A closed loop framework of decision-making and learning in primate prefrontal circuits.

Bhargav Teja Nallapu

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

Ecole Doctorale Mathématiques et Informatique

Informatique

Par

Bhargav Teja NALLAPU

A closed loop framework of decision-making and learning in primate prefrontal circuits

Using Computational Modeling and Virtual Experimentation

Sous la direction de : Frédéric ALEXANDRE
(co-directeur : Thierry VIEVILLE)

Soutenue le 05, Décembre, 2019

Membres du jury :

Mme. NIKOLSKI, Macha	CNRS	Président
Mme. CANAMERO, Lola	University of Hertfordshire, U.K	Rapporteur
M. BOURET, Sebastien	CNRS	Rapporteur
M. KHAMASSI, Mehdi	CNRS	Examinateur
M. CHAKRAVARTHY, Srinivasa	I.I.T Madras, India	Examinateur

Titre : Un cadre en boucle fermée sur la prise de décision et l'apprentissage dans les circuits préfrontaux des primates, par modélisation computationnelle et expérimentation virtuelle.

Resumé :

Cette thèse tente de construire un cadre de travail au niveau des systèmes informatiques qui aiderait à comprendre l'organisation des systèmes du cortex préfrontal (PFC) et des ganglions de base (BG) et leurs interactions fonctionnelles dans le processus décisionnel et le comportement ciblé chez les humains. Environnement de jeu vidéo avec un agent artificiel, Minecraft est utilisé pour concevoir des expériences visant à tester le cadre dans un environnement qui pourrait être plus complexe et réaliste, si nécessaire. Malmo, une plateforme développée par Microsoft, permet de communiquer avec le jeu vidéo Minecraft pour concevoir les scénarios dans l'environnement et contrôler le comportement de l'agent. Le cadre, avec l'expérimentation virtuelle forme une architecture en boucle fermée pour l'étude du comportement animal de haut niveau. Il est souligné que les principes génériques qui sous-tendent les comportements animaux flexibles donnent également un aperçu du développement de l'intelligence artificielle (I.A.) qui est plus générale et autonome dans la nature de l'apprentissage, en plus des systèmes actuels d'I.A. qui sont spécialisés dans une tâche particulière.

Le comportement, d'un humain ou d'un animal, est un ensemble de réactions à un certain stimulus (physique ou abstrait). Une réponse est essentiellement un choix parmi plusieurs options possibles ou simplement une décision entre faire un choix parmi les options disponibles ou non. Les corrélats neuronaux de la prise de décision chez l'homme sont une question très recherchée dans de multiples domaines allant de la psychologie du comportement, de la neuroéconomie et à l'intelligence artificielle (I.A.). En particulier dans le domaine de la neuroéconomie et de l'I.A., il y a une recherche énorme pour comprendre les fondements de la prise de décision dans le cerveau. Avec l'intérêt croissant pour la compréhension des substrats neuronaux de la prise de décision, de l'apprentissage et du comportement, du moins chez les mammifères d'ordre supérieur comme les rongeurs, les primates non humains et les humains, plus de recherche mène à des questions plus profondes sur notre compréhension du processus décisionnel lui-même. Ce n'est pas si surprenant, étant donné qu'une espèce, dans une certaine mesure, dépend des mécanismes de sélection des actions ou de prise de décision pour sa survie dans un environnement incertain. L'homme est sans doute le décideur le plus souple et le plus adaptable qui peut apprendre la structure sous-jacente du monde, même si cette structure est cachée, et il peut adapter rapidement son comportement. Le cortex préfrontal (PFC) est à l'avant-garde de cette faculté et on croit qu'il a facilité cette évolution vers un répertoire plus large de comportements qui émergent des mécanismes sous-jacents de sélection des actions primitives. Il est souligné que l'étude de la prise de décisions complexes et réalistes dans des scénarios écologiques nécessitera des méthodes d'expérimentation plus sophistiquées que les simulations numériques classiques utilisées. Les expériences conçues dans Minecraft peuvent être utilisées pour tester le cadre dans un environnement qui pourrait être plus

complexe et réaliste, si nécessaire. La valeur ajoutée majeure d'un environnement virtuel et d'un agent qui y interagit est que les caractéristiques corporelles de l'agent peuvent être soulignées (comme les besoins) et leur rôle dans la prise de décision basée sur la valeur peut être discuté. Par la suite, le cadre, avec l'expérimentation virtuelle forme une architecture en boucle fermée pour l'étude du comportement animal de haut niveau.

Le cadre des systèmes neuronaux dans ce travail repose sur la dynamique des réseaux entre les sous-systèmes de PFC et BG. On croit que le PFC joue un rôle crucial dans les fonctions exécutives comme la planification, l'attention, le comportement ciblé, etc. Les BG sont un groupe de noyaux sous-corticaux qui ont fait l'objet d'études approfondies dans le domaine du contrôle moteur et de la sélection d'action. Différentes régions du PFC et structures au sein des BG sont anatomiquement organisées, en association avec une région corticale sensorielle respective, en boucles parallèles et séparées (chacune d'entre elles étant appelée ici une boucle CBG). Ces boucles peuvent être, à un niveau élevé, divisées en 3 types : les boucles limbiques, les boucles associatives et les boucles sensori-motrices. Imaginez un animal interagissant avec des stimuli dans un environnement. Voici quelques-unes des questions les plus pertinentes relatives à l'état actuel de l'animal en ce qui concerne les stimuli présents : (i) *Quel* est (la valeur de) ce stimulus ? (*Préférence*) (ii) *Pourquoi* ce stimulus est-il pertinent pour mes besoins internes actuels ? (*Besoin*) (iii) *Où* est ce stimulus situé par rapport à ma référence dans l'environnement actuel (*Orientation*), et (iv) *Comment* atteindre le stimulus 'souhaité' (*Approche*). Les boucles limbiques répondent aux questions *Quoi?* et *Pourquoi?* Les boucles sensori-motrices sont concernées par les questions *Où?* et *Comment?*. Les boucles associatives forment une association multimodale de l'information sur l'état actuel, par exemple quel stimulus dans les boucles limbiques est représenté à quelle position dans les boucles motrices. En outre, dans chacune de ces boucles, comme la sous-région de la PFC représente l'objectif choisi, le processus de réalisation de l'objectif par une activation soutenue entre la sous-région de la PFC et la région corticale sensorielle correspondante est décrit. L'expérimentation, en particulier virtuelle, permet de mettre en évidence ce phénomène en faisant preuve de souplesse dans l'adaptation du plan d'action une fois l'objectif choisi.

Tout d'abord, un cadre global avec les boucles parallèles susmentionnées est mis en œuvre. Les quatre boucles sont mises en œuvre de manière algorithmique, décrivant les influences mutuelles entre chacune des sous-régions préfrontales. Il est important de noter que, bien qu'il n'y ait pas de hiérarchie explicite établie dans le système entre les boucles, deux niveaux de hiérarchie pourraient implicitement apparaître. Premièrement, bien que les boucles motrices soient libres de prendre des décisions dans l'espace d'action, avec suffisamment d'apprentissage dans l'espace limbique, les décisions dans n'importe laquelle des boucles limbiques pourraient conduire les décisions dans l'espace sensori-moteur à travers la boucle associative. Deuxièmement, on suppose que la motivation fondamentale de l'animal est l'homéostasie interne, c'est-à-dire de maintenir ses besoins internes dans des limites acceptables. Ainsi, dans certaines situations, la motivation interne peut conduire la dynamique dans les boucles limbiques, avec la boucle *Pourquoi?*. Les entrées pour les boucles CBG sont fournies par la perception sensorielle du cadre qui communique les informations fournies par Malmö à partir de l'environnement de jeu

vidéo aux représentations correspondantes dans le cadre. De même, la sortie du cadre est transformée en représentations Malmo appropriées des commandes d'action qui entraînent l'agent dans l'environnement. Le cadre cognitif étant décrit par plusieurs contraintes biologiques, plusieurs adaptations ont été apportées à l'utilisation de la plate-forme de Malmo, en termes de perception sensorielle de l'environnement et de contrôle moteur de l'agent.

Ensuite, nous utilisons ce cadre pour étudier de plus près le rôle des boucles *limbiques* dans la prise de décision guidée par les valeurs et le comportement ciblé. L'accent est mis sur les boucles limbiques. Les boucles associatives et sensori-motrices sont donc modélisées de manière algorithmique, à l'aide de la plate-forme d'expérimentation pour le contrôle moteur. Comme pour les boucles limbiques, le cortex orbitofrontal (OFC) est la partie d'une boucle pour les préférences et le cortex cingulaire antérieur (ACC), pour les besoins internes. Ces boucles sont formées par leur contrepartie limbique en BG, striatum ventral (VS). Le VS etait fait l'objet de nombreuses études et on a signalé qu'il encode divers substrats de valeur, faisant ainsi partie intégrante de la prise de décisions fondées sur les valeurs. Des scénarios simplistes sont conçus dans l'environnement virtuel en utilisant l'agent et certains objets et des récompenses appétissantes dans l'environnement. Les boucles limbiques ont été mises en œuvre selon les modèles informatiques existants de prise de décision dans les BG et l'amygdale. Ainsi, le cadre et la plate-forme expérimentale servent de banc d'essai à des modèles informatiques de processus spécifiques qui doivent s'inscrire dans une perspective plus large.

Parmi les boucles limbiques, le rôle de l'OFC a été étudié de près. Au fil des décennies, l'OFC a été impliqué dans presque tous les aspects de la prise de décision - représentation de l'état, prédiction des résultats, sélection des actions, évaluation des résultats et surtout, l'apprentissage. En outre, les déficits ou les lésions de l'OFC ont été argués pour causer des déficiences comportementales multiples telles que l'inhibition de réponse pour ne plus récompenser le stimulus, l'apprentissage quand les contingences de récompense sont inversées, etc. Avec des techniques de lésions plus avancées et une analyse plus fine, plusieurs de ces observations ont été rejetées. Néanmoins, le rôle d'OFC dans la prise de décision et l'apprentissage fondés sur les valeurs est souligné à maintes reprises, alors que l'on ignore encore la manière exacte dont il affecte le processus. Dans le cadre de cette thèse, plusieurs observations remarquables sur le rôle d'OFC dans le comportement ont été résumées en consolidant de nombreuses preuves expérimentales et revues. En voici quelques exemples : la prise de décision perceptive et la prise de décision fondée sur les valeurs ; au sein d'un même épisode de prise de décision (l'essai), différents types de participation à une étape différente (présentation des options, sélection des actions, prestation des résultats, etc.) ; les associations des stimuli et des résultats d'apprentissage (Pavlovien) et d'actions-résultats (instrumental). On a constaté que les neurones d'OFC présentent une corrélation frappante avec la valeur des résultats, exprimant de façon plus intéressante un phénomène d'adaptation de l'intervalle, s'adaptant à l'évolution de l'intervalle des valeurs. L'OFC est censé apprendre une représentation spatiale d'état de l'espace de travail pour pouvoir accéder à des informations partiellement observables en vue d'une décision. L'hétérogénéité structurelle d'OFC ajoute à la complexité sous-

jacente inhérente à l'étude du rôle d'OFC dans la prise de décision, l'apprentissage et le comportement ciblé. Cette question a été étudiée au cours des dernières années, avec des études axées sur la dissociation des rôles des sous-parties latérale et médial de l'OFC. Souvent, le cortex préfrontal ventromédial (vmPFC) est pris en compte dans le cadre d'OFC médial. Bouret et al 2010, Noonan et al 2010, Rudebeck & Murray 2011 sont quelques-unes des rares études approfondies qui ont clairement plaidé en faveur de rôles distincts pour l'OFC latéral et médial.

Enfin, pour expliquer les résultats des différents rôles des régions latérales et médianes de l'OFC, l'architecture informatique existante des boucles CBG, l'apprentissage pavlovien dans l'amygdale et les multiples preuves des interactions amygdales-OFC-VS sont réunies dans un modèle unique. Les règles d'apprentissage du renforcement ont été adaptées pour tenir compte de l'attribution de crédits appropriée (résultat correct pour corriger le stimulus choisi) et de la différence de valeur des options de choix. Par conséquent, plusieurs résultats d'expériences sur des animaux étudiant les rôles séparables ont été reproduits. En particulier dans le contexte des différents rôles de l'OFC latéral et médial dans la prise de décision en fonction de la différence de valeur entre les options, des rôles distincts et dissociés des régions latérale et médiale ont été observés. L'OFC médial semblait plus crucial pour le choix entre deux options proches l'une de l'autre, alors que les lésions de l'OFC médial ne semblaient pas affecter la performance de l'animal lorsque la différence entre les valeurs des deux options est suffisamment éloignée. Au contraire, de manière surprenante, l'OFC latéral s'est avéré crucial lorsque les décisions sont faciles à prendre alors que les lésions de l'OFC latéral ne semblaient pas affecter les choix difficiles où les valeurs des options sont proches les unes des autres. Des résultats similaires ont été trouvés dans les performances des singes avec des lésions a l'OFC latéral et celles avec des lésions à l'OFC médial. Des rôles dissociables dans le transfert instrumental pavlovien ont également été observés.

Nonobstant les architectures neuronales détaillées et les descriptions neuronales de base utilisées dans certaines parties de ce travail, les mécanismes neuronaux de tous les paradigmes comportementaux ont été discutés à un niveau très simpliste. Tout au long du travail, seul le comportement appétitif a été décrit, alors que la plupart des processus décrits dans ce travail sont également connus pour expliquer les comportements aversifs comme éviter les punitions. En outre, le rôle de la dopamine en tant que neurotransmetteur facilitant l'apprentissage a été extrêmement simplifié. De plus, avec les multiples systèmes d'apprentissage de renforcement impliqués dans le cadre, il exige un rôle détaillé sur la façon dont la dopamine pourrait avoir un effet différentiel sur ces systèmes. L'un des éléments les plus importants du comportement qui n'est pas pris en compte dans le cadre est la mémoire. En fait, en complétant le cadre par un compte rendu informatique existant d'un modèle de mémoire de travail minimale, les mécanismes des activités soutenues pour maintenir les objectifs jusqu'à ce qu'ils soient atteints, des aspects comme l'abandon si l'objectif n'a pas été atteint depuis longtemps, etc. peuvent être explorés davantage. L'ajout d'une mémoire explicite pour stocker un minimum d'informations spatiales et épisodiques permettrait au cadre d'expliquer des comportements plus flexibles comme des comportements purement ciblés ou opportunistes. Cependant, cela nécessiterait des

implémentations très sophistiquées de boucles de moteur où l'on peut naviguer dans une position désirée.

Néanmoins, les recherches sur les preuves observées autour de l'OFC permettent de mieux comprendre le processus même de la prise de décision et le calcul de la valeur en général. En s'aventurant dans un domaine d'apprentissage adaptatif bio-inspiré dans un agent virtuel incarné, décrivant les principes de motivation, de sélection d'objectifs et d'auto-évaluation, il est souligné que le domaine de l'apprentissage par renforcement et de l'intelligence artificielle a beaucoup à gagner à étudier le rôle des systèmes préfrontaux dans le processus décisionnel.

Mots clés : prise de décision, apprentissage, cortex préfrontale, expérimentation virtuelle, comportement vers un but, modélisation computationnelle

Title : A closed loop framework of decision making and learning in primate prefrontal circuits: Using computational modeling and virtual experimentation

Abstract : This thesis attempts to build a computational systems-level framework that would help to develop an understanding of the organization of the prefrontal cortex (PFC) and the basal ganglia (BG) systems and their functional interactions in the process of decision-making and goal-directed behaviour in humans. A videogame environment with an aritficial agent, Minecraft is used to design experiments to test the framework in an environment that could be more complex and realistic, if necessary. Malmo, a platform developed by Microsoft, allows to communicate with the videogame Minecraft to design the scenarios in the environment and control the behavior of the agent. The framework, along with virtual experimentation forms a closed-loop architecture for studying the high-level animal behavior. It is pointed out that the generic principles behind the flexible animal behaviors also give insights into developing artificial intelligence (A.I) that is more general and autonomous in the nature of learning, in addition to the current A.I systems that are specialized in a particular task.

Behavior, of a human or an animal, is a pattern of responses to a certain stimulus (physical or abstract). A response is essentially a choice among several possible options or simply a choice between whether or not to make a choice from the available options. The neural correlates of decision-making in humans is an extensively sought after question across multiple fields ranging from behavioural psychology, economics to neuroscience and artificial intelligence (AI). Especially in the field of neuroeconomics and AI, there is a huge pursuit to understand the underpinnings of decision-making in brain. With rapidly growing interest in understanding the neural substrates of decision-making, learning and behaviour, at least in higher order mammals like rodents, non-human primates and humans, more research is leading to deeper questions about our understanding of decision-making itself. It is not so surprising because, given that any species, in some degree or the other, depends on the mechanisms of action selection or decision-making for its survival in an uncertain environment. Humans are presumably the most flexible and adaptive decision-makers who can learn the underlying structure of the world, even if the structure is hidden, and rapidly adapt their behaviour. The prefrontal cortex (PFC) has been at the forefront of this proposition and is believed to have facilitated this evolution towards a wider repertoire of behaviours that emerge from underlying primitive action selection mechanisms. It is highlighted that studying complex realistic decision-making in ecological scenarios will require a more sophisticated experimentation methods than the regular numerical simulations used. The experiments designed in Minecraft can be used to test the framework in an environment that could be more complex and realistic, if necessary. Major value addition of a virtual environment and an agent interacting in it is, that the bodily characteristics of the agent can be emphasized (like needs) and their role in value-based decision making can be discussed. Subsequently the framework, along with virtual experimentation forms a closed-loop architecture for studying the high-level animal behavior.

The neural systems framework in this work rests on the network dynamics between the subsystems of PFC and BG. PFC is believed to play a crucial role, in executive functions like planning, attention, goal-directed behavior, etc. BG are a group of sub-cortical nuclei that have been extensively studied in the field of motor control and action selection. Different regions in the PFC and structures within BG are anatomically organized, including a respective sensory cortical region, in parallel and segregated loops (each of them referred here as a CBG loop). These loops can be, on a high level, divided into 3 kinds : limbic loops, associative loops and sensori-motor loops. Imagine an animal interacting with stimuli in an environment. Some of the most pertinent questions to the current state of the animal with respect to the stimuli present are : (i) *What* is (the value of) this stimulus? (*Preference*) (ii) *Why* is this stimulus relevant to my current internal needs? (*Need*) (iii) *Where* is this stimulus located with respect to my reference in the current environment (*Orientation*), and (iv) *How* do I reach the 'desired' stimulus (*Approach*). Limbic loops address the questions *What?* and *Why?*. Sensori-motor loops are concerned with the questions *Where?* and *How?*. Associative loops form a multi-modal association of the current state information, for instance which stimulus in the limbic loops is at which position represented in the motor loops. Furthermore, in each of these loops, as the subregion of PFC represents the chosen goal, the process of achieving the goal by sustained activation between the PFC subregion and the corresponding sensory cortical area is described. Especially virtual experimentation helps highlight this phenomenon by demonstrating flexible adjustments to action plan once the goal is selected.

First, a comprehensive framework with the above mentioned parallel loops is implemented. All the four loops are algorithmically implemented, describing the mutual influences between each of the prefrontal sub-regions. It is important to note that, although there is no explicit hierarchy built in the system among the loops, there are two levels of hierarchy that could implicitly arise. First, although the motor loops are free to make decisions in the action space, with sufficient learning in the limbic space, the decisions in any of the limbic loops could lead the decisions in the sensori-motor space. through the associative loop. Secondly, it is assumed that the fundamental motivation of the animal is internal homeostatis, that is to maintain its internal needs in acceptable bounds. Thus, in certain situations, the internal motivation might lead the dynamics in the limbic loops, with the *Why?* loop for internal motivation biasing the *What?* loop which might be more stimulus-driven, when there is no pressing internal need. The inputs for the CBG loops is provided by the sensory perception of the framework that communicates the information provided by Malmo from the videogame environment to the corresponding representations in the framework. Similarly the output of the framework is transformed to appropriate Malmo representations of action commands that drive the agent in the environment. Since the cognitive framework is described by several biological constraints, several adaptations have been made in the way the Malmo platform is used, in terms of sensory perception of the environment and the motor control of the agent.

Next, we use this framework to study more closely, the role of *limbic* loops in value-guided decision making and goal-directed behavior. The emphasis rests on the limbic loops. Therefore the associative and sensori-motor loops are modeled algorithmically,

taking help of the experimentation platform for motor control. As for the limbic loops, the orbitofrontal cortex (OFC) is the part of a loop for preferences and the anterior cingulate cortex (ACC), for internal needs. These loops are formed through their limbic counterpart in BG, ventral striatum (VS). VS has been widely studied and reported to be encoding various substrates of value, forming an integral part of value-based decision making. Simplistic scenarios are designed in the virtual environment using the agent and some objects and appetitive rewards in the environment. The limbic loops have been implemented according to existing computational models of decision making in the BG and amygdala. Thus the framework and the experimental platform stand as a testbed to computational models of specific processes that have to fit in a bigger picture.

Of the limbic loops, the role of OFC has been closely studied. Ranging over diverse studies across decades, OFC has been implicated in almost all aspects of decision-making - state representation, outcome prediction, action selection, outcome evaluation and primarily, learning. Furthermore, deficits or lesions of OFC were argued to cause multiple behavioral impairments such as response inhibition for no longer rewarding stimulus, learning when reward contingencies are reversed etc. With more advanced lesion techniques and keener analysis, several such observations were turned down. Nevertheless, the role of OFC in value-based decision making and learning is underlined time and again, while the exact ways in which it affects the process are still unknown. As part of this thesis, several outstanding observations about the role of OFC in behavior have been summarized by consolidating numerous experimental evidences and reviews. To highlight a few, OFC is implied in : perceptual decision making and value-based decision making; within a single decision-making episode (trial), different kinds of involvement at a different phase (option presentation, action selection, outcome delivery etc.,); learning stimuli-outcome (pavlovian) and action-outcome (instrumental) associations. The neurons in OFC were found to vividly correlate with the value of the outcomes, more interestingly expressing a phenomenon of range adaptation, adapting to the changing ranges of values. OFC is believed to learn a state space representation of the task space to be able to access partially observable information for a decision. The structural heterogeneity of OFC adds to the inherent underlying complexity about studying the role of Orbitofrontal Cortex (OFC) in decision making, learning and goal-directed behavior. This has been studied in the recent years, with studies focused on dissociating the roles of lateral and medial subparts of OFC. Often, ventromedial prefrontal cortex (vmPFC) is considered under medial OFC. Bouret et al., 2010, Noonan et al., 2010, Rudebeck & Murray 2011 are some of the few comprehensive studies that clearly argued for separate roles of lateral and medial OFC.

Lastly, to explain the findings of different roles of lateral and medial regions of OFC, existing computational architecture of CBG loops, pavlovian learning in amygdala and multiple evidences of amygdala-OFC-VS interactions are put together into a single model. The learning rules of reinforcement have been adapted to accommodate the appropriate credit assignment (correct outcome to correct chosen stimulus) and the value difference of the choice options. As a result, several findings from animal experiments studying the separable roles, were replicated. Particularly in the context of different roles of lateral and medial OFC in decision making as a function of the value difference between options,

distinct and dissociate roles of lateral and medial were observed. Medial OFC seemed to be more crucial for the choice between two options that are close to each other, whereas lesions to medial OFC did not seem to affect the animal's performance when the difference between the values of the options are sufficiently apart. On the contrary, surprisingly lateral OFC appeared to be crucial when the decisions are easy to make whereas lesions to lateral OFC did not seem to affect the difficult choices where the values of the options are close to each other. Similar results were found in the performances of the monkeys with lesions to to lateral and those with lesions to medial OFC. Dissociable roles in Pavlovian Instrumental Transfer were also observed.

Notwithstanding the detailed neural architectures and basic neuronal descriptions used in certain parts of this work, the neural mechanisms of all the behavioral paradigms were discussed at a very simplistic level. Throughout the work, only appetitive behavior has been described, whereas most of the processes described in this work are also known to account for aversive behaviors like avoiding punishments. In addition, the role of dopamine as the neurotransmitter facilitating learning has been extremely simplified. Furthermore, with multiple systems of reinforcement learning involved in the framework, it demands for a detailed role of how dopamine could have a differential effect on these systems. One of the most important elements of behavior that is not accounted for in the framework is memory. In fact by complementing the framework with an existing computational account of a minimal working memory model, the mechanisms of sustained activities to maintain goals until achieving, aspects like giving up if the goal hasn't been reached for a long time etc, can be explored further. Adding an explicit memory to store minimum spatial and episodic information would allow the framework to explain more flexible behaviors like pure goal-directed or opportunistic behaviors. However, that would require much sophisticated implementations of motor loops where a desired position can be navigated. Nevertheless, the investigations into the observed evidences around OFC offer great insight into understanding the very process of decision-making, value computation in general. By venturing into a realm of bio-inspired adaptive learning in an embodied virtual agent, describing the principles of motivation, goal-selection and self-evaluation, it is highlighted that the field of reinforcement learning and artificial intelligence has a lot to gain from studying the role of prefrontal systems in decision-making.

Keywords : decision-making, learning, prefrontal cortex, virtual experimentation, goal-directed behaviour, computational modeling

INRIA Bordeaux Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talence, France
LaBRI, Université de Bordeaux, Bordeaux INP, CNRS, UMR 5800, Talence, France
IMN, Université de Bordeaux, CNRS, UMR 5293, Bordeaux, France

Acknowledgements

I would first like to express my sincere gratitude to my directeur de thèse Monsieur Frédéric Alexandre, firstly for sending me that short email with one of his classic puns, proposing this thesis to me. Secondly, well, clearly for leading me through this incredibly interesting research topic. You are one of the most influential 'teachers' I ever had. Please let me thank you for believing in me, encouraging me by giving me the right mix of freedom and guidance, and for constantly supporting me to teach, to attend conferences and and most of all, enlightening me with "Les Fables de la Fontaine".

I would like to thank Thierry Vieville, my "co-directeur de thèse" for his "virtual" availability, when ever I needed. Your feedback is always constructive and thanks for all the time you gave in the middle of the million other things you deal with. I believe I was at a great advantage to have you alongside Fred in a perfectly complementary role.

I express my sincere gratitude to the jury members - Mme. Macha Nikolski (CNRS, Université de Bordeaux, France), Mme. Lola Cañamero (University of Hertfordshire, U.K), M. Sébastien Bouret (CNRS, France), M. Mehdi Khamassi (CNRS, France) and M. Srinivasa Chakravarthy (I.I.T Madras, India) - for graciously accepting to take part in the review of this thesis. I strongly believe, this work could benefit a great deal from the feedback given your esteemed expertise in the topics related to this thesis.

I would also like to thank Nicolas P. Rougier, researcher in the team and my previous mentor for discussions and feedback, Thomas Boraud, Arthur Leblois and Nicolas Mallet, neuroscientists at IMN for not kicking me out every time I knocked on your office door (if I did) for "one little" question every time about "what would a monkey or a mouse do?" It was an invaluable experience of working on the same floor to be able to interact with you all.

I'd like to thank my dearest friends, but for reasons related to research, Pramod and Hari, for being there any time I needed and for all the discussions we had.

I'd like to specially thank Thalitha for helping me all along the 3 years, for sharing a friendly office space, technical expertise, french-life-hacks and especially for every administrative thing you have helped me with! Thanks for being the *stackexchange* for my French life.

I'd like to thank the entire team of Mnemosyne, including the previous Ph.Ds with whom I got a chance to spend great time, it was a great experience to be a part of the same team with each one of you. And a special thanks to Chrystel, who made sure each and every official formality went smooth throught my time at INRIA.

The following people, though not directly related to my research, played a huge role in my personal life at a level that ensured my well-being to go ahead with my research the way I did.

Erika, Alexandre, Celine - Vous avez été l'une des raisons les plus fortes pour lesquelles je suis revenu à Bordeaux pour cette thèse, merci beaucoup d'avoir été une si grande partie de ma vie personnelle en France. Un très grand merci !

Priyatham, Varun, Marte and Ateeth, thank you for being there for me at all times,

you guys have all been an integral part of my life for the past couple of years.

Miguel and Remya, thank you for being such friendly colocs, and friends more so. It mattered a lot, to come back to a living space with someone as peaceful as you guys.

A special and sincere thanks to Amma, Nana and Akka, for making me who I am today and for your unconditional love.

Dedicated to my parents... for every sacrifice they made to give me the privilege of good education!

...the difference in mind between (wo)man and the higher animals, great as it is, certainly is one of degree and not of kind...
- Darwin, 1871
paraphrased

Who said you cannot compare apples to oranges? There is a whole scientific field explaining that
- Nallapu, 2019

Key Abbreviations

PFC	Prefrontal Cortex
$LPFC$	Lateral Prefrontal Cortex
$VLPFC$	Ventrolateral Prefrontal Cortex
$DLPFC$	Dorsolateral Prefrontal Cortex
$MPFC$	Medial Prefrontal Cortex
$VMPFC$	Ventromedial Prefrontal Cortex
$DMPFC$	Dorsomedial Prefrontal Cortex
OFC	Orbitofrontal Cortex
$lOFC$	Lateral Orbitofrontal Cortex
$mOFC$	Medial Orbitofrontal Cortex
ACC	Anterior Cingulate Cortex
BG	Basal Ganglia
STR	Striatum
VS	Ventral Striatum
$NAcc$	Nucleus Accumbens
$core$	Nucleus Accumbens Core
$shell$	Nucleus Accumbens Shell
DMS	Dorsomedial Striatum
DLS	Dorsolateral Striatum
GPI	Globus Pallidus Externa
GPE	Globus Pallidus Interna
SNr	Substantia Niagra pars Reticulata
THL	Thalamus
STN	Subthalamic Nucleus
BLA	Basolateral Amygdala
CeA	Central nucleus of Amygdala
VTA	Ventral Tegmental Area
SNc	Substantia Niagra pars Compacta
DA	Dopamine
RL	Reinforcement Learning
TD	Temporal Difference

Contents

Chapter 1

Introduction

"A robot walks into a bar, doesn't get the joke"

This casual internet meme has a deep rooted sarcasm as well as objectivity to it. It basically highlights the fact that Artificial Intelligence (AI), considered in which ever variety there is out there, doesn't understand humour. Then there comes a whole different question - does A.I *understand* anything? Oh, then there could be a follow up question, does AI *need* to understand anything?

When it was first put together as an official field of study in 1956, the motivation behind A.I was straight forward in the form of an assertion - "every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it". The founding members of the field were so positive about how the goals can be achieved within no longer than a generation from then. Arthur Samuel's checkers program, developed in the early 60s, eventually achieved sufficient skill to challenge a respectable amateur. Marvin Minsky said in 1970 - "In from three to eight years we will have a machine with the general intelligence of an average human being". Fundamentally this idea of matching or exceeding human level intelligence in any given task (or in everything?) has become central to the idea of A.I. To this date, beating humans in a complex game is one of the hallmarks of measuring a performance of an A.I (Silver et al. 2016).

Then came the road-block, the famous A.I winter. In the 1970s, A.I research hit several setbacks with newly discovered complexities of the problems at hand. Few of the most pressing concerns were limited computer power, combinatorial explosion, then existing structure of logic and most importantly reasoning. More than solving theorems, there was a pushing need to solve vision, robotics, language, so that a robot can cross a room without bumping into anything. Identifying things needed the knowledge of the world in the same way that a child has, which is basically a vast amount of information. The combinatorial explosion problem that the "toy" version of the problems that A.I was solving would rapidly grow in complexity and demand exponential computational time also meant that it would require unimaginable amounts of computer power to scale-up into useful systems. Clearly, we have come a really long way from there, updating the structure of logic itself, rapidly advancing in the computational power there is available out there, the amount of data that can be handled and processed. The famous AlphaGO algorithm by Google DeepMind that defeated world champions in GO game, although using some expensive hardware ($25 million), opened avenues for new insights into learning in machines, propelling forward to much advanced versions of itself, where there was no human involved in the process of learning (Alpha GO Zero).

However, besides the above mentioned advancements, one of the most influential ideas that advanced A.I is its union with neuroscience. A.I formed its bases on neuroscience and psychology although for a brief period the limits got magnified owing to the later development of the latter fields, thus partly losing the interaction. Nevertheless, neuroscience and A.I, if not ever before, are at the forefront of interdisciplinary research garnering each of their individuals interests which are - to understand how brain works and to recreate the understanding, respectively.

Besides powerful processors and ever-growing data, A.I owes its successes in part to its fundamental underlying concept of the artificial neural network (ANN). Neuroscience and cognitive science provide a rich source of inspiration for new types of algorithms. Visual processing architecture in brain has laid foundations of image processing, reinforcement learning has pioneered understanding motor skilled learning in robotic bodies,

focused attentional mechanisms are suggesting computationally more efficient algorithms of information processing Fu et al. 2017. On the other hand, modern day A.I and its sub-fields of machine learning and ANNs have proved useful for studying the brain. With its ability to process vasts amounts of data and to identify subtle patterns in sparse data, A.I has rendered itself extremely beneficial for heavy medical imaging data and sparse single neuron recordings, speeding the research in these fields. A.I techniques come in handy not just as a tool for handling data, but for making models and generating ideas. A.I is providing primarily testbeds for various models of cognitive science about how the brain performs computations.

Nonetheless, modern day A.I and ANNs have proved useful for studying the brain. Neural data are extremely complicated, often being either too heavy (e.g, functional magnetic resonance imaging, fMRI) or too sparse as in the case of neural activities of hundreds of neurons that are believed to be few among millions of neurons that control precise limb movement. Machine learning, a branch of A.I, has its main strength in recognizing patterns that might be too subtle or too buried in huge data sets for people to spot. This kind of subtle pattern recognition applied on sparse neuronal data, can be applied to derive fine-grained set of instructions for the control of prosthetic arm movement. A.I techniques come in handy not just as a tool for handling data, but for making models and generating ideas. A.I is providing primarily testbeds for various models of cognitive science about how the brain performs computations. Most importantly using a machine to analyse these models and data is speeding up the research in the fields of neuroscience and cognitive science.

