.10).

4.2.2 Protocol B: Formation of habits

Figure 4.12 contains the results of both states; *i.e.* with active and inactive GPi. As we can see, the model is able to express previously acquired habits. In the habitual condition, it makes the optimal choice in every trial ($100.0\% \pm 0.0\%$). Additionally, the model begins every session in the novel condition with random choices ($60.2\% \pm 18.6\%$), but it succeeds to learn the new associations by the end of the session ($99.5\% \pm 1.7\%$). Furthermore, habits are still expressed after inactivating GPi, but not perfectly as before ($95.3\% \pm 3.40\%$); a small tendency for exploration exists. On the other hand, the model is unable to deduce new associations and remains to a random mode (from $56.0\% \pm 14.1\%$ to $60.6\% \pm 20.9\%$).

Finally, the decisions are faster for pre-learned pairs of choices compared to unknown, but also when BG contribute to final decision than otherwise.

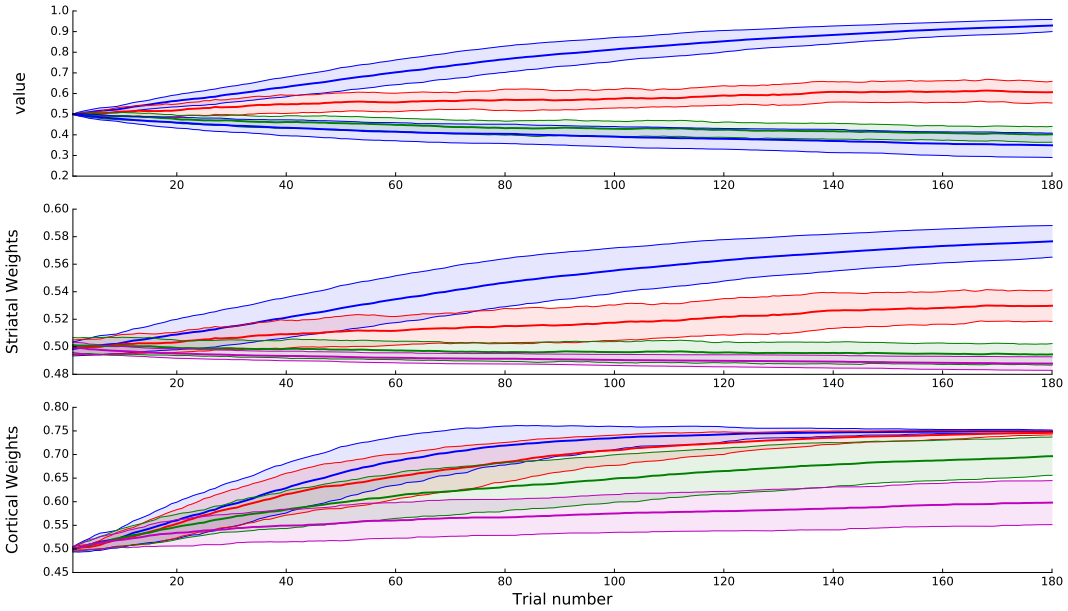


Figure 4.11: **Evolution of weights and values** The average cortico-striatal and cortico-cortical weight, and the state value for the 4 cues are presented with the different colored lines: $P(R) = 1$ blue, $P(R) = 0.66$ red, $P(R) = 0.33$ green, and $P(R) = 0$ magenta

The mean reaction time is $258.58 \pm 4.53ms$ during a session in habitual condition but significantly increases when GPi is inactivated to $1014.05 \pm 62.30ms$. Analogous significant rise is observed in the novel condition with mean reaction time of $330.35 \pm 50.22ms$ with GPi On and $1129.45 \pm 57.98ms$ with GPi Off.

4.2.3 Protocol C: Storage of habits

As expected based on the results of protocol B on Day 1, the model is unable to learn new associations without feedback from BG (beginning of sessions $51.5\% \pm 11.0\%$, end $53.0\% \pm 17.1\%$) [Figure 4.13]. However, after the reactivation of GPi the model reaches instantly optimal performances ($94.2\% \pm 6.6\%$ in 25 first and $99.5\% \pm 1.3\%$ 25 last trials), confirming our theory of value acquisition in BG and their ability of leading cortical decision.

After GPi is inactivated once more, the model does not return to the chance level (i.e. random choices), but it shows to have learned the optimal choice, although not perfect yet ($84.6 \pm 10.3\%$ in 25 first and $87.4 \pm 9.4\%$ 25 last). That indicates that habit acquisition at cortical level,but also highlights the necessity of BG as his teacher during the acquisition. All these results are summarized in Figure 4.13.

The results from Day 2 confirm our theory of implicit value acquisition

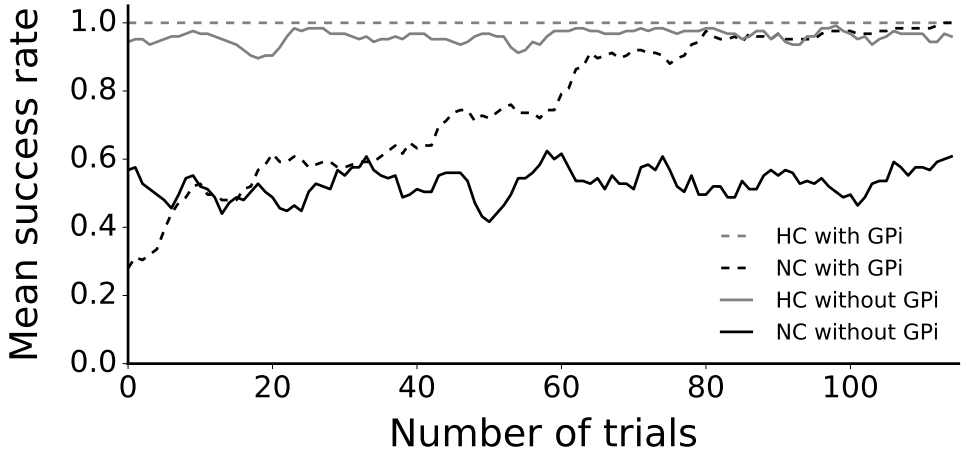


Figure 4.12: Mean success rate across successive trials in the two conditions, routine (gray) and novelty condition (black), and in the two states with (dashed line) and without (solid line) GPi. The curve is smoothed using a moving average filter of 10 consecutive trials.

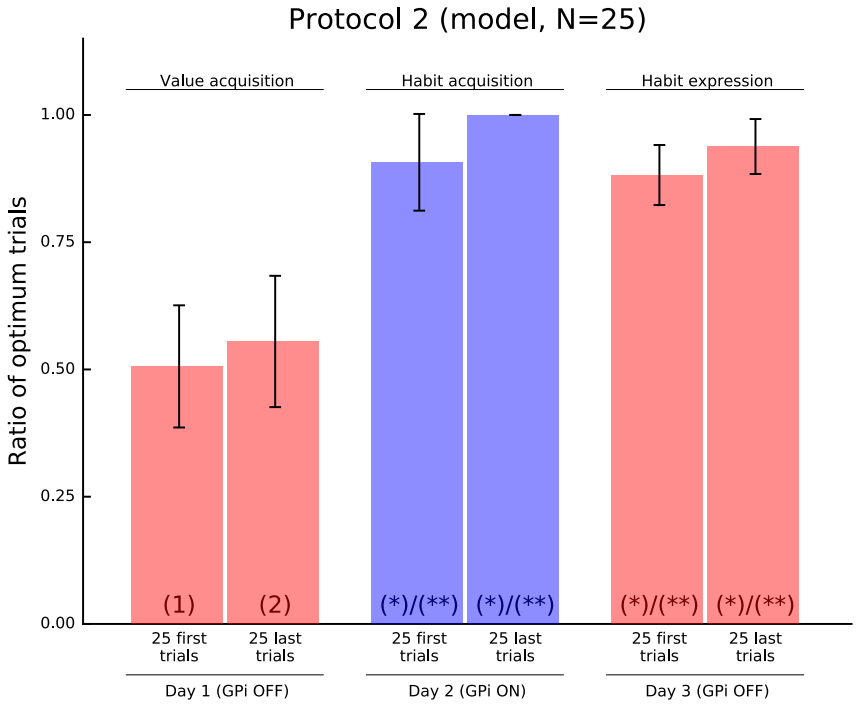


Figure 4.13: D1 corresponds to the first day of the experiment where GPi is suppressed (removal of GPi-Thalamus connection). D2 corresponds to the second day where the suppression of the GPi is removed. During D3, GPi is suppressed again.

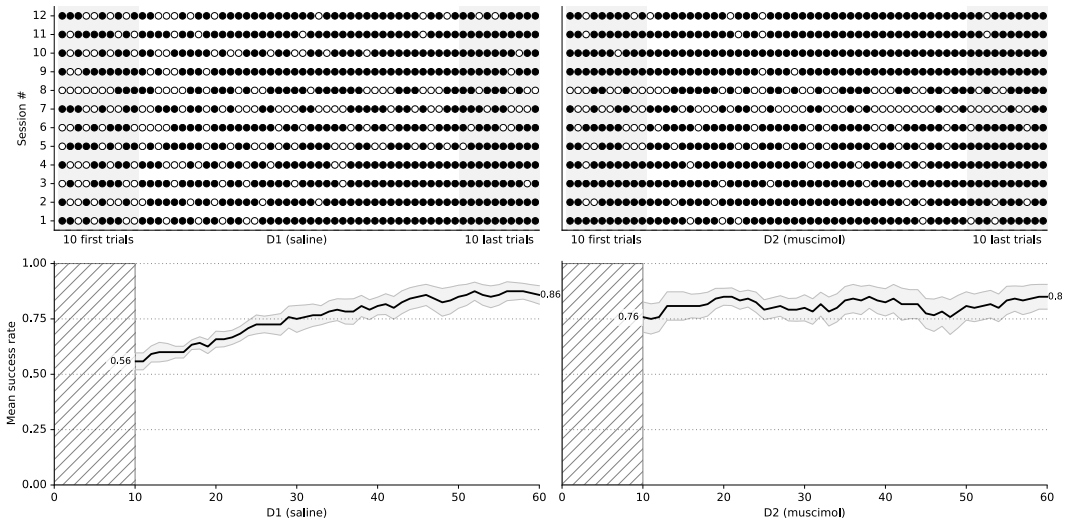


Figure 4.14: On the top figures, all the trials of the 12 sessions are presented. The black dot represents an optimal trial, meaning that the model chose the cue with the highest reward probability, contrary to the white dot which represents a trial that the cue with the lowest reward probability has been chosen. The bottom figures show the mean success rate of the session with a window of 10 trials. D1 corresponds to the first day of the experiment where GPi is active in the model, and D2 to the second day where there is suppression of the activity of GPi.

in BG during Day 1, even though GPi is inactive at this state. Also, on Day 2 is demonstrated the BG ability of leading cortical decision before habit acquisition. Finally Day 3 indicates that cortex needs BG supervision for the acquisition of habits, a procedure parallel to goal-directed learning.

To be consistent with the monkeys experiments, we tested the model also to a protocol equivalent to the monkeys. On Day 1 (active GPi), the model starts naive ($56.0 \pm 3.8\%$), but, during the session, it is able to learn the new presented contingencies ($86.5 \pm 4.0\%$) [Figure 4.14]. As expected, from the results of the three days protocol, on Day 2 (suppressed GPi) the model performs well ($76.3 \pm 9.1\%$) from the beginning of the session , and continues to improve until the end ($92.6 \pm 6.7\%$).

4.2.4 Protocol D: Characterizing habits

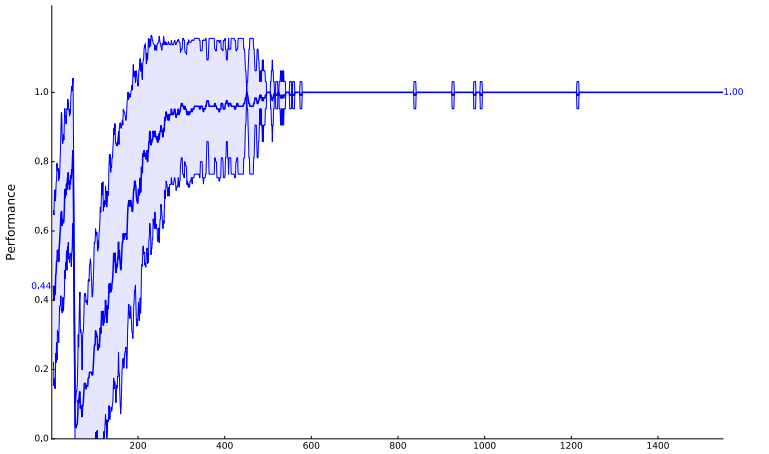
Figure 4.15 shows three experiments with reverse of reward probabilities among the cues at t = 50, 450, 950. At the first case, the model has reached performances over 80% before the reverse occurs, so it is still sensitive to changes resulting to an immediate exploratory behavior after it. This is not the case in the second experiment, where the model adheres to old routines for more

trials than before the reverse, but eventually it's able to learn which is the optimal choice after the reverse. However in the last experiment, the model is unable to modify its strategy, continuing choosing the same choice for the whole session.

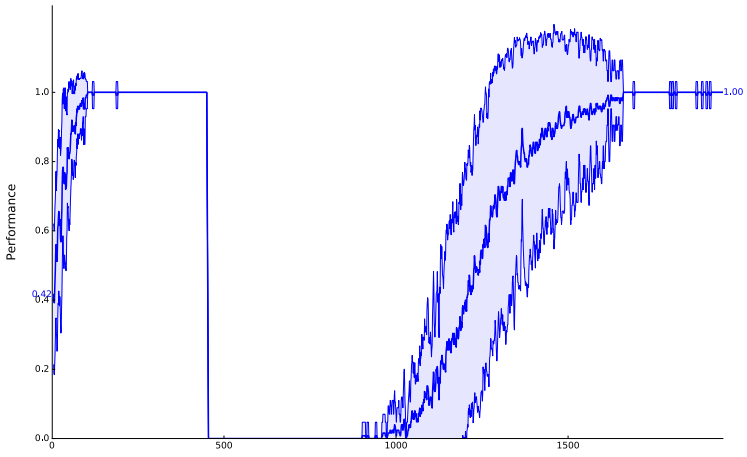
Many researchers argue that the subject has acquired habits when there is no strategy change after reward devaluation [Yin and Knowlton, 2006; Graybiel, 2008]. But for how many trials did they tested after the modification? What would happen if they had included more trials in their experiments? For example, if we had stopped our sessions in less number of trials when the reverse occurs after 450 trials, we would have conducted the same conclusions. However, the model shows that the network just needs more training before to start learning again.

The experiments that I presented previously, are only a part from a series of experiments that we conducted. The model behavior was observed on the same protocol for 18 different cases; with the reverse occurring from 50 to 90 trials per 50. In all of these cases, we defined the amount of trials that the model needed to start exploring after the reverse (mean success rate over 0%) and to start exploiting (mean success rate over 90%). All these results are gathered and presented in Figure 4.16 where we can see that the pre-training and the exploration of new strategies are linearly dependent. In other words, the more pre-training occurs, the more trials are needed to start reducing the expression of old strategies. However, the model, after a finite number of trials (500), is unable to express goal-directed actions again. Unfortunately, we don't have any experimental results on monkeys to compare with the results from our model. However, based on these results, we suggest that a habit has been acquired when in order to alter this action, the subject requires more training than for the acquisition.

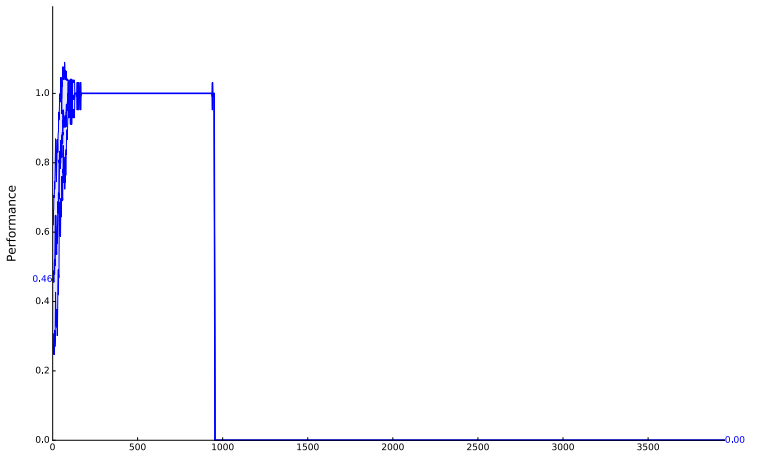
As human beings we are able to modify our strategies based on previous failures on relative fast pace. Why the model cannot? One reason is that the involvement of other areas is necessary, such as ACC, to compare the consequence of the last action with the previous one. On the other hand there are also computational limitations. Figure 4.22 contains the progress of the sessions in two experiments (t = 50, 950), displaying their performances and the evolution of learning. From figure 4.17a, we observe that the exploration of alternative strategies starts when the strongest striatal weight decreases its strength at almost the level of the other ones. At this point the cortical learning of the new strategy also begins. Our network does not contain any anti-Hebbian or forgetting rule in cortical level, resulting finally to equally learning for all inputs. In this case, if BG were inactive, cortex would not be able to express any habits and would choose randomly as in the beginning of



(a)



(b)



(c)

Figure 4.15: Averaged performances over 25 sessions of the model during three experiments with reverse at (a) $t = 50$, (b) $t = 450$, and (c) $t = 950$

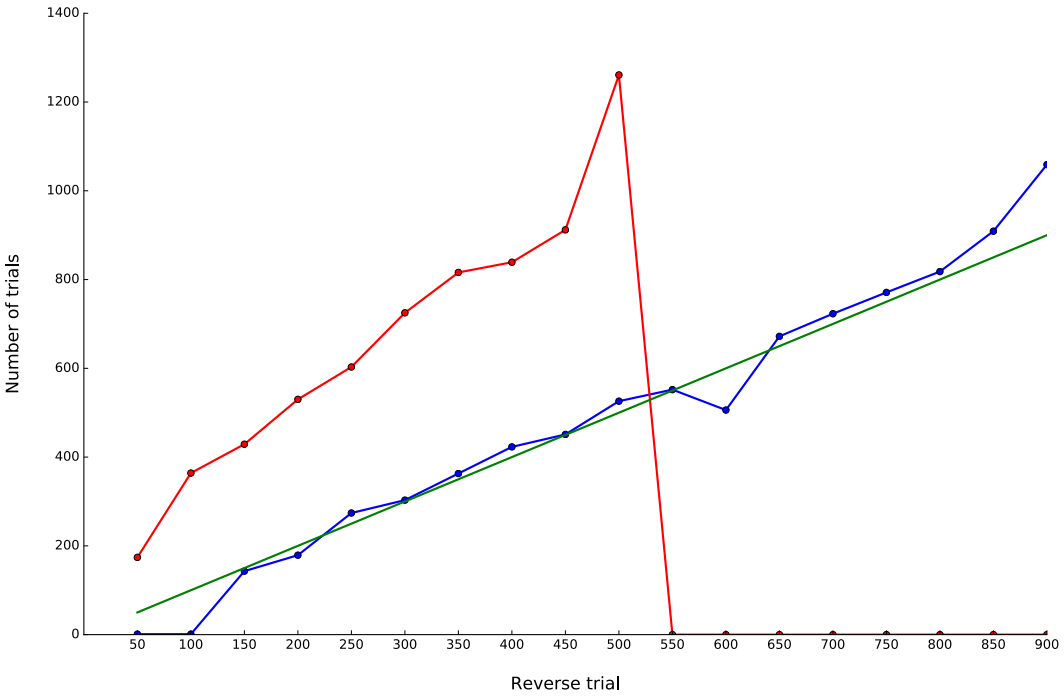


Figure 4.16: The y-axis donates how many trials are needed for the model to start exploring (blue line) or exploiting (red line) after a reverse occurring at a specific trial, donated by the x-axis. The green line is given by the equation $y = x$, and it is displayed to emphasize the fact that the exploration line is almost linear. For more details please refer to the text.

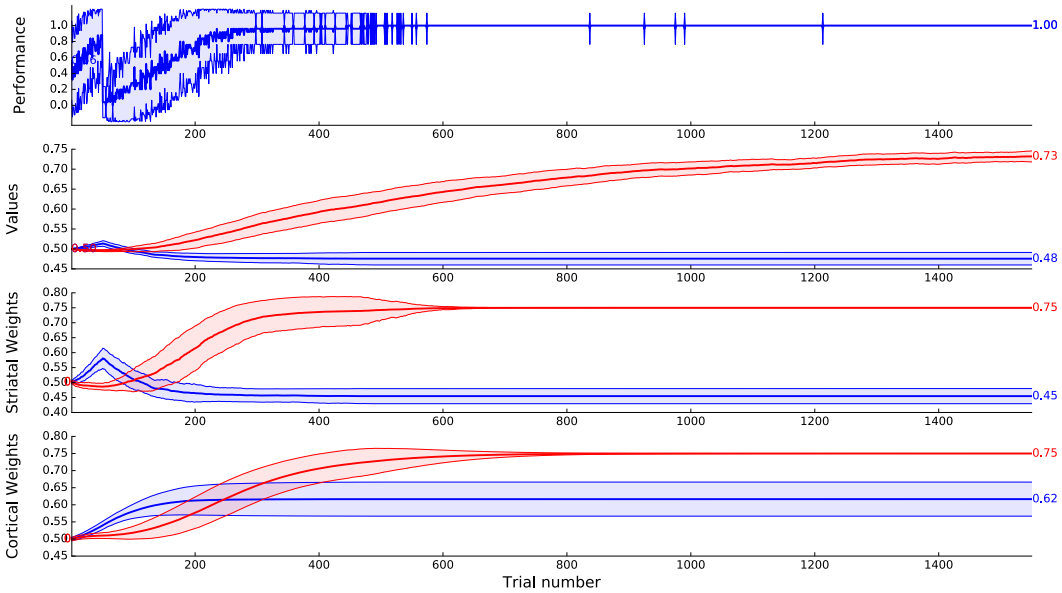
a session.

Furthermore, in the latter experiment, the strongest striatal weight begins to decrease after the reverse, but after some trials, it increases again (Figure 4.17b). This is a mathematical deficit of our implementation.

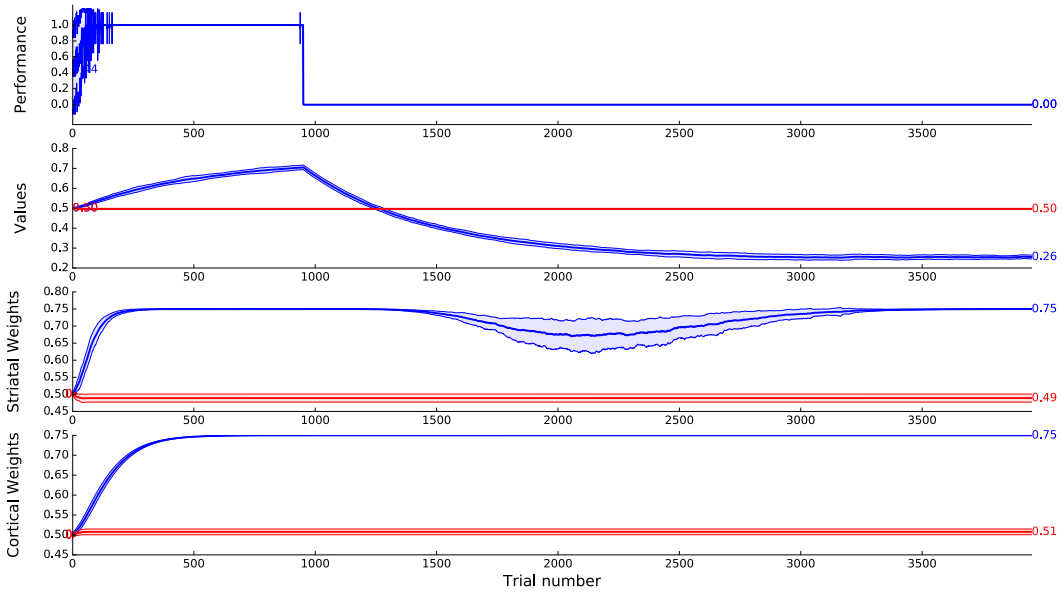
When a weight reaches the maximum level before the reverse, then after the reverse, the value of the input decreases too fast. Learning occurs when the prediction of the reward is different from the actual value. In other words, if the prediction error is zero, there is no reason for learning to follow. And, when the value of the input reaches the minimum, there is no prediction error. Consequently, when the value v in equation 3.11 reaches 0.25, then:

$$PE = R - v = R - 0.25 = \begin{cases} +\frac{3}{4}, & \text{when} \quad reward = 1 \\ -\frac{1}{4}, & \text{when} \quad reward = 0 \end{cases} \tag{4.1}$$

If we now compute the amount of change in synaptic weights for the two cases using the PE values, we derive to :



(a)



(b)

Figure 4.17: The first panel displays the performances of the model in a specific task, and the other three the evolution of the values used in RL, the cognitive cortico-striatal and cortico-cortical weights during the session. Reverse experiments at (a) $t = 50$, and (b) $t = 950$.

$$\Delta W_{A\to B} = \begin{cases} \frac{3}{4} \times \alpha_{LTP} \times U_B, & \alpha_{LTP} = 0.010, \quad \text{when} \quad reward = 1 \\ \frac{1}{4} \times \alpha_{LTD} \times U_B, & \alpha_{LTD} = 0.008, \quad \text{when} \quad reward = 0 \end{cases} \tag{4.2}$$

Now, if only the bad choice is selected, then it will be rewarded at average among trials once per four trials. In total of four trials with only one rewarded, the weight will change as follows:

$$\begin{aligned} 1 * \Delta W_{A\to B,reward=1} \quad - \quad & 3 * \Delta W_{A\to B,reward=0} = \\ & = 1 \times \frac{3}{4} \times 0.010 \times U_B - 3 \times \frac{1}{4} \times 0.008 \times U_B \\ & = (0.010 - 0.008) \times \frac{3}{4} \times U_B \\ & = 0.002 \times \frac{3}{4} \times U_B > 0 \end{aligned} \tag{4.3}$$

According to the results of 4.3, in average of four trials the model in this case of four trials will increase its weight, as we observe in Figure 4.17b. Another aftermath from 4.3 is that the greater the difference between LTP & LTD the faster the model becomes unable to recover from a reward probability change.