At the heart of this remarkable intersection of the fields of A.I and neuroscience, there remains one, if not the final frontier, that has eluded philosophers, psychologists, neuroscientists and now AI scientists all alike - *reasoning*. As the ball keeps rolling on the side of specificity of abilities of what A.I can do, growing interest surrounds on the side of generality of A.I, one of the hallmarks of human abilities. Time and again, we have been asking ourselves the question - "how we do what we do?", except that in this current context, the question, as complex as *reasoning*, will require a deeper understanding of

human behavior, a fundamental approach to questioning what makes us as intelligent as we think we are.

Behavior, of a human or an animal, is a pattern of responses to a certain stimulus (physical or abstract). A response is essentially a choice among several possible options or simply a choice between whether or not to make a choice from the available options. Humans exhibit a wide range of behaviors in day to day life. We close our eyes without deliberation when there is a sudden wind onto the face. We walk to our home automatically after getting down at the bus stop without verifying the street names. We take the same route to work everyday that is proven to be efficient but we conveniently come up with an alternative route when we are informed that the usual route is blocked. Beyond such routine behaviors, we make many complex decisions. A decision that Warren Buffet has to make about two investing options might cost him millions of dollars. If you are the leader of a country like India, making a budget allocation decision between a space program and stable electricity to villages is certainly not straight forward. Chesley Sullenberger made a decision in under 2 minutes to safely land the U.S. Airways plane carrying 155 passengers in the Hudson river. And very well on the other hand, we also spend considerable amount of time in deciding which restaurant to eat in with a group, which smart phone to choose and most importantly what to wear for a special occasion. Animals have complex decisions to make too. A lion chasing a pack of 3 zebras needs to decide which one to chase when they separate into two paths, one zebra in one and the others in a different direction. A monkey foraging on a tree branch, when encountered with a split on the branch with one branch certainly leading to a berry and the other likely leading to a bunch of berries.

In all these decisions, essentially there is an invariant irrespective of how complex the decision might seem. The animal (human or a monkey alike) tries to make a choice from several alternatives on the basis of a subjective value that it places on them. This inherent subjectivity of value-based decision-making that renders it a challenging aspect to understand. One could make a best guess as to what I will choose based on your past experience of my choices and their knowledge of my current goals, but without knowing

my precise internal state, this still remains far from an objective evaluation. Essentially, value-based theory of decision-making states that a rational choice would be the one that has the largest combination of expected success (probability) and the subjective value. For Mr. Buffet the subjective values on both the investments might be in terms of money. For a leader the value of a satellite investment that would progress the country's agriculture as well as electricity to villages both mean welfare of the citizens. For a common person trying choose between two smartphones might be essentially looking commonly at 'value for money' (maximum capacity within what is needed at a minimum price). Then there are in-commensurable decisions. Do I go for a movie or relax at home? Do I go on a vacation or buy a new gaming console? The computations that a human brain carries out in order to make these value-based decisions, as well as the neural implementations of those computations are extensively studied topics in the fields such as Psychology, Neuroscience, Neuroeconomics and Computer Science.

But going beyond these rather modern concepts of money, technology and society, what is the ecological meaning of human behavior? How is it related to the high level cognitive behavior described above? A question rather tangential to the topic but related is What is it that sets apart humans from other animals? How close are we to other animals in terms of these behaviors and how different are we to them? To ruminate on these questions, as a small detour, we can use the tool that Aristotle provided through his series of books Nicomachean Ethics - called *ergon*. *Ergon* of something, in simplified terms, can be defined as its 'characteristic activity'. We can describe *Ergon* of different living things in the following way:

- *Ergon* of plants can be described as growth, nutrition and reproduction.
- *Ergon* of animals, in addition to the basic capabilities of plants, can be described as movement and sensation. With sensation, through whichever sensory modalities the animal has, comes the ability to perceive. Perception results in the ability to feel pleasure and pain and thus appetite and aversion. Voluntary movement is in turn connected to perception.

- *Ergon* of certain high-functioning animals like primates can be described as *reasoning*, in addition to those capabilities that describe an animal.
- *Ergon* of humans can be described as *reasoning* and *language* in addition to those capabilities of a normal animal.

In fact, it can be objectively said that plants are flourishing well, if they can exercise their capabilities well, their *ergon* of growth, nutrition and reproduction. Similarly animals can be thought to be doing well if they could, in addition to growth, nutrition and reproduction, exercise their capabilities of voluntary movement and sensation in order to maximise their pleasure and minimize their pain. Similarly, for a human or a high functioning animal to do well, all the vegetative and animal capabilities must be in a satisfactory condition, in order for the higher capabilities of reason and thought to be exercised. The exertion of these unique capabilities with the help of language allows the high level cognitive functions of humans such as social structures, contemplation about universe, and about human life itself.

Coming back to the questions mentioned above before the detour, principles described in terms of the capabilities that define humans distinctly, still render humans closer to animals. Ecological meaning of any animal's behavior refers to the motivational and emotional bases that have to be taken into account. Within the capabilities of sensation and movement, the ability of the animal to maintain its organization and retain its capabilities to perceive pleasure and minimize pain despite the changes in the environment, remains crucial from a single-celled organism to animals with higher cognition (Maturana and Varela 1991; Varela 1992). Most importantly, this ability to adapt arises autonomously within an animal (Varela 1991). This ability of adaptation depends fundamentally on the very capabilities that the animal is disposed with (in terms of sensations and movements) and further depends also on the ability of the animal to learn from previous experience.

While many biological systems have limited generalization capability and learning performances, humans employing their capability of reasoning at their best capacity, demonstrate flexible and complex, even novel behaviors, by generalizing well and adapt-

ing to the changing environment. Similar differences can be seen across other species in terms of behavioral flexibility as a function of their ability to reason (for example, from macaques to chimpanzees and gorillas). However, these emotional and motivational bases as a part of higher order cognition, which have been argued to be the basis for intelligent behavior (Canamero 1997), are seldom addressed in computational neuroscience and in cognitive science, and they still elude the field of robotics and A.I.

In the context of studying complex human behaviors, this thesis urges the need to describe a global cognitive architecture in which any cognitive operation such as decision-making, planning or learning is studied with a mandatory reference to the relation between the body and the environment. It is argued that understanding the constituents of high-level cognitive functions depends on the organization of their fundamental characteristics and properties deep rooted within the brain-body system.

.

Central Objective

Within the available capabilities of voluntary behavior, humans, besides some other primates and mammals, can exhibit the most flexible and complex behavior. As a function of evolution of brain development across species, one of the major structural changes that makes humans stand out across all of primates, carnivores, and rodents is the size of a rather recent brain region, Prefrontal Cortex (PFC). This increased PFC is believed to facilitate most of the high-cognitive functions (executive functions) we involve in. One of the crucial building blocks of flexible behavior is flexible decision-making. Beyond perceptual decisions, which are more classifying than evaluating, the interactions we make with the environment shape the contingencies we associate with each object, action and interaction with respect to that which is appetitive for us or aversive. Inevitably, we learn these associations and take them into account for future interactions. While this is true for many animals, the amount of reasoning we employ in this learning defines the sophistication of our behavior. Understanding these value-based decision-making and

learning mechanisms within PFC provides a basis to understand the reasoning and more flexible behavior. While such mechanisms are many, one of the regions of PFC called the Orbitofrontal Cortex (OFC) is specifically known for learning emotional associations between objects in the environment and an appetitive value thus attributing an emotional value to an object. Further these associations shape the decision-making and learning processes across several other PFC regions. The aim of this thesis is to study the processes within OFC in the context of the rest of PFC and to analyse how these processes shape behavior. As OFC is involved in these processes in conjunction with several other sensory and prefrontal cortical, and sub-cortical structures of the brain, it is first needed to look at the possible organization of these brain regions that interact wit OFC. Only by identifying the global place of OFC in the brain organization as well as the landscape of decision-making and learning, will it be possible to come any closer to understanding detailed processes within OFC.

Since the scope of this study is not necessarily to outline the neuroscientific correlates of intelligent animal behavior, but rather to outline its general principles, the approach taken will be computational and algorithmic modeling of the known brain pathways by constantly finding supporting evidence from relevant neuroscientific studies. On the other hand, this effort into understanding the process related to behavior in the scope of OFC will inevitably open numerous avenues to study the same questions from the point of view of Computer Science, Robotics and Artificial General Intelligence. Therefore this work primarily aims at :

Road Map for the thesis

Behavioral and Computational theories

On the side of behavioral theories, a few classic theories like Pavlovian conditioning, instrumental learning, reinforcement learning (RL) are explained. On the computational side, RL is elaborated with formal mathematical notations and implementations. The

importance of visiting these classical theories is that many of the attributes that make up these theories like *stimulus, reward, error* form the basis of discussing many decision-making paradigms.

Global organization of behavior in the brain

Major brain regions that play a crucial role in voluntary behavior are indicated in 1.1, as compiled in great detail in Alexandre 2016. The regions are color-coded according to the information flow of a certain behavioral scenario. Although, not all of the regions highlighted in the figure 1.1 are discussed in this thesis, it gives a good picture of the parallel and segregation across the brain regions in terms of the information they process. The regions that are implied globally are :

1. **Frontal cortex** : PFC, occulomotor, premotor and motor cortices
2. **Sensory cortex** : Temporal, Insular and Parietal cortices
3. **Basal Ganglia** : Dorsolateral Striatum (DLS), Dorsomedial Striatum (DMS) and limbic part, also called Nucleus Accumbens (NAcc) with a shell and a core; output structures like the internal Globus Pallidus and the substantia nigra pars reticulata (GPI-SNr) and the ventral pallidum VP ; with dopaminergic regions - ventral tegmentum area (VTA) and substantia nigra pars compacta (SNc)
4. **Extracortical structures** - Amygdala, Hypothalamus, Superior Colliculus (SC) and Cerebellum
5. **Hippocampus**

However, within the scope of this work, the most relevant regions to OFC are the prefrontal cortex in general, BG, Thalamus, Amygdala and Hypothalamus. First, to place OFC among the other prefrontal regions, a general anatomical description of PFC, then the organization of cortical regions with the BG in the form of parallel and segregated loops, followed by a description of Anterior Cingulate Cortex (ACC) in particular, is

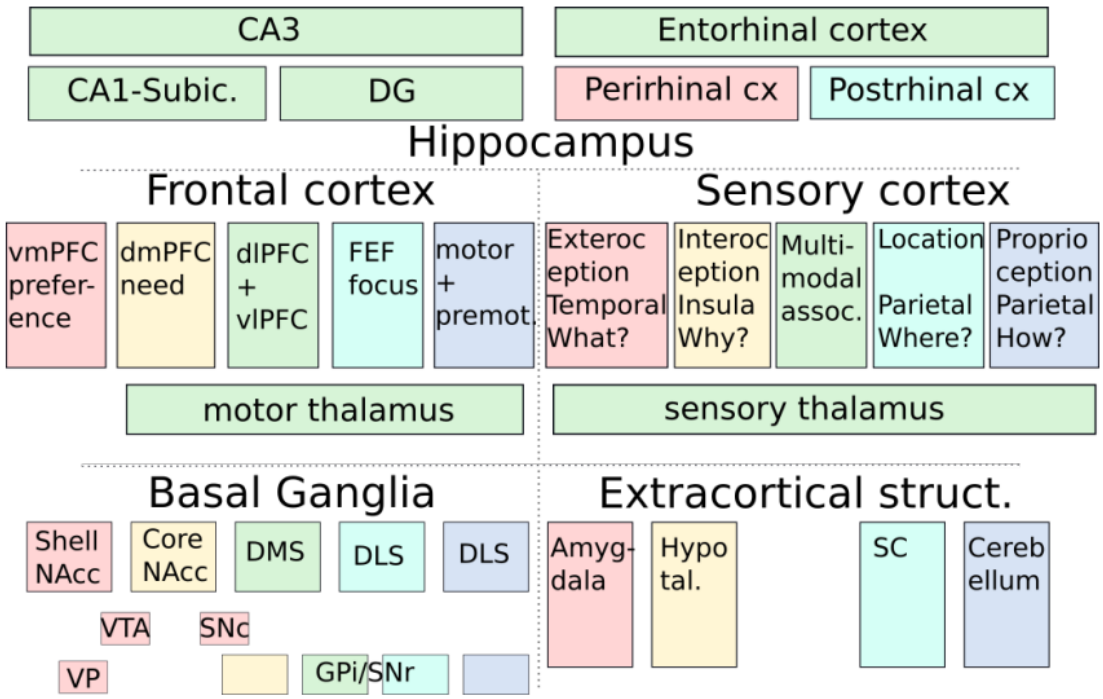


Figure 1.1: A general high-level segregation of several brain regions - prefrontal and sensory cortices, the basal ganglia, extra-cortical structures like amygdala and hypothalamus. The parallel and segregated loops are highlighted in different colors. As compiled in Alexandre 2016.

given in chapter 3. The chapter 4 gives an extensive review of literature on OFC, its role in value-based decision-making and several theories about how the dissociation of OFC sub-regions play distinct role in behavior.

Cognitive architecture of voluntary behavior

Functional loops associating cerebral structures including the basal ganglia in the brain of most species along the evolution are highlighted (Alexander 1986). They are dedicated to the organization of behavior under the constraint of reinforcement, in their simplest expression, corresponding to the selection of action for survival.

First, considering an artificial agent with internal motivation in an external world,

voluntary behavior is expressed in terms of four parallel loops each answering a particular questions about the environment.

1. The *Why* loop selects the current motivation from the interoception of agent.

2. The *What* loop selects the goal according to the preferences. The goal object can be consumed if it is directly available, otherwise it will become the goal for the spatial and temporal organization of the behavior.

3. The *Where* loop considers the spatial location of the goal and selects the orientation behavior relevant to face it.

4. The *How* loop supports the latest postural adjustments when the goal is attainable, by simply reducing the distance or possibly manipulating the object before consuming it.

The parallel organization of cortical regions into feedback loops with the BG (Alexander 1986), as described in chapter 3 is then imposed on the simple architecure of the parallel loops of questions described above. Essentially, with respect to the information flow shown in figure 1.1, the aspects that are taken into consideration for this work are now represented in figure 1.2. This functional description highlights that the generic processing of response selection by the BG is ascribed in a generic loop, also associating the frontal cortex, subcortical and cortical sensory structures.

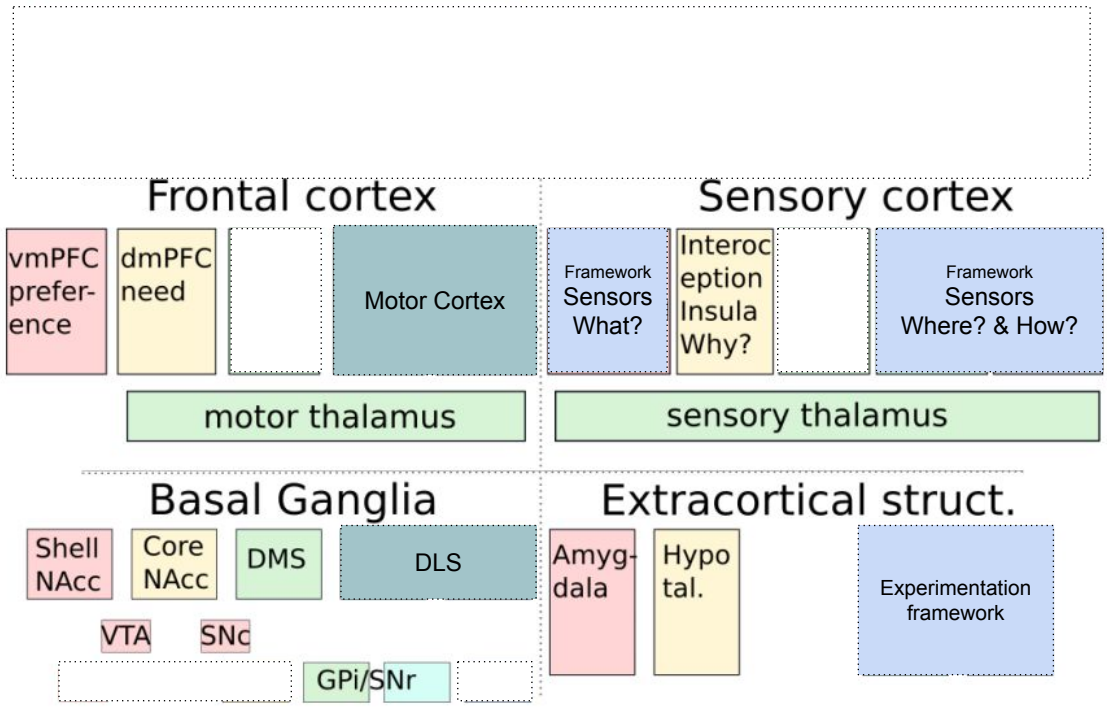


Figure 1.2: Simplified version of the loops in the figure 1.1. The neural counterparts that are modelled as a part of the framework.

Experimentation

Many recordings of neural activity in people come from the brains of those with epilepsy who are due to have brain tissue removed as researchers are limited by ethical considerations in terms of how much they can intervene in processes in the healthy human brain. On the other hand, animal models enable researchers to use more invasive procedures, but it still limits in the case of high-level intelligent behaviours that cannot be replicated in other species. AI systems that can mimic human behaviour and be modified will provide scientists with extra tools for exploring how the brain works: researchers could teach a network to reproduce certain behavior, and then impair some of the involved attributes to observe what happens, for instance.

A video game environment, Malmo is used as a virtual experimentation platform to design experiments that simulate scenarios closer to ecological situations. Malmo is dedi-

cated to support research in various AI related areas and it allows to incorporate various models of reinforcement learning, planning and related problems into the Minecraft game environment. The models that can be incorporated range from a basic Q-learning algorithms on a single agent to more collaborative and competitive strategies among multiple agents. Malmo has been used for various kinds of specific experimentations like learning to navigate in the Minecraft world (Matiisen et al. 2017) and computational models of animals living in block world (**Strannegard2018b**). In most of such cases, very specific feature of Malmo is used (either the 3D space in the environment or the block nature of the world or the agent to perform a task). We exploit, simultaneously, various features of Malmo like the agent's internal body attributes, external constraints like sensory perception, and progress in the execution of action after decision. Subsequently it allows to reproduce several behavioural experimental scenarios similar to those which are studied using animals but more difficult (like studying decision-making in a freely-moving animal).

Model of information processing within OFC

With a stable framework of CBG loops in place, existing computational models explaining different behavioral paradigms like Pavlovian conditioning or Instrumental learning are combined into a single model of OFC and the limbic system. Furthermore, several hypothesis about the dynamics of reward-guided learning have been implemented in the model of OFC and several behavioral studies on animals have been replicated with the model. The model and the results are analyzed to understand the underlying principles behind the hypothesis that were derived and tested.

Note :

I personally do not have a strong affinity towards putting 'labels' on things. More so when it comes to the brain and neuroscience in general, I would refrain myself from restricting and compartmentalizing a certain brain region or structure to a certain behavioral phenomenon.

That said, however, as demanded by the larger goal of this thesis, to discuss flexible animal behavior with a high-level systemic view of the brain, I took the liberty of 'confining' the scope of each of the brain regions discussed to a certain degree of specificity, as needed for the argument, but with a thorough scrutiny in the literature. For instance, on a high-level, it would appear as if :

- Basal Ganglia (BG) : action selection mechanisms and coordinating learning with the prefrontal cortex
- Prefrontal Cortex (PFC) : abstraction of not-so obvious information from the environment and using it for behavior (in conjunction with BG or alone), thus flexibility and adaptability.
- Amygdala : Learning emotional valence of otherwise neutral elements of the environment.

Of course, while not discussed in depth, a wide range of other implications of each of these brain regions/structures in behavior are certainly surveyed and taken into account

Chapter 2

Behavioral and Computational Theories Of Decision Making and Learning

Sommaire

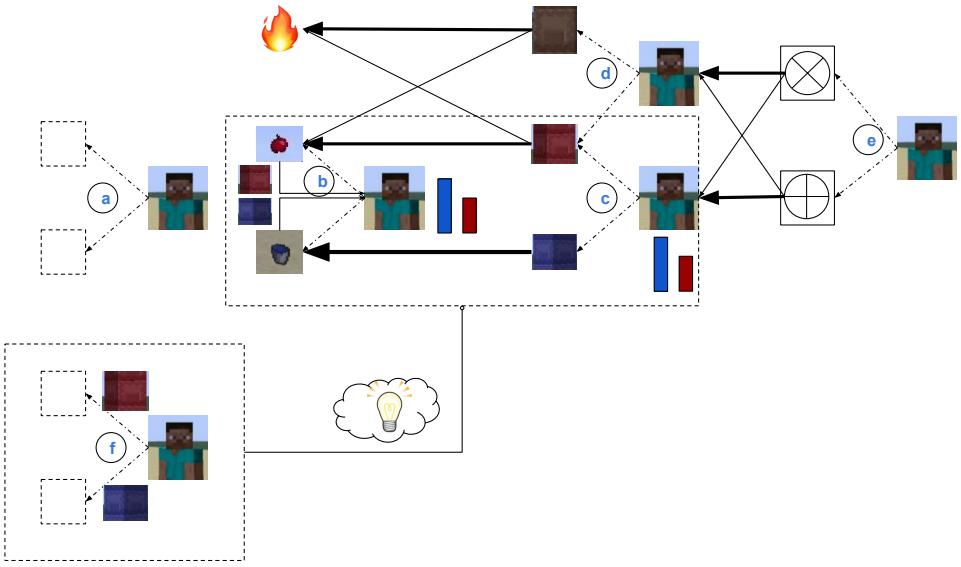


Figure 2.1: An illustration of different choice situations an agent can face with. (a) A simple choice between actions right and left with no added criteria. (b) A choice between two appetitive stimuli, each relevant to different motivation, requiring to do specific actions (right or left). (c) A choice between two otherwise neutral stimuli, colored blocks, each leading to an appetitive stimulus. (d) A choice between two neutral stimuli, one leading to an appetitive stimulus and the other to an aversive stimulus, needed to avoid. (e) A choice between two neutral stimuli, symbols, where one of them could most likely lead to another choice, a totally appetitive choice but also a little bit likely to lead to risky choice, and vice-versa, the other could lead more likely to the risky choice and a bit likely to an appetitive choice. (f) A simple choice like in (a) except that there is more information which is similar to the appetitive choice made elsewhere.

Most of the animals rely on their expectations about the appetitive and aversive outcomes of their actions in the environment, to continue a behavior that guarantees their well-being. On formal grounds, these expectations are referred to as representations of "value". Such a "value" plays at least two crucial roles in behavior. Firstly, *value* can drive choices. Values as the expectations of outcomes of the available choices, render the choices comparable on some abstract common grounds. Depending on the relation of the outcome being appetitive or aversive, the best or the least "value" can be chosen. Secondly, *values* support learning. The expectations are measured against reality and they

are constantly revised, pushing the future expectations closer towards reality. What follow in this chapter are the concepts of value, its role in decision-making, learning and several theoretical constructs around them, and a computational theory called Reinforcement Learning (RL) which has formalized this behavioral theory with a strong computational foundation.

2.1 Value

In what was referred to as a behaviorist account of human behavior, and what is one of the earliest theories of choice proposed, Skinner describes human behavior as merely an association between a response and an evoking stimulus, by virtue of *earlier contingencies of reinforcement* Skinner 1938. Alternatively, psychological and economic theories (Bandura 1997; Von Neumann and Morgenstern 1944) highlighted *value*, precisely *subjective value* or *expected value*, is at the heart of choice process. Such an *expected value* could possibly represent several dimensions of the possible outcome to the response in the choice, for example the magnitude of the outcome (on some scale) and the very probability of the outcome. In contrast, Skinner, in what can be described as an *associative model* (Skinner 1965), argues that *values* are not quantifiable mental representations of the agent, but are just inferred by the observer from the behavior. Such an associative model of behavior, sufficiently explains several behaviors like foraging, or in classic economic choices where responses can be learned by moving towards maximizing a certain *utility*. However, humans are known to make complex choices, even in novel situations, that cannot be described by simple conditional responses. Modern theories underpin this ability to — value — whereas an *associative* point of view attributes it to intrinsic noise or randomness. Experimental evidences over decades in monkeys (Padoa-Schioppa et al. 2006; Sugrue et al. 2004) as well as humans (Ravlin and Meglino 1987; Verplanken and Holland 2002) showed the transitive nature of value and hence it cannot be noise. Most importantly, contrary to the main prediction of *associative* accounts animals when faced with new situations, learn to make choices gradually. it was shown that it is possible for

animals to effectively make choices in novel encounters by relying on internal construct, which is often referred to as *value representations* (Christopoulos et al. 2009).

On the other hand, prominent economic theories also highlighted how humans do not necessarily behave rationally all the time, as one would expected in a pure value-based framework (Kahneman and Tversky 1984). Many more constructs like biases and heuristics have been introduced. Notwithstanding these diverse theories, the quantitative and qualitative expectations about the outcomes constitute important considerations in studying decision making.

2.2 Decision Making

Perceptual decision-making refers to the process of arriving at a decision using almost solely sensory processing than any other abstract representation on top of it. Most common scenarios include classification and identification situations. It is like the baggage screener at the airport security effectively identifying that liquid bottle more than 100ml you hid in the bag. Although computational explanations of these problems still rely on an objective value (like confidence or similarity) and ultimate comparison on that value (for instance, neural networks used for image classification), whereas in reality these judgements can be explained by adaptive sampling of sensory information (Cassey et al. 2013).

Value-based decision-making In case of humans, non-human primates and many other animals, behavior depends not only on (the representation of) the sensory input, but also on past memory, internal motivations and goals, and predictions about possible outcomes. In a fundamental ecological context, an animal is presented with decisions that need to be evaluated at several levels of abstraction. An animal interacts with the environment by choosing from a set of available actions that are biased by the expectations of outcomes. An appropriate choice between available courses of action can be viewed to rely on value representations on these actions based on their consequences. If the possible outcomes are proximal in time, but not apparent from the current sensory state, the animal should rely

on the ability to retrieve relevant past experiences from memory, and subsequently use them to make the current choice. On the other hand, in the modern day societal context, humans are faced with economic decision-making, based on the valuation of the choices in various dimensions, added by the intrinsic *risk* involved in making those decisions. Value-representations, although subjective in general, provide an objective framework to take into account these varying dimensions, and many more related to the animal's internal state like preferences, needs and the actions costs. Much of this thesis revolves around value-based decision making and numerous interpretations provided with the help of comprehensive decision-making studies will be discussed in the following chapters.

Foraging In both naturalistic animal environment or modern-day human societal context, choices might very well be comprised of distal outcomes (outcomes possible in future provided the current choice is made). A pure value-based representation of decision-making would not account for a distinct system for foraging. However, animals in the context of encounters experienced sequentially, take into account different contextual information about the environment. Thus the question becomes not only how to make a choice among the available options, but also whether or not to engage in the current choice in order to find something else somewhere else. Several accounts have shown that besides some key neural decision-making mechanisms dealing with a limited-number of choices, there could be distinct mechanisms that account for a more ecological seemingly binary choice, that is foraging and the specific *value* constructs that would comprise foraging (like costs of moving away from the current choice) (Rudebeck et al. 2006).

2.3 Learning

In this thesis, a connection is developed from the sort of computations that happen in the dopamine system to how behavior is influenced through those circuits. The study of animal conditioning is broadly divided into two main areas: Classical or Pavlovian and instrumental or operant conditioning.

2.3.1 Pavlovian / Classical conditioning

Pavlovian conditioning is described when the occurrence of an innate bodily response is shifted from its natural trigger to an unrelated stimulus (Fearing et al. 1929). By repeatedly associating the unrelated stimulus to the original stimulus that elicits the natural response, eventually the response is triggered just by the occurrence of the unrelated stimulus. Famously studied by a Russian psychologist Ivan Pavlov, in this experimental paradigm, the neutral stimulus is often referred as a Conditioned Stimulus (CS) and the stimulus that elicits a natural response is referred as an unconditioned stimulus (US) and the natural response that is elicited upon US is called unconditioned response (UR). A dog's natural response of salivation (UR) upon seeing or smelling food (US) can be, by repeatedly pairing the sound of a bell (CS) to the food. Thus, after sufficient pairing, the sound the bell elicits salivation even in the absence of food. Similar association applies for punishments too. When the US is something like an electric shock that causes physical harm, by repeatedly pairing a tone to the shock, the tone alone can elicit fear response in the animal. Importantly, the animal learns to make these associations without having to perform any action. It is just the co-occurrence of the stimuli that forms the association. This form of conditioning is modelled (Rescorla and Wagner 1972) in this thesis, as it is a crucial part of animal behavior in learning the emotional values of neutral cues. Such an association with a CS will further suggest the animal, upon perceiving the CS, to perform a preparatory course of action to expect a US.

2.3.2 Respondant / Operant Conditioning

Respondant / Operant Conditioning, also called instrumental learning, involves animals doing voluntary actions that results in an outcome. That is by associating a response of the animal to an outcome, the likelihood that the animal would elicit that action is altered. In one of the earliest original experiments, Edward Thorndike observed that cats increased the probability of their lever-pressing response upon associating this with a escape and reward, as opposed to doing it only by chance before. This gave rise to

Thorndike's "Law of effect" which states "responses that produce a satisfying effect in a particular situation become more likely to occur again in that situation, and responses that produce a discomforting effect become less likely to occur again in that situation" (Thorndike 1898). Further this behavioral theory was demonstrated in more detail by the famous Skinner's Box (1948), where rats learned to press the lever more after realising that it is associated with a food pellet.

While Pavlovian conditioning causes an animal to learn to expect motivationally significant events, it instrumental learning that allows the control over these events in the form of learned actions in order to satisfy the motivations (Balleine and Dickinson 1998). Instrumental behavior can enable the animal to demonstrate habitual or goal-directed behaviors. Habitual behavior learns about the instrumental contingency between the action and outcome, whereas Goal-directed behavior consists of the acquisition of action and incentive value of the outcome. The primary difference is that after learning to associate an action to a food outcome for instance, in case the animal is satiated for this food, goal-directed system immediately updates the value of the action with the current value of the food outcome (which is nearly none because of satiation) where as habitual system still has the same value for the action as it learned when the food outcome was valuable (before satiation).

2.3.3 The Pavlovian Instrumental Transfer (PIT)

Pavlovian Instrumental Transfer is a transfer paradigm when separately learned stimulus-outcome associations and action-outcome associations interact and give rise to forming stimulus-action associations, which are not explicitly taught (see Cartoni et al. 2013 for a good review). Demonstrated widely in rats as well as humans, subjects are first taught to associate a stimulus to an outcome without having to perform any action (Pavlovian). Later, the subjects are allowed to choose from different actions and learn which action is associated to the outcome (Instrumental). Finally, the subjects are allowed to choose from the same actions, tested both in the presence and absence of the stimulus that was

learned before (Testing). Results show that the subjects perform the action that gave the outcome more when the stimulus is present than when it is absent. This is the case even if the stimulus and the action were never presented together before the testing phase. This transfer of learning from the Pavlovian system to the instrumental system can be studied under two forms (for e.g Corbit et al. 2016).

Specific PIT When a conditioned stimulus (CS) associated with an outcome makes an instrumental action linked to the same reward more likely, it is referred as specific PIT. It can be viewed as a chain of stimulus-outcome and outcome-action associations resulting in stimulus-action association.

General PIT General PIT, the presence of CS enhances an action linked to an outcome different from the one CS predicts. That is, the presence of a stimulus which is linked to an outcome generally increases the motivation of performing all the actions that are linked to any outcomes, if there was no particular action linked to this particular outcome that the stimulus was linked.

Table 2.1: An example PIT task

Pavlovian	**Instrumental**	**Transfer**
S1 -≻ O1	R1 -≻ O1	S1 : R1 vs R2
S2 -≻ O2	R2 -≻ O2	S2 : R1 vs R2
S3 -≻ O3		S3 : R1 vs R2

2.4 Neurotransmitters and Behavior : Dopamine

Dopamine as a scientific topic is one that has protruded arguably the most into common vernacular, becoming almost a proxy word for addictions. Dopamine, likewise serotonin, noradrenaline and acetylcholine are neurotransmitters or neuromodulators that are exchanged from specific sites in the brain projecting to most of the brain regions. Neurons in specific regions have specialized receptors at different sites for one or more of these neurotransmitters. Essentially, in tightly and widely interconnected anatomy of brain, the information broadcasting in the form of neurotransmitters can affect the intrinsic

properties and the functioning of neurons across extended widespread networks (Dayan 2012). Each of these neurotransmitters have intrinsic properties of excitation or inhibition, sometimes both depending on the associated processes. Several experimental and computational accounts have proposed mechanisms for Acetylcholine in context learning and expected uncertainity (Calandreau et al. 2006; Yu and Dayan 2005), Noradrenaline (or Norepinephrine NE) in unexpected uncertainty in the case of reversal of expected contingencies (Aston-Jones et al. 1997; Bouret and Sara 2005; Yu and Dayan 2005). While GABA is generally considered as an inhibitory and sensitive to quantitative aspect of outcomes (Eshel et al. 2015), Dopamine and Serotonin have been strongly debated for varying implications including appetitive and aversive reward prediction (Cools et al. 2011; Daw et al. 2002), behavioral inhibition and impulsivity (Doya et al. 2002), and risk (Balasubramani et al. 2014).

Dopamine has been at the forefront of widespread popular beliefs, controversies and eluding inquiry into pleasure and addictions. Part of it can be attributed to one of the earliest experiments done by James Olds and Peter Milner in the 1950s by allowing the rats to self-stimulate Dopamine release into their striatal regions in brain to the extent of foregoing eating and drinking, thus concluding that release of dopamine is the cause of pleasure (Olds and Milner 1954). However this popular theory was overturned by Kent Berridge and colleagues, who highlighted a distinction between the motivation to seek rewards and the pleasure obtained from them. That is, essentially separating 'liking' to do something from 'wanting' to do something. For example, someone who is addicted to smoking but trying to quit, might step outside with a compulsion to smoke but it is highly unlikely that the individual would describe this smoking experience as pleasurable. It was shown by Berridge and Robinson 1995 that while a different chemical subsystem, the μ-opioid system regulates seeking pleasure ('liking'), dopamine regulates the drive to do something ('wanting').

One of the most commonly agreed view on Dopamine is that its release signals, what is described as, a reward prediction error (RPE) (Schultz et al. 1997). As mentioned earlier regarding the value-representations, an action is elicited with an expectation of

outcome. RPE is basically the difference between what was predicted and what was actually experienced. In a simplified sense, if what was experienced was better than predicted, the RPE is positive and otherwise it is negative. Therefore RPE turns out to be crucial in learning since the future actions depend on updating these expectations depending on the experience. This aspect of Dopamine's connection to RPE is utilized in this thesis, as a core learning mechanism through neuromodulation in the cognitive architecture.

2.5 Reinforcement Learning (RL)

Reinforcement learning (RL) is a computational theory that formulates the behavioral and psychological understanding of how learning occurs through interaction with the environment for specific goals and the reinforcement of the interaction. As opposed to explicit supervision and teaching, the learning occurs by taking actions using certain criteria (termed as *policies*) and further taking into account the consequences of these actions to update the *policies* or likelihood of taking the same actions. From a computational perspective, RL differs from traditional supervised and unsupervised machine learning algorithms in the sense that the feedback after a transition is more explicit in the former and none at all in the latter.

In RL, *state* - the observation space of an environment, and *action* - what an agent can do to cause change in the environment, are two most fundamental building blocks. The goal in RL is defined by a *reward function*. Roughly speaking, reward function maps each perceived *state* (or *state-action* pair) of the environment to the intrinsic desirability of that *state*. A *value function* represents the total amount *reward* the agent can accumulate over time. Broadly, RL methods are often discussed in two categories which optimize value functions and policy estimations in totally different approaches : *model-free* and *model-based*. In the context of a choice, besides the selection mechanisms, action values (AV) or object values (OV) can be acquired or modified either by actual experience or by predictions derived from the internal model of the environment, as described in *model-free*

and *model-based* reinforcement learning, respectively.

2.5.1 Model-based and Model-free RL

Model-based algorithm is an algorithm that uses a transition function, which is a model of the environment, and the reward function in order to estimate the optimal policy. Model-based learning separately updates the model of environment and the control of the model on reinforcement learning, which facilitates the outcome prediction (or a simulated experience) Doya et al. 2002; Sutton and Barto 1998. Model-based system learns the state transition structure of the environment and searches through this model to generate predictions about future reward.

A *model-free* algorithm either estimates a *value function* or the *policy* directly from experience (that is, the interaction between the agent and environment), without using neither the transition function nor the reward function. Model-free system is computationally efficient as at every choice point, it is only an incremental adjustment of the expected values, albeit only upon experience (outcome). However, it doesn't track the transition and underlying structure of the environment, thus requiring a lot of experience to learn reliable predictions. Any momentary inconsistency would be marked as a prediction error, and is used to modify the plasticity that allow more accurate learning.

Essentially, both the systems have complementary strengths and weaknesses, with a trade-off between statistical and computational efficiency Dayan 2009. This has been highlighted by the recent observations that humans performing a 2-stage task based on state transitions, showed a mixture of both the systems in terms of their behavior Daw et al. 2011; Gläscher et al. 2010, and by common model-based policy optimizations with model-free tunings Ha and Schmidhuber 2018.

RL happens to be a central component of this thesis as the very definition of a reinforcement learning problem lies in identifying an agent within an environment which is trying to maximize the rewards the environment has to offer. One of the most relevant parts of RL to this thesis is the Temporal-Difference (TD) learning, which is the closest

formulation to the role of Dopamine in learning in behavioral scenarios.

2.5.2 Temporal Difference RL

Often regarded as the central idea of RL, temporal-difference learning methods do not need the model of the changes in the environment and learning of the expectations happens by parts, without having to wait for the final outcome, to learn. This makes it particularly relevant to the natural scenarios where the outcomes that an animal would get from its actions are usually distant in time. Simplest mathematical representation of a TD-learning algorithm is shown in the equation 2.1, where δ_t is termed as the temporal-difference, $\hat{V}_t$ and $\hat{V}_{t+1}$ are the expected rewards at time t and $t+1$ respectively and R_t is the actual reward received at time t. γ is called the *Discounting Factor*, which basically controls the present value of future rewards.

$$\delta_t = R_t + \gamma \hat{V}_{t+1} - \hat{V}_t \tag{2.1}$$

RL, especially TD, was in part inspired by earlier computational formalisms that were based on psychological principles which related the concept of secondary reinforcers (like a CS) and its reinforcing properties (Minsky 1954). Most importantly, the neural correlates of temporal difference learning were identified in the way Dopamine fires in connection to the reward prediction errors (Schultz et al. 1997), making equation 2.1 a concrete tool to understand learning in biological behavior.

Most of these behavioral and computational theories have been extensively studied and continue to be active areas of research in the field of neuroscience, to investigate the neural correlates of these paradigms. Similarly in the field of robotics, these paradigms have been addressed extensively in the context of autonomous agents. Especially by endowing bodily attributes like pleasure to artificial agents and thus empowering them to autonomously define their goals, several robotic architectures have described the possible transfer of neuroscientific understanding processes like value and motivation based decision making into situated autonomous agents (Canamero 1997; Lewis and Cañamero

2016; Schrodt et al. 2017; Verschure 2012). Such frameworks provide an extensive account of the external sensorimotor processes with respect to affordances (Cisek 2011) in the environment and the internal homeostatic processes, together resulting in the agent behavior (Cos et al. 2010). They in turn inspired novel reinforcement learning (RL) algorithms such as multi-objective Q-learning (Strannegård et al. 2018) and homeostatic RL (Hulme et al. 2019; Keramati and Gutkin 2011, 2014). As both the fields of neuroscience and robotic cognitive architectures approach to an interacting point, a better integration of experimental and computational methods with real-world robotic systems permits a validation of the proposed architectures of autonomous agents in ecologically valid scenarios (Adams et al. 2012; Pezzulo et al. 2011).

However, the core interest of this thesis lies in identifying, understanding and describing the brain systems involved in flexible behavior. A high-level cognitive architecture will be described in a simple manner that accommodates the interactions between the proposed brain systems and allows to the expression of behavior in an external environment. Regarding the neural correlates found in the brain, especially the Prefrontal cortex (PFC) and other relatively older brain structures like basal ganglia (BG) and amygdala, the neural processes behind several of the above discussed paradigms in these brain regions will be explored in the following chapter.

Chapter 3

The Prefrontal Cortex (PFC)

Sommaire

The prefrontal cortex (PFC) is a large part of the frontal lobe in mammalian brain. In humans and most primates, PFC has been implicated in various higher order mental faculties like planning complex cognitive behavior (Luria 2012; Shallice 1982), attention (Shallice 1988), goal-directed behavior (Hunt and Hayden 2017), working memory (Goldman-Rakic 2011), expectations based on actions and their outcomes, and in social conduct (Yang and Raine 2009). In psychological terms, most of these functions are termed as *executive functions* and PFC is believed to play a crucial role in executive functions, executive control and top-down modulation of bottom-up processes (Frith and

Dolan 1997). The cognitive control exerted in these executive functions employs several mechanisms like working memory (Baddeley et al. 1986), conflict resolution (both in abstract thoughts and concrete choices), outcome evaluation and response inhibition (in novel situations where past knowledge no longer holds). Several sub-regions within the PFC have been implicated in these functions, although not that specific sub-region maps uniquely to each of them. In fact, these executive functions are distributed across multiple interacting regions within the PFC as well as the sub-cortical structures that they are connected to. Arguably, it is the expansion of the PFC in humans, compared to other primates, that contributed to human-unique higher order cognitive functions. In non-primate mammals, PFC is known to support "conditional motor behavior" (Passingham 1993), where the behavior of the animal takes external context and internal state into account.

Further in this section, a brief anatomical definition of PFC will be described, predominantly with respect to primates, and a short note on the equivalence of PFC in rodents. The striking organization of cortical regions in the form of closed feedback loops with the Basal Ganglia (BG) will be described, as it forms the basis of modeling descriptions in this thesis. Few noteworthy sub-regions of PFC and their implicated roles in behavior, their organization within PFC and across sub-cortical or sensory areas will be highlighted. Understanding this organization of PFC with sub-cortical structures like basal ganglia (BG) and amygdala helps studying the central theme of this thesis, the Orbitofrontal cortex and decision making (as will be seen in the next chapter 4), and the complementary role of the anterior cingulate cortex (later in this chapter).

3.1 Brief anatomy of PFC

Referred generally as the "granular frontal cortex" in primates, PFC is anterior to the premotor and motor cortices and is distinct in terms of cytoarchitecture. This well-developed granular part of the frontal cortex is unique to primates and cannot be found in other mammals (Krettek and Price 1977). Evidence from neurophysiological and neuropsy-

chological studies of the PFC in macaques highlights its additional role in learning and applying abstract concepts and rules (Miller et al. 2003; Wallis et al. 2001). However, considering the lower order behavioral flexibility, it pushes for a rather more inclusive definition of PFC to accommodate in non-primate mammals like rodents to account for behaviors like context-dependent action selection. One of the approaches that allows to define PFC anatomically across species is the connectivity with the dorsomedial nucleus of the thalamus (Kolb 2007; ROSE and WOOLSEY 1948; Seamans et al. 2008). This kind of approach includes dysgranular (discontinuous layer IV) and agranular (layer IV absent) areas like dorsal part of anterior cingulate cortex (ACC), Brodmann area 13 (as will be seen in section 4.2).

Several distinct sub-regions of the PFC can be identified using the anatomical directional terminology. The area that is above the orbit of the eyes is referred as the Orbitofrontal cortex (OFC), the lateral and medial views of the PFC, lPFC & mPFC show a distinct connectivity patterns and implicated with different functional roles. Medial PFC can be further looked at distinctly as ventromedial prefrontal cortex (vmPFC, usually implied closely with the OFC) and dorsomedial prefrontal cortex (dmPFC, part of which is ACC). Extensive evidence suggests that OFC & vmPFC, often complemented with the dorsal part of ACC & ventral striatum form a robust value system that feeds the decision making processes. Lateral PFC is connected to wide range of secondary sensory regions like Frontal Eye Fields (FEF), secondary visual cortex, parietal cortex, supplementary motor cortex and pre-motor cortex. Especially dorsolateral prefrontal cortex (dlPFC) in primates has been found crucial for the most flexible, complex, and expectation-oriented behaviors that need to be organized, planned and produced.

Short note on rodents : Notwithstanding the fact that several physiological and behavioral studies are conducted on rats using complex task structures, the equivalent of PFC in rat brains is a debated argument. Based upon topological definitions and connectivity similarities (ROSE and WOOLSEY 1948) to, for instance, mediodorsal thalamaus (MD), equivalent areas to the caudal and ventromedial areas of the primate prefrontal cortex were recognized in rats, with some limitations (Preuss 1995; Uylings et al. 2003).

Furthermore, more specific comparisons have been made in rats to identify more in rest of the nonprimate mammals, an equivalence of the PFC could be established, with areas with contextual inputs from higher sensory areas and about current needs from amygdala, influencing action through pre-motor cortex.

3.2 Cortico-Basal Ganglia (CBG) loops

It is the overwhelming complexity of behavior and the nature of decision making processes that underline the compelling need to study individually (and together) different subsystems of the Prefrontal Cortex (PFC). It is difficult, if not impossible, to precisely define the sub-regions and to describe a reliable information flow through the PFC. However, thanks to the anatomically distinct and parallel loops in the frontal lobe, it is possible to sketch a rough path ; from acquiring, processing and representing sensory information in the lateral and medial PFC to the other parts of PFC that plan the required responses for the expression of desired behavior, through other cortical regions like pre-motor and parietal cortices. These loops originate in the frontal lobe, involve specific subregions of the Basal Ganglia (BG; striatum and pallidum) and certain nuclei of Thalamus, and ultimately affect areas of the frontal lobe.

The prefrontal regions learn the sensori-motor associations, the contingencies between the stimuli and actions, in the form of extensive antero-posterior cortico-cortical connections resulting in habitual (Topalidou et al. 2018) and goal-directed behaviors (Valentin et al. 2007). However, the association of (interoceptive or exteroceptive) sensations with (internal or external) responses is sometimes straightforward and a simple sensori-motor structure is enough to trigger the response. But the involvement of BG is essential when the selection of the response (e.g., goal or action) is based on ambivalent or uncertain criteria (Floresco 2015). Moreover, there is no unique prefrontal cortical region that integrates all of contextual, reinforcement information with the corresponding sensorimotor representations. Rather, it is the extensive network that prefrontal regions form with several of the basal ganglia (BG) nuclei, that facilitates a closed-loop framework of decision mak-

ing, reinforcement of choice and subsequently adaptive behavior. It has been highlighted that this network (hereafter referred as CBG network) comprises several feedback loops (CBG loops) which are described as parallel and segregated (Alexander 1986) because they correspond to distinct and related territories of the structures involved. Because they are structurally similar (in terms of involved neural populations and connectivity), suggesting that the same kind of processing is applied generically to different information. It should be noted that other sub-cortical structures like amygdala, hypothalamus and other sensory regions also play a crucial role alongside these loops.

These parallel loops are classified into three major classes : *limbic*, *sensori-motor* and *associative*, shown schematically in Fig. 3.1. The *limbic* loops originate in the orbito-medial prefrontal cortex (generally comprised of OFC and the Anterior Cingulate Cortex (ACC), through amygdala, hypothalamus and the subdivisions of ventral striatum (nucleus accumbens) and end back in the medial PFC. The *limbic* loops (figure 3.1, blue), besides processing external information, are based on interoceptive information. They are organized around the selection of the goal of the behavior, according to its motivational value, in response to perceived needs or according to its hedonic value. Individual subregions of medial prefrontal cortex form the feedback loop through different nuclei of ventral striatum (Kringelbach 2005; Niv et al. 2007), together receiving information from the temporal and insular cortices. The *sensori-motor* loops (figure 3.1, red) originate in the sensorimotor and premotor cortices, process exteroceptive information. They are organized around the motor behavior allowing to reach the goal, according to its spatial position (orientation) or according to the physical characteristics involved (handling). The occulomotor and motor cortices form the feedback loop through the dorsolateral striatum, together receiving the required motor information from the parietal cortex (Alexander 1986; Sommer and Wurtz 2004)). Finally, the *associative* loops, involve the lateral prefrontal cortex and dorsomedial striatum (and corresponding specific nuclei of thalamus). *Associative* loops (figure 3.1, green), also called cognitive loops, are implied in cognitive control (Koechlin et al. 2003), related to the ability to manipulate abstract rules. The lateral prefrontal cortex forms an associative loop with the dorsomedial striatum, receiving

multimodal information from the associative regions of the posterior cortex.

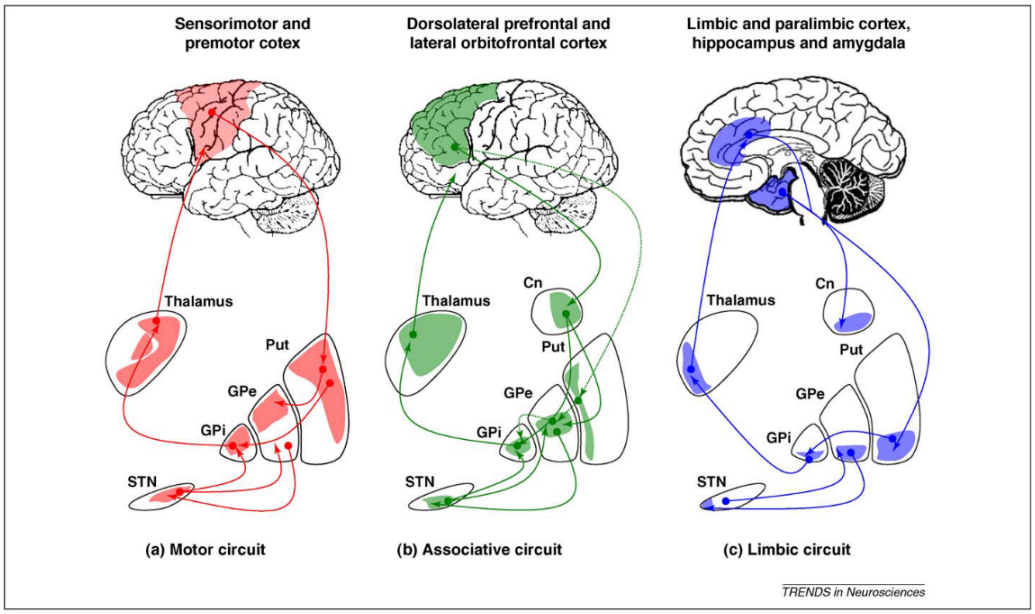


Figure 3.1: Schematic diagram illustrating the main cortico-basal ganglia–thalamocortical circuits within human brain. Red : *motor* loops, Green : *associative* loops, Blue : *limbic* loops (as illustrated in Krack et al. 2010).

3.2.1 Basal Ganglia

In a scenario of competing action space, when an animal has to select the most appropriate action among conflicting ones, the Basal Ganglia (BG) are believed to play the role of a central switch that selects the action, for instance to resolve conflicts over access to limited motor resources (Redgrave et al. 1999). BG are a group of nuclei comprising striatum, globus pallidus (internal and external), that are extensively connected to cerebral cortex and the thalamus. In vertebrates, BG are interconnected with the frontal cortex and thalamus. Across species, many aspects of motor function like movements, learning and habituation of actions are believed to be modulated by the processes in BG (Chakravarthy et al. 2010; Gaffan 1996; Graybiel 1995; Marsden and Obeso 1994; Robbins and Brown 1990). Numerous computational accounts explained the possible processes within BG and their interaction with cortex and thalamus that drive action selection (Doya 1999; Gurney

et al. 2001a; Guthrie et al. 2013; Leblois 2006). The role of BG and the thalamo-cortical network they form, has initially been highlighted in the selection of motor programs (Marsden 1982; Mink 1996). And the selection by competition in the CBG network to break the symmetry between similar actions has been elaborated in several computational accounts, for instance using a bias input to the competition mechanism (Leblois 2006). Furthermore it was shown that intrinsic noise is sufficient to generate a motor response in choice situations without external bias (Guthrie et al. 2013).

The cortical populations to drive the competition process and resolve for a selective choice, rely on the process of mutual lateral inhibition (Wickens et al. 2007). It has been postulated that the competition mechanisms for action selection that occur through in the BG-thalamo-cortical loop circuits also complement the cortical mechanisms (Boraud et al. 2018). The mutual lateral inhibition in cortical populations is attributed to the dense network of inhibitory neuron population. However, this kind of lateral inhibition gets weaker towards distant cortical populations. Alternatively, the projections of STN to GPi are divergent and the influence of cortical populations on GPi is convergent. Thus, together through thalamus, the BG network allows a much farther spread of the inhibitory mechanisms facilitating competition resolution in farther cortical areas.

Contrary to the idea that reward-based choice is driven by distinct component processes that are sequential and functionally localized, it has been argued that choices emerge from repeated computations that are distributed across many brain regions (Hunt and Hayden 2017). As a part of this thesis, an algorithmic model of these parallel CBG loops will be implemented to be able to study the dynamics between the loops (chapter 6). It will be demonstrated how mutual modulation within the feedback-loops and implicit hierarchical organization of timescales for information processing across the loops can drive behavior. The computational models that have already accounted for several processes within the BG pathways will be combined in the framework to study more detailed dynamics of decision and behavior within the limbic loops of the framework. Furthermore, focusing on the role of OFC in these CBG networks, several neuroanatomic and behavioral evidences with respect to organization within OFC will be considered to

understand its possible mechanisms of driving behavior.

However, before entering the modeling part of the work, a short review of the anatomical and functional organization within the PFC, especially medial PFC, followed by that of ACC will be presented in this chapter. Subsequently in the next chapter, an in-depth review of OFC and its subregions and their dissociable contributions to cvalue-based decision making is presented.

3.3 Functional organization within the PFC

The prefrontal cortex, at a very high level, can be viewed as lateral PFC (lPFC), medial PFC (mPFC) and the orbital PFC (usually referred as OFC). Here OFC is included under mPFC.

3.3.1 Lateral PFC

The lateral prefrontal cortex, involving the associative regions of the posterior and prefrontal cortex, is mainly engaged in cognitive control (Koechlin et al. 2003), related to the ability of the prefrontal cortex to manipulate abstract rules when the selection criteria is required to be more elaborated. When the selection is not trivial and requires memory, context, and abstract rules combining them, lPFC complements additionally the other PFC systems and the downstream selection mechanisms (Badre 2008). Quite remarkably with directional connections, lPFC stands as a site of association between the more limbic medial PFC regions, secondary sensory cortices (visual, auditory), premotor and supplementary motor cortices and parietal cortex.

Dorsolateral prefrontal cortex (dlPFC) has been implied in flexible, complex, and future-oriented human behaviors and those of other mammals. DlPFC has been argued to maintain flexible encoding of rules by showing that individual neurons in the region fired categorically different responses depending on the specific rule used (Wallis et al. 2001). It was also shown that dlPFC plays an important role in short-term memory (Funahashi

et al. 1989) and that it is involved principally with spatial locations (GOLDMAN-RAKIC 1995), possibly allowing dlPFC to elaborate more complex rules as a sequential arrangement of actions.

Ventrolateral PFC (vlPFC) is viewed to be more involved with the visualization with more precision (GOLDMAN-RAKIC 1995; O'Reilly et al. 2010), possibly owing to its connections with the sensory areas in the inferotemporal cortex and the auditory superior temporal gyrus. Although vlPFC has been regarded to play a role in learning stimulus-outcome associations as does OFC, an interesting dissociation has been pointed out that OFC is necessary for updating associations that signal desirability (palatability), whereas the VLFC is necessary for updating associations that signal availability (probability) (see review Murray and Rudebeck 2018). The vlPFC also has outputs to DLPFC, and it is possible that vlPFC governs the processes in the DLPFC, transforming the information from stimulus to behavior. In addition, a well explored role of VLPFC is in behavioral inhibition (Anderson and Weaver 2009; Depue et al. 2007).

Besides suggested differences between dlPFC and vlPFC in terms of the content they represent, it was also shown that they could be mediating distinct processes as monitoring actions and active maintenance of information in working memory respectively (Petrides 1996).

3.3.2 Medial PFC

As mentioned earlier, what is referred as medial PFC here includes the ventral part - ventromedial PFC (vmPFC) along with the orbital PFC (OFC), and the dorsal part - dorsomedial PFC (dmPFC) along with the anterior cingulate cortex (ACC). Quite generally put, information about sensory stimuli is conveyed to the OFC, where representations of the values of various options may be represented. Value signals, and much other information, then flow rostrally to the other parts of medial PFC, where information that influences decision making like action values comes into play (Elliott 2000; Knutson et al. 2005; Schultz et al. 1997). In other words, OFC exerts *emotional control* and ACC ex-

erts *motivational control* over behavior. The resulting signals then flow dorsally to other lateral PFC regions that use this information to plan possible responses and propagate to the premotor and parietal cortices to give rise to behavior finally through the parietal and the motor cortical regions. OFC is believed to play a wide variety of roles in order to exert *emotional control* over behavior. Since OFC forms a central part of this thesis, a critical review is presented in the next chapter (ch.4). However, Anterior Cingulate Cortex (ACC) is found to be in important complementary roles with OFC and remains an important region to understand in order to gain a comprehensive understanding of the role of OFC. Therefore, a short description of ACC follows here, before moving on the review on OFC.

3.3.3 Anterior Cingulate Cortex (ACC)

Anterior Cingulate Cortex (ACC) is one of the regions that is frequently implied in value computation, maintenance and comparison, alongside the ensemble of OFC, vmPFC and ventral striatum. ACC has well positioned to serve this role since it has multiple inputs conveying information from a variety of systems, including perception, emotion, attention, and memory. In the same context of value-based decision making that OFC will be mostly described in the following chapter, ACC has been implied in some interesting roles. In fact, in several studies, direct dissociation of roles of OFC and ACC have been pointed out (Rudebeck et al. 2008). The differences in OFC and ACC function could be the result of differences in their relative connectivity with sensory and motor systems. ACC has fewer connections to highly processed sensory information, particularly visual information in contrast to the high accessibility of such information to OFC (Carmichael and Price 1996; Kondo et al. 2005; Van Hoesen et al. 1993). On the other hand, ACC plays a crucial role in action selection via its direct connections to premotor and motor cortex (Croxson et al. 2005; Dum and Strick 1993; Wang et al. 2004). A few important roles of ACC will be described below.

Response selection

In reward guided action selection tasks, while the activity in lateral OFC was shown to be reflecting the association of choice and reward type information, ACC was found to be using this reward type information for guiding action (Noonan et al. 2011). There have been several implications that ACC is involved in response selection, especially when higher effort is required for greater reward. ACC lesions in rats resulted in lesser willingness to exert more effort to gain the high reward (Rudebeck et al. 2006; Walton et al. 2003). Activity in ACC neurons reflected the cost as well as the benefit of a course of action (Kennerley and Wallis 2009).

Foraging

Making a decision between the currently available options or foraging for different options is itself a decision to make. Such a behavior requires to maintain the value of the options that might be present during foraging and the possible cost of foraging itself if any, besides the expected values of the present options. In contrast to vmPFC that encodes specific well-defined options in a choice, ACC is found to be playing a considerable role in the decision processes of foraging. In a decision making task in humans where the subjects have to either engage in the current choice of known value or search among a set of potential alternatives also of known value, ACC was found to encode the average search value, i.e, the value of foraging environment and also the cost of foraging (Kolling et al. 2012). This can be possibly attributed to the connections to ACC from memory systems as well as ventral striatum, which allows the values to be represented irrespective of the frame of foraging or current choice.

Reinforcement Learning

ACC has shared anatomical connections with brain structures critical for reward processing and reinforcement learning such as the amygdala and ventral striatum (Morecraft et al. 2007; Porrino et al. 1981). ACC has been shown to have as much role as OFC, if not

more, in reward-guided learning owing to reward prediction errors (Kennerley et al. 2011) and surprise (Alexander and Brown 2011; Bryden et al. 2011). This notion is supported by several findings that ACC is involved in evaluating the course of actions on the basis of current feedback. ACC and the adjacent dmPFC were found to be involved choosing actions based on feedback (Fellows 2011) and switching behaviours after surprising outcomes (Alexander and Brown 2011; Kolling et al. 2012; Quilodran et al. 2008). One of the key ideas that could be the reason behind these observations is a difference in the way reward prediction error (RPE), which usually drives the reward-guided learning, is processed.

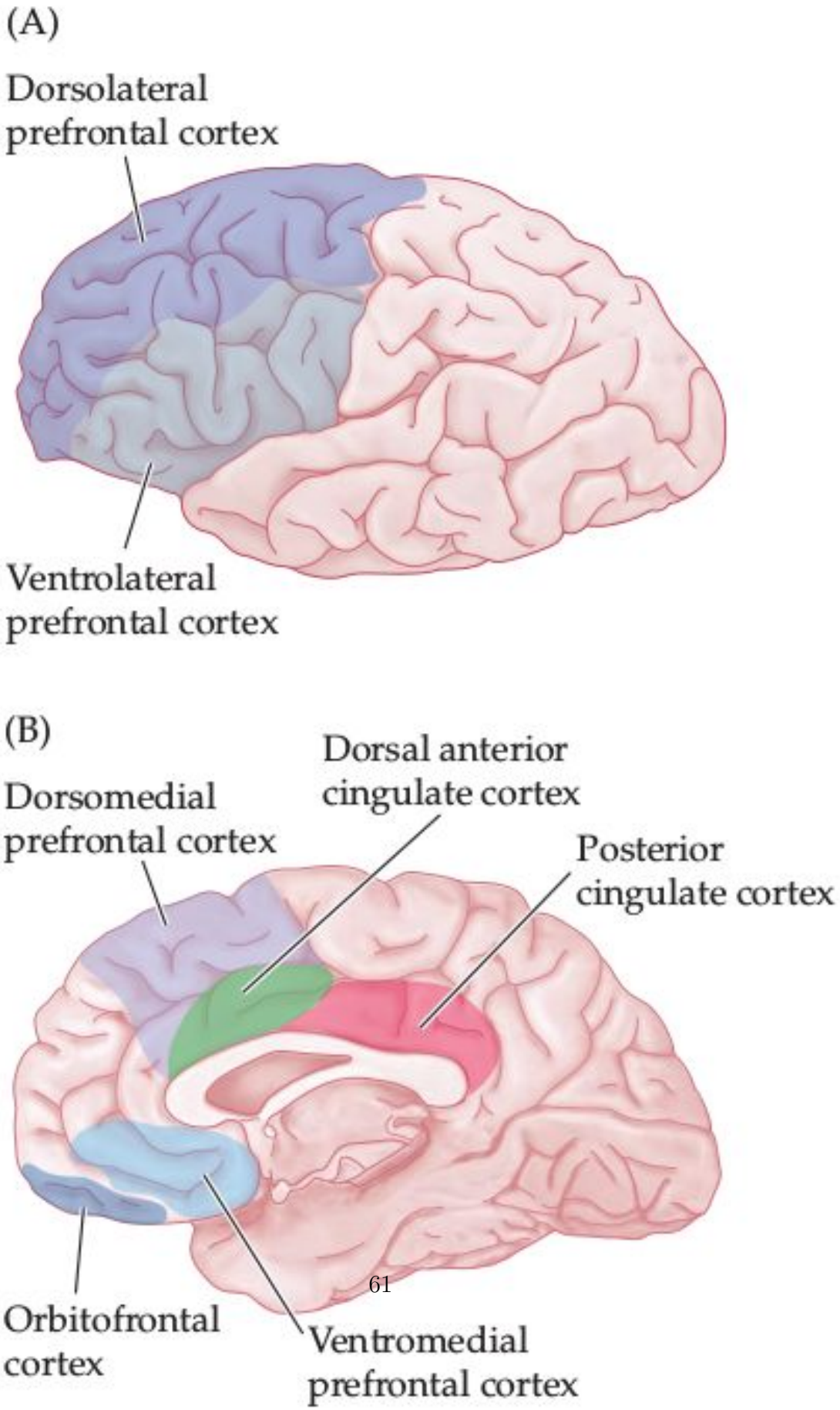
Although PFC in general is a major recipient of dopaminergic input which signifies the prediction error (Williams and Goldman-Rakic 1993), it was found that frontal neurons encode a signal similar to a RPE during the tasks that leads to rapid adaptation of behaviour in response to changes in reward contingencies (Matsumoto and Hikosaka 2007; Seo and Lee 2007). This could give rise to a possibility that ACC weakens the role of RPE in learning so that no outcome trials affect less, depending on the task contingencies (Kennerley and Wallis 2009). Several studies have pointed out how ACC may modulate how much influence individual outcomes and prediction errors should be given in guiding the adaptive responses (Behrens et al. 2007; Kolling et al. 2012).

Thus, although it might appear that reward representations seem to be present in both OFC and ACC (Grabenhorst and Rolls 2011; Rolls 2009; Rolls and Grabenhorst 2008; Rushworth et al. 2007), it could be the case that value of rewards is projected from the OFC to ACC. ACC would then be in a suitable position to combine the information about specific rewards with information about actions and action costs, thus associating actions with the value of their outcomes leading to the selection of correct action that will lead to a desired reward (Rushworth et al. 2007; Walton et al. 2003). In different studies, it has been found that ACC encodes the action-outcome learning and a related signal to encoding the recent history of rewards with respect to the actions taken (Luk and Wallis 2009; Seo and Lee 2007). In addition, there was no evidence that ACC encoded choice mechanism between rewards.

3.4 Dopamine : neural correlates of RPE

Ventral tegmental area (VTA) of the midbrain and the substantia nigra pars compacta (SNPc) contain the most completely studied dopamine neurons. These neurons project to various regions of the brain, but especially to the PFC and ventral striatum, where they are found to regulate neural activity (Figure 3.3). If two neurons fire in sequence, and if dopamine is also present, their connection may be strengthened. If dopamine is released when the outcome is better than expected, it would strengthen connections that are active right before that release occurred. Thus, dopamine may strengthen connections when the environment is better than expected and learning is favored. Effectively, a systematic change in the response of dopamine neurons briefly to reward signals RPE. When the reward is unexpected, baseline firing of the neurons is increased. When the reward is paired with a cue (like in the case of Pavlovian learning), the cue elicits a response following learning, while the neural response is no longer affected by the reward itself. Finally, when the cue is followed unexpectedly by a failure to provide a reward, the dopamine neurons briefly pause their firing, thus carrying a negative RPE signal (Schultz et al. 1997). In this work, these principles are used to model RPE that facilitates learning all across the model.

As mentioned earlier, OFC stands as a crucial nexus both from within the prefrontal sub-regions point of view as well as the CBG loops point of view. In addition, OFC forms a crucial part of the Pavlovian system with amygdala as well. Taken together with the amount of sensory projections OFC receives, it pushes for a closer look at the anatomical and functional positioning of OFC. Before entering the next chapter which is an extensive review of numerous implications on OFC, it should be noted that, ventromedial PFC (vmPFC) is adjacent to medial region of OFC and is often referred together with medial OFC. Hence, to keep the discussion tractable, the term medial OFC (mOFC) refers to both medial subregions of OFC as well as vmPFC. The detailed anatomy of these subregions will be sufficiently highlighted in the review.