4.3 Experimental results

4.3.1 Protocol B: Formation of habits

After saline injections, animals were able to maximize their reward in the Habitual condition and to learn new values in the Novelty. The mean success rate (for the last 25 trials) was $99.4 \pm 3.3\%$ (Figure 4.18 A,B), [$98.8 \pm 0.6\%$ for monkey F (Figure 4.18C,D) and $100.0 \pm 0.0\%$ for monkey Z (Figure 4.18E,F)]. The difference in success rate between the two animals was not significant (unpaired t-test). In the *NC*, both animals learned progressively the difference between the two shapes (Figure 4.18A,C and E). At the beginning of training, their choices were made randomly. Although, at the end of the session the animals displayed a preference for NC_1, the target associated with the highest reward probability (mean $53.8 \pm 4.4\%$ for the first 25 trials and $93.0 \pm 2.5\%$ for the last 25 trials, Figure 4.18B). Mean reaction time in *NC* was significantly higher than in the HC (respectively $447.6ms \pm 5.6ms$ and $418.8ms \pm 4ms$, $P < 0.01$ unpaired t test, Figure 4.19A).

After muscimol injections, the success rate did not decrease significantly (mean $97.0\% \pm 1.8\%$, 4.18A,B) when compared with saline. On the other hand, in the *NC*, at the end of the session, the animals did not display any preference for either of the 2 targets after the muscimol injections (mean $42.4\% \pm 4.5\%$ to $52.0\% \pm 7.0$, $F1, 72 = 2.13, P > .05$, 4.18B). Muscimol injections in the GPi significantly increased the reaction time in both condition *HC* ($452.5ms \pm 4.2ms$) and *NC* ($495.7ms \pm 6.5ms$) when compared with the saline injections (2-way ANOVA: $F1, 76 = 47.42, P < .01$ between the 2 conditions

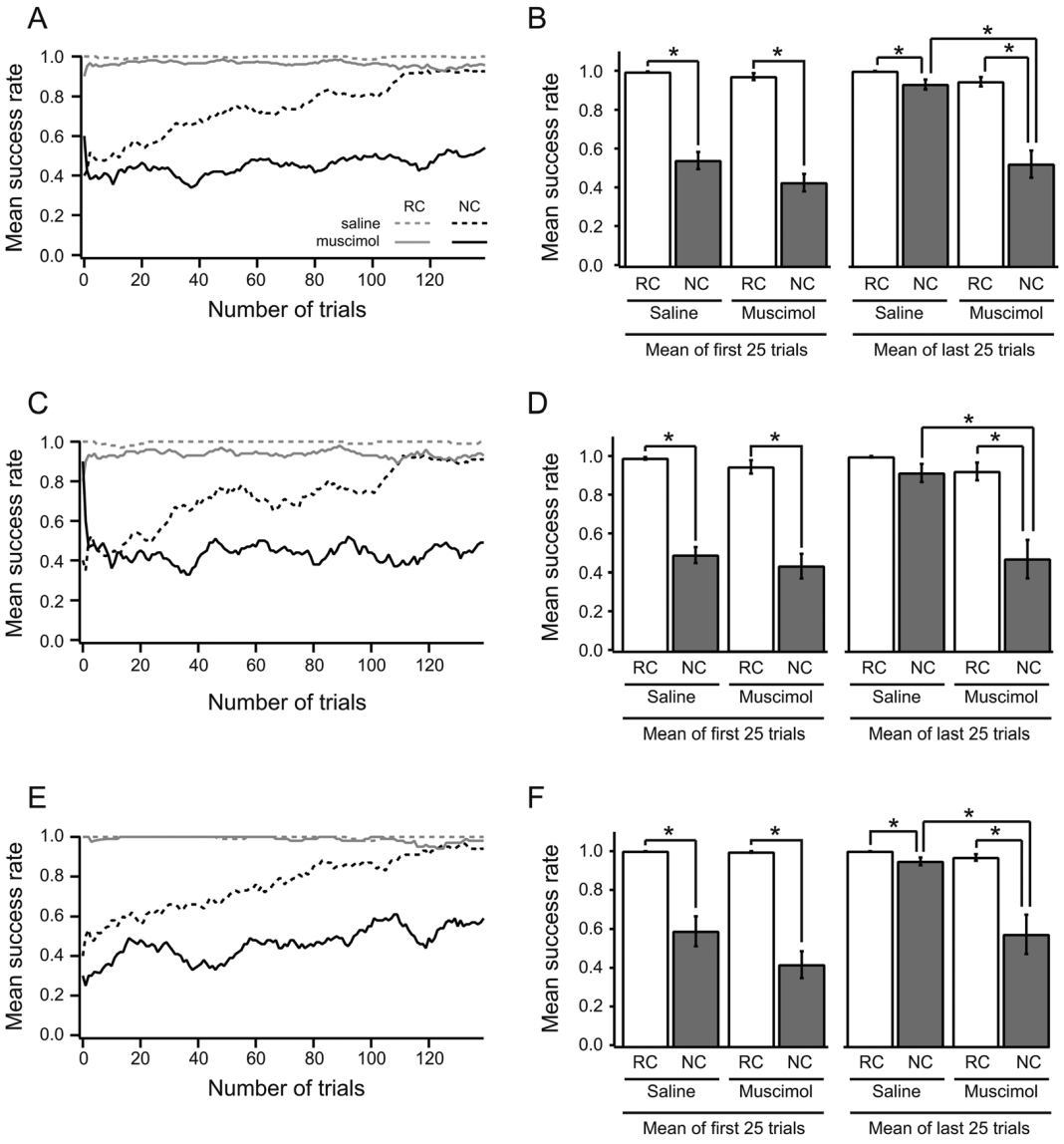


Figure 4.18: As in Piron *et al.* [2016]. Mean success rate of the monkeys. Please for more details refer to the text or the original paper.

and $F1, 76 = 61.24, P < .01$ between saline and muscimol, 4.19A). Overall these results are compatible with the results of the model.

4.3.2 Protocol C: Storage of habits

We observed similar behaviors between the model and the monkeys in this protocol. On Day 1 of the first experiment, the monkeys were injected with

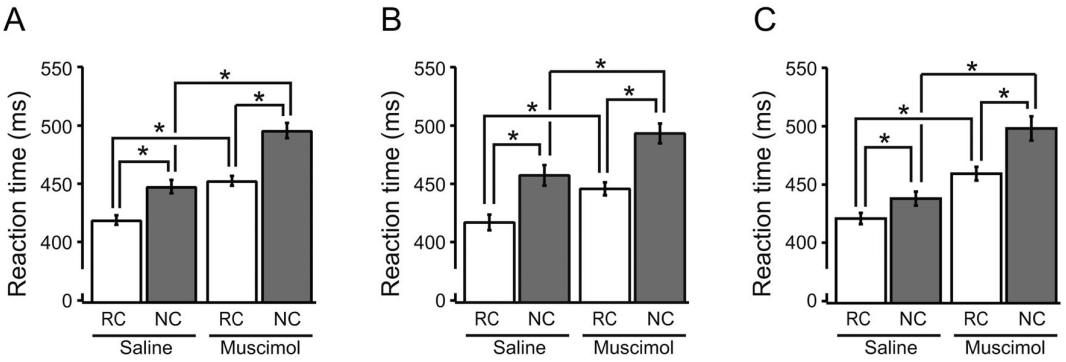


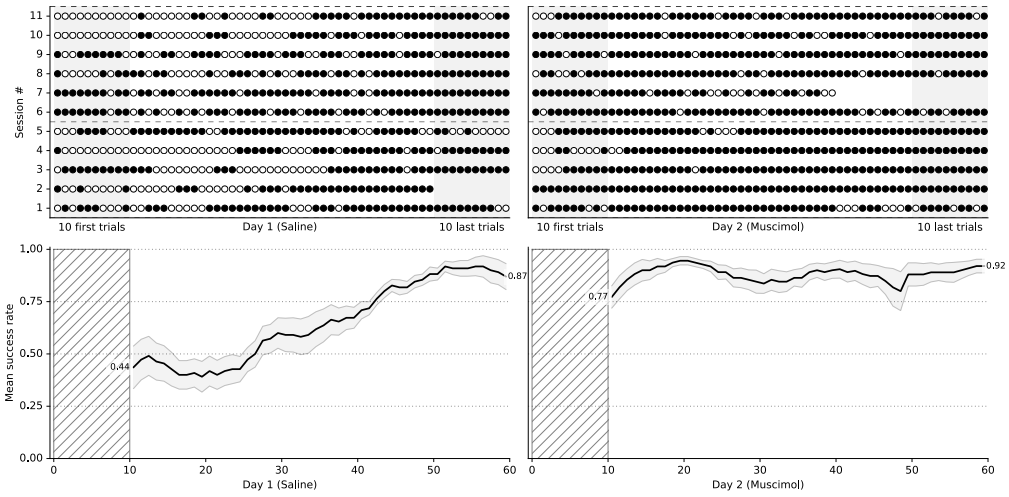
Figure 4.19: As in Piron *et al.* [2016]. Reaction time of the monkeys. Please for more details refer to the text or the original paper.

muscimol, and, as in protocol B, were unable to learn new contingencies as their performances showed [0.39 ± 0.12 in 25 first trials to 0.41 ± 0.14 in the 25 last (Figure 4.20b)]. However, the next day and under saline, they reached instantly optimal performances (0.97 ± 0.05 in 25 first and 1.0 ± 0.0 25 last trials for both).

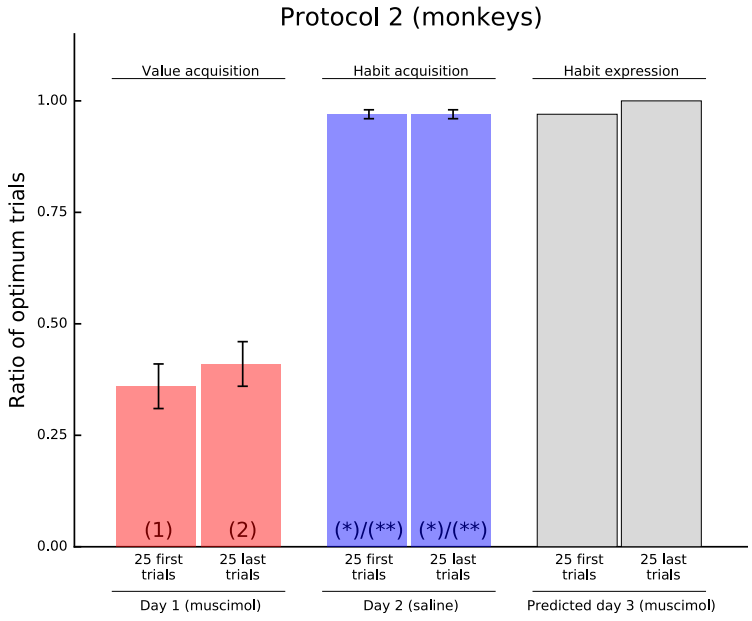
In the second experiment, the monkeys on Day 1, after saline injection, were tested on novel pair of shapes, and reached good performances at the end of the sessions (0.87 ± 3.6). On Day 2 under muscimol, the monkeys started at the same level of performances as at the end of Day 1 (0.77 ± 5.3), and improved until the end of the sessions (0.92 ± 4.5) [Figure 4.20a]. It is worth to notice that even when GPi is inactive, there is an improvement in performances, which support the theory that cortex learns based on the statistics. To explain it better, take for example the beginning of the session. From the performances of the monkeys derives that there is 72% probability of cortex to choose optimally. Say it differently, in 100 trials the subject will choose the best target 72 times and 28 the other one, so it will learn about 3 times more the best one over the other thanks to the nature of cortical learning. So in the end of these 100 trials the optimum choice will be stronger learned at the cortical level, resulting this choice to be even more probable to be chosen in later trials.

4.4 Overall interpretation of the results

The results of all protocols described in the previous section show that the model is able to select an action among available choices, and learn the one with the highest probability to be rewarded. Furthermore, the model is able to express similar behaviors with the monkeys in protocol B, where cortex is



(a)



(b)

Figure 4.20: As in Topalidou *et al.* [2016]. (a) In Protocol 1, D1 corresponds to the first day of the experiment where the habits are believed to have been acquired by the end of the day. D2 corresponds to the second day where GPi output is suppressed in the monkeys (muscimol injection). (b) In Protocol 2, D1 corresponds to the first day of the experiment where GPi output is suppressed in monkeys (muscimol injection). D2 corresponds to the second day where the suppression of the GPi output is removed. D3 (suppression of GPi) results is only a prediction based on model results and monkey protocol 1 results. They have not yet been confirmed.

able to express habits without the feedback from basal ganglia, but not to learn new contingencies. These experiments provided a framework to make prediction of the monkey behaviors in different tasks.

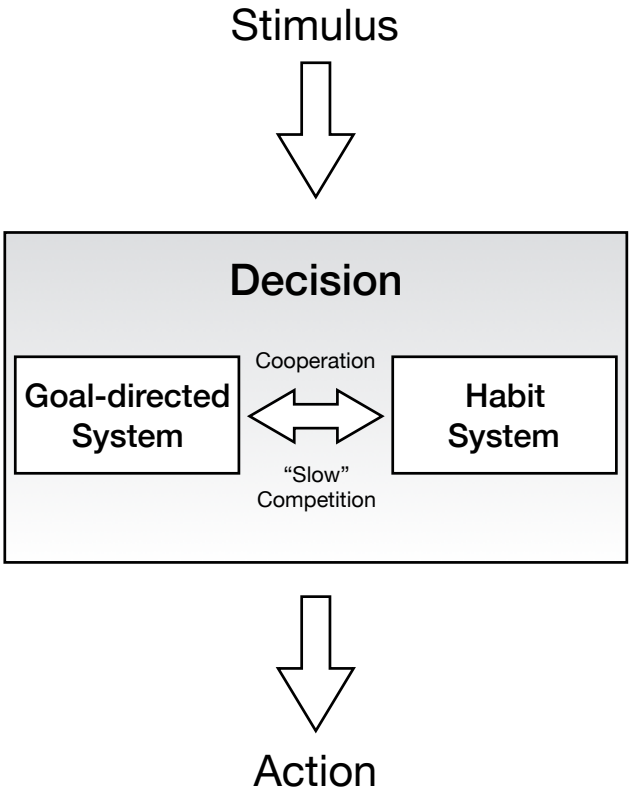


Figure 4.21: Illustration of our proposal about how a goal-directed action or a habit is chosen to be executed.

The theoretical and experimental results of protocol C confirm our prediction, by demonstrating that Hebbian and reinforcement learning can be explicitly dissociated by inactivating the output of the basal ganglia while preserving the normal function at striatal level, which we interpret as the learning of the stimuli value. When GPi is inactive the Hebbian learning at cortical level is equal for all choices, as a result of the randomness of the choice among the options in any trial; the learning there is independent of the reward. On the contrary, at striatal level the learning, which depends on the reward, following reinforcement learning rules is analogous to the outcome of the choice in each trial. Although, BG are able to learn properly in this case, they are unable to interfere into the cortical decision, resulting to chance level performances through the whole session. However, when GPi is active again, BG are able to use the learning of the previous session to teach cortex. This is possible, because BG answer is faster than the cortical, giving them time to influence

the cortical decision. Finally, after an adequate training, cortical learning is strong enough to make optimal decisions without the interference of BG (with inactive the GPi). A habit has been formed. Based on these theoretical results and in light of experimental results in the monkey, we can predict that equivalent processes underlay the habit formation in the frontal cortex of the primates. Our hypothesis is that there are two actors and one critic that participate in the acquisition of habits (Figure 4.22a). However, when the output of BG is inactive, covert learning exists inside the basal ganglia that is ready to be expressed when GPi becomes active again (Figure 4.22b).

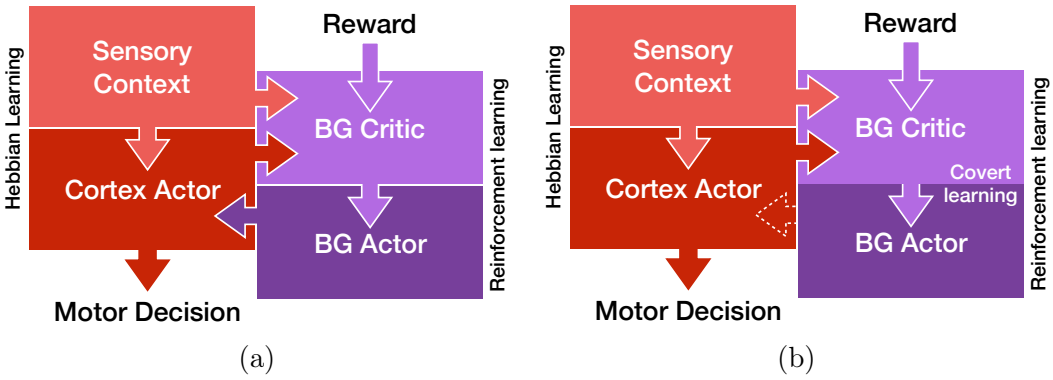


Figure 4.22: Illustration of our hypothesis that cortex and BG implement a 2 actor - 1 critic system with (a) active and (b) inactive GPi.

Even though the term habit has a long history in many different fields of neuroscience and psychology, there is still a large degree of uncertainty around the exact definition. This characterization may further vary across fields depending on the species, tasks and methodologies such that in the end, it is difficult to assess if a given behavior emanates from a habit or from another process. Our results suggest a new explanation for the primates where a behavioral decision results from both the cooperation (acquisition) and competition (expression) of two distinct but entangled memory systems. Finally, our last protocol indicates that the existing criteria for defining the type of a behavior (habit or goal-directed) should be revised to take also in account the capabilities of the species.

In conclusion, based on the theoretical and experimental results, we support the hypothesis that the two types of instrumental behavior are implemented by two distinct mechanisms. However, contrary to the hypotheses that I described in Chapter 2 (Daw *et al.* [2005]; Dezfouli and Balleine [2013]; Ashby *et al.*

[2007]), we propose that the goal-directed and habit mechanisms interact, and cooperate or compete in order the optimal action to be chosen as a result of a specific stimulus (Figure 4.21). Unlike to previous theories, when they compete each other, there is not a unique factor that indicates who will be the winner. For example, Daw *et al.* [2005] proposed that the winner is always the mechanism with the strongest salience. In our theory, this is true before the habits have been formed. In this case, the goal-directed system influences a lot the habit, and in the end is the one that leads the final decision. By contrast, after the acquisition of habits, the habit mechanism is faster that the goal-directed, resulting in being the leader of the decision, as in [Ashby *et al.*, 2007]. An outcome of our theory is that habits and goal-directed actions are not totally independent. Contrary, habits are a graded phenomenon that has as a basis goal-directed actions (Figure 4.23).

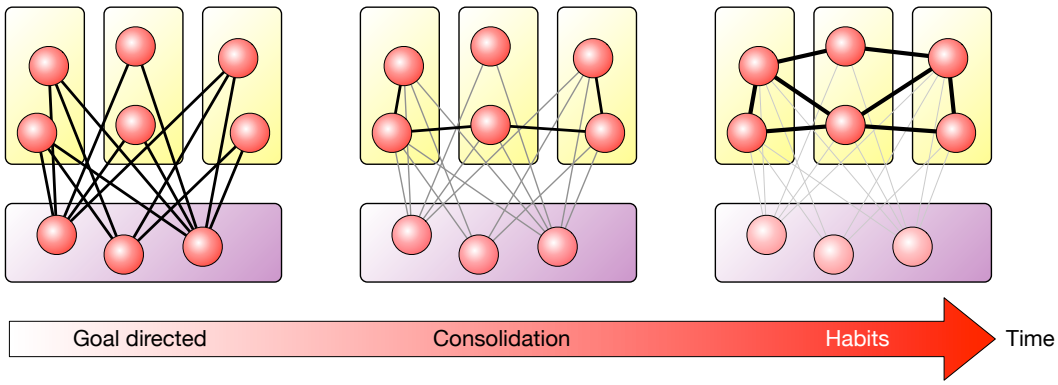


Figure 4.23: Illustration of my proposal about how a goal-directed action or a habit is chosen to be executed.

4.5 Comparison of our model with existed models

The model by Gurney *et al.* [2001a,b] and its extension by Girard *et al.* [2008] contain almost all the known structures and connections participating in the cortico-basal loop, and is undeniably more complete than ours. Its choices are driven by a 'winner-take-all' mechanism that selects the neuron with the most salient input. However, two identical inputs cannot produce a single action. In GPR model none of the actions will be chosen, and in CBG all of them will, but in both cases that is what each of them wanted to obtain. The explanation for the first model is that none of them is better than the others, so no action

should be performed, whereas for the second, they assume that it is valid more than on action to be selected, as in multitasking paradigms (walking and talking simultaneously). In our model, we assume that the proposed actions are opponents, so only one can be produced. That is the reason we added noise in our system, to break the symmetry, and ensure action selection. Another difference is that in our model we assume that the indirect pathway plays a secondary role, contrary to the hyperdirect at which we assign a primary role. Wanting to keep the complexity as simple as possible in order to focus in the investigation of the habit mechanism, we implemented the least of the structures and connections needed to obtain action selection.

On the other hand, [Ashby *et al.*, 2007] proposed an even simpler model than ours to describe the acquisition of motor habits associated with a sensory input. For this reason, they implemented a sensory cortex which projects directly to premotor area, but also indirectly through the cortico-basal-thalamic loop. Through this loop, thalamus receives input from sensory cortex and projects to premotor area. That results that at the beginning of training the selection of the move occurs at striatum, which by disinhibiting thalamus increases the cortical activity leading to the execution of the action. In our model the procedure of selection starts and ends in the cortex thanks to the closed loops with basal ganglia and thalamus. Despite the differences between the models that mostly derive from the nature of the tasks and the examined cortical areas, we both support the idea that basal ganglia lead decision in a naive model, but through training cortico-cortical connections are strengthen, and so the cortical action selection precedes the basal-ganglia one. Further, it is suggested that both the cortico-cortical pathway and the cortico-basal-thalamo-cortical pathway process the input in pre and post-learning periods, however the later pathway is faster in the beginning of training and gets the chance to influence the other one, and in the end the cortico-cortical decision is much faster that the second have no time to interfere and modify it. By an ecological point of view, it is reasonable since the first one contains only one connection of two populations, and the second four.

Baldassarre *et al.* [2013] followed the same route, proposing that habits are stored in cortico-cortical connectivity providing a repetition bias. However, through the properties of the model it is predicted that even without the help of basal ganglia the cortex can form habits, of course much more trials will be needed until it will finally learn. This is due to the cortical learning rule that they implemented, which considers that the learning occurs only when dopamine is released as a consequence of reward. This contrasts the existent bibliography suggesting that dopamine in cortical level stays for several minutes, so all the consequential actions are learned and not only the one that triggered dopamine release.

The theoretical model by Daw *et al.* [2005] propose that goal-directed actions and habits are processed separately and parallel by the two parts of dorsal striatum (associative and sensorimotor respectively). For each decision, both systems process the input, make a choice and then compete each other for expression. Through this theory a question had arisen: Which is the purpose of associative striatal learning? Our answer is that in this loop the most appropriate strategies emanated from prefrontal cortex are reinforced in order the total input from this loop to the sensorimotor, which are not totally segregated, to become stronger and influence its decision.

Finally, Dezfouli and Balleine [2013] suggested that habits are learned action sequences, triggered by a goal. They proposed that decision follow a hierarchical model that first defines if a habit or a goal-direct action will be executed, and if it is a habit, then the selected sequence will be insensitive to the received feedback among the different stages. However, they do not precise which area or loops are responsible for each mechanism, goal-directed or habits, or how habits are formed in first place.

In summary, with this work we provide a new framework to explore the interaction of the two types of learning at cortical (Hebbian) and striatal (einforcement) level. Our new hypothesis about the formation and expression of habits opens the road to reconsider the role of basal ganglia in these procedures.

"Our habits define us!
Therefore, by choosing them, we can redefine ourselves, and become whoever we have dreamed of!"