(A)
Dorsolateral prefrontal cortex
Ventrolateral prefrontal cortex
(B)
Dorsomedial prefrontal cortex
Dorsal anterior cingulate cortex
Posterior cingulate cortex
Orbitofrontal cortex
Ventromedial prefrontal cortex

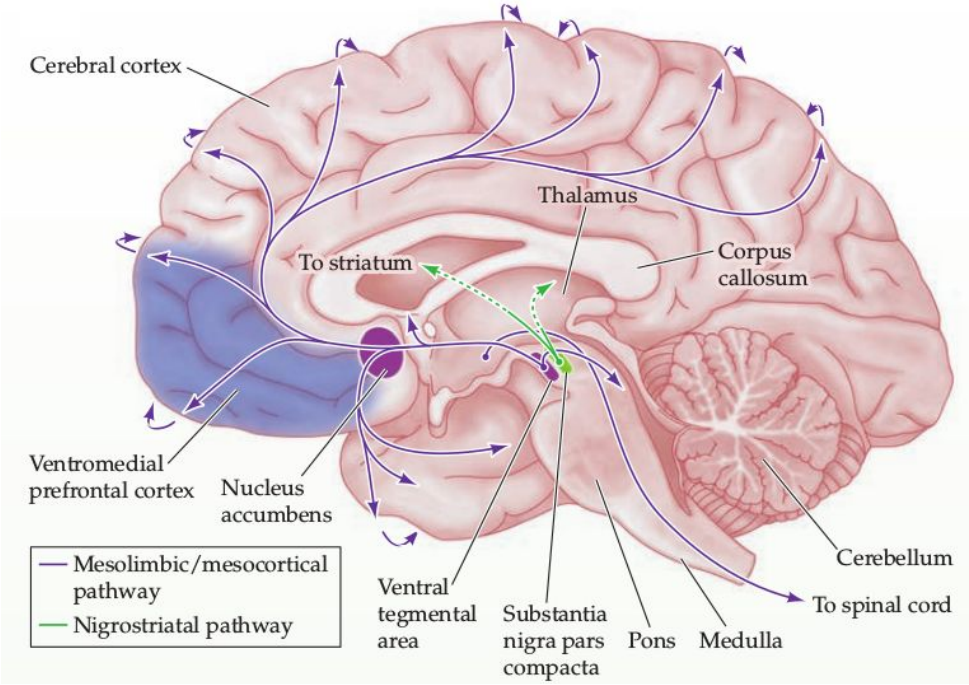


Figure 3.3: Dopaminergic pathways to different regions in the brain. Image from : Purves .D et al, 2018

Chapter 4

The Orbito Frontal Cortex : Lateral (lOFC) and Medial (mOFC)

Orbitofrontal Cortex (OFC) is one of the most heterogeneous regions in the Prefrontal Cortex (PFC), in terms of cytoarchitecture as well as functional anatomy. It is attempted to have a close look at the anatomy of OFC in the beginning of this chapter, however it is extremely difficult to mark clear boundaries owing it to the variations and complications in the methods, techniques and inferences of the reported studies. (i) The neuroscientific studies that explicitly claimed their regions of interest to be either lateral or medial parts OFC are discussed respectively. (ii) Studies that either observed both lateral and medial, or that lesioned them both, or those that generally referred OFC without specifying the anatomy - have been discussed under the general section before the discussion of lateral and medial dissociation begins. (iii) Arguably, at least so far as the humans studies are considered, ventromedial PFC is referred alongside medial OFC. (iv) Although the homologous part of OFC in rats has often been debatable, the studies are still mentioned as long as they discuss some common faculty that can extend to monkeys and humans. Nevertheless, such mentions should be taken with caution.

Sommaire

4.1 Introduction

Value-based decision making and resulting animal behavior, encompass a plethora of constructs like external stimuli and their features, their valuation possibly with respect to anticipated outcomes, value of those outcomes and the value of actions that would result in the outcomes; furthermore generally identifying patterns in this process and learning from it for similar future situations, to mention the least. On the other hand, the prefrontal cortex in primates (and equivalent regions in rodents to some extent) has been known to be responsible for some high level animal behaviors often contributing to the aforementioned faculties. Specifically, Orbitofrontal cortex (OFC) is one of the prominent regions of the prefrontal cortex, identified in primates as well as rodents (at least corresponding to some homologous parts), that is believed to play a crucial role in animal behavior. The very aspects of the OFC - its heterogeneous cytoarchitecutre, vast connectivity and vast variety of seemingly crucial functions it has been implicated in - joined together with the very conceptual challenges of understanding decision making as

a high-level process, pushes for an extremely complicated study that is the organization of information about the environment in the OFC and its role in value-based decision making and behavior. Therefore, we first introduce the anatomical background of OFC across species, thus setting a pretext of different functional roles it is believed to be performing. Further we specify the key features of the OFC regarding its heterogeneity in terms of anatomy as well as functional roles and discuss the challenges involved in studying such a complex yet pivotal region in the prefrontal cortex.

Fundamental decision making and learning are observed to be possible even in the absence of the OFC, thanks to some sub-cortical structures like Amygdala and the basal ganglia (BG). But what is intriguing in studying OFC in the context of decision making and learning is, as the tasks become more and more abstract, and the environment being only partially observable, the behavior becomes more and more impaired in the absence of OFC. However, it is established that decision making even as complex as in humans, builds on some fundamental principles conserved across species, being likely appetitive for primary rewards, and aversive for simple punishments, thus encouraging the whole range of methods in behavioral and systems neuroscience to be employed to develop that complete understanding.

4.2 OFC Anatomy in general

As the name suggests, the orbitofrontal cortex (OFC) in primates is a large area in the frontal lobe directly above the orbit of the eyes, topologically comprising the ventral surface of the prefrontal cortex and including parts of the medial wall between the hemispheres. There is a considerable amount of variance in the specific anatomical description of the Orbitofrontal Cortex (OFC), more so when its homologous regions in non-human primates and rodents are discussed. It is not so surprising, given a vast variety of roles that OFC has been proposed to play in crucial faculties of decision making and behavior.

So, as we move towards a closer look at its organization (structural and functional, within itself and across other prefrontal regions), we describe first, a broader picture

of what has been predominantly considered the Orbitofrontal Cortex (OFC) in humans (and its homologous regions in non-human primates and rodents). Quite generally, OFC in primates can be topographically referred to as the ventral part of the frontal lobe. However, as we'd note later in this section, this definition is not consistent across studies, where some of them include parts of medial prefrontal cortex in what they consider OFC. Importantly, solely topographical description doesn't account well for the understanding of the heterogeneity that is observed in the connectivity patterns and functional roles observed in different parts of OFC. Besides, it is the connectivity pathways with different cortical regions and sub-cortical structures that helped the formation of more recent maps of OFC in non-human primates and thereby the homologous regions in humans and rodents. Especially, connections with the mediodorsal thalamic nucleus (MD) provided a good basis for finding the similarities across species (Carmichael and Price 1996), also for the prefrontal cortex in general (see 3.1). Furthermore, cytoarchitecturally, OFC in primates comprises both granular (similar to sensory cortex) and agranular areas (similar to motor cortex).

In one of the first maps of the cerebral cortex labelled by Brodmann (Brodmann 2007), though the orbitofrontal cortex was not as detailed as the medial prefrontal cortex was, most of it was included under area 11. It is in the later maps by Walker (Walker 1940), including the granular area 10 delineated by Brodmann, areas 11, 12, 13 and 14 were marked as occupying the rostral, lateral, central and medial orbital surface respectively (Price 2007). While Brodmann's maps provide reference to the most recent studies of medial prefrontal cortex, it is the Walker's areas that provide reference to the orbitofrontal cortex in monkeys. It is important to note that most of these divisions of orbitofrontal cortex go alongside the descriptions of medial prefrontal cortex, and while it is a reasonable generalization to consider Walker's areas 10, 11, 12, 13 and 14 make up most of the OFC in monkeys, a much finer architectonic descriptions by Carmichael and Price (Carmichael and Price 1994) is recommended for a larger perspective. Ongur et al (Öngür et al. 2003) reported that all of the areas in the macaque orbitofrontal cortex have counterparts in humans, with minor distinctions. However, notably, the area 12 in

the lateral part of orbital cortex in Walker's maps of macaques seemed to correspond to area 47 in Brodmann's human maps. Petrides and Pandya (Petrides and Pandya 2002)) later named the area 47/12 to emphasize the correlation between the area in humans and in macaques. Accordingly, Ongur et al (Öngür et al. 2003) gives a broader description of OFC (and other cortical regions) with a direct comparison to that of Carmichael and Price (Carmichael and Price 1994) in monkeys.

As we'd note in the later sections, several studies to understand different roles of orbitofrontal cortex were performed on rats as well, implying certain homologous areas to primate orbitofrontal cortex. As mentioned earlier, despite several debates, an equivalent area to PFC has been implied in rats (see Sec 3.1). Especially, Krettek and Price (Krettek and Price 1977) showed that all but the rostral parts of primate OMPFC (area 11) can be identified in rats, particularly, medial and lateral orbital areas (MO and LO) which may be comparable to areas 14 and parts of area 13, respectively, in monkeys.

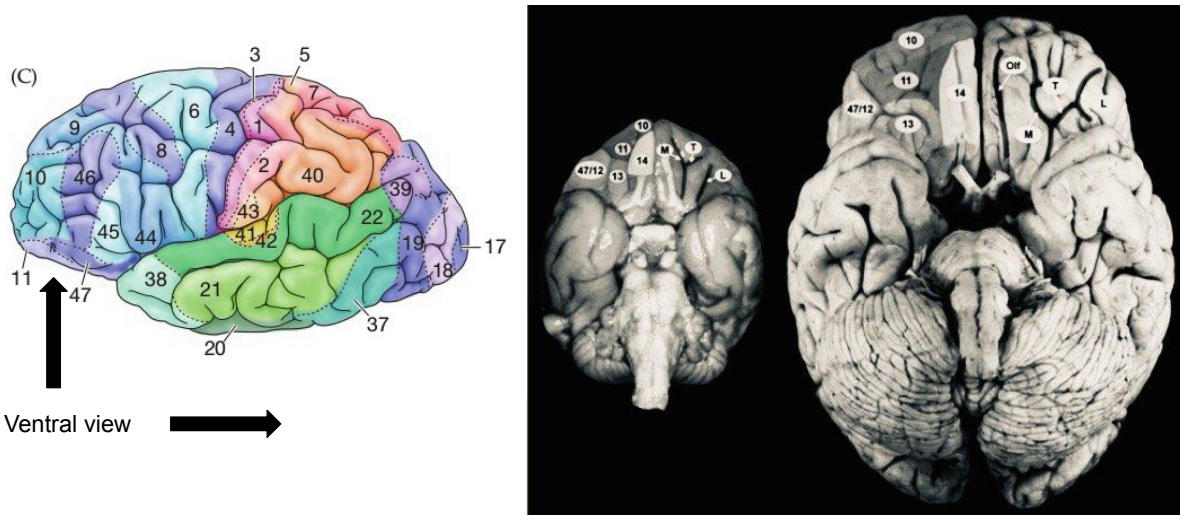


Figure 4.1: Brodmann Areas, right: rhesus monkey brain on the left, human brain on the right Image adapted from : Purves .D et al, 2018

Orbitofrontal cortex stands as a remarkably densely connected prefrontal cortical area with connections (sometimes bidirectional) to all sensory domains, learning and memory structures like striatum, amygdala and hippocampus as well as several frontal regions. Most of the studies on the functional roles of the orbitofrontal cortex and its subdivisions

in humans and nonhuman primates base their definition of OFC on some or all of the Walker's areas 10, 11, 47/12, 13 and 14 (Glasser et al. 2016; Price 2007).

In a nutshell, all the discussion about varying roles of the Orbitofrontal Cortex in behaviour would revolve around some classic aspects like value, reward and decision making in general. There are some good reviews that give a good structure to these concepts, which will be used extensively in the review to follow.

- (Rangel et al. 2008) - A framework to study neuro-biology of decision-making
- (Dolan and Dayan 2013) - Goal-directed vs Habitual behaviour
- (Schultz 2015) - Reward

4.3 OFC Function in general

The Orbitofrontal Cortex (OFC) is a classic example against pinning down a brain region to a unique and specific functional role. There is a wide repertoire of faculties that OFC is believed to be responsible for, thanks to its heterogeneous anatomy and wider connectivity with several other specialized brain structures. From facial processing to subjective valuation of stimuli (not just visual, but also gustatory, odours) to driving flexible and adaptive behavior emerging from a value-guided choice : the range of implications is huge. As extensive the range of underlying sub-processes in decision-making and learning is, so are the roles that are attributed to the OFC. In humans, non-human primates and rodents, the neural activity in the OFC is generally believed to play central role in the cognitive aspects of decision-making. In particular, it is widely implied in affective decision-making, representing *emotion* and in economic decision-making, representing *re-*

ward. Most of such implications collectively highlight that the OFC is pivotal, in one way or the other, in almost all of the fundamental processes involved in value-based decision making : Representation, Valuation, Action selection, Outcome evaluation and Learning (Rangel et al. 2008).

One of the earliest common findings about the deficits following OFC damage is the ability to switch the behavior when, in a choice situation, the learned associations between cues and outcomes are reversed. Generally coined as *(Object / Visual Discrimination) Reversal Learning* , it is different from spatial reversal learning, where dorsolateral prefrontal cortex is implicated (Goldman-Rakic 2011). Object discrimination reversals in choice situations have been studied extensively employing a wide variety of procedures in humans (Bechara et al. 1997; Fellows 2003; Hornak et al. 2004; Rolls et al. 1994; Tsuchida et al. 2010), non-human primates (Izquierdo et al. 2004; Jones and Mishkin 1972; Meunier et al. 1997) and rodents (Chudasama and Robbins 2003; Kim and Ragozzino 2005; Riceberg and Shapiro 2012). Often this deficit has been associated to the hypothesis that OFC is responsible for inhibiting responses after the reversal begins (Izquierdo and Jentsch 2012; Kringelbach and Rolls 2004). Alternatively, it was proposed that the OFC maintains flexible stimulus-outcome associations and hence the deficit in reversal learning due to OFC damage (Fellows 2003; Hornak et al. 2004). Further, it was also proposed that OFC is sensitive to shift behavior upon a negative reinforcement (after the reversal of already learned contingency)(Roberts 2006). However, if the deficit in reversal learning in the case of OFC damage is because OFC would otherwise inhibit (well learned) response, it should be expected that the response inhibition would take place even in the absence of reversal, at the beginning of the task, such as inhibiting some kind of default response. To the contrary, (Schoenbaum et al. 2002, 2003) showed that OFC damaged subjects are almost unimpaired in the initial acquisition phase. Similarly (Walton et al. 2010) showed that the OFC damaged monkeys, after the reversal, in fact switched their responses more. Furthermore, it would be redundant for OFC to specifically maintain flexible associations while other regions are found to be doing the same more effectively (Cohen et al. 2012; Morrison et al. 2011; Paton et al. 2006). (Schoenbaum et al. 1998, 1999) also show in

single-unit studies that the basolateral amygdala maintains the stimulus-outcome association with greater reliability than OFC, in the earlier phase of learning.

(Walton et al. 2010) also showed that the choice behavior of the animals with OFC-lesion was largely unimpaired before the reversal, consistent with other studies that have argued that the OFC is only important following reversals in outcome associations (Clarke et al. 2008; Schoenbaum et al. 2002). However, (Walton et al. 2010) argues that (i) OFC primarily plays a role in contingent learning, that is the reward received should be correctly credited to only the chosen option, but not to other alternatives in the current trial (ii) and in the absence of OFC, the animal might rely on a different learning mechanism, approximating the stimulus-outcome association based on the history of reward (to the stimulus that most recently rewarded or most frequently rewarded) or to the stimulus of the following trial after the received outcome.

Such is the complexity of understanding how exactly OFC plays a role, even when a "loss-of-function" study shows a possible deficit in any mental function after OFC damage. Nevertheless, we briefly discuss some common findings and their arguments about how OFC affects different sub-processes involved in decision making (Rangel et al. 2008).

Representation : For an animal to make decisions in an environment, one of the crucial aspects is the way (the sensory aspects of) the environment is represented. The subjective representation of the environment, with respect to the animal's exteroception and internal state, is what forms the basis for appropriate estimation of risk and outcomes, eventually supporting the frameworks of learning from the outcomes. When encountered again with such representations, the animal could perform similar actions that earned rewarding outcomes or to avoid aversive outcomes. One of the key aspects attributed to OFC in the context of decision making, is that the OFC abstracts a more comprehensive representation of the current environment of choice or task scenario, relating it to possible outcomes. It is the extensive connectivity of OFC with multiple sensory areas, subcortical structures such as the amygdala and lateral hypothalamus, that makes it a suitable candidate for such comprehensive representation. In fact, an OFC damaged animal may still be able to decide and learn, as long as the task at hand is not abstract or with com-

plex contingencies. In an experiment to study the effects of the lesions of different OFC subregions in monkeys that were satiated on a particular food, given a choice between two food items (not their reinforcers, like boxes), there was virtually no difference in the behavior of control monkeys and OFC lesioned (lateral or medial) monkeys (Rudebeck and Murray 2011).

However most tasks are complex , where the *current state* could be only partially observable, for example, that which requires information from working memory. In such cases, the *current state* depends not only on the perceptually detected (possibly high-dimensional) features at the moment but also on the history of previous choices and outcomes, and their possible link to the current state. Increasing accounts propose that the role of OFC in decision-making is to provide this sophisticated representation that includes partial, not immediately observable information of the environment (Bradfield et al. 2015; Fellows 2011; Schuck et al. 2018; Wilson et al. 2014). Animal behavior, however basic or complex, depends on the evaluation of outcomes and associating them to the choices and the sensory states that represented the choices. Reinforcement Learning (RL) is long known to be one of the likely mechanisms through which animals evaluate outcomes and adapt their behavior (Joel et al. 2002; Schultz et al. 1997; Sutton and Barto 1998). RL systems generally describe a task structure as a link between *states* (often, through *actions*), and therefore *state representation* forms a crucial part of such systems. As it has been termed, what OFC provides is the *cognitive map of task space* (Wilson et al. 2014) or the State space representation (Schuck et al. 2018) to the other RL systems in brain, which explains the classical deficits observed in OFC damage like *reversal learning*. More importantly, this kind of representation may be important not only across the duration of a task, but also in the context of new tasks.

This theory has been studied recently in both humans and rodents using various paradigms. Through a multivariate analysis in imaging studies on humans by (Chan et al. 2016) and (Schuck et al. 2016), in probabilistic environments where participants were required to make inferences about the hidden states, it was shown that a full posterior distribution over hidden states, that was separate from value, was correlated to the activity

in OFC. In an odour-guided choice task (Takahashi et al. 2011), dopamine neurons were recorded in OFC-lesioned rats and it was shown that OFC is crucial for representing the states, influencing the prediction error signals in VTA dopamine neurons and driving the value learning elsewhere, for instance through its projections to ventral striatum (Voorn et al. 2004). In a similar task from (Stalnaker et al. 2010), it was shown that when OFC is lesioned, the information representing the unobservable state of the task was degraded in the cholinergic interneurons in the dorsomedial striatum (Stalnaker et al. 2016). Notably, (Nogueira et al. 2017) demonstrated that the prior information is stored in OFC, only if it is behaviorally advantageous, but not when passively exposed without rewards.

Valuation : Psychological and economic accounts of human and animal decision making have placed a great emphasis in the concept of value (Cabanac 1992; Kahneman and Tversky 1984; Sugden et al. 1996). As described in section 2.1, value, although likely to be associated to behavior in several ways, holds to be a wide construct, that spans across multiple neural pathways (Fellows 2011). In economic choice situations, neurons in OFC seem to be potentially assigning underlying values to facilitate the choice. In one of the detailed single-neuron activity studies on the OFC and proximal vmPFC in monkeys performing a task related to value perception, (Bouret and Richmond 2010) showed that the activity of neurons in both the regions strongly correlated with reward size at cue onset, with action at the feedback. The monkeys also performed self-initiated trials and the activity of neurons also showed strong effect of reward size. Specifically, several studies (see review (Padoa-Schioppa and Conen 2017)) imply the role of OFC in encoding the *value* of the offered and chosen goods (Padoa-Schioppa and Assad 2006; Padoa-Schioppa et al. 2006). It might appear that there are other valuation systems in the brain (Kawagoe et al. 1998; Platt and Glimcher 1999; Shidara and Richmond 2002), but what makes OFC unique in valuation is that it encodes the value irrespective of the visuo-spatial and motor aspects.

Furthermore, (Padoa-Schioppa and Assad 2008) highlights a crucial relation between *transitivity* in choice behaviors and *menu invariance* in the neuronal encodings. *Transitivity* suggests that the choice preference remains consistent across different presentations,

i.e, if option A is chosen over option B, and option B over option C, then upon presentation of options A and C, A should be chosen. *Menu invariance* is typically (though not always) observed across multiple value encoding neuron types in OFC, where the activity of a neuron firing for a particular option is not affected by the value of other options in the choice. (Padoa-Schioppa and Assad 2008) describes the neuronal activities in OFC and their nature of *Menu Invariance* reflect the transitivity in behavior. Wherein the preferences depend on the menu but not invariant, transitivity could be violated at the behavioral level.

However, humans make choices that are largely varying in their subjective values. One could make a choice, on the same day, between €10 and €14 mobile subscription and later between a €1000 and €1400 worth laptop. (Padoa-Schioppa 2009), and later (Conen and Padoa-Schioppa 2019), showed that individual value encoding neurons in OFC partially adapt their activities to appropriately represent the available value range of the options. Such an adaptation, the authors discussed, features an increased minimum neural firing in response to certain value ranges, which if modulated, ensures choice variability in the animal's behavior (Conen and Padoa-Schioppa 2019). Then follows the question, when presented in a choice scenario, whether such wide range of values are represented in terms of their absolute value or relative to each other, for the comparison mechanisms. Imaging studies in humans, rating their subjective pleasantness of odors, highlighted that there are simultaneous yet distinct representations of absolute value and relative value in different subregions of OFC (Grabenhorst and Rolls 2009). Furthermore, it was also shown that, should such a subjective 'value' (pleasantness) be compared to another from different sensory modalities (pleasantness of odor vs pleasantness because of warmth in a cold room), these otherwise incomparable values are represented in a some common neural scale in the ventral prefrontal cortex, predominantly the OFC, that might facilitate a comparison process on such a common scale for decision making(Grabenhorst et al. 2010), as suggested by standard economic theories. Furthermore, substantial emphasis was also put on whether the process of *common scaling*(Rolls et al. 2008) is qualitatively distinct from that of converting different rewards into a *common currency* (Cabanac 1992;

Montague and Berns 2002). It is argued that converting different type of rewards into a common currency, probably by the same neurons, would mean losing the reward identity in the process of conversion. To the contrary, it was proposed, common scaling of the activity of neurons representing different rewards would facilitate the comparison process retaining the identity of the rewards, imposing that the output of the decision making process maintains the specificity of the goal (Grabenhorst and Rolls 2011).

In contrast to this predominant view of linking value information to the stimuli, it should be noted that, in certain behavioral situations where the value information might correspond to both stimuli as well as actions, there might be parallel, dissociate processes at work through the frontal lobe. (Fellows 2011) reports distinct differences in the performance of patients with OFC lesions as opposed to that of patients with the dorsomedial frontal lobe (DMF) lesions on two different stimulus-value and action-value tasks.

Outcome-prediction OFC, across rodents and primates has been shown to be responding to action outcomes by encoding outcome predictions (Boorman et al. 2009; Gottfried et al. 2003; Jones et al. 2012; McDannald et al. 2011) . One way outcome prediction could be seen differently from outcome valuation, is due to the fact that predictions are made even in the absence of experience whereas valuation usually arises from experience of an outcome. Imagine that two stimuli are associated by presenting them together (without any outcome). Now if one of the stimuli (A) is associated with an outcome, by the internal structure that has been learned, it is possible that the other stimulus (B) (that has not been associated to the outcome separately), would predict the outcome that A has been associated to. OFC is believed to have a role in forming these implicit associations and thereby predict outcomes. By testing rats in a sensory preconditioning and blocking task, (Jones et al. 2012) found that OFC lesioned rats failed to predict the outcome when exposed a cue that was not particularly paired but associated with another cue that was paired. Control rats were able to perfectly predict the outcome even when only one cue was paired to the outcome and the unpaired cue was presented, as long as both the cues were associated (presented) together before, without any outcome.

Similar results were found in monkeys and humans as well implying that OFC might use outcome predictions to estimate action outcomes (Hampton et al. 2006; Noonan et al. 2010). Similar arguments were extended to purely instrumental situations as well, implying that OFC predictions outcomes specific to choices and actions (Rudebeck and Murray 2014). Such an ability to represent specific prospective outcomes and thereby being able to predict them, plays a critical role in being rapidly sensitive to devaluation of those outcomes (Fellows 2011; O'Doherty 2011; Valentin et al. 2007). Similar results were also shown in rats, using optogenetics techniques to selectively inactivate neurons in lateral OFC (Gardner et al. 2017). Such an inactivation of lateral OFC neurons dis not affect economic choice behaviour per se, but disrupted devaluation-sensitive behaviour.

In fact, deficits of not only OFC, but those of Amygdala are also known to impair the outcome expectation of an otherwise conditionally reinforced cue (Parkinson et al. 2001; Pears et al. 2003). Furthermore, damage to basolateral amygdala (BLA) also causes deficits in devaluation tasks, similar to those caused by OFC damage (Málková et al. 1997). This raises a question of considerably overlapping function of Amygdala and the OFC. One way the question can be dissociated is that the cue-outcome associations are learned in the BLA whereas the information can be utilized to guide behavior only through OFC. It was found that the BOLD signal in BLA correlated with the value of the items shown. However the activity of OFC was correlated with a value only when it was used to make a decision (Arana et al. 2003).

Action selection The cognitive theory of action selection has always highlighted a dichotomy at different levels. The one most commonly discussed is that of goal-directed vs habitual nature of action selection (or decision making). As will be seen in 3.2.1, numerous accounts have been made on how the sub-cortical nuclei Basal Ganglia (BG) facilitate the process of action selection through competition. What amplifies the sophistication of this process across the mammalian species is the top-down influence of prefrontal and other cerebral cortical influence on the processes in BG. OFC is one of the prefrontal regions that is found to be extensively connected to ventral striatum (a part of the BG),

alongside amygdala, and is believed to play a major role in the interplay of action selection, forming a closed feedback loop with the BG systems and the thalamus. Largely due to its implications in the subjective valuation of options and outcome prediction through inferred state representations (Jones et al. 2012), OFC interacts widely with the striatal structures to pass on such relevant information to the choice mechanisms, for instance in dorsomedial striatum (DMS), to affect the choice. The exact mechanisms by which OFC influences the down-stream structures are yet to be explored, i.e, whether by sequential or parallel process, for instance in conjunction with ACC (Fellows 2011). Nevertheless, it has been established through single neuron studies in rats that OFC modulates the cholinergic interneurons in DMS during selection, which still responded in the absence of OFC as long as the state information was not needed for the choice (Stalnaker et al. 2016).

In decision making situations with uncertainity, OFC is believed to drive the value-based exploitation behavior of choice as opposed to the exploratory influence from other cortical substrates (frontopolar cortex)(Daw et al. 2006). In fact, even in a perceptual selection task, where the choices are not value-based but preference based, like choosing between a red watch and a blue watch (and many such neutral stimuli), patients with OFC damage were unable to make choices that are consistent with their own preferences (Fellows 2011). This deficit extends to value-based choices where overall pattern of choices maximizes the reward payoff in the end (Samuelson 1938).

Outcome evaluation The role of OFC has been widely discussed in evaluating a positive outcome and reinforcing the stimulus or action or even the hidden state that led to the outcome. This will be visited in the next subsection under *Learning* (4.3.1). However, it is crucial to evaluate an outcome in the context of uncertainty and risk that was estimated before making the choice. In the intracranial recordings in with high temporal definition, (Li et al. 2016) observed different phases of making a choice starting from cue presentation to outcome evaluation, through reward anticipation phase. Both lateral OFC and vmPFC showed encoding of risk during the reward anticipation phase. Similar results have been

shown in non human primates, particularly implicating that OFC leads the risk estimation network and sends the information downstream to ventral striatum, hippocampus and mid-brain dopamine systems. (Fiorillo et al. 2003; O'Neill and Schultz 2010). While the evidence of risk encoding in human OFC during the outcome anticipation and its effect during the outcome evaluation is sparse, it would be important to use techniques that are temporally sensitive to highlight just causal effects between risk estimation and outcome evaluation.

OFC is also implied in representing *regret* for an unchosen option. Increasing BOLD activity in human OFC, beside other regions, was observed (Coricelli et al. 2005) not only representing regret when the outcome of unchosen gamgle was revealed, but also just before making next choice implying the anticipation of experienced *regret*. Similar evidence was later emphasized in rats in an interesting task paradigm in which rats encountered wait or skip choices for delayed delivery of different outcomes (Steiner and Redish 2014). The neural ensembles from the OFC (and ventral striatum) showed that they represented missed action, that might have produced a more valued outcome, implying the OFC is active during the expressions of *regret*. This implication provides a basis for studying the psychological and economic theories of human decisions being risk-aversive as well as regret-aversive.

4.3.1 Role in Learning

In humans as well as nonhuman primates, one of the consistent findings reported about OFC lesions is that it causes deficits in reversal learning, that when the identity of the high rewarding option is switched (Fellows and Farah 2007; Kringelbach and Rolls 2004; Walton et al. 2010). However, the reason for this impairment has been long associated either to OFC's role in adjusting subsequent behavior after a specifically negative feedback (Fellows and Farah 2007; Kringelbach and Rolls 2004) or to the fact that OFC lesions cause animals to make persevered choices. In contrast, (Walton et al. 2010), through all OFC lesion in monkeys performing a 3-armed bandit task with a locally variable rewarding schedules

for each bandit, argued that OFC-lesioned monkeys rapidly adapting locally fluctuating values of a high rewarding stimulus as good as the controls. It was proposed, that is not perseverance, but in contrast the increased switching behavior between alternatives that underlies the reversal deficits in OFC-lesioned monkeys. Furthermore, OFC deficits could cause the the inability to associate a to particular outcome to the specific choice made. Similar evidences were pronounced through studies in humans with damage to OFC that reversal deficits might be a result of neither persevered selection nor response inhibition and insensitivity to feedback neither. Rather, specific role of OFC in contingency learning could be assigning the credit of an outcome to an appropriate stimulus among several options (Fellows 2011). This kind of deficit further may result in not only assigning the outcome backward to recent or frequent history of choices, but also forward, to the choices made after an outcome is received (Walton et al. 2010). Another important aspect of learning from outcomes is that it is not just the current choice situation and the outcome that is relevant for learning the overall value of the choices. Maintaining relevant information about the state - stimuli, context and outcomes - across trials and across contexts, is key to flexible behavior and OFC has been implied to dynamically modulate activity representing this maintenance (Rich et al. 2018).

In essence, the common notion is that impaired OFC function in an animal does still demonstrate fairly normal behavior, to learn and perform basic tasks using using reinforcement learning (RL). Rather that OFC damages causes the inability to distinguish perceptually similar states, hence restricting the animal to behave only based on purely observable, stimulus-bound information. Computational of RL and several animal behavioral theories have long implied two schemes of learning - *model-based* and *model-free*. In the context of a choice, learning occurs either by actual experience or by predictions derived from the internal model of the environment, as described in *model-free* and *model-based* schemes, respectively. Although computational formalisms for both the theories are sound and there has been much emphasis that a combination of both is observed in human behavior. However, understanding how such a model-based system might be implemented in brain, the precise neural correlates as understood in the case of a more general, model-

free (for instance, in pavlovian or instrumental conditioning), is still an open question. There is a growing evidence on how OFC could be a central player in such a model-based learning, owing to its ability of task space and state space abstraction (Hampton et al. 2006; McDannald et al. 2011; Wilson et al. 2014) or its closer interaction with working memory systems. While the computational theories have already been depending on such implications for the evolution of more efficient RL methods (like episodic or meta-RL) (Wang et al. 2018), it is of high importance to investigate the neural underpinnings of such a model-based behavior in OFC, if there are any.

Details of studies on OFC †- probable, unclear from article					
Article	**OFC**		**Species**	**Technique**	**Remark**
	lOFC	mOFC			
Padoa-Schioppa2017	areas 13 m/l and 11l	vmPFC (area 14)	Monkeys	Single-cell	Review
Gottfried2003	area 11/13, lateral †	–	Humans	Imaging	predictive reward encoding
Kringelbach2003	lateral †	–	Humans	Imaging	Subjective Pleasant-ness
Ostlund2007a	lateral	–	Rats	Single-cell	outcome endocing in pavlo-vian, not instrumen-tal

Rushworth2007a	general OFC		Primates	–	Review, OFC vs ACC
Rudebeck2008b	areas 11/13	area 14†	Monkeys	Lesion	Comparison with ACC
Izquierdo2004	Walker's areas 11, 13, 14, and 10		Monkeys	Lesion	Reward contingencies
Glascher2009	–	medial OFC, vmPFC†	Humans	Imaging	Action values.
Hare2008	–	vmPFC†	Humans	Imaging	Goal values
Rudebeck2011b	Areas 11/13	Area 14	Monkeys	Lesion	–
Noonan2011a	lateralOFC	mOFC/ vmPFC	Humans	Imaging	Also ACC
Rushworth2011	Areas 11/13 (lateral)	mostly Area 14	Primates	–	Good dissociation - Review
Nogueira2017a	lateral	–	Rats	Single-cell	Perceptual, novel task

In the course of a closer look at the functions carried out by the OFC, it is worthwhile to consider its role in conjunction with two major sub-cortical counterparts - amygdala and ventral striatum. However, ventral striatum encompasses more complex subregions, their implications and closer interaction with vmPFC and ACC, at least in humans. Whereas, amygdala, at varying degrees of indulgence in rats to primates, seems to exert strong

influence on OFC's ability to guide flexible behavior. There is extensive evidence in rats on a more detailed view on the bidirectional influences between the basolateral amygdala (BLA) and OFC have on each other. While BLA is believed to be crucial in acquiring initial significant emotional information and transfering to OFC, OFC is known to persist this information independent of BLA to guide behavior (Pickens et al. 2003). Similar findings were also found in nonhuman primates (Baxter and Murray 2000; Paton et al. 2006) and humans (Bechara et al. 1999; Gottfried et al. 2003). Also in case of reversal deficits, it has been implied that in fact it is in fact BLA that is sensitive to negative feedback and hence the deficits. In addition, a difference in the latency in learning in BLA and OFC has been observed, in a single neuron study in monkeys. It was found that the aversive learning to negative feedback was faster in amygdala than OFC. In contrast, the appetitive learning happens faster in OFC than in amygdala (Morrison et al. 2011). Furthermore the role of ventral striatum, particularly the nucleus accumbens (NAcc) has been equally pronounced to complement that of OFC. In a study of five food flavours, (O'Doherty et al. 2006) demonstrated predictive responses in human ventral striatum that directly reflected subjective preferences for the flavours. This suggests a sensitivity to reward value within this region, alongside OFC. There is a growing literature on how these different OFC circuits, like those with BLA and NAcc (Schoenbaum et al. 2006), interact contributing to flexible decision making. However, more such studies are needed to be done in humans and primates to complement the evidences reported from studies in rats (Groman et al. 2019).