Conclusion

Habits are an important element for the production of complex and high-level behaviors that can be found in superior vertebrates. They are actions executed fast, with the minimum amount of effort (optimum use of muscles and energy), and free of attention in order to be used for more important functions; e.g. look for predators. Most of our knowledge today about habits comes from experimental studies on a variety of species, although the different species don't always share the same capabilities. For this reason, it is important to take in account this diversity of capacity among the species when we generalize the conclusions of a study on a specific species. The dominant view, nowadays, supports that habits are goal independent, which was proposed mostly from experiments on rodents. Primates, on the other hand, do not express this characteristic in non-pathological situations. For example, monkeys conduct the experiments only because they are kept under hydric deprivation to motivate them during the task. So, if they are exposed to liquids before the experiment, equivalent to the protocols for reward devaluation applied on rodents, the monkeys will have no motivation for completing the task. Concluding from the existing bibliography, the features that characterize an action as habit, in non-human and human primates, are: automaticity, unconsciousness, inflexibility, incremental acquisition, goal-oriented.

The dominant view of the 20^{th} century supported the idea that cortex is responsible for flexible thinking, and consequently for goal-directed actions. In contrast, habits are automatic, so they were thought to be generated by a less complex, subcortical structure such as the basal ganglia. Basal ganglia are a group of subcortical structures existing in all vertebrates. Also, it has been shown to participate in action selection, an ability that all vertebrates share, in contrast to invertebrates where instinct triggers the behavior. The evolution of computational modeling accelerated the research of the role about basal ganglia in instrumental tasks, as well as their underline properties. A variety of models have been developed in order to examine different aspects of basal ganglia. For example, some models study the interaction of the pathways inside BG during action selection, and others study the cortico-basal

loops investigating their role in habits and goal-directed actions, but both put a piece in the puzzle of understanding the anatomy, the physiology and the functional role of basal ganglia. In the same time, machine learning, and more specifically reinforcement learning, provided critical insight and theories of the mechanisms governing instrumental behavior. Combined evidence from theoretical and experimental studies led to a new theory, which supports the existence of two distinct, parallel cortico-basal loops responsible for habitual and goal-directed behavior. Also, this theory suggests that habits are stored in basal ganglia. Nowadays, there are many studies suggesting that habits are stored in cortex, which nevertheless needs basal ganglia during their formation. We also provided evidence supporting this theory, showing that if the main output of BG is inactivated then habits are still executed, but no new associations between a stimulus and an action can be learned. At this point, I want to highlight that the acquired knowledge we investigate in this work, derives only from instrumental learning; learning by feedback.

In this work we explored the nature of instrumental behavior through a computational, biological inspired model of cortico-basal-thalamic closed loop that we developed. We kept the complexity as simple as possible so the model can perform action selection, in order to focus on the mechanisms underlying the acquisition and expression of habits. One hypothesis derived from our model was that habits are developed parallel to the acquisition of proper action-outcome association necessary for achieving better performances in goal-directed actions. A lot of evidence suggests that cortical learning is based on Hebbian learning rules, *i.e.* every choice is learned independently to its outcome, and basal learning on reinforcement learning rules, *i.e.* each choice is evaluated depending on its output and reinforced according to this evaluation. Therefore, we argued that cortex learns by the statistics provided implicitly by early basal ganglia action selection. To explain it better, when basal ganglia have learned the appropriate action for a stimulus, they influence cortical decision. In this way, the optimal choice is over-represented at cortical level; *e.g.* if the optimal cue is chosen 100 times and another one only 10, then cortex will learn the optimal 10 times more. However, when cortical learning is strong enough, the decision in cortical level is fast enough to prevent the interference of basal ganglia. To test our hypothesis, we designed a protocol of a two arm-bandit task, which contained three parts. At the first part, new contingencies, associated with reward probability, were presented to the system, after the inactivation of GPi (one of the basal ganglia output). The model was unable to explicitly select the stimulus associated with the highest probability of reward. Its answers were random. Because of the properties of the model, we know that this occurred because basal ganglia could not influence the decision anymore, resulting in cortical learning of each performed action to be learned at cortical level, because cortical learning is independent

of the action outcome. However, we hypothesized that the value of the stimuli were covertly learned inside the BG (cortical-striatal connections) even though BG was unable to influence the decisions. To test this theory, in the second part of the experiment we reactivated GPi and presented the same input. As expected, the model was able to make optimal choices from the beginning of the session, proving that, during the first part, striatum had already learned the values of the stimuli. Finally, the aim of the last part was to verify that basal ganglia were able to train cortex and not just lead the decision during the second part. For this reason, we inactivated once more GPi, and present again the same options. We observed that from the beginning of the session the model chose nearly optimally. This experiment additionally has been conducted on monkeys, which expressed the same behaviors in the three parts as the model. The experimental results confirmed our prediction, which suggests that at the beginning of learning, BG are responsible to train cortex by leading cortical decision, but after learning has occurred BG role is to accelerate cortical decision without being necessary anymore for the decision per se. Furthermore, this protocol provides a way to dissociate cortical Hebbian learning from striatal reinforcement learning that are normally entangled.

Indubitably, our model has limitations. First of all, a lot of details in anatomical and physiological level are missing, as for example the indirect pathway, striatal interneurons or connections from GPi or thalamus back to striatum. Now that we have an idea for the role of basal ganglia as an entity, these additions could expose the intrinsic mechanisms. Also, the notion of context is not included. This means that the model would express habits whatever the context. For example, at the sight of a switch, the action to turn the light on is triggered only when it is combined with a dark room. However, our model would produce the action even if the simulated room was not dark, because it can only associate the sight of the switch with the appropriate action, independent of the context. Furthermore, the model cannot relearn after extensive training, contrary to humans or primates. This is due to our implementation of basic reinforcement and Hebbian learning rules. One reason is that at the cortical level there is no forgetful element. That means that even if after reward devaluation of the pre-learned choice the model learns the new optimal choice owing to BG, it will be unable to express habits without the BG, because the two choices will have been learned equally at cortical level. Another limitation of the model is that the salience of a stimuli can take precedence over its value. This means that the model might be unable to choose the best stimulus (in terms of probability of reward) if the worst stimulus is made salient enough. If a stimulus salience is strong enough, the model will choose this one, even though another stimulus could have bigger value. Also, the choice between two stimuli presented in a different timing depends on the amount of the interval timing. If the difference of presentation is small,

then the model will choose based on the values. Whereas if the interval is big enough, then the model will choose the first presented. Same behaviors can be observed also in humans. For example, lets say that a piece of cake is given, and after some time a delicious ice-cream. Even if the subject has started to reach the piece of the cake, he will change his mind and go for the ice-cream, if it is given after few seconds. However, if the ice-cream is presented to the subject after an hour, it will be irrelevant for his first and only choice at that moment, so he will reach the piece of cake.

Finally with this work, we provided a new framework to explore the interaction of the two types of learning, Hebbian and reinforcement, and set the foundation for new experiments. For example, a way to investigate the relative strength of these two types of learning on instrumental behavior would be to display a single stimulus (forced choice) with a given reward probability. By controlling the number of time a specific stimulus has been presented versus the associated reward probability, we could measure the relative influence of reinforcement learning versus Hebbian learning. If one stimulus with high probability is presented with low frequency, but another is associated with low probability and presented with high frequency, then the model predicts that the choice between the two stimuli will depend on the probability if a session is small, but on the frequency if it contains many trials. Furthermore, the model has been extended in the framework of a neuroeconomic task that requires a two steps decision. This means the reward is obtained only after the second action. By slightly modifying the semantics of each input structure (motor, cognitive, associative cortices), the model has been shown to be able to solve the task. However, after the learning of the action sequences, the model is unable to separate the actions and execute them in similar tasks. It actually uses a strategy-based rather than a model-based decision and this limits strongly the scope of the model. Finally, in our model we hypothesized that habits are stored in cortex, although today we don't have solid evidence for this assumption. In order to assess if Hebbian learning is actually responsible for the storage of habits, it would be necessary to inactivate associative learning in dorsolateral prefrontal or orbitofrontal cortex and check whether monkeys would complete the task without developing habits. This is ongoing work but there is no data at the time of writing.

Despite this promising start, our model needs further experiments to be confirmed. Nevertheless it is noticeable that it reverses the old idea that automatism is a sub-cortical feature. The fact that automatic input/output association occurs at cortical level, bypassing a long sub-cortical journey and therefore saving cognitive resources is a strong ecological argument. If our model is confirmed by further experiments, it opens new questions such as: i) is it a mammal specificity? ii) a primate specificity? iii) how such automatisms

are implemented or even are they implemented in other vertebrates?

In the end, my dog learned the new path, so habits can be reversed!

Appendix A

Parameters Table

A.1 Guthrie *et al.* [2013]

A Model Summary	
Populations	Twelve: Cortex (motor, associative & cognitive), Striatum (motor, associative & cognitive), GPi (motor & cognitive), STN (motor & cognitive), Thalamus (motor & cognitive)
Topology	–
Connectivity	one to one, one to many (divergent), many to one (convergent)
Neuron model	Dynamic rate model
Channel model	–
Synapse model	Linear synapse
Plasticity	Reinforcement learning rule
Input	External current in cortical areas (motor, associative & cognitive)
Measurements	Firing rate

B Populations					
Name	**Elements**	**Size**	**Threshold** (h)	**Noise**	**Initial state**
Cortex motor	Linear neuron	1×4	-3	5.0%	0.0
Cortex cognitive	Linear neuron	4×1	-3	5.0%	0.0
Cortex associative	Linear neuron	4×4	-3	5.0%	0.0
Striatum motor	Sigmoidal neuron	1×4	0	5.0%	0.0
Striatum cognitive	Sigmoidal neuron	4×1	0	5.0%	0.0
Striatum associative	Sigmoidal neuron	4×4	0	5.0%	0.0
GPi motor	Linear neuron	1×4	-10	5.0%	0.0
GPi cognitive	Linear neuron	4×1	-10	5.0%	0.0
STN motor	Linear neuron	1×4	-10	5.0%	0.0
STN cognitive	Linear neuron	4×1	-10	5.0%	0.0
Thalamus motor	Linear neuron	1×4	-40	5.0%	0.0
Thalamus cognitive	Linear neuron	4×1	-40	5.0%	0.0
Values (V_i)	Scalar	4	–	–	0.5

C Connectivity					
Source	**Target**	**Pattern**	**Weight**	**Gain**	**Plasticity**
Cortex motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	0.1	-
Cortex cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	0.1	-
Cortex motor	STN motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
Cortex cognitive	STN cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Cortex motor	Striatum motor	$(1, i) \rightarrow (1, i)$	0.5	1.0	-
Cortex cognitive	Striatum cognitive	$(i, 1) \rightarrow (i, 1)$	0.5	1.0	(F1)
Cortex motor	Striatum associative	$(1, i) \rightarrow (*, i)$	0.5	0.2	-
Cortex cognitive	Striatum associative	$(i, 1) \rightarrow (i, *)$	0.5	0.2	-
Cortex associative	Striatum associative	$(i, j) \rightarrow (i, j)$	0.5	1.0	-
Thalamus motor	Cortex motor	$(1, i) \rightarrow (1, i)$	1.0	0.4	-
Thalamus cognitive	Cortex cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	0.4	-
GPi motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	-0.3	-
GPi cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-0.3	-
STN motor	GPi motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
STN cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Striatum cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum motor	GPi motor	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum associative	GPi motor	$(*, i) \rightarrow (1, i)$	1.0	-2.0	-
Striatum associative	GPi cognitive	$(i, *) \rightarrow (i, 1)$	1.0	-2.0	-
Cortex motor	Cortex motor	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex cognitive	Cortex cognitive	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex associative	Cortex associative	$(i, j) \rightarrow (*, *)$	1.0	-0.5	-
Cortex motor	Cortex associative	$(1, i) \rightarrow (*, i)$	0.5	0.03	-
Cortex associative	Cortex cognitive	$(i, *) \rightarrow (i, 1)$	0.5	0.03	(F2)
Cortex cognitive	Cortex associative	$(i, 1) \rightarrow (i, *)$	0.5	0.03	-
Cortex associative	Cortex motor	$(*, i) \rightarrow (1, i)$	0.5	0.03	-

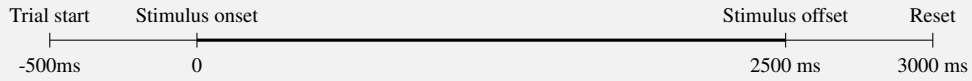
D1 Neuron Model	
Name	Linear neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = max(V, 0)$

D2 Neuron Model	
Name	Sigmoidal neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = V_{min} - (V_{max} - V_{min}) / \left(1 + e^{\frac{V_h - V}{V_c}}\right)$

E Synapse	
Name	Linear synapse
Type	Weighted sum
Output	$I^B_{syn} = \sum_{A \in sources}(G_{A \to B} W_{A \to B} U_A)$