4.4 Dissociation of roles within the OFC :

Long story short : With a gross generalisation and to be taken with *pinch of salt* : the lateral OFC has been proposed to be important for learning stimulus-value associations driven by external factors, reward identities that are critical for motivating actions towards rewards whereas medial OFC / vmPFC may be concerned with internal motivation driven evaluation, value-guided decision-making and maintenance of choices over successive decisions. Courtesy : among many others, (Bouret and Richmond 2010; Noonan et al. 2010; Rudebeck and Murray 2011)

There is an evident underlying complexity about studying the role of Orbitofrontal Cortex in decision making, learning and goal-directed behavior. Simply skimming through the collection of articles that report findings on the OFC (or any of its distinct subregions), it can be quickly remarked that the roles that are attributed to OFC together form a superset of several high-level behavioral paradigms that are so intricate in terms of conception across scientific fields and interpretations of behavior across species; From a bird's-eye view, OFC is implied in : perceptual decision making and value-based decision making; within a single decision-making episode (trial), different kinds of involvement at a different phase (option presentation, action selection, outcome delivery etc.,); learning stimuli-outcome (pavlovian) and action-outcome (instrumental) associations.

How then, can one underpin a unified understanding of what the role of OFC in animal and human behavior is, and how it is achieved? To begin with, OFC, as marked by extensive anatomical studies (Öngür et al. 2003), Monkeys, Rats), is a largely heterogeneous region, unlike the rest of PFC which is homogeneously granular. The heterogenity is multi-fold : neurons are found to be differently encoding different aspects in a single task context, cyto-architecturally different areas (granular and agranular), remarkably

distinct connectivity pathways through different brain structures.

In one of the hallmark studies that showed how neurons in the OFC encode economic value (Padoa-Schioppa and Assad 2006), most of the neuronal recordings were focused on area 13 of OFC, part of commonly referred lateral OFC. Yet within one area, different groups of neurons are shown to encode different attributes of the decision process and furthermore, it is argued that these groups of neurons can sufficiently generate decisions. Further across different areas that constitute OFC, topologically different subregions have been implied to be playing functionally distinct roles. One such distinction is between the anterior(Choi et al. 2004) and posterior(Hebscher et al. 2016; Rudebeck et al. 2013) OFC. However, the topological distinction that is most extensively reported to imply strikingly different functional roles is the one between lateral and medial parts of OFC. Quite often in studies, neighboring region ventromedial prefrontal cortex is referred as a part of or in addition to medial OFC. However, it has been argued that the reward-related vmPFC/mOFC region in humans is homologous to mOFC (Walker's areas 10 and 14) in macaques (Cavada 2000b; Mackey and Petrides 2010). Therefore by saying in the remainder of this volume, both medial OFC and vmPFC are implied, especially w.r.t the studies in humans.

It has been observed that lOFC and mOFC have clear divergent connections to different networks as opposed to very few regions showing specific connectivity to anterior or posterior OFC (Zald et al. 2014). Both in monkeys and humans (Cavada 2000a; Kahnt et al. 2012) lateral OFC is reported to receive extensive projections from diverse sensory modalities through the insular cortex, and also heavy projections from amygdala. Where as medial OFC has strong projections from hippocampus, hypothalamus, ventral striatum, relatively less projections from amygdala, and strongly connected with the cingulate cortical areas (23a, 23b and 23v). Such a constellation of distinct connectivity patterns in OFC highlight several possibilities in which it modulates emotional and motivational behavior. Although numerous studies individually reported specific implications of either lateral OFC and medial OFC (or some of vmPFC), there are only handful of studies that tried to dissociate the complementary roles, if any, of lOFC and mOFC in a similar be-

havioral context. Moreover several accounts that derived their observations in the name of OFC, effectively were studied in only one of the two regions.

Experimental methodologies employed to study the dissociable roles or individual roles of lateral and medial OFC in learning and decision-making vary across species. It is generally single-neuron recording studies or lesion studies in macaques or rats (Murray and Izquierdo 2007) predominantly on the better accessible lateral OFC than the medial OFC, and BOLD signal correlation from fMRI studies in humans. Few behavioral studies on frontal damage patients are also discussed (Fellows 2003, 2011; Fellows and Farah 2005, 2007) and specifically one study in humans reported the findings on spatio-temporal neural dynamics of expected value, risk and experienced value signals using the LFPs from intracranial EEG recordings (Li et al. 2016). Consequently, it raises a question on few contradictory findings that used different techniques and methodologies. In the studies that reported the encoding of risk-related signals during the anticipation of rewards, few fMRI studies and electrophysiological recording studies in monkeys agree on the role of lOFC whereas there are other fMRI studies which implied mOFC for the same (O'Neill and Schultz, 2010). Such accounts need a much closer and careful look to separate the difference in task methodologies and limitations of the technique used, if any.

One of the earliest studies to dissociate the roles of lOFC and mOFC in decision making and behavior was done on monkeys by (Noonan et al. 2010), partly because of the lesions done in mOFC. (Noonan et al. 2010) was essentially a double dissociation study where the task which the monkeys performed in several combinations of reward probabilities of the options - highest (V1) and second highest (V2) values being distinct but very close, V1 and V2 difference begin medium, and the difference between V1 and V2 being highest (easiest). And since the studies were done on both lOFC and mOFC lesioned animals separately, the authors could make a wide range of observations highlighting distinct functional roles of lOFC and mOFC. Striking of the observations, in the case when the difference between V1 and V2 was close, mOFC lesioned monkeys showed impaired performance while the controls and lOFC lesioned animals fairly performed well as V1 and V2 were distinct. Such an impairment argued for mOFC's role to be more sensitive

to the value difference between the options. Conversely, when the difference between the values was more, quite surprisingly lOFC lesioned animals were impaired where as the controls and mOFC lesioned animals could steadily perform optimal choices. While it was an interesting observation to see how mOFC could not compensate even when the difference between the values is high (meaning it is an easy choice), it highlighted the role of lOFC in appropriate credit assignment, i.e. assigning the reward to the appropriate choice made in the current trial rather than to the previous or even the succeeding choice or even to the choice that rewarded the most historically.

In a different study with single-cell recordings in macaques, (Bouret and Richmond 2010) also dissociated the roles of OFC (most probably the lateral part) and vmPFC (which generally goes with same considerations with mOFC) in the context which processes drive the decision making in each of the subregions. With well-defined task sets including cued-trials, passive trials and self-initiated trials, (Bouret and Richmond 2010) first established from the recordings that both OFC and vmPFC closely represented the perceived value of task events. Strikingly, it was shown that while OFC predominantly emphasized the value information arising from external factors like visual cues, vmPFC neurons seemed to represent the value information more with respect to internal motivational processes like satiety levels, inline with the anatomical knowledge that vmPFC is more connected to regions related to autonomic regulation. By comparing the neuronal activities between cued-trials and self-initiated trials, it was also observed that neurons in OFC represented very less influence of response animals needed to perform to get the reward. Furthermore, with respect to the feedback received, OFC was observed to more sensitive to the action value after receiving the feedback whereas vmPFC was more sensitive to the action value before the feedback, underscoring some common theories that vmPFC probably has more role in making the choice before acting and OFC being more crucial for the credit assignment after the feedback.

There have been several independent accounts about the representation of absolute value of an outcome as opposed to its relative value compared to another outcome within the medial and lateral subdivisions of orbitofrontal cortex. A consistent narrative seems

to be that there is activity in the medial OFC that represented the absolute value of an outcome such as monetary, olfactory, taste etc (Grabenhorst et al. 2008; O'Doherty et al. 2001, 2003; Rolls et al. 2008). In addition, (Grabenhorst and Rolls 2009) provided a dissociative account of relative pleasantness of odors in the anterolateral orbitofrontal cortex and absolute pleasantness in the medial orbitofrontal cortex, area 10. However, in a study which investigated only relative value of monetary outcomes, (Elliott et al. 2008) reported activations in the vmPFC related to the same perceptual stimulus, greater when it predicted higher monetary reward than when it predicted the less valuable. Furthermore, it has been reported the there is a dissociation of absolute and relative loss representation in the lateral orbitofrontal cortex that is dependent on personality. Fujiwara2008 showed that lateral OFC encodes relative loss in the case of neuroticism whereas absolute loss in the case of introversion. However, it is not entirely clear if the absolute and relative subjective value in terms of pleasantness (or gains) is comparable to those in terms of unpleasantness (losses), as it has been observed in several studies with olfactory or monetary outcome that the unpleasantness or losses were represented in anterior insula (Abler et al. 2005; Kringelbach et al. 2003). Notwithstanding such complementary as well as contrasting findings, separable representations of absolute values and relative values seem to be a key dissociation between the lateral and medial orbitofrontal cortex. Future studies that could effectively dissociate the factors such as salience of the stimuli, internal motivation and thereby investigate the difference between absolute and relative value encoding , should provide more insights into whether lateral and medial OFC play dissociable role to represent them.

Another interesting theory about the dissociating role of lOFC and mOFC, probably requiring closer look, is that although value comparison might take place in the mOFC, whereas mOFC might represent the values of options in a context where there is no choice to be made, lOFC doesn't represent the value in a choice-free context. (Nogueira et al. 2017) analyzed single-cell recordings from rats performing outcome-coupled perceptual decision-making task where reward was delivered not depending on the value acquisition from the stimulus but rather on the outcome of previous trials.

There have been more studies on varying roles of lateral and medial OFC at various stages and several aspects of decision making and behavior. In one of the first intracranial EEG recordings of OFC in humans, (Li et al. 2016) recorded Local Field Potentials (LFPs) from intact OFC in patients with partial epilepsy performing a probabilistic reward learning task (Vanni-Mercier et al. 2009). As the authors in (Li et al. 2016) could clearly distinguish the recording site of the electrodes in reference to the separation of medial orbital sulcus dividing lateral and medial OFC (Zald et al. 2014), it was found that both lOFC and mOFC showed activations encoding risk and reward probability. Interestingly, a significant role in lOFC was shown to be encoding experienced value in the reward delivery phase. Above all, the work highlights the temporal progression of the emergence of reward and risk signals w.r.t the task phases, thus underlining the importance of studying the spatio-temporal aspects of decision making to dissociate different roles of lOFC and mOFC.

Table 4.2: Dissociable functions of lateral and medial OFC

✓- positive correlation ∗ ✗- lesion, impairment				
Study	**Function**	**lOFC**	**mOFC/vmPFC**	**Remarks**
Bouret and Richmond 2010	Motivation	✓- externally driven	✓- internally driven	cued vs self-initiated trails
	Neurons representing action compared to reward	✓- less	✓- similar	cued vs self-initiated trails
	Sensitivity to action w.r.t feedback	✓- after feedback	✓- before feedback	cued trails
Noonan et al. 2010	Switch choices after outcome	higher for both	higher for both	higher for errors than rewards

	Option value difference	✗when more	✗when less	Value comparision in mOFC/vmPFC
	Rapid stimulus-reward learning	✗impaired	✗- increased (like controls)	When value : best >> avg(all the stimuli)
	Credit Assignment	✗- heavily impaired	✗- no or little effect	lOFC lesion : recent or frequent choice
Lebreton et al. 2009; Nogueira et al. 2017	Valuation in choice-free context	✗Nogueira et al. 2017	✓Lebreton et al. 2009	Value comparision in mOFC/vmPFC
Rudebeck and Murray 2011	Reinforcer Devaluation	✗- heavily impaired	✗- no or little effect	lOFC : stimulus - outcome mapping
Howard and Kahnt 2017	Selective Satiety	✓- change w.r.t sated outcome identity	✓- both reward representations modulated	Ambiguous : reward value or reward identity
Grabenhorst et al. 2010	Subjective value of presented options	✓- relative value	✓- absolute value**	**-contrast to other studies (relative value)

4.4.1 Challenges in studying dissociation of lateral and medial OFC

In the few studies that have described the nature of dissociation between lateral and medial OFC, certain results were difficult to explain and there appears to be several possibilities for such results. For instance, when it is observed that neither of the individual lesions of lateral and medial OFC impair the animal in Reversal Learning while the lesion of OFC as a whole does, it could be possibly mediated by other sub regions of

OFC (central, anterior or posterior) or there might very well be a possible mechanism through which they partially compensate for one another interacting with other parts of the brain. In consecutive reinforcer-devaluation tests, while monkeys with lOFC lesion showed significant impairment, the ones with mOFC were less consistent in their choices across sessions contrasting overall performance with the controls and shift in different scores (DS) with those of lateral lesions (Rudebeck and Murray 2011).

As it has been established thoroughly, the roles that OFC as a whole has been implicated in - stimulus values, outcome expectations, risk, outcome values, action values, task contingencies, relevance to internal needs, to name a few - are some of the most fundamental conceptions that form the basis of decision-making and behavior in primates and rodents. Henceforth, it is extremely challenging yet inevitable to design experimental paradigms that capture such constructs and methodologies and allow to infer appropriate neuronal correlate. Only then we might get a glimpse into studying distinct subregions of OFC, their interactions within and across cortical and subcortical counterparts.

Most of the studies mentioned above have studied the roles of lateral and medial OFC in dissociation. However, the role of any possible interaction between the both, given their connectivity through the medial orbital sulci, has not been explored much. However, even if it is the case that the lateral OFC represents identity specific rewards and vmPFC represents general, scaled reward signals, it is unclear how these two signals could be linked to subserve goal-directed behavior. To this extent, there is not much evidence except one study, an fMRI analysis in humans performing a food-odour task with tested with satiety. Besides individual activities in lOFC and vmPFC, (Howard and Kahnt 2017) analyzed the connectivity between lOFC and vmPFC to show that the functional connectivity was predictive of satiety related changes in choice behaviour.

As extensive is the range of faculties that are attributed to the role of OFC, so is the set of challenges involved in studying OFC for understanding of underlying processes. The anatomical description implicitly highlights the difference in the accessibility of lateral and medial OFC for experimentation procedures. The same applies to imaging studies, which happen to be a major contribution of studies in humans, that the regions highlight

by BOLD signals cannot be precise enough within the scope of subregions. Another challenge is the homologies of the OFC among humans, nonhuman primates and rodents. Since a good part of the literature on OFC is almost equally contributed by the studies in all three species, it would be major task at hand to be wary of the similarities and the differences among what is defined as OFC in each of these studies. Functionally, it is also not straightforward to identify whether a difference in an ability of one species (say humans) to demonstrate a faculty and that of another (say rats) is a difference in kind or a difference in degree. Although, it would be fairly possible to extend the conclusions from one species to another depending on what is being studied (for example, the findings related to action-outcome contingencies or basic behavior in rats might extend well to primates beyond which more flexible representations might emerge).

Well within one species, reported counter-intuitive findings also might encourage the need to dissociate the roles of lateral and medial OFC. Quite often the ambiguity arises from the task structure which doesn't sufficiently dissociate closely related aspects of a certain behaviour. Especially in humans, the most accessible experimentation at the moment being functional imaging restricts the task structures and the nature of data sets. But as the techniques in lesion and single cell studies in animals evolved, more feasible techniques like intracranial EEGs have been emerging (Li et al. 2016). This used to be a problem in animal studies, when they were done using aspiration lesions, where not only the intended region, but also connecting and passing fibers and axons get lost, resulting in possibly inappropriate conclusions. Part of the literature on initial accounts of OFC being crucial primarily in response inhibition or reversal learning were done using aspiration lesions (Iversen and Mishkin 1970; Jones and Mishkin 1972) Whereas recent studies using fiber-sparing excitotoxic lesions overturned those ideas (Rudebeck et al. 2013).

Table 4.2 does not include the studies that have individually studied lateral and medial OFC but essentially showed no impairment in the function. For instance, (Noonan et al. 2010) and (Noonan et al. 2012) studied the possible dissociation in function of lateral and medial OFC both in macaques and humans, but only to find out that there are no contrasting effects on the sensitivity to punishment and reward.

4.4.2 Lateral OFC

Valuation Within a trial, lateral OFC lesions disrupted the assignment of precise values to stimuli (Walton et al. 2010). Across trials, one of the key aspects that highlights the role of OFC over other sub-cortical decision systems is the possible maintenance of historical choice scenarios (offers, choices as well as outcomes), either by internal mechanisms or with the help of working memory. Even in the context of perceptual decision making where there is no stimulus-value involved in decision, information from previous experiences could influence the current decision. Consistent with other theories on OFC that represents a partially observable current state in a comprehensive manner with previous information (Bradfield et al. 2015; Schuck et al. 2018; Wilson et al. 2014), a single-cell study in rats reported that populations in lOFC together encoded a second-order prior (related to relevant choice after an incorrect response), previous choice and previous outcome (Nogueira et al. 2017). Neuronal activities were recorded in rats performing a decision-making task where reward in each trial is coupled to previous outcome received, hence requiring the rats to integrate prior information, while stimulus is an ambiguous inter-tone time interval (ITI). This study highlights, unlike many other implications of OFC, an important role in choice anticipation before stimulus onset in a trial setting. One of the striking remarks that the authors in (Nogueira et al. 2017) make is that the activity in lOFC represents this prior only when there is a choice to be made but not in passive trials. This seems consistent with the earlier accounts that proposed state-space

information interacting with decision-making processes in OFC when it is behaviorally relevant(Schuck et al. 2018). Single-cell recordings in monkeys also highlighted the representations of values in lateral OFC, and further showed that the neurons in lateral OFC exhibit a phenomenon called **range adaptation**. The same neurons can inform decisions made between stimuli associated with two small rewards of only slightly differing sizes and, on a different occasion, between stimuli linked to very large rewards (Kennerley et al. 2011; Kobayashi et al. 2010; Padoa-Schioppa 2009).

Outcome evaluation and Learning Although both the medial and lateral OFC are known to encode risk and reward probability, lateral OFC plays a predominant role in encoding experienced value of an outcome. Contrary to many previous accounts of OFC lesions causing perseverance of same options in case of no reward or error outcomes, recent studies highlighted that in fact, the animals are more likely to switch responding rapidly between choices following reversals (Noonan et al. 2012; Walton et al. 2010).

The role of lOFC after choice feedback in a decision trial has been underscored in several studies (Jones et al. 2012; Rudebeck and Murray 2014; Takahashi et al. 2011; Walton et al. 2010; Wilson et al. 2014). The role of lOFC, but not mOFC, has been highlighted in evaluating externally driven motivation for decision through evaluating stimulus features (Bouret and Richmond 2010). Subsequently studies pointed out the role of lOFC in assigning a particular outcome to a particular stimulus (Noonan et al. 2010; Walton et al. 2011) . Furthermore, it has been proposed that, lOFC, owing to its finer sensitivity towards stimulus features, does not only assign credit to the correct stimulus, but also associates the stimulus to the specific reward type and more so in the same way for both rewarding and error outcomes(Noonan et al. 2011). In addition, several studies showed that lOFC may be involved in updating information about the transition between stimulus and outcome identities (Boorman et al. 2016; McDannald et al. 2011).

Thus lOFC accounts for several explanations of its role in adaptive valuation of the stimuli, learning precise stimulus-outcome associations based on feedback thus facilitating more accurate valuation, and more comprehensive representation of the environment

(stimuli, choices, outcomes; identities and values; history) throughout the task process, that is behaviorally relevant. However, as it could be noted, the role of lOFC in the value comparison process, the final choice between the options has not been discussed. Rather it is the complementary subregion mOFC that has been widely implicated in the process of choice while lOFC is believed to be supporting the choice process by its extensive learning processes.

4.4.3 Medial OFC / Ventro Medial PFC

Medial OFC (along with the neighboring vmPFC in humans), is the region where often value-related signals in decision making have been identified (Boorman et al. 2009; Knutson et al. 2005; Kolling et al. 2012; Tom et al. 2007; Walton et al. 2010). However, as complicated the construct of value is, the neural correlates are usually found in multiple brain areas (lateral OFC, ventral striatum). Moreover, there has been debate over the precise role of this region in value-guided choice (Kable and Glimcher 2009; Noonan et al. 2010). Some fMRI studies showed that vmPFC has been found to signal a difference between chosen and unchosen values (Boorman et al. 2009; Serences 2008) whereas in others it has appeared to signal the overall value of available reward (Blair et al. 2006), and even possibly the value of just the chosen option (Kable and Glimcher 2007).

Value integration : Not emphasizing any particular kind of value vmPFC might actually represent during a choice, it has been proposed to integrate the values of rewarding options over time as well as multi-dimensional information throughout presentation (Milosavljevic et al. 2010). It was shown in monkeys that, in contrast to the consistency of reward identity mappings in the lateral OFC, mOFC/vmPFC activity seems to reflect the expected values of outcomes and occurrence of positive outcomes, irrespective of where the same outcomes were consistently mapped to the choices (Noonan et al. 2011). (Bouret and Richmond 2010) reported from recordings in monkeys how neurons in vmPFC encoded internal factors such as internally-driven motivational processes during the valuation pro-

cess. Possible computational explanations about how vmPFC might be achieving this integration process have been proposed, for instance drift-diffusion models (DDMs) and attentional DDMs (Bogacz et al. 2006; Krajbich and Rangel 2011; Milosavljevic et al. 2010). Furthermore, it has also been suggested that vmPFC also encodes a higher order state information in addition to just recent reward history for an action. (Hampton et al. 2006) found the activation in vmPFC appearing to be making a state-based inference similar to a bayesian markovian model. This finding supports numerous other accounts that have been implying prefrontal circuits to be facilitating the part of model-based valuation in a rather hybrid (along with model-free valuation) behavior found in humans and rats (Daw et al. 2011; Dezfouli and Balleine 2019; Lee et al. 2014).

Value comparison : VMPFC activity has been linked to a wide range of valuation signals that processes information about temporal delay (Mcclure2004, Kable2007, Peters2009), uncertainty (Levy2010), or even social advice (Behrens2008), and reward outcome (Rolls2008). This variety of implications pushes for a possible different hypothesis that vmPFC might be, through representing different types of values probably across different sub-populations, driving the choice using these representations (Grabenhorst and Rolls 2011). (FitzGerald et al. 2009) showed using a binary choice paradigm, the activity in medial OFC correlated to the value difference between the options which are essentially incomparable objects (money and objects), developing on the existing ideas of possible wide range of value representation and adaptation to different range of values (Padoa-Schioppa and Assad 2006, 2008). Supporting the view that the relative difference of the presented options is represented in vmPFC, multiple value comparison mechanisms have been proposed. This value difference signal further allows vmPFC to perform a value comparison to facilitate the choice through principles of mutual lateral inhibition (Grabenhorst and Rolls 2011; Rolls et al. 2010; Strait et al. 2014; Wang 2008).

Differing from a notion that vmPFC is important for value comparisons under risk or uncertainty, (Fellows and Farah 2007) showed that vmPFC damaged patients were inconsistent in maintaining value preferences in general value-based decision making while

retaining optimal performance in perceptual decisions. In addition, it was observed that there was no impact on response time, meaning that the vmPFC damaged patients made inconsistent decisions as quickly as control subjects.

Furthermore for objective value evaluation and comparison, it is also important to consider mOFC's role in taking into account the anticipation of high-level emotions that are possibly tied to the decisions. In an imaging study on subjects where information was provided about the unchosen option (Coricelli et al. 2005), a pronounced regret modulated behavior of the subjects strongly correlated with the BOLD activity in mOFC (besides ACC and hippocampus). In addition, mOFC showed similar activity before making choices in an increased regret-aversive behavior, alongside amygdala.

Action selection and learning: An array of fMRI studies in humans showed supporting evidence that part of vmPFC encodes action-outcome associations (Daw et al. 2006; Hampton et al. 2006; Kim et al. 2006; Tanaka et al. 2016). The observations were underlined by the correlation of activity in vmPFC to the expected outcome value derived from a model of reinforcement learning which attributed past rewards to past actions in a multi-arm bandit task (Daw et al. 2006). Multiple studies have demonstrated how mOFC or vmPFC could be encoding this kind of outcome expectation and thereby action value estimation, in humans as well as monkeys (Hampton et al. 2006; Noonan et al. 2010; Rudebeck and Murray 2011). However often a multi-arm bandit task paradigm as in (Daw et al. 2006) would still leave an open question whether such value prediction of an outcome is rooted in a goal-directed behavior or rather habitual, although there have been arguments that vmPFC contributes to more goal-directed action selection (O'Doherty 2011). In addition to value comparison, mOFC is also believed to be crucial for maintaining a choice over successive decisions. Lesions in the mOFC caused monkeys to lose their normal predisposition to repeat previously successful choices (Rudebeck et al. 2013). Conversely, lesions to Walker's area 14 (medial OFC) impaired the ability of monkeys to learn to stop responding to a previously rewarded object, while there was no such impairment in the case of lesions to areas 11/13 (Rudebeck and Murray 2011).

Often when Outcome Devaluation paradigms are used in experiments, the extent of training has an important implication as to whether the learning that results would goal-directed (limited training) or habitual (overtraining) (Daw et al. 2005; Killcross and Coutureau 2003). Valentin et al., 2007 reported an fMRI analysis performed in humans selecting instrumental actions for food rewards with a moderate training and further devaluing an outcome (overfeeding). The authors could isolate the activity in mOFC to show sensitivity to action-outcome associations, showing the difference in activities of devalued action and valued action, before and after satiety. In effect, Valentin et al., 2007 reports no correlation to the habitual component of the learning process by comparing response profiles during test on trials involving choice of the valued and devalued actions to those of neutral condition (tasteless outcome) thus eliminating the possibility of stimulus-response associations in learning. However, the nature of the task doesn't exclude the possibility of stimulus-outcome associations playing a role in the observed goal-directed behavior.

More precise evidence eliminating the possible encoding of stimulus-outcome associations in mOFC in the context of goal-directed learning came from (Tanaka et al. 2008) and (Gläscher et al. 2009), employing tasks where the subjects have to choose from motor responses in the absence of explicit discriminative stimuli, still reporting the evidence for the activations in vmPFC in encoding action–outcome-based value signals specifically.

In an interesting proposition, (Hunt et al. 2012) analyzed source-reconstructed magneto-encephalography (MEG) data while the subjects performed a value-based decision task, where they needed to integrate stimulus information between two risky options (where each represented an amount of reward and the probability of getting it). It was reported that activities in vmPFC were even more remarkably distinct between more deliberative situations with slower reaction times as opposed to trials towards the end of the experiment or even no brainer trials (highly probable high reward versus the opposite). Moreover, they reported that the involvement of value-difference signal in vmPFC consistently decreased towards the later trials of the task.

4.5 Discussion

The Orbitofrontal cortex and its inner mechanisms that facilitate the flexible animal behaviour have been of great interest for long time now. There has been an evolution of testing paradigms, experimental techniques, more detailed internal anatomical and functional differences through this time. Especially recent studies that are similar to many previous studies overturned some of the implications from older findings. The theories that once stated that the OFC's primary function is response inhibition or that OFC (lateral or medial specifically) is more sensitive to negative feedback, have been contested and have been argued to the contrary.

The growing interest in the functional subdivisions of OFC has added to this complexity. Rather surprisingly, although numerous studies demonstrated the effects of whole OFC lesions in reversal learning (Fellows and Farah 2005; Izquierdo et al. 2004; Jones and Mishkin 1972), lesions of neither lOFC nor mOFC in monkeys showed any impairment in reversal learning (Kazama and Bachevalier 2009; Rudebeck and Murray 2011). Furthermore, studies like (Kazama and Bachevalier 2009) showed that lateral OFC deficits did not impair reversal learning in monkeys but those of medial OFC did, and to the contrary (Rudebeck and Murray 2011) reported that not only lateral OFC deficits do not impair reversal learning, neither do the lesions of medial OFC. Findings like these highlight that OFC is certainly much more complicated than a dissociation between its lateral and medial subdivisions, or of any other kind of subdivisions (anterior and posterior for example). More recent fMRI analyses have reported functional dissociation between anterior and posterior OFC, that the former specifically encoded secondary rewards (money) than primary rewards (taste and smell) and vice-versa for the latter (Li et al. 2015; Sescousse et al. 2013). Subsequently, it has been proposed that probably the anterior OFC codes more abstract rewards while the posterior OFC represents more concrete or tangible rewards (Bechara and Damasio 2005). Essentially the specificity with which a certain study on OFC is being carried out is crucial for implicating OFC to a certain function to any degree.

Table 4.3: Important reviews on the role of OFC

Article	Emphasized Role of OFC
Murray and Rudebeck 2018	Reward guided decision-making
Padoa-Schioppa and Conen 2017	Economic decisions
Stalnaker et al. 2015	Invalidation of common notions
Wilson et al. 2014	A cognitive map of task space
Wallis 2012	Cross-species value-based decision-making
Noonan et al. 2012	Reward and reinforcement
Kringelbach 2005	Functional neuroanatomy

It can be noted that only the core theoretical aspects of value-based decision making and goal-directed behaviour are considered in this chapter, implicating OFC to be playing a crucial role. Despite the paradigms becoming more sophisticated in their analysis, the concept of value is often conflated with risk, uncertainty, or even when it is about different types of reward. The previous famous notion was that all the values must be computed in terms of a "common currency" somewhere in brain (Montague and Berns 2002) whereas more and more evidence is now growing on the theories that the values must be assigned separately in a choice to retain the reward identity information and further compared somewhere else on a common neural scale (Grabenhorst and Rolls 2011; Kringelbach 2005). Besides, it is completely plausible that both are possible with independent valuations as well as combining with other factors and further computing in a "common currency" serving the behavioral choices. Nevertheless, a full understanding of the representations of reward value in OFC and their link to behaviour depend on a comprehensive distinctions among various circumstances of valuation. Moreover, extending beyond value-based decision making, there are numerous studies which investigated OFC in different behavioral and emotional contexts. Subsequently it can be imagined how the role of OFC in the aspects described here might translate to high level behavioral traits. Lateral OFC has been reported to play a role in justified and unjustified violence (Domínguez D et al. 2018). VMPFC damage was observed to show insensitivity to private counterfactual value signals but not to social (others' choices) counterfactual value signals (Bault et al. 2019).

In terms of the experimental techniques, it can be noted that all the studies discussed so far range from BOLD signal analyses of fMRI in humans to single-cell recordings in rats. It might be then, debatable as to how far these studies could be discussed at the same level of agreement. More so, when the debate around the anatomical homologues between primates and rodents is still active, this question becomes more pertinent. But we highlight, that besides notable differences between species, there are also comparable similarities between species that have been established by extensive anatomical findings. However, it is still inevitable for the studies across different species to agree on some common ground about the experimental paradigms they follow so that the findings remain in the acceptable range of comparison or at least complementary (Wallis 2012). Given the high level faculties that are studied in the context of OFC, it could be beneficial to view the relevant regions and subregions involved from a more functional perspective than a strictly anatomical perspective across species. More similar tasks are reproduced across humans, non-human primates and even rodents (for example (Padoa-Schioppa and Assad 2006) in monkeys and (Gardner et al. 2017) in rats), although carefully modifying the tasks according to the species. This would support a great deal to take forward the enormous effort to understand the underpinnings of the role of OFC.

1. OFC most certainly stands out a critical region of the Prefrontal cortex that plays a crucial role in adaptive and flexible behavior. It is interesting to note that in a general sense, lack of OFC doesn't hamper basic behavior or decision-making abilities of the animals. Animals are found to demonstrate adequate behavior with the help of brain circuits involving sub-cortical structures like Amygdala and the Basal Ganglia (BG). In more abstract scenarios however, when the state of the world is more complex than it is directly seen and more sophisticated decision-making is required, OFC seems to be very crucial for the behavior.

2. Precise representations and mechanisms within OFC that contribute to behavior in complex task structures still remain as active questions of interest. First functional positioning of OFC both as a part of the generic prefrontal circuits that facilitate

expression of voluntary behavior and as a part of the limbic system with amygdala and the BG needs to be described in a comprehensive way. Then the idea of anatomical and functional dissociation, or even possible interaction within the sub-regions of OFC, at least lateral and medial besides other possible sub-divisions, needs to be explored.

3. In the context of voluntary behavior that is a function of value-based decision-making, a closed-loop experimentation structure that includes external environment and the internal bodily state would allow to study the resulting behavior, not only quantitatively, but also in terms of qualitative behavior, as observed in animals. Further, the available computational accounts of each of the individual behavioral paradigms that involve other sub-cortical systems in conjunction with OFC can be leveraged to build a systems level model in which the representation and dynamics within the subregions of OFC can be explored.

4. Taking these observations into account, precise objectives of this thesis work will be defined in the next chapter that would address several of these issues. The objectives include tools used to build the experimental framework, algorithmic representations of possible organization in the prefrontal cortex with respect to behavior, computational modeling of specific pathways involving OFC which represent the existing anatomical and experimental knowledge, tasks that are used to evaluate the framework and the model, and finally the analyses of the implications of a dissociate view of OFC, along with the circuits it is part of, on flexible behavior.

Chapter 5

Objectives

Sommaire

The core interest of this thesis is two-fold - *(i)* To build an experimental framework that will allow to study a behavioral architecture that describes the key components of voluntary behavior of animals (or an agent) in a changing environment. This behavioral architecture is based on the neuroanatomical findings about PFC and the distinct feedback loops between its subregions and the structures of BG (referred as parallel CBG loops from chapter 3) *(ii)* Within such a framework, having identified that OFC plays a crucial role (chapter 4), use computational descriptions and modeling to study the possible role of OFC in voluntary and flexible behavior. This part of the study relies on OFC's well known involvement with the downstream brain regions to form specific decision, valuation and learning systems, and recent light on its anatomical and functional heterogeneity

The above stated goals are divided into several precise objectives. In the following chapters, each of the objectives is described in detail, relevant experiments are presented

and the results are analyzed. The objectives can be presented across two major axes.