F1 Plasticity	
Name	Reinforcement learning
Type	Delta rule
Delta	$\Delta W_{A \to B} = \alpha \times PE \times U_B$
	$PE = Reward - V_i$
	$\alpha = 0.008$ if $PE < 0$ (LTD), $\alpha = 0.01$ if $PE > 0$ (LTP)
	$\Delta V_i = \beta \times PE, \beta = 0.002$

F2 Plasticity	
Name	Hebbian learning
Type	Hebb rule
Delta	$\Delta W_{A \to B} = \alpha \times U_A \times U_B, \alpha = 0.001$

G Input	
Type	Cortical input
Description	A trial is preceded by a settling period (500ms) and followed by a reset period. At time $t = 0$, two shapes are presented in cortical cognitive area ($I_{ext} = 7$ at $\{i_1, i_2\}$) at two different locations in cortical motor area ($I_{ext} = 7$ at $\{j_1, j_2\}$) and the cortical associate area is updated accordingly ($I_{ext} = 7$ at $\{i_1, i_2\} \times \{j_1, j_2\}$).
Timing	

H Measurements	
Site	Cortical areas
Data	Activity in cognitive and motor cortex
	Cortico-striatal weights

I Environment	
OS	OSX 10.11 (El Capitan)
Language	Python 3.5.1 (brew installation)
Libraries	Numpy 1.10.2 (pip installation)
	Cython 0.23.4 (pip installation)
	Matplotlib 1.5.0 (pip installation)

A.2 Topalidou *et al.* [2016]

A Model Summary	
Populations	Twelve: Cortex (motor, associative & cognitive), Striatum (motor, associative & cognitive), GPi (motor & cognitive), STN (motor & cognitive), Thalamus (motor & cognitive)
Topology	–
Connectivity	one to one, one to many (divergent), many to one (convergent)
Neuron model	Dynamic rate model
Channel model	–
Synapse model	Linear synapse
Plasticity	Reinforcement learning rule
Input	External current in cortical areas (motor, associative & cognitive)
Measurements	Firing rate

B Populations					
Name	**Elements**	**Size**	**Threshold** (h)	**Noise**	**Initial state**
Cortex motor	Linear neuron	1×4	-3	5.0%	0.0
Cortex cognitive	Linear neuron	4×1	-3	5.0%	0.0
Cortex associative	Linear neuron	4×4	-3	5.0%	0.0
Striatum motor	Sigmoidal neuron	1×4	0	5.0%	0.0
Striatum cognitive	Sigmoidal neuron	4×1	0	5.0%	0.0
Striatum associative	Sigmoidal neuron	4×4	0	5.0%	0.0
GPi motor	Linear neuron	1×4	-10	5.0%	0.0
GPi cognitive	Linear neuron	4×1	-10	5.0%	0.0
STN motor	Linear neuron	1×4	-10	5.0%	0.0
STN cognitive	Linear neuron	4×1	-10	5.0%	0.0
Thalamus motor	Linear neuron	1×4	-40	5.0%	0.0
Thalamus cognitive	Linear neuron	4×1	-40	5.0%	0.0
Values (V_i)	Scalar	4	–	–	0.5

C Connectivity					
Source	**Target**	**Pattern**	**Weight**	**Gain**	**Plasticity**
Cortex motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	0.1	-
Cortex cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	0.1	-
Cortex motor	STN motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
Cortex cognitive	STN cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Cortex motor	Striatum motor	$(1, i) \rightarrow (1, i)$	0.5	1.0	-
Cortex cognitive	Striatum cognitive	$(i, 1) \rightarrow (i, 1)$	0.5	1.0	(F1)
Cortex motor	Striatum associative	$(1, i) \rightarrow (*, i)$	0.5	0.2	-
Cortex cognitive	Striatum associative	$(i, 1) \rightarrow (i, *)$	0.5	0.2	-
Cortex associative	Striatum associative	$(i, j) \rightarrow (i, j)$	0.5	1.0	-
Thalamus motor	Cortex motor	$(1, i) \rightarrow (1, i)$	1.0	0.4	-
Thalamus cognitive	Cortex cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	0.4	-
GPi motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	-0.3	-
GPi cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-0.3	-
STN motor	GPi motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
STN cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Striatum cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum motor	GPi motor	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum associative	GPi motor	$(*, i) \rightarrow (1, i)$	1.0	-2.0	-
Striatum associative	GPi cognitive	$(i, *) \rightarrow (i, 1)$	1.0	-2.0	-
Cortex motor	Cortex motor	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex cognitive	Cortex cognitive	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex associative	Cortex associative	$(i, j) \rightarrow (*, *)$	1.0	-0.5	-
Cortex motor	Cortex associative	$(1, i) \rightarrow (*, i)$	0.5	0.03	-
Cortex associative	Cortex cognitive	$(i, *) \rightarrow (i, 1)$	0.5	0.03	(F2)
Cortex cognitive	Cortex associative	$(i, 1) \rightarrow (i, *)$	0.5	0.03	-
Cortex associative	Cortex motor	$(*, i) \rightarrow (1, i)$	0.5	0.03	-

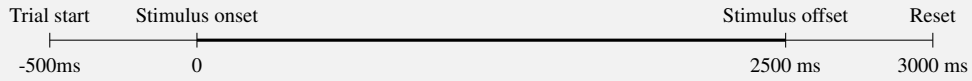
D1 Neuron Model	
Name	Linear neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = max(V, 0)$

D2 Neuron Model	
Name	Sigmoidal neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = V_{min} - (V_{max} - V_{min}) / \left(1 + e^{\frac{V_h - V}{V_c}}\right)$

E Synapse	
Name	Linear synapse
Type	Weighted sum
Output	$I^B_{syn} = \sum_{A \in sources}(G_{A \to B} W_{A \to B} U_A)$

F1 Plasticity	
Name	Reinforcement learning
Type	Delta rule
Delta	$\Delta W_{A \to B} = \alpha \times PE \times U_B$
	$PE = Reward - V_i$
	$\alpha = 0.008$ if $PE < 0$ (LTD), $\alpha = 0.01$ if $PE > 0$ (LTP)
	$\Delta V_i = \beta \times PE, \beta = 0.002$

F2 Plasticity	
Name	Hebbian learning
Type	Hebb rule
Delta	$\Delta W_{A \to B} = \alpha \times U_A \times U_B, \alpha = 0.001$

G Input	
Type	Cortical input
Description	A trial is preceded by a settling period (500ms) and followed by a reset period. At time $t = 0$, two shapes are presented in cortical cognitive area ($I_{ext} = 7$ at $\{i_1, i_2\}$) at two different locations in cortical motor area ($I_{ext} = 7$ at $\{j_1, j_2\}$) and the cortical associate area is updated accordingly ($I_{ext} = 7$ at $\{i_1, i_2\} \times \{j_1, j_2\}$).
Timing	

H Measurements	
Site	Cortical areas
Data	Activity in cognitive and motor cortex
	Cortico-striatal weights

I Environment	
OS	OSX 10.11 (El Capitan)
Language	Python 3.5.1 (brew installation)
Libraries	Numpy 1.10.2 (pip installation)
	Cython 0.23.4 (pip installation)
	Matplotlib 1.5.0 (pip installation)

Appendix B

Articles

* M. Topalidou, D. Kase, T. Boraud, and N. P. Rougier, "Dissociation of reinforcement and Hebbian learning induces covert acquisition of values in the basal ganglia," under review in PLOS Computational Biology, 2016.

* C. Piron, D. Kase, M. Topalidou, M. Goillandeau, N. P. Rougier, and T. Boraud, "The globus pallidus pars interna in goal-oriented and routine behaviors: Resolving a long-standing paradox," Movement Disorders, 2016.

* M. Topalidou and N. P. Rougier, "[Re] Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study," ReScience, vol. 1, no. 1, 2015.

* M. Topalidou, A. Leblois, T. Boraud, and N. P. Rougier, "A Long Journey into Reproducible Computational Neuroscience," Frontiers in Computational Neuroscience, vol. 9, no. 28, 2015.

Bibliography

Aizman, O., Brismar, H., Uhlen, P., Zettergren, E., Levey, A.I., Forssberg, H., Greengard, P. and Aperia, A., 2000. Anatomical and physiological evidence for D1 and D2 dopamine receptor colocalization in neostriatal neurons. *Nature Neuroscience*, 3.

Albin, R.L., Young, A.B. and Penney, J.B., 1989. The functional anatomy of basal ganglia disorders. *Trends in Neurobiology*, 12(10):366.

Alexander, G., Cruthcer, M. and De Long, M., 1991. Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" fuctions. *Prog Brain Res*, 85:119–146.

Alexander, G.E., DeLong, M.R. and Strick, P.L., 1986. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci*, 9:357–381.

Arbuthnott, G.W., Ingham, C.A. and Wickens, J.R., 2000. Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy*, 196:587–596.

Ashby, G.F. and Ennis, J.M., 2006. The role of the basal ganglia in category learning. *Psychol. Learn. Mem.*, 46:1–36.

Ashby, G.F., Ennis, J.M. and Spiering, B.J., 2007. A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, 114(3):632–656. doi:10.1037/0033-295X.114.3.632.

Ashby, G.F., Turner, B.O. and Horvitz, J.C., 2010. Cortical and basal ganglia contriburions to habit learning and automaticity. *Trends in Cognitive Sciences*, 14(5):208–215. doi:10.1016/j.tics.2010.02.001.

Baldassarre, G., Mannella, F., Fiore, V.G., Redgrave, P., Gurney, K. and Mirolli, M., 2013. Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks*, 41:168–187. doi:10.1016/j.neunet.2012.09.015.

Balleine, B.W., Delgado, M.R. and Hikosaka, O., 2007. The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, 27(31):8161–8165. doi:10.1523/JNEUROSCI.1554-07.2007.

Balleine, B.W., Lijeholm, M. and Ostlund, S.B., 2009. The integrave function of the basal ganglia in instrumental conditioning. *Behav. Brain Res.*, 199:43–52.

Balleine, B.W. and O'Doherty, J.P., 2010. Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology REVIEWS*, 35:48–69. doi:10.1038/nnp.2009.131.

Bar-Gad, I., Morris, G. and Bergman, H., 2003. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71:439–473.

Barandiaran, X. and Di Paolo, E. A., 2014. A genealogical map of the concept of habit. *Frontiers in Human Neuroscience*, 8(522). doi:10.3389/fnhum.2014.00522.

Barto, A.G., 1995. Adaptive critics and the basal ganglia. In *J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), Models of Information Processing in the Basal Ganglia.* MIT Press, Cambridge, MA.

Belin, D., Jonkman, S., Dickinson, A., Robbins, T.W. and Everitt, B.J., 2009. Parallel and interactive learning processes with the basal ganglia: Relevance for the understanding of addiction. *Behavioural Brain Research*, 199:89–102. doi:10.1016/j.bbr.2008.09.027.

Benabid, A.L., 2003. Deep brain stimulation for parkisnon's disease. *Current Opinion in Neurobiology*, 13:696–706.

Bernacer, J. and J.I., Murillo, 2014. The aristotelian conception of habit and its contribution to human neuroscience. *Frontiers in Human Neuroscience*, 8(883). doi:10.3389/fnhum.2014.00883.

Bevan, M., Booth, P., Eaton, S. and Bolam, J., 1998. Selective innervation of neostriatal interneurons by a subclass of neurons in the globus pallidus of rats. *Journal of Neuroscience*, 18(22):9438–9452.

Boraud, T., 2015. *Matière à décision.* CNRS Editions, Paris, France.

Box, G.E.P and Draper, N.R., 1987. *Empirical Model-Building and Response Surfaces.* New York, John Wiley and Sons.

Bradfield, L.A., Bertran-Gonzalez, J., Chieng, B. and Balleine, B.W., 2013. The thalamostriatal pathway and cholinergic control of goal-directed action: Interlacing new with existing learning in the striatum. *Neuron*, 79:153–166. doi:http://dx.doi.org/10.1016/j.neuron.2013.04.039.

Brovelli, A., Nazarian, B., Meunier, M. and Boussaoud, D., 2011. Differential roles of caudate nucleus and putamen during instrumental learning. *NeuroImage*, 57:1580–1590. doi:10.1016/j.neuroimage.2011.05.059.

Calabresi, P., Picconi, B., Tozzi, A., Ghiglieri, V. and Di Filippo, M., 2006. Direct and indirect pathways of basal ganglia: a critical reappraisal. *Nature Neuroscience*, 17(8):1022–1030. doi:10.1038/nn.3743.

Calabresi, P., Pisani, A., Mercuri, N.B. and Bernardi, G., 1996. The corticostriatal projection: From synaptic plasticity to dysfunctions of the basal ganglia. *Trends in Neurosciences*, 19:19–24.

Carelli, R. M., Wolske, M. and West, M. O., 1997. Loss of lever press-related firing of rat striatal forelimb neurons after repeated sessions in a lever pressing task. *Journal of Neuroscience*, 17:1804–1814.

Cazorla, M., Delmondes de Carvalho, F., Chohan, M.O., Shegda, M., Chuhma, N., Rayport, S., Ahmari, S.E., Moore, H. and Kellendonk, C., 2014. Dopamine D2 receptors regulate the anatomical and functional balance of basal ganglia circuitry. *Neuron*, 81:153–164. doi:10.1016/j.neuron.2013.10.041.

Chomsky, N., 1959. Reviews: Verbal behavior by B. F. Skinner. *Language*, 35(1):26–58.

Cisek, P., 2007. Cortical mechanisms of action selection: the affordance competition hypothesis. *Phil. Trans. R. Soc. B*, 362:1585–1599. doi:10.1098/rstb.2007.2054.

Cohen, M.X. and Frank, M.J., 2009. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav. Brain Res.*, 199:141–156.

Cohen, N.J. and Squire, L.R., 1980. Preserved learning and retention of pattern-analyzing skill in amesi: dissociation of knowing how and knowing that. *Science*, 210:207–210.

Colwill, R. M. and Rescorla, R. A, 1995. Associative structures in instrumental learning. In *G.H. Bower (Ed.), The psychology of learning and motivation*, volume 20, pages 55–104. New York: Academic Press.

Cui, G., Jun, S.B., Jin, X., Pham, M.D., Vogel, S.S., Lovinger, D.M. and Costa, R.M., 2013. Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, 494:238–242. doi:10.1038/nn.3743.

Da Cunha, C., Gomez-A, A. and Blaha, C.D., 2012. The role of basal ganglia in motivated behavior. *Rev. Neurosci.*, 23(5-6):747–767. doi:10.1515/revneuro-2012-0063.

Daw, N. D., Niv, Y. and Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatl systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711. doi: 10.1038/nn1560.

Daw, N.D. and Doya, K., 2006. The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.*, 16:199–204.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. and Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69:1204–1215.

Dayan, P. and Berridge, K.C., 2014. Mode-based and model-free pavlovian reward learning: Revaluatio, revision, and revelation. *Cogn Affect Behav Neurosci*, 14(2):pp 473–492. doi: 10.3758/s13415-014-0277-8.

Delgado, M.R., Li, J., Schiller, D. and Phelps, E.A., 2008. The role of the striatum in aversive learning and aversive prediction errors. *Phil. Trans. R. Soc. B*, Theme Issue 'Neuroeconomics'. doi:10.1098/rstb.2008.0161.

DeLong, M.R., 1990. Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, 13:281–285.

Deniau, JM., Menetrey, A. and S., Charpier, 1996. Thelamellarorganizationofthe rat substantia nigra pars reticulata: segregated patterns of striatal afferents and relationship to the topography of corticostriatal projections. *Neuroscience*, 73:761–781.

Dezfouli, A. and Balleine, B.W., 2013. Actions, action sequences and habits: Evidence that goal-dierected and habitual action control are hierarchically organized. *PLOS Comput Biol*, 9(12). doi: 10.1371/journal.pcbi.1003364.

Dickinson, A., 1985. Actions and habits: the development of behavioural autonomy. *Phil. Trans. R. Soc. Lond. B*, 308:67–78.

Dickinson, A. and Balleine, B. W., 2010. Hedonics: The cognitive–motivational interface. In *M. L. Kringelbach and K. C. Berridge (Eds.), Pleasures of the brain*. Oxford: Oxford University Press., 74–84.

Dickinson, A. and Balleine, B.W., 1993. Actions and responses: The dual psychology of behaviour. In *Eilan, N. and McCarthy, R.A. (Eds), Spatial Representation: Problems in Philosophy and Psychology.*, page 277–293. MA: Blackwell Publishers.

Ding, L. and Gold, J.I., 2013. The basal ganglia's contributions to perceptual decision making. *Neuron*, 79. doi:http://dx.doi.org/10.1016/j.neuron.2013.07.042.

Dolan, R.J. and Dayan, P., 2013. Goals and habits in the brain. *Neuron*, 80:312–325. doi:http://dx.doi.org/10.1016/j.neuron.2013.09.007.

Doya, K., 1999. What are the computations of the cerebellum and the basal ganglia and the cerebral cortex? *Neuron Networks*, 12:961–974.

Feenstra, M.S. and Botterblom, M.H., 1996. Rapid sampling of extracellular dopamine in the rat prefrontal cortex during food consumption, handling and exposure to novelty. *Brain Res.*, 742:17–24.

Fodor, J., 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Foerde, K. and Shohamy, D., 2011. The role of the basal ganglia in learning and memory: Insight from parkinson's disease. *Neurobiology of Learning and Memory*, 96:624–636. doi:10.1016/j.nlm.2011.08.006.

Frank, M.J., 2005. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and non-mediacated parkinsonism. *J. Cogn. Neurosci.*, 17:51–72.

Friend, D.M. and Kravitz, A.V., 2014. Working together: basal ganglia pathways in action selection. *Trends in Neurosciences*, 37(6):301–303. doi:10.1016/j.tins.2014.04.004.

Fuster, J. M., 2001. The prefrontal cortex-an update: Time is of the essence. *Neuron*, 30:319–333.

Gale, J.T., Amirnovin, R., Williams, Z.M., Flaherty, A.W. and Eskandar, E.N., 2008. From symphony to cacophony: Pathophysiology of the human basal ganglia in parkinson disease. *Neuroscience and Biobehavioral Reviews*, 32:378–387. doi:10.1016/j.neubiorev.2006.11.005.

Gerfen, C.R., Engber, T.M., Mahan, L.C., SUsel, Z., Chase, T.N., Frederick, J., Monsma, J. and Sibley, D. R., 1990. D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, 250:1429–1432.

Gillan, C.M., Papmeyer, M., Morein-Zamir, S., Sahakian, B.J., Fineberg, N.A., Robbins, T.W. and de Wit, S., 2011. Disruption in the balance between goal- directed behavior and habit learning in obsessive-compulsive disorder. *Am. J. Psychiatry*, 168:718–726.

Girard, B., Tabareaua, N., Phama, Q.C., Berthoza, A. and Slotineb, J.-J., 2008. Where neuroscience and dynamic system theory meet autonomous robotics: A contracting basal ganglia model for action selection. *Neural Networks*, 21:628–641.

Graf, P. and Schacter, D.L., 1985. Implicit and explicit memory for new associations in normal and amnesic subjects. *J. Exp. Psychol. Learn. Mem. Cogn.*, 11:501–518.

Graybiel, A. M., 2005. The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology*, 15:638–644. doi: 10.1016/j.conb.2005.10.006.

Graybiel, A.M., 2008. Habits, rituals and the evaluative brain. *Anuu. Rev. Neurosci.*, 31:359–387. doi:10.1146/annurev.neuro.29.051605.112851.

Graybiel, A.M., Aosaki, T., Flaherty, A. and Kimura, M., 1994. The basal ganglia and adaptive motor control. *Science*, 265:1826–1831.

Graybiel, A.M. and Grafton, S.T., 2015. The striatum: Where skills and habits meet. *Cold Spring Harb Perspect Biol*, 7. doi: 10.1101/chperspect.a021691.

Gurney, K., Prescott, T.J. and Redgrave, P., 2001a. A computational model of action selection in the basal ganglia. I. a new functional anatomy. *Biological Cybernetics*, 84(6):401–410. doi: 10.1007/pl00007984.

Gurney, K., Prescott, T.J. and Redgrave, P., 2001b. A computational model of action selection in the basal ganglia. II. analysis and simulation of behaviour. *Biological Cybernetics*, 84(6):411–423. doi:10.1007/pl00007985.

Guthrie, M., Leblois, A., Garenne, A. and Boraud, T., 2013. Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study. *Journal of Neurophysiology*, 109:3025–3040.

Haber, S.N., 2003. The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26:317–330. doi:10.1016/j.jchemneu.2003.10.003.

Hammond, L. J., 1980. The effect of contingency upon the appetitive conditioning of free-operant behavior. *J. Exp. Anal. Behav.*, 34:297–304.

Hazrati, L.-N. and Parent, A., 1992 a. Convergence of subthalamic and striatal efferents at pallidal level in primates: an anterograde double-labeling study with biocytin and pha-l. *Brain Res*, 569:336–340.

Hazrati, L.-N. and Parent, A., 1992 b. Differential patterns of arborization of striatal and subthalamic fibers in the two pallidal segments in primates. *Brain Res*, 598:311–315.

Hélie, S., Ell, S.W. and Ashby, G.F., 2015. Learning robust cortico-cortical associations with the basal ganglia: An integrative revies. *Cortex*, 64:123–135. doi:http://dx.doi.org/10.1016/j.cortex.2014.10.011.

Herrnsteing, R.J., Vaughan, W., Mumford, Jr.D.B. and Kosslyn, S.M., 1989. Teaching pigeons an abstract relational rule: Insideness. *Perception & Psychophysics*, 46(1):56–64.

Hikosaka, O. and Isoda, M., 2010. Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. *Trends in Cognitive Sciences*, 14(4):154–160. doi:10.1016/j.tics.2010.01.006.

Hirsh, J., 1974. The hippocampus and contextual retrieval of information from memory: a theory. *Behav. Biol.*, 12:421–444.

Hollerman, J.R. and Schultz, W., 1996. Activity of dopamine neurons during learning in a familiar task context. *Soc. Neurosci. Abstr.*, 22:1388.

Hollerman, J.R and Schultz, W., 1998. Dopamine neurons report an erron in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309.

Hopfield, JJ., 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Natl Acad Sci U S A*, 81:3088–3092.

Isoda, M. and Hikosaka, O., 2011. Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behav-

iors. *European Journal of Neuroscience*, 33:2058–2069. doi: 10.111/j.1460-9568.2011.07698.x.

Ito, M. and Doya, K., 2011. Multiple representations and algorithms for reinforcement learning in the cortio-basal ganglia circuit. *Current Opinion in Neurobiology*, 2011(21):368–373. doi: 10.1016/j.conb.2011.04.001.

Joel, D. and 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., Kubota, Y., Connolly, C.I., Hillegaart, V. and Graybiel, A.M., 1999. Building neural representations of habits. *Science*, 286:1745–1749.

Kita, H., Tokyno, H. and Nambu, A., 1999. Monkey globus pallidus external segment neurons projecting to the neostriatum. *Neuroreport*, 10(7):1476–1472.

Lashley, K. S., 1950. In search of the ngram. *In Society of experimetnal Biology Symposium*, 4:454–480.

Leblois, A., Boraud, T., Meissner, W., Bergman, H. and Hansel, D., 2006. Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *Journal of Neurosciences*, 26:3567–3583.

Leisman, G., Melillo, R. and Carrick, F.R., 2013. Clinical motor and cognitive neurobehavioral relationships in the basal ganglia, basal ganglia - an integrative view. doi:10.5772/55227.

Liénard, J., 2013. *Models of the Basal Ganglia: Study of the Functional Anatomy and Pathophysiology using Multiobjective Evolutionary Algorithms*. Thèse de doctorat, Université Pierre et Marie Curie.

Liljeholm, M. and O'Doherty, J.P., 2012. Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends in Cognitive Sciences*, 16(9):467–475. doi:http://dx.doi.org/10.1016/j.tics.2012.07.007.

Ljungberg, T., Apicella, P. and Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67:145–163.

Maia, T.V., Cooney, R.E. and Peterson, B.S., 2008. The neural bases of obsessive-compulsive disorder in children and adults. *Dev. Psychopathol.*, 20:1251–1283.

Mannella, F., Gurney, K. and Baldassarre, G., 2013. The nucleus accumbens as a nexus between valus and goals in goal-directed behavior: a review and a new hypothesis. *Frontiers in Behavioral Neuroscience*, 7(135). doi:10.3389/fnbeh.2013.00135.

Merchant, H., Zainos, A., Hernandez, A., Salinas, E. and Romo, R., 1997. Functional properties of primate putamen neurons during the categorization of tactile stimuli. *Journal of Neurophysiology*, 77:1132–1154.

Mink, J.W., 1996. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50:381–425.

Mink, JW and Thach, WT, 1993. Basal ganglia intrinsic circuits and their role in behavior. *Curr Opin Neurobiol*, 3:950 –957.

Mirenowicz, J. and Schultz, W., 1994. Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.*, 72:1024–1027.

Mirenowicz, J. and Schultz, W., 1996. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379:449–451.

Miyachi, S., Hikosaka, O. and Lu, X., 2002. Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Exp. Brain Res.*, 146:122–126.

Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z. and Rand, M.K., 1997. Differential roles of monkey striatum in learning of sequential hand movement. *Exp. Brain Res.*, 115:1–5.

Montague, P.R., Dayan, P. and Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16:1936–1947.