Axis I : Closed-loop experimental framework of voluntary behavior

5.1 Experimentation Framework and Virtual environment

In this work to study animal behavior, the key aspects used are an agent characterized by a *body* (sensors, internal needs and motors), an environment in which the agent is a part of, and the agent's interactions with the environment (processing in the agent's *brain*). Such a description allows to study agent's voluntary behavior as a part of a closed-loop system where at any given point of time, the environment stimulates the agent and the agent acts on the environment. Agent's action in such a case is viewed as the consequence of the voluntary decisions the agent makes. Agent's actions in the environment carry mutual effect on each other, of the environment on the agent's state and of the agent's action on the environment.

Computationally, decision-making has been studied and explained in the form of several mathematical models and numerical simulations. In this work, as the interest spans to understand more flexible decision-making across different brain circuits, a virtual environment is used to demonstrate the experiments. A software platform Malmo has been chosen which provides a communication interface with a well-known 'survival' based video game Minecraft. Malmo acts as a software layer to communicate with the game, receiving the current state and modifying it through commands, making available several attributes of the game. Although later multiple video game-like environments have evolved to allow behavioral experimentation, the choice of Minecraft was rooted in the vast possibilities available to design in Minecraft and its closer resemblance to the scenarios of an animal in an ecological situation. The environment is adapted to simulate behavioral experiments on an artificial agent (like an animal).

Goal I : Describe an experimental framework to study voluntary behavior

Goal II : Adapt a virtual environment according to the behavioral framework

5.2 Behavioral Architecture of Parallel Generic Feedback-loops

Imagine an agent (or an animal, referred as *agent* hereafter) with a set of internal needs, exploring in an environment. There are stimuli that are appetitive corresponding to different needs of the agent. As the agent processes the information from the environment, there are several aspects that need to be considered. For instance, the identities of the stimuli; are they novel ? if not, are they known to be appetitive? what is their relevance to the current needs ? if none, what could be their relevance to future needs? their relative positions; action plans if should be pursued; Most of these aspects are often incommensurable and cannot be represented as a single common factor that drives the agent behavior.

In this work, first an algorithmic framework of behavior is described that would *organize* the information (external from the environment and internal from the body), *evaluate* it and further *act* if necessary. This framework is based on the distinct and parallel anatomical organization of the areas in frontal cortex, sensory cortex, the basal ganglia (BG) in addition to the other sub-cortical structures, as shown in figure 1.2 in chapter 1. Each of these loops is implemented with an underlying generic structure and function of processing respective information. Figure 5.1 shows a generic loop that involves one area/structure each from sensory cortex, frontal cortex and BG. Each loop, although representing different kind of information, is implemented in common terms (neural activities

for e.g.), so as to facilitate the modulation between the loops. That is to say, the resultant activity in one loop can affect the activity in the other.

The agent behavior at any given time is viewed as a closed-loop work flow : Evaluation, Selection, Execution and back to Evaluation. The emphasis in this kind of description of behavior is not on decision making per se, but evaluation before the decision and execution after the selection.

1. ***Evaluation*** As the information about the environment is passed to the frontal cortical region, more relevant representation of the information is abstracted, and built-in rules are applied to evaluate the information. The repertoire of rules grows in time with experience by other processes, which are assumed to be hard-wired.

2. ***Selection*** The BG structure is specialized for selection, to resolve the competition between ambiguous options. Evaluation from the frontal region provides a necessary bias according to the situation and history and leads to a beneficial choice.

3. ***Execution*** This is achieved by an important property of representation in the sensory cortical region. There are two representations in each sensory area : *desired* and *actual*. For each of the given repertoire of elements that make up a sensory state - stimuli, positions, needs - there is a distinction whether the element is *actual* as perceived currently, or *desired* as per the motivation of the agent. The action execution is achieved in the form of sustained activity until the actual state matches with the desired state, and that's when the action execution ends (discussed in detail in the following chapter cf. Sec 6.4 pg. 135).

And finally, depending on the resulting state, the *evaluation* begins again, especially taking into account any possible outcomes that follow the current execution. As will be seen in following chapters, *execution* is described algorithmically to demonstrate the principle and implementation whereas *evaluation* and *selection* are discussed in greater detail with supporting neuro-computational evidences and accounts.

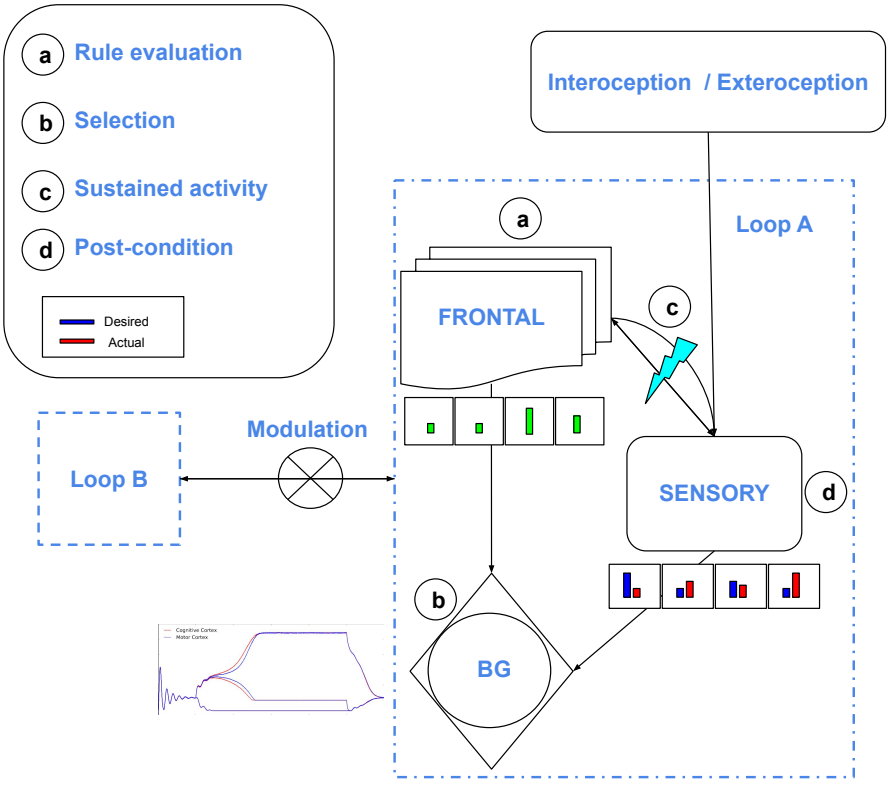


Figure 5.1: A generic loop of behavior. Involving a sensory cortical area, one frontal area and a structure of basal ganglia.

Goal III : Summarize global organization in Brain underlying flexible behavior

Goal IV : Integrate a biologically-inspired model of parallel loops in the virtual environment

Axis II : Representations and processes of value-based decision-making in OFC

In the behavioral framework described in the previous section, *evaluation* and *selection* make most of the cognitive architecture that guides the agent behavior. Therefore, first a cognitive architecture that sits within the behavioral framework is described as a combination of two fundamental paradigms - Pavlovian and instrumental - that drive *evaluation* and *selection*. Further, available computational models in each paradigm are described and leveraged to be a part of the architecture that drives the agent behavior. Finally, with the support from the evidence discussed in chapter 4 and more anatomical observations with respect to Pavlovian and instrumental counterparts in brain, a case for the dissociation between lateral and medial OFC, within the same cognitive architecture is proposed. The individual models are combined to interact with each other, with adaptations and extensions, making them able to perform multiple tasks with minor adjustments. Also the kind of learning used in these models is modified to account for more specialized information from the OFC. As a result, with such a systems-level comprehensive model that incorporates OFC in a distinct manner, a couple of neuroscientific studies performed in animals could be replicated.

5.3 Behavioral Paradigms - Neurocomputational models

Throughout most of this work, it is assumed that the animal is hardwired to identify a set of outcomes as appetitive, corresponding to an Unconditioned Stimulus (US). Therefore, to study a more complex behavior, the situations that are considered are those in which neutral objects form predictive relations with the known outcomes and thus guide behavior. These objects are usually referred as *reinforcers* or *predictors* or in this work hereafter, Conditioned Stimulus,***CS***. A paradigm that is generally used to study the effect of such CSs on behavior is the Pavlovian to Instrumental Transfer (PIT) paradigm. Described in multiple formats, PIT consists of an animal separately learning Pavlovian expectations (section 2.3.1) and instrumental actions (section 2.3.2). Subsequently, the

presence of the Pavlovian cues affect the animal's instrumental actions. First, detailed models of instrumental and Pavlovian conditioning, which have been implemented using a common neuronal dynamics, are presented and shown to reproduce previously replicated experimental evidences.

As highlighted in the figure 5.2, in the middle of the information flow from sensors and sensory representations (of both stimuli and positions) to the motor action, is the combined network of selection mechanisms, Pavlovian and instrumental. Although Pavlovian systems only elicit hardwired consummatory responses but not preparatory or planned actions, they influence the actions in the instrumental systems through the processes termed as Pavlovian Instrumental Transfer (PIT). Thus, several aspects of behavior will be developed using these behavioral paradigms and the modeling accounts that explore these paradigms.

Neural architecture for instrumental learning

In the behavioral framework presented in this thesis, the core action selection mechanism between multiple options - cognitive stimuli or motor actions, is implemented using an architecture of the basal ganglia (BG) similar to that has been described in classical descriptions of pathways in BG, summarized well in (Boraud et al. 2018). Figure 5.3 highlights the structure of a typical feedback loop that BG forms through thalamus and prefrontal cortex. Although specific implementation of different such loops will follow in the following chapter, the architecture presents a general idea of the connectivity between the input structures of BG - Subthalamic Nucleus (STN) and Striatum (STR) and the output structures - Globus Pallidus pars Interna (GPi) and Substantia Nigra pars Reticulata (SNr). Fig. 5.3 shows different connectivity pathways that are part of the CBG network described in the architecture - the direct pathway from the prefrontal cortex (PFC, here used generally, including OFC) via STR to GPi, and the hyperdirect pathway from CTX via STN to GPi. It has to be noted is that this is only one of several possible interpretations of action selection mechanism within BG, as it sufficiently explains

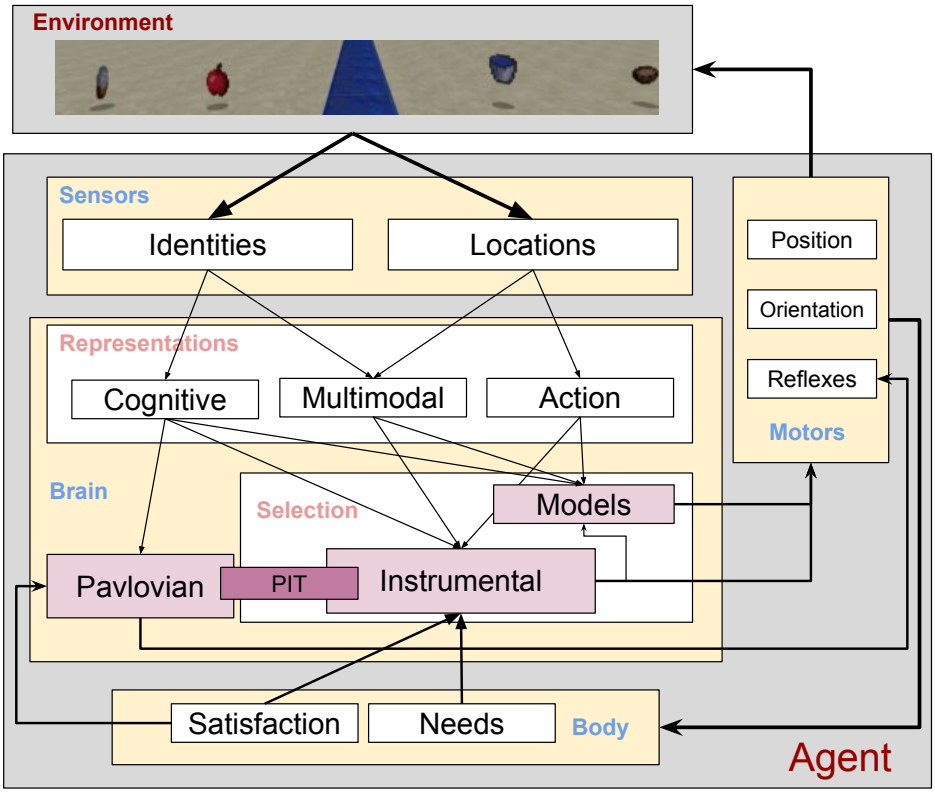


Figure 5.2: High-level description of the experimental framework and the cognitive architecture of the agent brain, as implemented. From an agent-centered view, therefore environment is seen independent.

using one excitatory and inhibitory pathway. Several other interpretations exist, for instance a rather "classical" view of CBG network describing an indirect pathway, which are not considered in this work. Indirect pathway involves STN, GPe (Globus Pallidus pars externa) and STR Gurney et al. 2001a; Gurney et al. 2001b.

Neural implementation of Pavlovian learning

Pavlovian conditioning, in the context of the environment described so far, is associating a neutral stimulus to the presence of an *outcome*. The way this association is learned has been explained to be within the basolateral amygdala with strong influence from the lateral OFC, by several computational accounts (Kaushik et al. 2017; Montague et al. 1996; Schultz et al. 1997; Vitay and Hamker 2014), basically using the formalism by

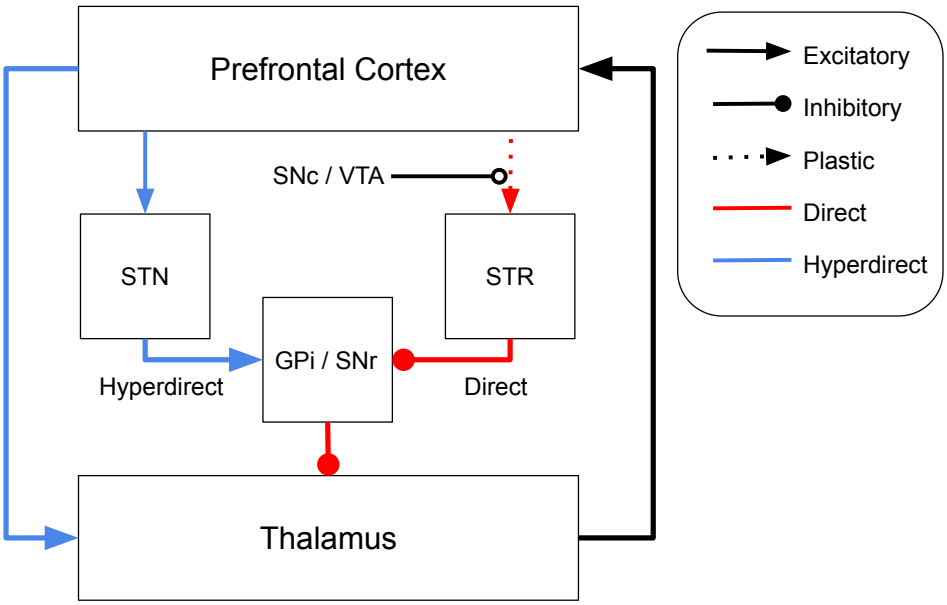


Figure 5.3: High level architecture of Thalamo-Cortical-BG Loops in primates. Classic BG connectivity : STN and STR as inputs, GPi/SNr as outputs. GPi: Globus Pallidus pars Interna; SNr: Substantia Nigra pars Reticulata; STN: Subthalamic nucleus; STR: Striatum. Image redrawn, inspired from (Boraud et al. 2018)

Rescorla-Wagner learning rule. It has been proposed that this learning is driven by the firing of neurotransmitter dopamine by corresponding neurons in the ventral tegmental area (VTA). The early accounts (Montague et al. 1996; Schultz et al. 1997) failed to explain certain neurophysiological observations of dopamine neurons such as cases where dopamine firing in VTA happens only once after learning, or when a reward is delivered earlier than learned. (Vitay and Hamker 2014) and (Kaushik et al. 2017) attempted to rectify these limitations and proposed striatal mechanisms that affect VTA and another neurotransmitter GABA accounting for the cancellation of dopamine peak at the time of reward delivery, after learning. More work has been in progress in this direction (Kaushik et al. 2017), dissociating the magnitude and timing aspects of reward and studying the role of ventral striatum in affecting the timing of RPE signalling by dopamine firing. In this work, in addition to the discussed generic feedback-loops, a simple case of Pavlovian conditioning is implemented, implying amygdala, with the magnitude and the timing of

the outcome being invariant.

Goal V : Implement known computational neuronal models in the framework of parallel loops.

Goal VI : Replicate previous numerical simulations as virtual experiments

5.4 A systems level description of OFC

Following the detailed review of the OFC in chapter 4, the lateral and medial dissociation of OFC suggests a possible explanation of distinct learning and choice systems. In addition, a closer look at the functional connectivity of lateral and medial subregions of OFC with different nuclei of Amygdala and different structures of the ventral striatum (VS), reveals more possibilities. The influence of OFC on Pavlovian learning in amygdala, the nexus of Pavlovian and instrumental learning between amygdala and VS, and the evidence of dissociation between the subdivisions of OFC, Amygdala and VS together - altogether are taken into account. Therefore, of several parallel systems highlighted in Fig. 1.1, we concentrate on the *limbic* loops, especially OFC, to closely study the value-based decision making, and simplify the *sensorimotor* loops either as the motor loops through the BG or to their generalized algorithmic implementations.

Basolateral vs Central nuclei of Amygdala Throughout the anatomical descriptions of OFC across species, basolateral nucleus amygdala is found to be in strong reciprocal connections with the (lateral) OFC (Price 2007). Although BLA learns the stimulus-outcome associations, lateral OFC might represent outcome expectation through these BLA learned associations in conjunction with recent outcome history. Likewise, by the virtue of lateral OFC's sensitivity to the sensory features of the rewarding stimuli, through

top-down bias, it could alter the plasticity within BLA which is otherwise purely weight-based learning. Supporting this view is the finding that OFC lesion causes slower stimulus-outcome learning in BLA (Stalnaker et al. 2007). On the otherhand, CeA, generally described as the output nucleus of amygdala, interfere less with the stimulus-outcome learning (Hatfield et al. 1996). Therefore the role of CeA is restricted only to express Pavlovian learning in the efferent areas like core (of the ventral striatum). Figure 5.4 shows an illustration of possible organization of behavior across several subregions and sub-structures within OFC, Nucleus Accumbens (NAcc, core and shell) and Amygdala.

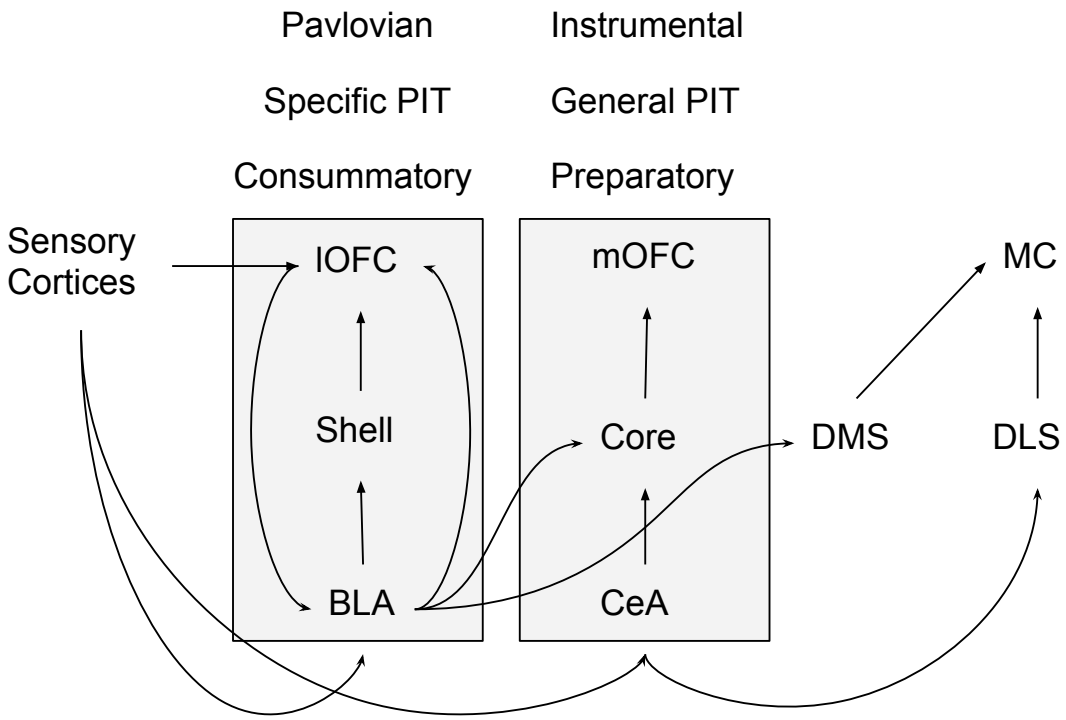


Figure 5.4: Possible functional dissociation among OFC, Nucleus Accumbens and Amygdala

Traditionally, in the context of behavior, learning and choice are intertwined such that choice is always viewed as a result of applying existing learning on an input state. This is the case, of course, when learning is permitted in the system and the system has

sufficiently learned from previous choices. Although the system could make a choice, once the learning occurs, it is purely the learning that drives the choice. This is the case even in the case of artificial neural networks in Computer Science. A network that is designed to provide an output given the input features, is modified by learning in the network. Thus, the mapping between a set of input features and an output is fixed as a result of learning and choice strictly follows learning.

One of the recent arguments is that learning and choice are separated in organization Miller 2018, such that the system can make a choice not only as a result of learning but in addition taking into account any moment-specific information. In this work, this implication is reinterpreted, suggesting that such a finding could be a result of dissociate network effect rather than the dissociate organization of learning and choice. This provides the possibility of suppressing the effect of learning on the current choice, temporarily if needed. This is well-known phenomenon in animal behaviors, describing how the decision systems in brain facilitate a balanced switch between exploring between options and exploiting the current knowledge.

Retaining the features of the instrumental learning models with cortico-striatal synaptic connections leading the decisions within the OFC loop, it has been attributed to lateral OFC in the model. Medial OFC, as found to reflect the value difference in many studies, is implemented as a value comparator. By the virtue of cortico-cortical influences, medial OFC and lateral OFC influence each other with value comparisons in the former and learning by assigning the reward correctly to the choice in the latter. One of the most recent implications noted from the detailed review of function of OFC in the previous chapter is that OFC might be important for abstracting the state space in the task from experience to support partially observable states.

Since the identity specific information is widely attributed to lateral OFC, a measure for state prediction error is implemented in lateral OFC, as it is suitable to label the states based on their identity. Since medial OFC has access to the ongoing estimated values from core, a task specific value pattern is learned in medial OFC which can possibly control the learning rates of lateral OFC. Further the emotional learning in BLA is transferred

to core to facilitate decisions in the case such as PIT.

Each of the computational model used in the underlying paradigms is explained in detail. Then the specific features implemented for lateral and medial OFC and simple effort-based representations in ACC are described. The model is tested on tasks that would highlight the importance of each of these features. A task that combines instrumental learning and effort is used to assess the effects of effort-based learning on instrumental learning. A 3-arm bandit task used to study differential effects of lateral and medial OFC lesions in monkeys is replicated, as it suits very well the goal of understanding the differential contributions of both the subregions to decision-making. Further, a variant of 2-arm bandit task, a 2-stage Markov task is also implemented to study the combined behavior of lateral and medial OFC together.

Goal VII : Implement ACC to interact with both lateral and medial OFC

Goal VIII : Implement credit assignment in lateral OFC and core

Goal IX : Implement value-comparison system between core and medial OFC

Goal X : Implement minimalistic reward history in both lateral and medial OFC

It is crucial to identify appropriate strategy to validate such a comprehensive systems level framework, as the number of inter-linked subsystems involved requires a more detailed

analysis of the dynamics. Within the scope of the thesis, the analysis is restricted to behavioral performance of the model as a starting step. Even at a behavioral level, as can be seen, there are multiple behaviors that can be explained, with different activations to the system. Most importantly, although the review strikingly pointed out dissociate roles of lateral and medial OFC, since it has not been explored a lot, the interaction between both the sub-regions also need to be studied. Therefore the experiments will be chosen accordingly so that the effects of lateral and medial OFC on the performance would be prominent.

Chapter 6

The Model

Sommaire

How about this? Let's define a simple framework of virtual experimentation with an artificial agent, discuss several fundamental aspects involved in studying behavior, including the agent's *Brain*. Eventually, once it is established how the rest of the framework works, *Brain* will be then described as the collection of subsystems, few of which will be described in detail, with neurobiological evidences. Especially, one of the fundamental goals of this thesis, understanding the organization of information within OFC, will be discussed with the help of existing models of those subsystems and modelling few experimental observations involving OFC in animals (primates and rodents).

Plan for this chapter :

6.1 The Experimentation Framework

Studying the behavior of an autonomous intelligent system, animal or agent, involves studying the relations between brain and body of the animal/agent and the environment. Firstly in this chapter, before introducing the actual video game environment that is used for this work, a general experimental framework is described that is constituted by an environment, an embodied agent and the agent's continuous interactions with the

environment. The characteristic description of such a framework can provide basis for any specific software platforms to study the relation between the agent and the environment in a more controlled way than hardware implementations. Figure 6.1 shows a schematic of the framework and several key elements involved.

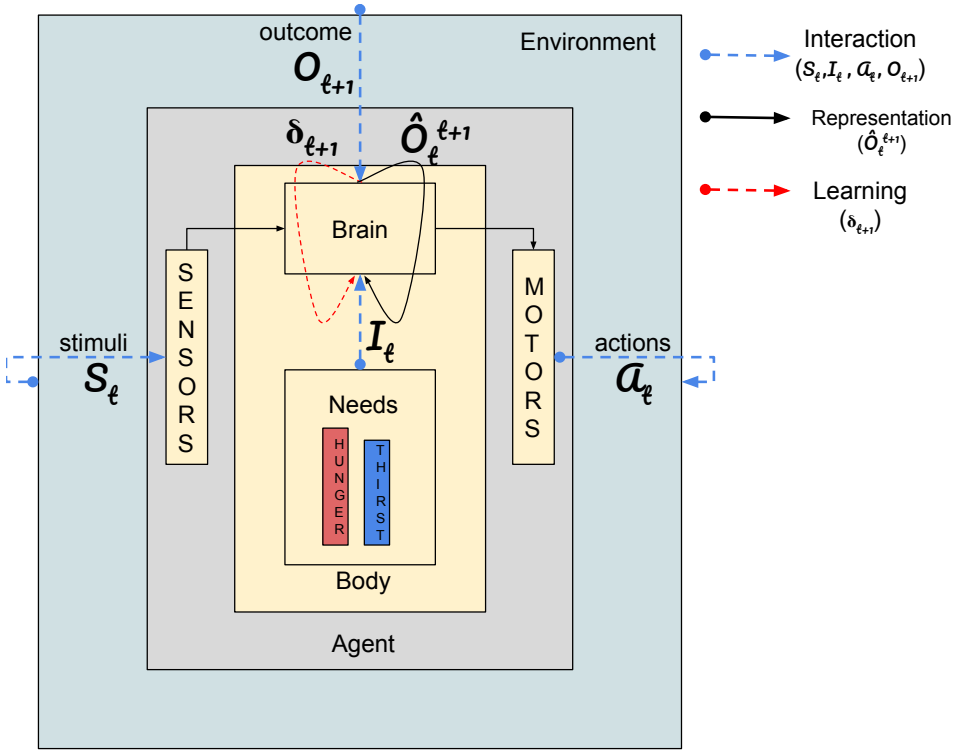


Figure 6.1: Behavioral Framework of an Embodied Agent. An agent is a part of an environment, besides some stimuli and outcomes (stimuli that directly affect the agent). At a given time `t`, the agent interacts with the environment (the stimuli S_t) and the interoception (bodily needs I_t), perceives the **representations** of stimuli (at time `t`) as well as the *expected* outcomes (at time `t+1`) $\hat{O}_t^{t+1}$, further interacts with the environment by eliciting *actions* (a_t). At time `t+1`, along side perceiving S_{t+1}, I_{t+1}, the agent also perceives the outcome O_{t+1} and incorporates learning , δ_{t+1} for future interactions.

Environment is a dynamically changing world that encompasses agents, stimuli, and the interactions of agents with stimuli. There could be a single agent or multiple agents, either independent or collaborating or competing. Stimulus is any object in the environment besides the agents. Stimuli can potentially affect the state of agents depending on their interaction with the stimuli.

Agent Each agent has a *body* and a *brain*. The *body* is characterized by a set of internal needs, which in turn shape motivation of the agent. In addition, an agent has *sensors* and *motors*. Sensors help the agent perceive the environment at any given time, generally by detecting certain features of the stimuli. Motors let the agent elicit an action in the environment, that would either alter the state of the agent or that of a stimulus in the environment. *Brain* facilitates the translation of the agent's perceived sensory state to a desired action. Such a translation could be a direct mapping between a sensory state and a motor action. Or it could be more sophisticated *processing*, like an intelligent animal/agent.

Stimuli (Fig. 6.1, S_t) are present in the environment alongside the agent carry certain relevance with respect to the agent and its bodily needs, either through the interactions or otherwise. Stimuli could be inherently *appetitive* - directly satisfy the needs of the agent, *aversive* - needs to be avoided for the overall wellness of the agent, *reinforcing* - being linked to different appetitive or aversive stimuli and lastly *neutral*.

Interactions The processes through which the agent's *brain* acquires the available information from the environment, transforms the desired state into observable effects in the environment are termed as *interactions*; they are of two major types, *Sensations* and *Actions* (Fig. 6.1, *blue* arrows) . *Sensations* are the information collected from the environment through the *sensors* in the agent's body. *Actions* are the means to manifest changes either externally in the agent's state or the environment through the activation of one or several *motors*, or internally within the agent's body, through bodily actions.

Representation Three kinds of representations are described in the framework (Fig. 6.1, *black* arrows) . *Perception* is the internal representation of acquired sensory information, in terms of which the subsystems of *brain* can process. *Evaluation* is understanding from the sensory representations if there is a choice to make and if yes,

with what motivation and how. And further, if a choice has to be made, evaluating subsequent results. *Action* is a high-level representation of the operation of single or a combination of *motors*.

Note on **stimuli** *:* Quite generally, the stimuli that are perceived to be a resulting state causally linked to previous sensation or action, are referred as *Outcomes* (Fig. 6.1, O_{t+1}) . Outcomes could be rewarding (appetitive) or punishing (aversive) or neutral.

Learning *Learning* is the process in which the outcome information (through subsequent *sensations*, often after an *action*, but not necessarily) is transformed into the knowledge that is used in all the subsequent future *evaluation*. Quite generally, learning is expressed as a function of difference (δ_{t+1}) between the expected ($\hat{O}_t^{t+1}$) and the actual outcome (O_{t+1}, Fig. 6.1). As a result of learning in the system, the expectation of an outcome gets updated. Such an updated expectation of the outcome is taken into account for the future evaluations.

Dynamics The frequency with which the entire system processes new information also affects the internal implementations. In the context of digital implementations, generally the system could operate either on a clock-based synchronisation information processed time unit that is predefined, or event-based information processed upon change of state of the environment (including the agent). In the figure 6.1, what is represented as `t` (*current* time cycle) and `t+1` (*subsequent* time cycle) could also be `n` (*current* event) and `n+1` (*subsequent* event). Figure 6.2 shows an example of a time cycle `t` in a clock-based processing. In a typical situation, the agent *perceives* the environment, *learns* about an outcome if one exists. Then *evaluates* the entire state (external and internal information) and *acts* if required. However, it is also possible within a time cycle, that there is nothing to perceive in the environment, hence nothing to learn, but still the agent could evaluate the internal state and act accordingly.

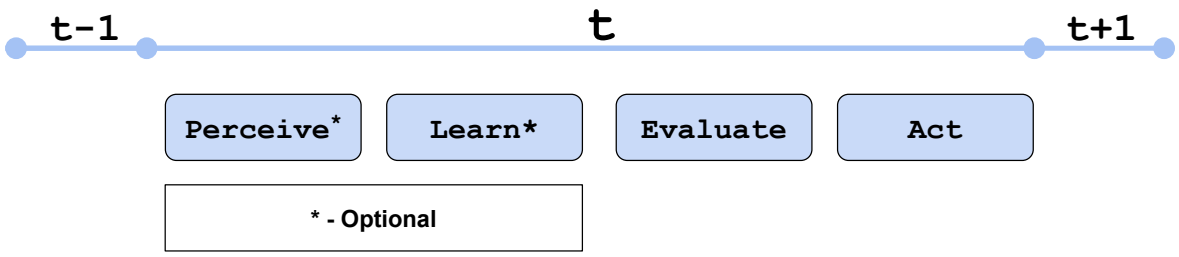


Figure 6.2: A typical clock cycle in a clock-based processing. The same applies to an event-driven processing as well, for an n^{th} event between the n-1^{th} and n+1^{th} event

6.2 Framework Implementation in a video game environment

In the context of the current modeling work, a well-known video game Minecraft is used as the environment with the help of a software platform Malmo. Minecraft allows virtual exploration, resource gathering, and survival task scenarios, for single or multiple players. It has been adapted before in our team, for a systemic neuroscience simulation platform (Denoyelle et al. 2016) called 'Virtual Enaction'. Malmo is a platform developed by Microsoft to interface with the video game Minecraft and to design, run and analyze an artificial agent in experiments related to computer science and artificial intelligence (Johnson et al. 2016). In the following sections, first the details of the implemented experimental framework will be presented in terms of aspects described earlier in this section. Malmo allows to programmatically construct a 3D world, providing access to all the resources of the video game environment Minecraft, often containing at least one agent, a number of stimuli that are either inherently neutral in their appetitive value (a block) or those which are appetitive in nature (an apple).

Importantly, there have been several adaptations that have been made on the top of what Malmo provides. It is either because a certain feature was not available (like distinct internal needs other than a common attribute *life*) or what Malmo provides is too explicit that it is not biologically realistic (like exact Cartesian coordinates). First *Environment*

and *Agent* are described in terms of what Malmo provides and later, *Adaptations* that have been done to the Malmo platform for the framework to account some biologically plausible implementations and later accommodate the neuro-computational models that will be discussed using this framework.

Table 6.1: The experimentation framework

Attribute		**Implementation**
Environment		Minecraft videogame
Agent		Player
Sensors	External	Vision
	Internal	Needs
Motors		Turn, Move

6.2.1 Environment

This is the world in which *Agent* (Sec.6.2.2) is free to move around and explore. One can procedurally place certain *Items* in the world, in the vicinity of the agent or elsewhere. *Items* can be attributed respective reward values that the agent is able to gain when it collects them. It is a simplified environment of the Minecraft world, designed to have a complete control on the external objects, and simple enough to understand the causal relationships with respect to the simulated agent behavior. It is 3 dimensional, allowing the items to be at a height above the ground and allowing the agent(s) to jump if necessary. The ground (*floor*) is defined in terms of blocks which have properties like texture, type and color. Such block properties like the color play the role of the environment *context*. In behavioral scenarios like fear learning or fear extinction, the *context* is a useful attribute because it adds an extra dimension to processing the stimulus information and attributes a preferential relevance to it (either from previous learning or from memory).

Stimuli

There are two kinds of stimuli used from what Malmo provides. *Blocks* are cubes of different material (and color) that can be configured by giving desired coordinates of the

diagonal of the cube and the type of material (which also often serves as different color). In this work, these blocks are used as cuboids of varying 'block' height and are referred as *Pillars*. For instance, a *Pillar* can be of height of two blocks or three. In addition, there is a list of *Items* that can be procedurally placed in the environment. When the *items* are in the configured vicinity of the agent, the positions and the orientations of the *items* are available for the agent. Each item can be configured with a certain scalar *reward* value at the beginning of the task. As a part of a task, the *reward* can be awarded to the agent, either for collecting the *item* or *discarding* it. There are several such items, from which we use *apple*, *bread*, *water_bucket* and *milk_bucket*. The distance within which the agent can *collect* the item can also be configured.

6.2.2 Agent

Malmo provides an agent, on which a full control can be exerted on its actions, either through the tool or through the model (of the agent's brain). The agent has specific sensors through which the *interactions* happen between the world and the *agent*. The state of the world is provided to these sensors that in turn feed the brain of the agent. In an abstract sense, the body of the agent also has 'sensors' of the level of internal needs (hunger and thirst) that are provided to the brain. The agent, after processing the 'state', decides on the action and executes it through the *motors* manifesting changes to itself and the world. Thus, more detailed representations of *Body* and *Brain* of the agent are adaptations on the top of what Malmo framework provides.

6.2.2.1 Sensors

Malmo framework provides the *state* of the world in terms of pixel information of the world as an image seen from agent's point of view. Additionally, the framework also provides information in a more symbolic format containing numerous aspects of the items present around the agent to the specified precision. For the sake of simplicity, the latter kind of information is used and provided to the agent's brain model. From external world,

the attributes that comprise the *current state* are : the agent's position and orientation in absolute coordinates with respect to the world, the items that are present in the chosen accessibility range around the agent, their positions and attributes. For the internal bodily state, the agent by default has an attribute *life* that is affected by the external world (e.g, when in contact with fire or attacked by other agents). The *life* stands as a crucial parameter in evaluating agent in survival kind of scenarios. It also has information about its own position and orientation with respect to the environment. Information about the item like its name, position and the reward it carries is also accessible to the agent. As explained earlier, *context* also is a part of the state, describing the type of the *floor* for a requested subset of blocks.

6.2.2.2 Motors

The agent has motors like *turn*, *move* and *jump*. Each motor can be controlled by a non-zero strength, zero being the motor at rest. For example, the motor *turn* with a strength ranging from -1 to 1, makes the agent turn clockwise at a fixed speed of turning normalized continuously between the strengths 0 and 1 whereas the agent turns anti-clockwise for the strengths -1 to 0. Any of the motors can be stopped by a strength 0. For example, a command `move 0.5` makes the agent move slower than the command `move 1`. Similarly `turn -0.5` turns the agent counter-clockwise direction slower than when the command `turn 1` makes the agent turn in the clockwise direction. Furthermore, Malmo provides commands to directly make the agent face a certain direction, `setYaw 45` instantly makes the agent face at an angle 45°. Similarly, there are commands available to 'teleport' the agent to a desired coordinate. `tpX 5` instantly moves the agent to $X = 5$ retaining the Y and Z coordinates of the agent position.

6.2.3 Adaptations

Malmo comes with numerous aspect that give a good control over the agent's behavior. However, in the context of building a cognitive architecture, and with a plan of integrating

computational neuronal models, several biological limitations as well as much desired constraints have to be taken into account for the experimentation framework. Therefore, several adaptations have been made as an extra software layer on the top of Malmo, so that the communication between the cognitive architecture of the agent's brain and Malmo is made in a biologically constrained way.

6.2.3.1 Body

In the framework, an agent has access to its vital variables like *life*, its current position and its current orientation with respect to the *World*. Along with these, the internal bodily needs are provided to the brain as a part of perception (as the information from the external *world*). Besides the framework, more internal needs have been implemented that play a role in the behaviour of the agent. Also, from a functional point of view, we adapted few aspects like *visibility* of the agent and the information about the *positions* (of items as well as the agent itself). These adaptations were important to add certain biologically plausible restrictions to the task.

6.2.3.2 Brain

The central part of this work, the model of OFC and the PFC-BG systems, represents the brain of the agent, where the information about the environment from the sensors and about the body is perceived, processed and the chosen actions are elicited through the *motors* to manifest changes in the environment as well as in the body (change of need levels). The model of each of the neural circuits discussed in this chapter are presented in greater detail in the later sections in an incremental fashion.

From the body of the agent : sensors of *hunger* and *thirst*. At any instant, the agent has information about its current levels of vital variables and how far they are from critical or fatal limits.

6.2.3.3 Needs

In the context of understanding decision-making and the effect of internal needs on the goal-directed behaviour of the agent, two vital variables have been implemented - *hunger* and *thirst*. Both these needs increase with time as well as with its efforts (meaning a *move* or *turn* action). Instead of a one dimensional *reward*, each *item* carries a value that is relevant to the *hunger* or the *thirst* level it would satisfy, and a value indicating the level of *preference* of the agent for this item. We use this internal need information in the scenarios where the need-based decisions are discussed. In the experiments discussed later, to analyze aspects like learning, only one kind of need, *hunger* will be referred.

6.2.3.4 Perception

The symbolic information of items from Malmo is transformed to corresponding representations in the *brain*. By design choice, the spatial map of the environment is not implemented in the model of the *brain*. Therefore, the sensor information about the location of agent and location of the relevant items are transformed into scalar quantities like the distance and orientation, which in turn are simplified into different zones of *visibility* and different positional arrangement (left, right or center). Instead of a one dimensional reward, each item carries a motivational index that is relevant to the *hunger* or the *thirst* level it would satisfy, and a level of *preference* with respect to the agent, expressing its emotional value. Also, from a functional point of view, there have been few adaptations in aspects like *visibility* of the agent and the information about the *positions* (of items as well as the agent itself). Few bodily, biological constraints are added to the agent. The information about the objects in the environment is also restricted depending on the distance from the agent within the field of vision. These adaptations were important to add certain biologically plausible restrictions to the task given that they would be modeled in detail as they don't fall in the scope of the goals of this thesis.

6.2.3.5 Visibility

Malmo provides information about the items all around the agent's vicinity of chosen range. However, the agent's 'Field Of Vision' (FOV) is restricted to a biologically plausible value (in this case, 150°), which is further divided into 3 different zones viz., *Appear*, *See* and *Reach*, depending on the distance from the agent.

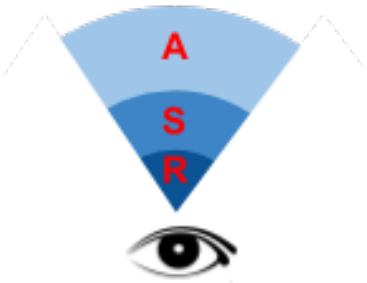


Figure 6.3: Zones of visibility in the field of agent's vision. Zone marked '**R**' is *Reach*, '**S**' is *See* and '**A**' is *Appear*

When the agent is moving and some *items* are present in the *Appear* zone, the agent has no precise information about the stimuli (the *items* that are perceived by the agent) such as the precise location of each, or their preference appetitive values. Rather, the agent has minimal information about the presence or absence of some *items* in some direction. When the stimuli are within the *See* zone, all the information about the stimuli is provided as inputs to the model. In the *Reach* zone, an additional information is provided, that the stimuli are accessible for the agent to *consume* (Figure 6.3).

6.2.3.6 Positions

Regarding the positions of the agent and the *items* in the environment, Malmo provides their exact coordinates, the absolute yaw details with respect to the environment and the agent. But this is not desired as the action execution after selection is an integral of the framework. To achieve this important feature within each loop of the model, using these exact position details has been actively avoided. Instead, the agent is taken as the origin, the relative distance and orientations of the items are converted into signals that regulate the activity within the loops of the model. It is usually these feedback signals relative to the desired state and the current state of the agent that sustain the execution of a selected goal. The desired and the current states not only correspond to the internal drives, for example, but also to positions and *items*.

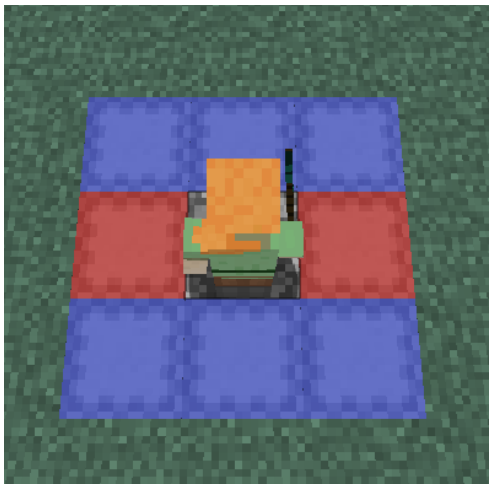
To begin with, the framework Malmo allows to define a grid of interest (GOI), in terms

of $\{x, y, z\}_{min}$ and $\{x, y, z\}_{max}$ coordinates, either absolute to the environment or relative to the agent. The elements, different block types for example, present in the coordinates in the environment that fall under the GOI, are returned every time cycle. When configured relative to the agent's position, $\{x, y, z\}_a$, $\{x, y, z\}_{GOI}$ is the super set of all the coordinates that are described in equation 6.1. For instance, for the kind of floor the agent is on as seen in the figure 6.4(b), if the GOI is configured as $\{x, y, z\}_{min} = \{-1, 0, -1\}$ and $\{x, y, z\}_{max} = \{1, 0, 1\}$ relative to the agent. According to eq.6.1, there are 9 total coordinates possible, an array of the identities of elements in these coordinates is returned, as shown in Fig. 6.4(a)

$$\{x, y, z\}_{GOI} = \begin{cases} \\ \cup\{(x_a + i, y_a + j, z_a + k)\}, \\ \\ \end{cases} \begin{array}{l} i \in \mathbb{N}, x_{min} < i < x_{max} \\ j \in \mathbb{N}, y_{min} < j < y_{max} \\ k \in \mathbb{N}, z_{min} < k < z_{max} \end{array} \tag{6.1}$$

Relative : min={-1, 0 , -1}, max={1, 0 , 1}		
blue	blue	blue
red	red	red
blue	blue	blue

(a) Grid data



(b) Environment

Figure 6.4: Example Observations from Grid in Malmo

However, the idea was to avoid exact information given by the platform but rather transform the information into a symbolic implementation that can be represented in a biologically plausible framework. From a robotics point of view, the agent has 4 visual sensors - *left*, *right*, *center* and *reach* - each of which detects information from left, right,

center and proximal positions respectively, about the *identity* of a pillar and the *strength* that depends on the range of the sensor . The list of pillar identities are assumed to be fixed for the experiments. That is the sensors are configured for four different types of pillars - blue, red, brown and white. When any of these pillars are present in the agent's GOI, they are considered as *salient* stimuli. Any other colored pillar present in the environment will not be considered salient by the agent's perception. So, at first the information from GOI obtained from Malmo is super-imposed on a fixed set of angles of each coordinate of the GOI with respect to the agent's current yaw. Then any block in the grid that has a relative yaw between $-\theta$ and θ ($2 * \theta$ being the FOV) is picked up the respective sensor, *left*, *right* or *center*. If the distance of the block from agent is 1 unit, irrespective of the relative yaw, the block is considered to be in the *Reach* zone of the agent and is picked up by the *reach* sensor. The range of θ in which the *left*, *right* and *center* sensors pick up respectively are given in the table 6.2. Also, each sensor picks up the information about a block with a certain *strength*. The *strength* is inversely proportional to the distance of the block from the agent. For example, the strength of a sensor for a block that is in *Reach*, *See* and *Appear* zone is 1, 0.75 and 0.5 respectively.

Sensor	Distance (d)			Coverage (θ)
	Reach	See	Appear	
Left	-	1-3	3-5	$-75° < \theta < -15°$
Right	-	1-3	3-5	$15° < \theta < 75°$
Center	-	1-3	3-5	$-15° < \theta < 15°$
Reach	1	-	-	$360°$

Table 6.2: Ranges of sensors

Figure 6.5 shows an example visibility zone for a chosen relative configuration of GOI, with $\{x, y, z\}_{min} = \{-5, 0, -5\}$ and $\{x, y, z\}_{max} = \{5, 0, 5\}$. The relative yaws with respect to the agent for each of the block in the GOI can be found in Fig. 6.5(a). For an example $\theta = 50°$ and the distance given in table 6.2, the respective zones of reach, See and Appear could be as highlighted in figure 6.5(b).

Reach See Appear

-45°	-38.6°	-30.1°	-21.8°	-11.3°	0°	11.3°	21.8°	30.1°	38.6°	45°
-51.3°	-45°	-36.8°	-26.5°	-14°	0°	14°	26.5°	36.8°	45°	51.3°
-59°	-53.1°	-45°	-33.7°	-18.4°	0°	18.4°	33.7°	45°	53.1°	59°
-68.2°	-63.4°	-56.3°	-45°	-26.5°	0°	26.5°	45°	56.3°	63.4°	68.2°
-78.7°	-76°	-71.6°	-63.4°	-45°	0°	45°	63.4°	71.6°	76°	78.7°
-90°	-90°	-90°	-90°	-90°	0°	90°	90°	90°	90°	90°

(a) Relative yaws

(b) Reach, See and Appear zones

Figure 6.5: Sensory processing from Malmo observations

6.2.3.7 Action execution

While executing an action, to verify whether the desired goal is achieved (or desired position is reached), instead of using the absolute information provided, the agent uses the sensory information at several instances during the execution, comparing it to the desired state, to realize if the goal is reached. Therefore, high-level actions are introduced like *left*, *right*, *ahead* and *consume* corresponding to the respective sensor. Instead of depending on the exact coordinates of the stimulus, sensor information about the stimulus is used. For example, a stimulus identified by the sensor *left* would fix the sign of the turn strength as minus $(-)$. In addition, the distance between the stimulus and the agent's current position is converted into an indirect signal as the linear velocity (eqn. 6.5) with which *move* command is executed. Similarly, the angle between the position of stimulus and the agent with respect to agent's current yaw, along with the sign, decides the angular velocity (eqn. 6.4) with which *turn* command is executed. Since both the commands can be executed simultaneously, it gives a smooth curved path from agent's current position to the position of stimulus, until both the strengths of *move* and *turn* (linear and angular velocity) approach 0. There is a constant slow increase in the internal needs of the agent with time. In addition, the needs also increase upon performing the actions (while performing *move*, *turn* or *jump*).

Assuming dz is the difference between the z-coordinates of target position and of the agent, and dx between the x-coordinates, the angular velocity (ω) for *turn* command and the linear velocity (υ) for *move* command are as follows :

$$\hat{\theta} = (\frac{arctan(dz,dx)}{\pi} * 180) - 90 \tag{6.2}$$

$$\Delta\theta = \hat{\theta} - \theta \tag{6.3}$$

$$\omega = \Delta\theta/180 \tag{6.4}$$

$$\upsilon = V * (1.0 - abs(\omega)) \tag{6.5}$$

6.2.4 Neuronal modeling

6.2.4.1 Processing

The dynamics of action selection according to the encountered sensory states is taken place during *processing*. For example, two *items* are found at time `t`, one in *see* zone and the other in *appear* zone, assuming both are of equal interest, a choice criteria could be choosing closest *item* and results in choosing the *item* in *see* zone. Similarly, *processing* can get more complicated as the relevance of items to the agent's current state becomes different and more dependent on multiple factors (relevance to current critical need, distance etc.). It is this aspect of the framework, where predominantly the computational models of brain circuits contribute to, that will be discussed in the following sections.

6.2.4.2 Learning

Figure 6.1 shows a red arrow on *Brain* to itself. The idea is that at any time instant `t`, an outcome that is observed upon action a_t contributes to the learning in the system. It is this learning that allows for an expectation of outcome at the next time step `t+1`, for the same set of sensory information and actions available as at `t`. It should be noted in Fig. 6.1 that, at time `t` depending on the sensory information received, S_t, there is an expectation of a possible outcome at time `t+1` ($\hat{O}_t^{t+1}$) which is considered as a part of the evaluation process that chooses an action a_t. Depending on the actual outcome

received at time `t+1` (O_{t+1}), the difference between O_{t+1} and $\hat{O}_t^{t+1}$ is often termed as reward prediction error (RPE) in computational theory of reinforcement learning (δ_{t+1} in fig.6.1). Multiple neural processes of learning will be discussed in sections further, when the computational modeling part of the *Brain* in the framework is discussed.

Most importantly, several aspects of learning have been selectively ignored or assumed to stay within the scope of cognitive aspects of decisions. For example, corresponding to the agent's internal needs of hunger and thirst, the amount of need that is satisfied by a particular outcome (reward, like *apple* for example) is assumed to be constant and built within the system. Similarly, when an action is chosen, for example approaching the stimulus on right, the precise motor planning to reach the stimulus in terms of shortest path, or the sense of direction to turn to have a minimum angle of turn etc.. are also assumed to be hardwired. Simply put, to reach a point on the right, the knowledge that *turn* strength should be positive so that the agent turns clockwise, is a hardwired information.

Lastly, a cognitive architecture is described in terms of existing behavioral and neuroscientific literature, based on which the agent's behavior is studied. This constitutes the *brain* of the agent. In the process of bridging both the descriptions, essentially a software platform is presented that interfaces together the videogame environment, the bodily attributes of the agent and the cognitive architecture of the agent's brain. The cognitive architecture of the *brain* will be derived from the description of the prefrontal cortex(PFC)-basal ganglia (BG) loops - *CBG loops* - and few other sub-cortical structures as described in the earlier chapters. Several demonstrations are shown by simulated scenarios in the Minecraft environment. Finally, as the position of the Orbitofrontal Cortex (OFC) is highlighted in the cognitive architecture, detailed considerations about the organization of information within the subdivisions of OFC are discussed. The model, involving OFC and the considered subdivisions, is compared against relevant experimental findings.

6.3 Behavioral vs Experimentation Framework

Behavioral	**Framework**	**Minecraft object (O) / command (C)**	**Details**
Conditioned stimuli (CS)	*Pillars*	O : Cuboids of *Blocks*	Different colors - blue, red, brown, white
Unconditioned Stimulus (US)	Reward	O : *items*	apple, bread, water, milk
Pavlovian Response	Consume	C : `jump`	when present next to the reward
Instrumental Action (IA)	Turn Right, Move Ahead and Turn Left	C : `move` and/or `turn`	If decision threshold is reached in the motor loops

6.4 An algorithmic model of parallel feedback-loops

As it has been highlighted in the previous chapters, a simplistic representation of segregated loops involving the prefrontal cortical, sensory cortical regions and basal ganglia structures will be demonstrated. The generic dynamics of information processing is similar in each of the loops. Each loop, at any given time, follows a flow of information processing - Evaluation, Selection and Execution. In this scope, primarily the dynamics within the loops for one time cycle will be discussed. The aspects of learning will be discussed in detail when concrete models within the loops are introduced in the later sections. The idea is to describe the workflow within the framework when a goal is being pursued. Figure 5.1 in the previous chapter described an example of each loop. Four dif-

ferent loops have been implemented based on the generic description. To help with clear description, the task is chosen to be simple. There are four possible appetitive objects for the agent to choose from corresponding to two of the agent's needs.

In that context, the two limbic loops can be defined as follows. The *Why* loop is responsible for the selection of the need. It receives sensory information about the levels of need through interoception and about the kinds of objects perceived by exteroception. Responses it can trigger correspond to the decision to go for food or drink, until the need is satisfied (by consuming upon reaching). The *What* loop is responsible for the selection of an object. It receives sensory information about the levels of preference through interoception and about the identity of the objects perceived by exteroception. Responses it can trigger correspond to the decision to select one object until the object is reached.

Similarly, the two sensorimotor loops can be defined, still using the same framework, as follows. The *Where* loop is responsible for the orientation of the agent in space. It receives the yaw of each object perceived by exteroception and when one is selected, triggers a movement of orientation which stops when the agent is facing the object. The *How* loop is responsible for the reaching of an object. It receives the distance to each object it is facing by exteroception and when one is selected, it moves forward until the object is reached.

Firstly, sensory cues corresponding to actual or desired sensations activate candidate actions in the frontal area. A primitive strategy is to trigger the action most often associated to these sensations. This corresponds to *habits*. Else, a selection process takes place to make a decision based on a deeper contextual analysis. This is attributed as one of the major roles of the basal ganglia. When an action is triggered, its expected sensory consequences are also activated to a specific desired level, representing the goal of the action. The action will be maintained until the expected sensory consequences (or other conditions for interruption, not developed here) are met. In some cases, triggering the action is not sufficient to reach the goal (e.g. deciding to eat is not sufficient to get some food) and the desired activity can itself trigger new actions in other loops (e.g. finding

some food). This process can recursively trigger other secondary goals, until some goal is immediately achieved, stopping the corresponding action.

From a more practical point of view, in each loop, the processing happens in the following stages, in a given small time interval - *information acquisition*, *action evaluation and selection* and *sustained activation by feedback control*. For each loop *(i)* acquire sensory information through exteroception and interoception, *(ii)* evaluate alternative responses, select the most appropriate one and set the corresponding goal, and finally *(iii)* sustain the activation of the response by a constant feedback until the goal has been achieved.

This generic mechanism of response maintenance and goal monitoring is an important aspect of the computational model, implemented in each loop. Selecting a response to be executed means defining a sensory state as desired and that must be achieved (the goal). Each of the sensory modules in the loops, namely *Sensory* cortex, *Insular* cortex and the respective *Parietal* Cortices, have two populations each - *desired* and *actual*. The *desired* populations have excitatory and inhibitory connections from the prefrontal regions with a threshold function as the activation function. That is, the 'desired' activation of the option *red* depends not just on the activity of the corresponding population in the frontal population (in this case, OFC), but rather on the difference of the activity of *red* population and that of the populations of rest of the options. And this difference has to be greater than a threshold implying that decision has been reached for the option to be *desired*. So for each population corresponding to option j in the Sensory module, the synaptic input $I^S_{syn,j}$ is a weighted sum of each of the ongoing activities in the population F, U^F_i with the weights w_{ij} being excitatory or inhibitory depending on whether $i = j$ or otherwise respectively (see eq.6.6, N number of populations of known stimuli).

In addition, there is a direct excitatory input from the Insular cortex to all Sensory modules (eq.6.7, n number of internal needs in the system, N number of populations of known stimuli each mapped to a relevant need), as per the simplification in the model that (anterior) Insular cortex has access to the body's current active needs and can transmit the information to all the loops. In reality, even if this might be true for at least the loop it forms with ACC, it is possible that this information is transferred to other secondary

$$I^S_{syn,j} = \Sigma_i w_{ij} * U^F_i, (i,j)\epsilon N \tag{6.6}$$

$$I^S_{ext,j} = \Sigma_i I^{Ins}_i, i\epsilon n, j\epsilon N \tag{6.7}$$

$$U^j_S = fn(V^j_S + \delta) \tag{6.8}$$

Equations : Activities of the *desired* populations of a sensory module in each loop.

sensory areas by other mechanisms. The idea is that the interoceptive information, having internal goals inform the secondary sensory areas for focused attention relevant to that goal. As a consequence, the rule of response execution is implemented as a sustained activation of the response which terminates, thanks to a feedback mechanism, when the goal is met. As it is elaborated in the section *Scenarios*, the goal is not always reached simply by activating the response, but sometimes requires other responses and secondary goals to be defined, still within the same generic mechanism. To implement this, we define a *desired* state of activation for goals that asks for additional responses until it becomes *actual*.

1. The *Why* loop selects the current motivation (satisfying hunger or thirst in our task) from the interoception of needs and possibly the costs of actions. The motivation is expressed in the anterior cingulate cortex (ACC) and the loop also associates the ventral striatum (the core of the nucleus accumbens), lateral hypothalamus and insula for interoception.

2. The *What* loop selects the goal according to the preferences (e.g., gustative preferences, quantity), innate or acquired and represented in the amygdala. Preferences are expressed in the orbitofrontal cortex (OFC) and the loop also combines the ventral striatum (the shell of the nucleus accumbens), amygdala and insula for gustative interoception. The goal object can be consumed if it is directly available, otherwise it would become the goal for the spatial and temporal organization of the

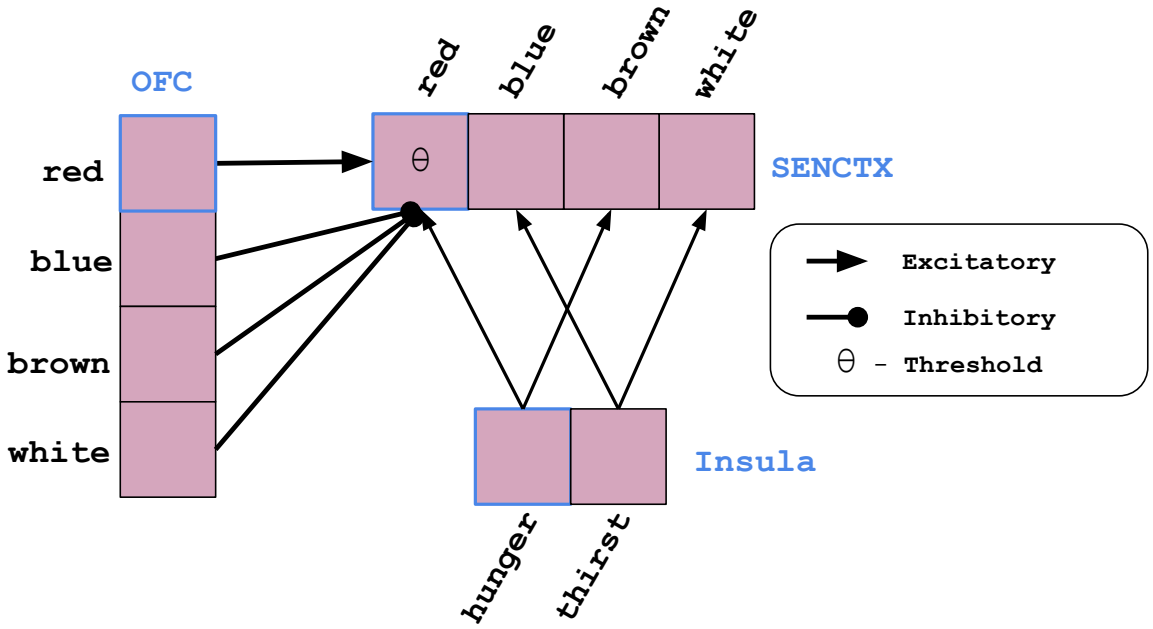


Figure 6.6: Example connections to a Sensory module from respective frontal module and Insula. For simplicity, connections are shown with respect one population ***j*** within SENCTX that corresponds to the stimulus 'red'.

behavior.

3. The *Where* loop considers the spatial location of the goal and selects the orientation behavior relevant to face it, which can concern eye movement as well as body orientation, as also observed in the superior colliculus. The orientation strategy is expressed in the Frontal Eye Field (FEF) in the frontal cortex and the loop also combines the dorsolateral striatum, the parietal cortex and the superior colliculus.

4. The *How* loop supports the latest postural adjustments when the goal is attainable, by simply reducing the distance or possibly manipulating the object before consuming it. This concerns the motor areas, the parietal cortex and the dorsolateral striatum.

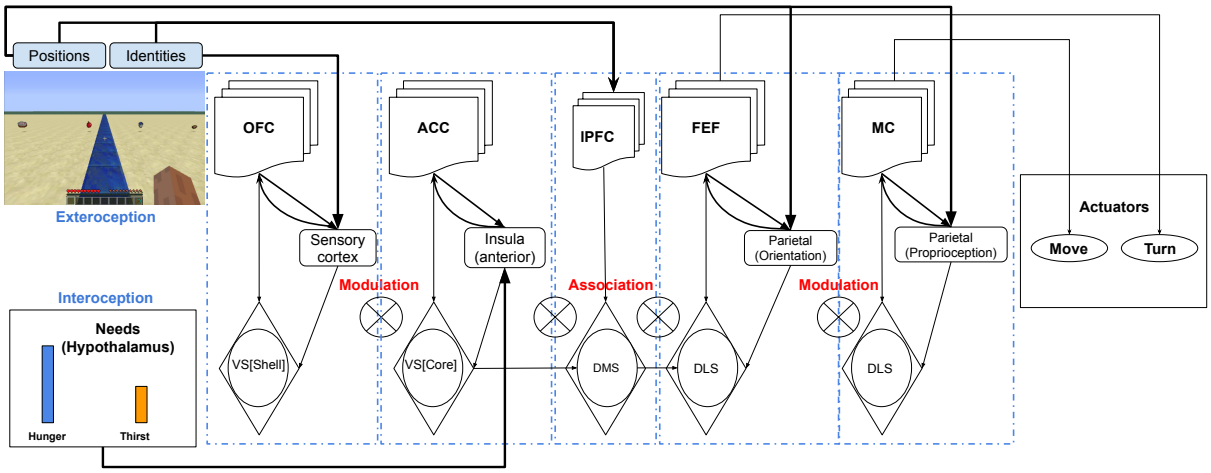


Figure 6.7: An algorithmic model of CBG loops with exteroception and interoception.

6.4.1 Exploratory behavior : default state of the agent

Exploration, in terms of the agent's motors, essentially is a combination of executing one of the action commands like *ahead*, *left* or *right*. Besides the internal needs of the agent, *hunger* and *thirst* that will be described as a part of the framework, another implicit need *curiosity* is implied that drives the agent's exploration behavior. Theoretically, the need *curiosity* can correspond to internally arising curiosity or even foraging kind of behavior, to obtain outcomes that can serve for later. In the context of this framework, it is simply a need that is active when there is no other active need like *hunger* or *thirst*. And when *curiosity* is active, it corresponds mostly to the motor movement *ahead*, and once in a while, random choice between *left* and *right*. This guarantees that the agent is constantly moving in the environment, exploring and encountering stimuli. In any case, once this kind of exploration kicks in, there is an inevitable increase in either of *hunger* or *thirst* over time, and then the behavior of exploration will be based on that active need. Therefore, *curiosity* is just an abstract need introduced to keep the agent exploring in the environment. Since during normal exploration, there is no specific position that the agent desires to be, the desired position is set as a position at a random distance (long) ahead of the agent or sometimes changed to a position either to the left or right of the agent at

a random yaw. Thus, depending on the situation, the desired position can be activated differently directing the agent where to explore.

During exploration, depending on certain criteria (see Fig. 6.8), the agent moves into *Decide* phase of the behavior where the *Evaluation* of the options takes place. Upon successful decision, the agent enters a *Pursue* phase, where the chosen action in the motor loops will be executed. It is to be noted that, in the *Pursue* phase, the winning action in the motor loops is executed, but the fact that it corresponds to the winning option in the limbic loops is guaranteed by the dynamics between the loops. the implicit hierarchy that arises due to the influence of limbic loops on the motor loops through the associative loop. However, there is a time limit for the *Decide* and *Pursue* to arrive at a decision (T_d) and complete the action (T_p) respectively. Arriving at a decision could involve decision criteria, for instance activity thresholds, for a valid decision. Completing the action means reaching the stimulus at the chosen position. Failing to achieve any of the two, the agent moves to *Give Up* phase where the current evaluation will be appropriately marked as a failure and continues to *Exploration*. Minimal memory is implemented to avoid revisiting the already visited stimuli.

The transition from *Explore* to *Decide* and the dynamics of the loops for *Evaluation* during the *Decide* phase depends on the state of the network of loops at that instant. This can be viewed differently for the descriptions of each of the cases of Stimulus-Driven (SD) and Goal-Directed (GD) behaviors (the following section).

- Explore : $\forall(need) > 0$
- Decide : $\forall(stimulus)$ in (See) zone of visibility
- Pursue : $T_d < T_g$; Time spent until reaching the decision criteria T_d, Threshold time to give up current phase, T_g.
- Giveup : $\forall(T_d, T_p) > T_g$; Time spent in *Pursue*, T_p

This basic behavior also forms an interesting basis to learn or update the contingencies in the environment or between characteristics of the environment and those of the agent.