Moro, E., Lang, A.E., Strafella, A.P., Poon, Y.-Y.W., Arango, P.M., Dagher, A., Hutchison, W.D. and Lozano, A.M., 2004. Bilateral globus pallidus stimulation for Huntington's disease. . *Annals of Neurology*, 56:290–294.

Moustafa, A.A., Bar-Gad, I., Korngreen, A. and Bergman, H., 2014. Basal ganglia: physiological and behavioral and and computational studies. *Frontiers in Systems Neuroscience*, 8(150). doi: 10.3389/fnsys.2014.00150.

Nadjar, A., Brotchie, J.M., Guigoni, C., Zhou, Q.L.S-B., Wang, G-J, Ravenscroft, P., Georges, F., Crossman, A.R. and Bezard, E., 2006. Phenotype of striatofugal medium spiny neurons in parkinsonian and dyskinetic nonhuman primates: A call for a reappraisal of the functional organization of the basal ganglia. *Journal of Chemical Neuroanatomy*, 26(34):8653–8661. doi:10.1523/JNEUROSCI.2582-06.2006.

Nakano, K., 2000. Neural circuits and topographic organization of the basal ganglia and related regions. *Brain Dev*, 22:S5–S16.

Nambu, A., Tokuno, H., Hamada, I., Kita, H., Imanishi, M., Akazawa, T. and Hasegawa, N., 2000 b. Excitatory cortical inputs to pallidal neurons via the subthalamic nucleus in the monkey. *Journal of Neurobiology*, 84:289.

Nambu, A., Tokuno, H. and Takada, M., 2002. Functional significance of the cortico /subthalamo /pallidal 'hyperdirect' pathway. *Neuroscience Research*, 43:111–117.

Nisenbaum, E. and Wilson, C., 1995. Potassium currents responsible for inward and outward rectification in rat neostriatal spiny projection neurons. *J Neurosci*, 15:4449–4463.

Niv, Y., Joel, D. and Dayan, P., 2006. A normative perspective on motivation. *Trend in Congitivie Sciences*, 10(8):375–381. doi: 10.11016.j.ticx.2006.06.010.

Nordlie, E., Gewaltig, M. and Plesser, H.E., 2009. Towards reproducible descriptions of neuronal network models. *PLoS Computational Biology*, 5(8):e1000456. doi:10.1371/journal.pcbi.1000456.

Packard, M.G., 2009. Exhumed from thought: basal ganglia and response learning in the plus-maze. *Behav. Brain. Res.*, 199:24–31.

Parent, A. and Hazrati, L.-N., 1995a. Functional anatomy of the basal ganglia. i. the cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, 20:91–127.

Parent, A. and Hazrati, L.-N., 1995b. Functional anatomy of the basal ganglia. i. the place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Research Reviews*, 20:128–154.

Parent, A., Sato, F., Wu, Y., Gauthier, J., Levesque, M. and Parent, A., 2000. Organization of the basal ganglia: the importance of axonal collateralization. *Trends Neurosci*, 23:S20–S27.

Paspathy, A. and Miller, E., 2005. Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433:873–876.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C. and Boraud, Th., 2007. Shaping of

motor responses by incentive values through the basal ganglia. *J Neurosci*, 27:1176–1183.

Pavlov, I., 1927. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* Translated by G. V. Anrep. [London] Oxford University Press: Humphrey Milford.

Pawlak, V. and Kerr, J., 2008. Dopamine receptor activation is required for corticostriatal spike-timing dependent plasticity. *J Neurosci*, 28:2435–2446.

Pezzulo, G., Verschure, P.F.M.J., Balkenius, C. and Pennartz, C.M.A., 2014. The principles of goal-directed decision making: from neural mechanisms to computation and robotics. *Phil. Trans. R. Soc. B*, 369. doi:http://dx.doi.org/10.1098/rstb.2013.0470.

Piron, C., Kase, D., Topalidou, M., Goillandeau, M., Orignac, H., N'Guyen, T-N, Rougier, N.P. and Boraud, T., 2016. The globus pallidus pars interna in goal-oriented and routine behaviors: Resolving a long-standing paradox. *Moving disorders*.

Prescott, T.J., Montes-Gonzalez, F., Gurney, K., Humphries, M. D. and Redgrave, P., 2006. A robot model of the basal ganglia: Behavior and intrinsic processing. *Neural Networks*, 19:31–61.

Prescott, T.J., Redgrave, P. and Gurney, K., 1999. Layered control architectures in robots and vertebrates. *Adap Behav*, 7:99–127.

Redgrave, P., Prescott, T.J. and Gurney, K., 1999. The basal ganglia: a vertebrate solution to the selection problem? *Neurosci*, 89:1009–1023.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., Delong, M.R. and Obeso, J.A., 2010. Goal-directed and habitual control in the basal ganglia: implications for parkinson's disease. *Nature Reviews, Neuroscience*, 11:760–772. doi:10.1038.nrn2915.

Reiner, A., Hart, N.M., Lei, W. and Deng, Y., 2010. Corticostriatal projection neurons–dichotomous types and dichotomous functions. *Frontiers in Neuroanatomy*, 4(142). doi:10.3389/fnana.2010.00142.

Reynolds, J. N. J. and Wickens, J. R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507–521.

Robinson, M. J. F. and Berridge, K. C., 2013. Instant transformation of learned repulsion into motivational "wanting. *Current Biology*, 23:282–289. doi:10.1016/j.cub.2013.01.016.

Romo, R., Merchant, H. and Hernandez, A., 1997. Categorical perception of somesthetic stimuli: Psychophysical measurements correlated with neuronal events in primate medial premotor cortex. *Cerebral Cortex*, 7:317–326.

Romo, R., Merchant, H., Ruiz, S., Crespo, P. and Zainos, A., 1995. Neuronal activity of primate putamen during categorical perception of somaesthetic stimuli. *NeuroReport*, 6:1013–1017.

Romo, R. and Schultz, W., 1990. Dopamine neurons of hte monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *J. Neurophysiol.*, 63:592–606.

Rudebeck, P.H., Behrens, T.E., Kennerley, S.W., Baxter, M.G., Buckey, M.J., Walton, M.E. and Rushworth, M.F.S., 2008. Frontal cortex subregions ply distinct roles in choices between actions and stimuli. *The Journal of Neuroscience*, 28(51):13775–13–785. doi:10.1523/JNEUROSCI.3541-08.2008.

Sandstorm, MI. and Rebec, GV., 2003. Characterization of striatal activity in conscious rats: contribution of nmda and ampa/kainate receptors to both spotaneous and glutamate0driven firing. *Synapse*, 47:91–100.

Schneider, R.M. and Shiffrin, W., 1993. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory , journal = Psychol. Rev. . 84:127–190.

Schroll, H. and Hamker, F.H., 2013. Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy. *Frontiers in Systems Neuroscience*, 7(122). doi: 10.3389/fnsys.2013.00122.

Schultz, W., 1998. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.

Schultz, W., 2002. Getting formal with dopamine and reward. *Neuron*, 36:241–263.

Schultz, W., 2015. Neuronal reward and decision signals: from theories to data. *Physiol Rev*, 95:853–951. doi:10.1152/physrev.00023.2014.

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

Sealfon, S.C. and Olanow, C.W., 2000. Dopamine receptors: from structure to behavior. *Trends in Neurosciences*, 23.

Seamans, J.K. and Robbins, T.W., 2009. Dopamine modulation of prefrontal cortex and cognitive function. *In K. A. Neve (Ed.), The dopamine receptors (2nd ed.)*, pages 373–398.

Seger, C.A., 1994. Implicit learning. *Psychol Bull.*, 115:163–196.

Seger, C.A., 2008. How do the basal ganglia contribute to categorization? their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews*, 32:265–278.

Seger, C.A. and Spiering, B.J., 2011. A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, 5(66). doi:10.3389/fnsys.2011.00066.

Shepherd, G.M.G., 2013. Corticostriatal connectivity and its role in disease. *Nature Rievews, Neuroscience*, 14:278–291. doi:10.1038/nrn3469.

Sherrington, C. S., 1906. Observations on the scratch reflex in the spinal dog. *Journal of Physiology*, 34:1–50.

Shiffrin, W. and Schneider, R.M., 1977. Controlled and automatic human information processing: 1. detection search, and attention. *Psychol. Rev.*, 84:1–66.

Shriki, O., Hasel, D. and Sompolinsky, H., 2003. Rate models for conductance-based cortical neuronal networks. *Frontiers in Systems Neuroscience*, 15:1809–1841.

Skinner, B. F., 1950. Are theories of learning necessary? *Psychological Review*, 57:193–216.

Squire, L. R. and Zola-Morgan, S., 1991. The medial temporal lobe memory system. *Science*, 253:1380–1386.

Squire, L.R. and Zola-Morgan, S., 1988. Memory:brain systems and behavior. *Trends Neurosci.*, 11:170–175.

Staines, W., Atmadja, S. and Fibiger, H., 1981. Demonstration of a pallidostriatal pathway by retrogradde transport of hrp-labelled lectin. *Brain Research*, 206:446–450.

Sutton, R.S. and Barto, A.G., 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tepper, J.M. and Bolam, J.P., 2008. Functional diversity and specificity of neostriatal interneurons. *Current Opinion in Neurobiology*, 14(6):685–692.

Thorndike, E.L., 1898. Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied*, page 277–293.

Tolman, E. C., 1948. Cognitive maps in rats and men. *Psychological review*, 55(4):189–208.

Topalidou, M., Kase, D., Boraud, T. and Rougier, N.P., 2016. Dissociation of reinforcement and hebbian learning induces covert acquisition of values in the basal ganglia. doi:10.1101/060236. In: *BioRxiv preprint*, submitted.

Topalidou, M., Leblois, A., Boraud, T. and Rougier, N.P., 2015. A long journey into reproducible computational neuroscience. *Frontiers in Computational Neuroscience*, 9(30). doi:10.3389/fncom.2015.00030.

Topalidou, M. and Rougier, N.P., 2015. [re] interaction between cognintive and motor cortico-basal ganglia loops during decision making: a computational study. *ReScience*, 1(1).

Utter, A.A. and Basso, M.A., 2008. The basal ganglia: An overview of circuits and function. *Neuroscience and Biobehavioral Reviews*, 32:333–342. doi:10.1016/j.neubiorev.2006.11.003.

van der Meer, M.A.A. and Redish, D.A., 2010. Expectancies in decision making, reinforcement learning, and ventral striatum. *Frontiers in Neuroscience*, 4(1):29–37. doi:10.3389/neuro.01.006.2010.

Wang, K.S., McClure, J.P.Jr., Alselehdar, S.K. and Kanta, V., 2015. Direct and indirect pathways of the basal ganglia: opponents or collaborators? *Frontiers in Neuroanatomy*. doi:10.3389/fnana.2015.00020.

Wendler, E., Gaspar, J.C.C., Ferreira, T.L., Barbiero, J.K., Andereatini, R., Vital, M.A.B.F., Blaha, C.D., Winn, P. and Da Cuncha, C., 2013. The roles of the nucleus accumbens core, dorsomedial striatum, and dorsolateral striatum in learning: Performance and extinction of pavlovian fear-conditioned responses and instrumental avoidance responses. *Neurobiology of Learning*

and Memory, 109(2014):27–36. doi:http://dx.doi.org/10.1016/j.nlm.2013.11.009.

Wickens, J., 1990. Striatal dopamine in motor activation and reward-mediated learning: Steps towards a unifying model. *Journal of Neural Transmission: General Section*, 80:9–31.

Wickens, J., 1993. A theory of the striatum. *New York: Pergamaon Press*.

Wickens, J., 1997. Basal ganglia: structure and computations. *Comput. Neural Syst.*, 8:R77–R109.

Wickens, J.R, 2009. Synaptic plasticity in the basal ganglia. *Behavioural Brain Research*, 199:119–128. doi:10.1016/j.bbr.2008.10.030.

Wilson, Cj. and Groves, PM., 1981. Spontaneous firing patterns of identified spiny neurons in the rat neostriatum. *Brain Res*, 163:67–80.

Wilson, H.R. and Cowan, J.D., 1972. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys J*, 12:1–24.

Yin, H., Ostlund, S.B., Knowlton, B.J. and Balleine, B.W., 2005. The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22:513–523. doi:10.1111/j.1460-9568.2005.04218.x.

Yin, H.H., 2008. From action to habits: Neuroadaptations leading to dependence. *Alcohol Res Health.*, 31(4):340–344.

Yin, H.H. and Knowlton, B.J., 2006. The role of the basal ganglia in habit formation. *Nature Reviews, Neuroscience*, 7:464–476. doi:10.1038/nrn1919.

Yin, H.H., Knowlton, B.J. and Balleine, B.W., 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt

habit formation in instrumental learning. *European Journal of Neuroscience*, 19:181–189. doi:10.1046/j.1460-9568.2003.03095.x.