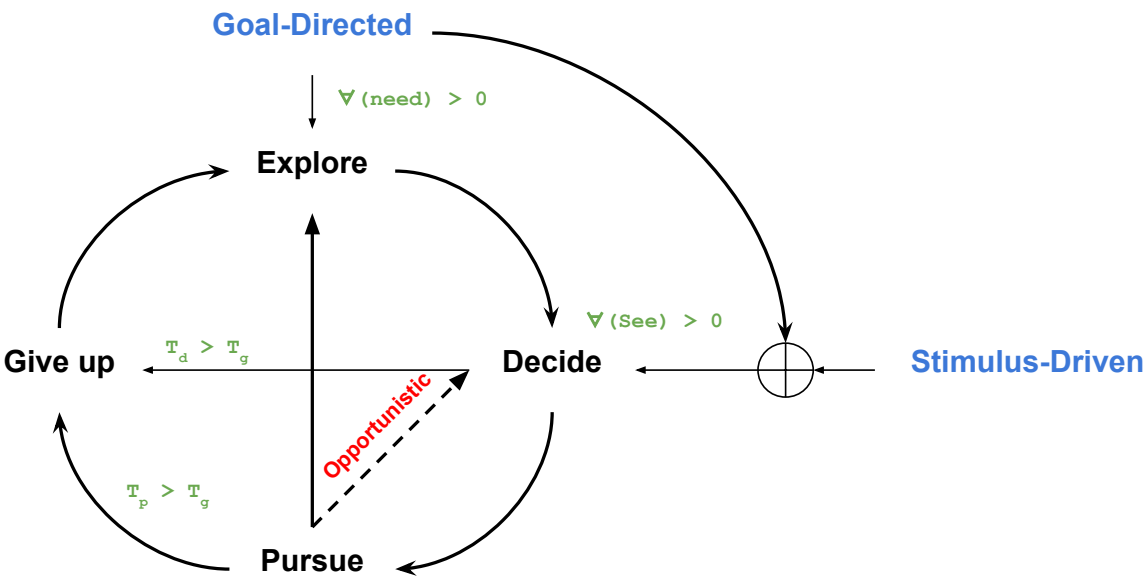


Figure 6.8: The sequence of phases across the agent's behavior in the environment.

Particularly, in the limbic loops, this can contribute to set the values of the preferences and help connect some items to the needs they can satisfy. In the sensorimotor loops, this can help calibrate the movements of the agent and learn the consequences (in terms of modification in the perception) of their activation.

Constraints

There are several constraints that the framework is subjected to in order to contain the dynamics of the system. Although, there is no explicitly imposed hierarchy in the system, certain constraints give rise to an implicit hierarchy. Most important of them all, is that the homeostasis of the internal needs of the agent is the central motivation. Thus the biases will be controlled between the systems such that when there is a pressing need, there is a higher bias on the feature-based processing systems and when there is no pressing need, system is free to explore more or make risky or uncertain choices.

1. Agent, while exploring, stops if more than one objects are in sight (according to the Visibility limits), to decide.

2. While deciding, no action will be performed unless the motor decisions reach a threshold

3. The modulation between the What? and Why? loop is controlled such that the biases contain the activities of the system within acceptable limits (Fig 6.9)

6.4.2 Implementation of Stimulus Driven Behavior (SD)

The behavior can be described as purely stimulus-driven when the agent is *exploring* with no pre-activated internal motivation, that is with no stimulus as desired. When there is a stimulus in the visibility zone of the agent in the range *Appear*, the desired position as the part of *Explore* phase is set in the direction of the stimulus, so that the agent can move forward in that direction and can get more details of the stimulus once it comes under the *See* range. If there are multiple stimuli in the *Appear* range and *See* range together, the direction of the desired position of the *Explore* phase is set to be the mean yaw of the stimuli. Subsequently, when there is one or more stimuli in the *See* range, the agent is explicitly brought to stop so that the *Decide* phase begins. During the *evaluation* of the *Decide* phase, the OFC loop processes the stimuli considering any previous learning about them and their value and any other momentary information that needs to be taken into account. Although there was no pre-activation in the ACC loop by an internal need, the loop still processes the choice taking into account the action costs involved in each of the choices. In a simplistic scenario, it is possible that the bias by the ACC loop on the OFC loop is quite minimum as there was no preactivated need nor the action costs would be so different as the stimuli are both in the *See* range. In that case, the decision in the OFC loop could bias the ACC loop, and in turn that would influence the motor loops, through the associative loop.

6.4.3 Goal-Directed Stimulus-Driven Behavior (GD-SD):

The Goal-Directed Stimulus-Driven Behavior similar to pure SD behavior except that during the *Exploration*, the agent has one or more of the internal needs identified as

critical and the stimuli known to be associated to these needs pre-activated as desired. This can influence the behavior in different ways. One way is the behavior remains same as SD behavior until the *Decide* phase and then the existing pre-activations of the stimuli relevant to the current need gets selected in the ACC loop, thus driving the decision also in the OFC loop and subsequently in the motor loops. Depending on how strong the motivation is and how fatal the level of need is, multiple constraints can be added to the framework. For example, in the *Explore* phase, if there are two stimuli ***A*** and ***B***, one in *See* range and the other in *Appear* respectively, and if stimulus ***A*** corresponds to the currently critical need, instead of exploring to get closer to B and make a decision between ***A*** and ***B***, *Decide* phase can be triggered directly to choose ***A*** as it would satisfy the critical need.

6.4.4 Modulation and Hierarchy in the Loops

In addition, this selection made locally in a specific loop is also modulated by the selection made in a different loop. For example, the activity strength of selection in *What* loop modulates the activities competing for selection in the *Why* loop. Similarly, the *Why* loop modulates the *Where* loop which in turn modulates the *How* loop. To keep the modulation simple and tractable, a simple biasing factor b_{ij} is implemented from unit i in one loop to the corresponding population j in another loop, as a function of the source activity a_i as shown in equation 6.9.

$$b_{ij} = a_i^{\tan \sigma_{IJ} \frac{\pi}{4}} \tag{6.9}$$

In equation 6.9, σ_{IJ} is a bias strength parameter that is specific between two populations I and J. For instance, the bias that the preference-based choice might have on the need-based choice could be less than the bias that the overall limbic choice has on that of any of the sensorimotor loops. Figure 6.9 shows the resultant bias as a result of population activities for different values of σ_{IJ}. This kind of interactions between different cortico-basal ganglial loops in animals (including primates) is a question of wide interest in the

field of neuroscience (Haber 2003). In the case of survival tasks like the one we demonstrate, we can particularly wonder how to model the functional interaction between these loops and if different forms of survival strategies (e. g. goal-driven or stimulus-driven) can be performed on this basis. The latter question forms one of the open problems in computational neuroscience (Daw et al. 2005), thus motivating our digital experiments.

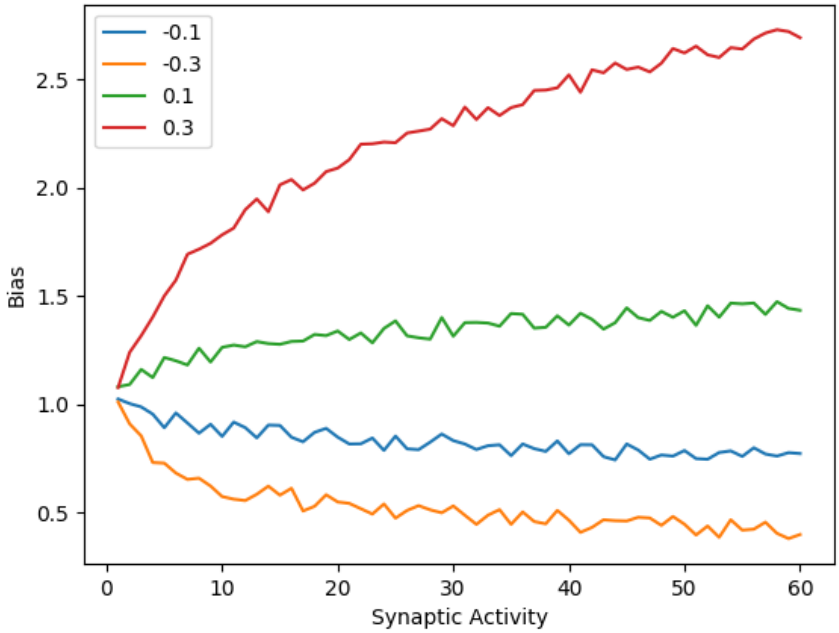


Figure 6.9: Evolution of effective output bias as a result of synaptic activity, for various bias strength σ

Hierarchy

There is no explicit hierarchy in the way loops are configured. Respective information is provided to each loop simultaneously upon sensory processing of the options : the information about identities to the limbic loops (OFC and ACC), the information about positions to the motor loops (FEF and MC), the combined identity-vs-position information to the associative loop (lPFC). However, as a result of the constraints the framework is subjected to, several implicit hierarchies can arise in the dynamics of the loops. For ex-

ample, in the beginning of exploration, decision within limbic loops could be independent of the decision in the motor loops. However, upon possible learning in the limbic loops that would preactivate the populations, the decision could occur faster than that in the motor loops and thus bias them through the associative loop. Similarly, hierarchy can be induced by choosing biasing strengths between the loops appropriately. For instance, as it was established that the central goal of the agent would be to maintain homeostasis of bodily needs, a fatal level of a need in the ACC loop could strongly bias the OFC loop to choose an option related to that need. In addition, as long as there is no strong influence of one loop on the other, the framework can also trigger an exploration/exploitation strategy (Humphries et al. 2012) and at any moment interrupt the current behavior to explore. The exploration behavior is also particularly important at the beginning of the task.

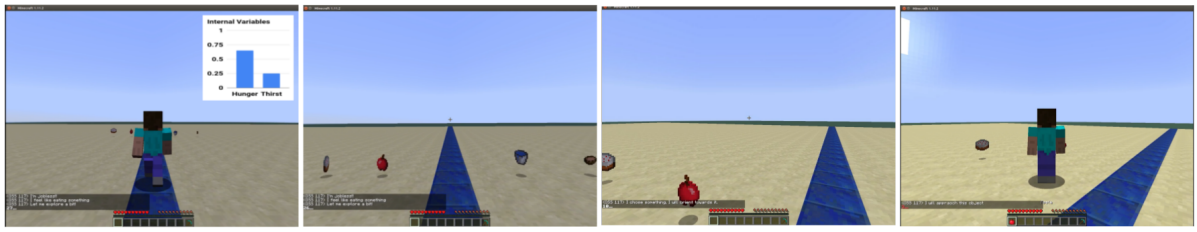
6.5 Action Execution by Sustained Sensory Activation

The GD-SD behavior is driven by a desired sensory state followed by a stimulus driven choice. Figure 6.10 shows several moments from an episode in the task implemented in Malmo, primarily concerned with different questions each loop in the model addresses. In a basic scenario, the agent starts exploring the environment (figure 6.10.a) with a *desired* activation for a particular item that (known from previous experience) would satisfy the current major *need* (figure 6.10.a, inset). This is a result of the internal state processing in the *Why* loop. When the agent perceives multiple stimuli (figure 6.10.b), along with the appetitive relevance of each of the stimuli, the action costs, depending on their positions, are also provided (implemented as a negative signal from the sensory-motor loops). Furthermore, the choice on the pre-existing preferences towards the stimuli corresponding to the selected *need* is made in the *What* loop and it modulates the selection in the *Why* loop.

Once the decision has been made in the *Why* loop and the goal has been set, the execution of goal involves two steps in the two sensorimotor loops. Once a stimulus is

chosen, the goal is to orient towards it. The *Where* loop is responsible for the agent to start *turning* towards it (see *apple* in figure 6.10.c) until the selected stimulus is in the sight of the agent. And finally, owing to the processing in the *How* loop, the agent moves to reach the stimulus that it has oriented towards and *consumes* it (an imaginary action which we equate to the agent reaching the *item* and updates the corresponding *hunger* and *thirst* values).

Sustaining the selection of goal until it is achieved is at the core of the processing in each loop. The *Where* loop, after choosing the orientation to turn, sustains the activity until the object is in sight. And the *How* loop sustains its activity from the point of orienting to the point where it has reached closer to the stimulus, to be able to *consume* it. This would now cascade back to the *limbic* loops which have been sustaining their selected responses. The *Why* loop, which has been active since selecting the current need, is sustained until the need levels are modified by the consumption of the stimulus. Similarly, the *What* loop, which has been sustaining activity since selecting a preferred stimulus, continues until verifying the consumed stimulus has the expected value.



(a) Why? (b) What? (c) Where? (d) How?

Figure 6.10: Snapshots at different stages in the task. (a) Internal needs monitored in the *Why* loop (inset). (b) Processing information about the stimuli in the *What* loop. (c) Orienting towards the selected stimulus using the *Where* loop. (d) Reaching the selected stimulus using the *How* loop, once oriented. *Note:* The different third person and first person views in (a) and (b) are only chosen to show the change in proximity of stimuli to the agent, but these views have no effect on the task. They can be switched while watching the task. Similarly, for (c) and (d)

Figure 6.10 shows a sequence of snapshots from a GD-SD behavior, as implemented in Malmo. In this basic scenario, *hunger* is the most urgent need the agent has (in figure

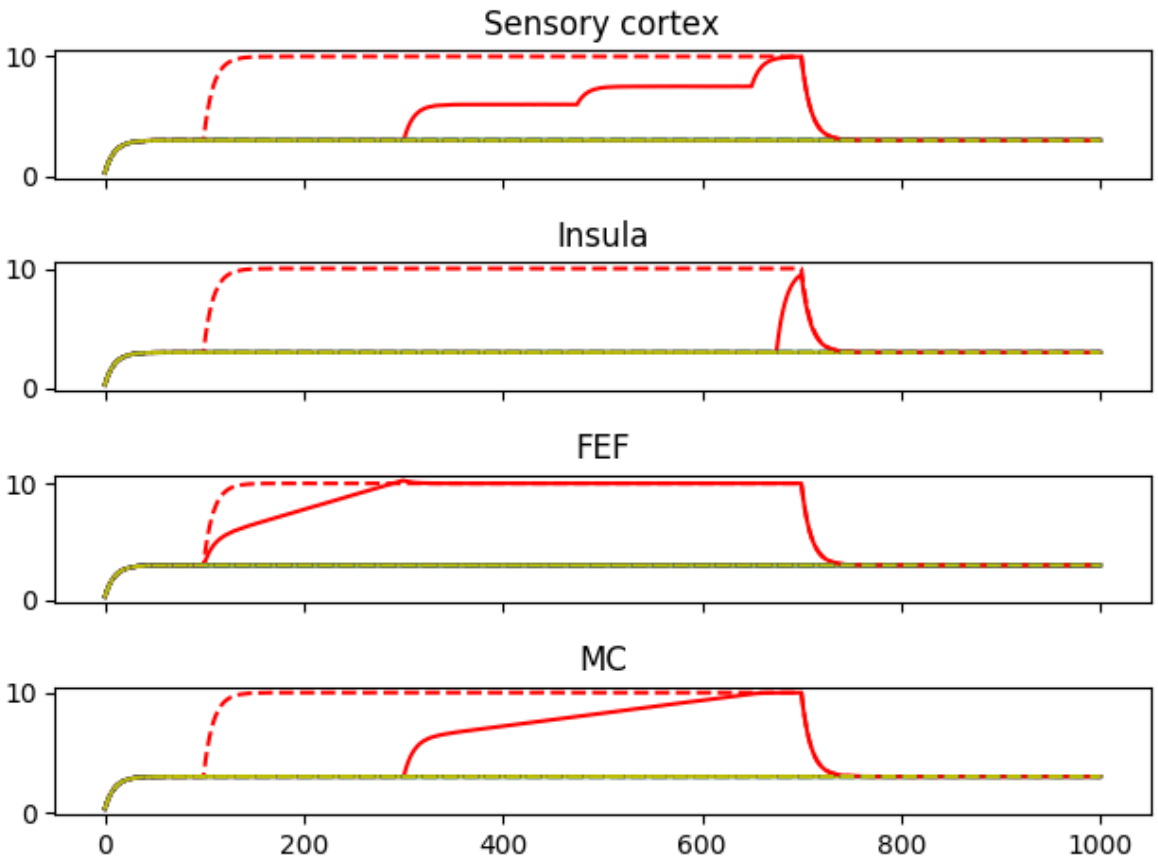


Figure 6.11: An example demonstration of the activities in the *desired* populations of sensory cortices are sustained until the corresponding goal in each loop is reached.

6.10(b) inset). The satisfaction of this need now becomes a *desired* state as a primary goal, which remains active in the sensory part of *Why* loop (Insula, as referred in the framework) until the need is satisfied. The *Why* loop triggers the *desired* state of the stimuli known to satisfy the need (*apple* and *cake*), in the *What* loop. Figure 6.10(a) illustrates this exploration behavior, where the agent starts to move with a *desired* activation for *apple* and *cake*.

When encountered with multiple desired *items*, the agent has to *decide* the suitable choice among the *items*. Supposing from previous learning, *What* loop has higher preference for *apple* over *cake* (as illustrated in figure 6.10(b)), once the decision has been

made, only the selected stimulus now remains to be *desired* in the sensory part of *What* loop (Sensory cortex, as referred in the framework) and remains active until it is reached. The decision in *What* and *Why* loops subsequently drive the decision in the motor loops, choosing the action to reach *apple*.

Once the *item* is chosen, *apple* becomes the desired 'line of sight' population in the sensory part of the *Where* loop (Parietal Orientation, as referred in the framework) and desired object in reach zone in the *How* loop (Parietal Proprioception as referred in the framework) also becomes *apple*. The agent starts *turning* towards *apple*, deriving a feedback signal to the *Where* loop to sustain the act of *turning* until the agent is oriented towards *apple*, as illustrated in figure 6.10(c)&(d). When *apple* is in line of sight of the agent, the desired goal of the *Where* loop is achieved. Then, the agent can move towards the target to reach it. And once reached, the desired goal in the *What* loop is achieved. In our current implementation, the internal action of consumption is automatically triggered when an item is reached. In this case, the goal in the *What* loop (sustained from the initial selection of the goal) is considered achieved. This consumption will also modify the level of need and similarly, the goal in the *Why* loop is also considered satisfied. This terminates the behavior as this was the primary goal of the scenario illustrated.

Before moving towards detailed discussions on the limbic loops, specifically the loop involving the OFC, a single CBG loop for motor action selection is described. As most of the network architecture and the parameters of population dynamics would be similar across different loops, the motor CBG loop is described in detail using a simple action selection scenario. Further, the additional loops are discussed in more complex tasks, in an incremental manner describing the organization of multiple loops together.

6.6 [Preview] A computational model of distinct OFC subregions among frontal regions and BG structures

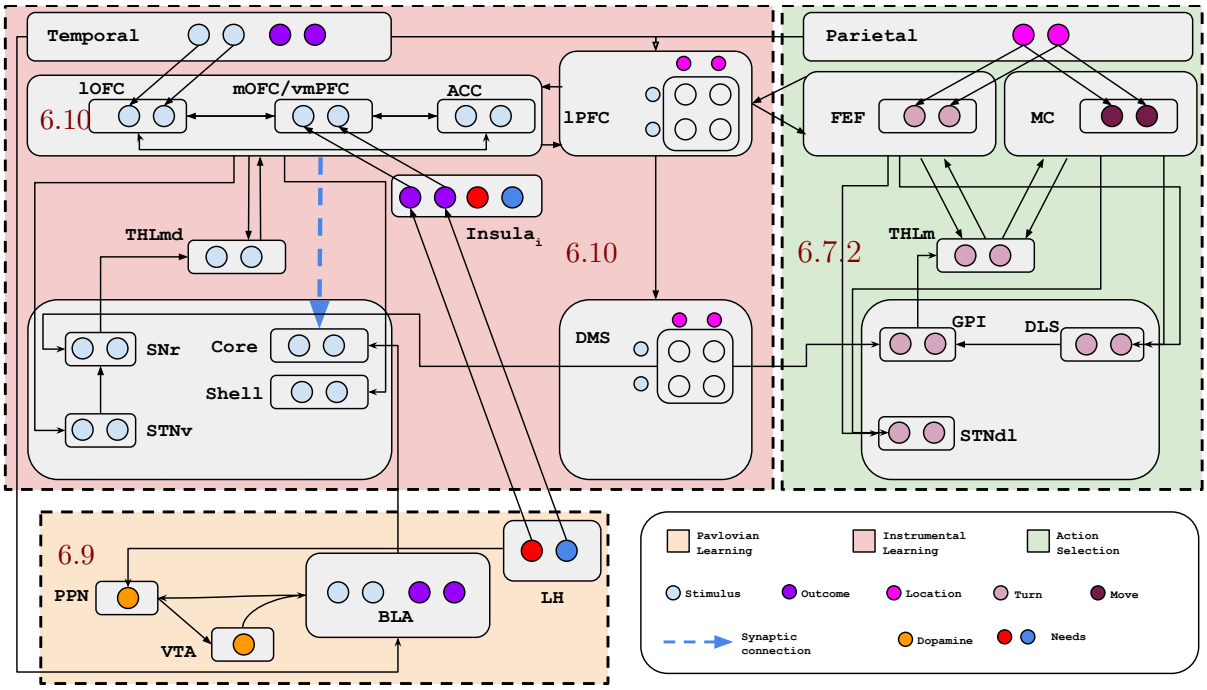


Figure 6.12: A schematic of the model of CBG loops, dissociating lateral and medial OFC

Figure 6.12 is a layout of the computational model of the role of distinct subregions of OFC, lateral (lOFC) and medial (mOFC/vmPFC), that has been implemented as part of this thesis . The model is from the point of view of the OFC being a crucial part of key interactions between the Pavlovian systems (with BLA and Shell of NAcc), Instrumental systems (Core of NAcc) and the Prefrontal regions and Frontal-Basal Ganglia (CBG) loops. Before describing the primary features of the model regarding the lateral and medial subdivisions of OFC, the detailed implementations of rest of the subsystems involved in the model are described in relation to OFC in general. At a high-level, the model can be discussed in 3 parts (color-coded in the figure 6.12). They are as follows :

1. A detailed implementation of the dynamics within a CBG loop, especially the sensori-motor loop. This will be discussed with the help of a simple motor task

of choosing between two equivalent actions with little or no effect of object features on the choice. The implementation details described for this loop are similar to the other CBG loops which will follow, particularly limbic loops.

2. A detailed implementation of parallel CBG loops, one limbic, one associative and one sensori-motor loop. This will be discussed with the help of a probabilistic reward-based decision making task. The details of implementation at single loop level remain similar to those of the sensori-motor loop and other specific features of the combined parallel loops that drive decision-making and learning will be described.

3. A computational model of Pavlovian conditioning - learning stimulus-outcome association within the basolateral amygdala (BLA) and related structures. This model will be discussed with a fairly simple task where there is no action to perform, but the agent repeatedly is exposed to a stimulus and paired together with an outcome.

6.7 A computational model of a single CBG loop (motor loop)

Imagine an agent exploring in an environment. For the sake of simplicity, let's assume the agent moves in a straight line. In an otherwise empty world, every once in a while, the agent comes across a point, where two salient stimuli are visible, one on either side of the agent, on left and right. To avoid complex object recognition, both the stimuli are identical in their appearance (both blue blocks). Now, in the first experience, to choose an action between right and left, the agent has no criteria to bias either of the actions. If these actions are represented in an action selection mechanism that can make a choice, there is no factor that could help resolve the competition between both the options since both the actions activate the selection mechanism equally. It is assumed that the objects that are on top of the three blue blocks are not visible to the agent, unless the agent approaches the blocks and jumps.

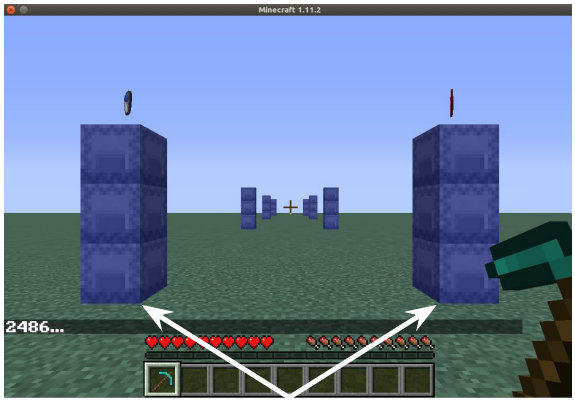


Figure 6.13: A simple choice of actions : left or right, when all other factors related to the actions are identical; same visual stimuli, equal distance (meaning equal cost of action).

However, as mentioned earlier, the key interest lies in studying limbic loops. In addition, the precise dynamics within the motor loops during action selection and execution is a wide field of interest in neuroscience. Various detailed representations within the motor cortex, the role of dopamine in refined execution of actions are some of the open questions in the field. Therefore, in studying several scenarios (sections 6.7 or 6.8), we merge the both Where? How? loops to be represented by the parietal cortex (PC) representing the actions. And a represented position can be reached using the available information from the sensors (see ***Sensors*** in section 6.2.2) and by transforming the exact coordinates of the agent and the objects into signals that derive linear and angular velocity of the agent (eqns. 6.5 and 6.4). The primary advantage of this transformation is that the action is driven by the motor output of the system and when desired, more detailed models of execution of goals, tracking progress and sustained activity during the execution could be studied.

6.7.1 Network

Following the architectural design described above, similar to what was implemented previously by (Leblois 2006) for the resolution of competition in BG, a biologically plausible neural network of representing populations has been implemented. A cortical layer, in this case, to represent Parietal Cortex (PC) is added to the input of the BG subsystems, dorsolateral striatum (DLS). The loop is completed as a feedback loop via the corresponding thalamus nucleus through GPi, the output of BG. Figure 6.14 shows a schematic of both direct and hyperdirect pathways. Each structure (PC, DLS etc.,) is a collection of

neuron populations where each population represents one particular option. In this case, each neuron population in a structure represents one of four possible actions - moving *left*, *right* or *ahead* and *jump*. The direct pathway is all One-To-One connections between structures, meaning one population in a structure is connected to the corresponding population the target structure (fig. 6.14, top, red arrows). These connections could be either excitatory (PC->DLS, THLm->PC, arrows with sharp ends) or inhibitory (DLS->GPI, GPI->THLm, arrows with round ends). The hyperdirect pathway, shown below in the figure 6.14, runs through the dorsolateral STN (STNdl), where the connections diverge to all the representing populations in GPI from each population in STNdl. Thus hyperdirect pathway is formed by the excitatory connections PC->STNdl, STNdl->GPI and THLm->PC and the inhibitory connections GPI->THLm. The diverging connections from STNdl reach all the populations in GPI (blue sharp arrows in the figure 6.14, bottom).

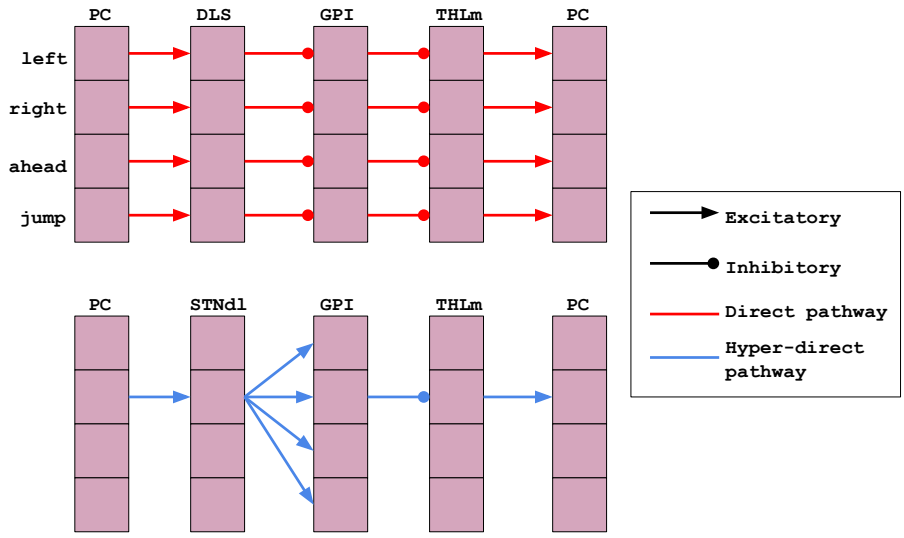


Figure 6.14: Detailed connectivity of an example CBG Loop, involving Parietal Cortex (PC) and Dorsolateral Striatum (DLS). Above : Direct pathway, through PC->DLS->GPI->THLm->PC. Below : Example connectivity for one population channel (representing one action), through PC->STNdl->GPI->THLm->PC. Note the divergent connections from one population of STNdl to all the populations of GPI.

As mentioned in the section 6.2.2, the *motors* consist of *move* and *turn*, that control the

movement of the agent. However, in this section, to elaborate on the network description more, high-level actions *left*, *right*, *ahead* and *jump* are represented in the motor loop. In the later sections, it will be described further how these high-level motor decisions will be converted to precise motor commands in terms of *move* and *turn*.

6.7.2 Population Dynamics

The dynamics of a neural population unit that is used in all the structures of the network is described in equation 6.10 as in (Leblois 2006). Assuming each population unit represents an ensemble tuned towards a particular option : I_{ext} is the external input representing the salience of the option, I_s is the input to the unit from its connections (synaptic input) and τ is the decay time constant of the synaptic input and V is the resultant activity of the unit. External input, I_{Ext} is provided only in cortical structures (for the other structures $I_{Ext} = 0$), and T is the threshold of a neuron, depending on the population. Also, symmetry breaking is generated by Gaussian noise δ to the activity of each ensemble at each time step.

$$\tau \frac{dV}{dt} = -V + I_s + I_{ext} - T \tag{6.10}$$

$$U = fn(V + \delta) \tag{6.11}$$

Structure	Threshold (T)	Noise (δ)
PC	-3	1.0%
DLS	0	0.1%
GPI	-10	3.0%
STNdl	-10	0.1%
THLm	-40	0.1%

Table 6.4: Parameters of CBG Motor Loop

The striatal population that is silent at rest (Sandstrom and Rebec 2003),requires concerted coordinated input to cause firing (Wilson and Groves 1981), and has a sig-

moidal transfer function due to both inward and outward potassium current rectification (Nisenbaum and Wilson 1995). This is modeled by applying a sigmoidal transfer function to the activation of cortico-striatal inputs in the form of the Boltzmann equation:

The activation function *fn* in Eq. 6.11 is a clamping function for all the structures except striatum, which bounds the activation value between a minimum (0) and a maximum activation value. Striatal projection neurons are generally silent at rest (Sandstrom and Rebec 2003), require coordinated input to cause firing (Wilson and Groves 1981). That is, there is a non-linear relationship between the input current and the membrane potential (Nisenbaum and Wilson 1995). This is modeled, as shown in equation 6.12, applying a sigmoidal transfer function to the activation of cortico-striatal inputs in the form of a Boltzmann equation.

$$V_{out} = V_{min} * \left(\frac{V_{max} - V_{min}}{1 + e^{(V_h - V_{in})/V_c}} \right) \tag{6.12}$$

where V_{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. The parameters used in the model for the function are listed in table 6.5

Parameter	Value
V_{min}	0
V_{max}	20
V_h	16
V_c	3

Table 6.5: Parameters of sigmoidal transfer function

The synaptic input to a unit j, I_s^j, which is the input as a result of the connections from units of other structures (say i), depends on the connections weights (w_{ij}) between units i and j , as shown in the equation 6.13. The plastic synaptic connections that change across the task depending on reward reinforcement are denoted by dashed arrows in Fig. 5.3). Except these plastic connections, the rest remain to be constant connection weights

chosen at the beginning, within the range of 0.25 and 0.75, generally chosen around 0.5.

$$I_s^j = \Sigma_i w_{ij} * \hat{G}_{ij} * m_i \tag{6.13}$$

Also, there is a fixed gain parameter that characterizes the strength of interaction between the two populations to which i and j belong. For example, for any pair of connections ij between CTX(i) and STR(j), the gain $\hat{\mathrm{G}}_{CTX\_STR}$ is fixed. A positive or negative $\hat{\mathrm{G}}$ defines the connection as excitatory or inhibitory respectively. In the "direct" pathway, as a result of two inhibitory and one excitatory connection, it is referred as a positive feedback loop. In the "hyperdirect" pathway, as a result of two excitatory and one inhibitory connection, it is referred as a negative feedback loop. (Leblois 2006) further did a theoretical analysis on the network dynamics and the interaction between the direct and the hyperdirect pathways. It was highlighted that for intermediate values of external activation in CTX, there exists a linear steady state where all populations are active. Using a reduced model, (Leblois 2006) demonstrates the non-oscillatory and oscillatory instabilities. Should the external input in one CTX population increase, the break of otherwise symmetrical populations is described as a function of net gains of the positive and negative feedback loop. If the product of gains in the positive feedback loops is denoted as G_+ and the product of gains in the negative feedback loop is denoted as G_-, it is the relation between G_+ and G_- that results in the symmetry breaking between otherwise strongly similar competing activations (Fig 6.15). The detailed reduced model and the dynamical model is not discussed here, but as a minor demonstration, the CTX-STR gain is altered in the network according to required G_+ and G_- that guarantee a motor choice is made in the network. The dynamics given by the equations 6.10 and 6.13 are phenomenological and are not constrained to a specific neuronal architecture.

Positive feedback loop	
Connection	Gain (G, $\pi_G = G_+$)
CTX -> STR	+1.0
STR -> GPI	-2.0
GPI -> THL	-0.5
THL -> CTX	+1.0

Negative feedback loop	
Connection	Gain (G, $\pi_G = G_-$)
CTX -> STN	+1.0
STN -> GPI	+1.0
GPI -> THL	-0.5
THL -> CTX	+1.0

Table 6.6: Connection gains within a motor CBG loop

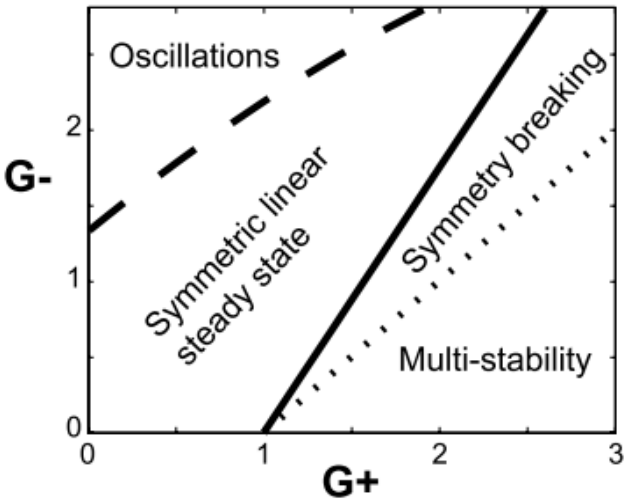
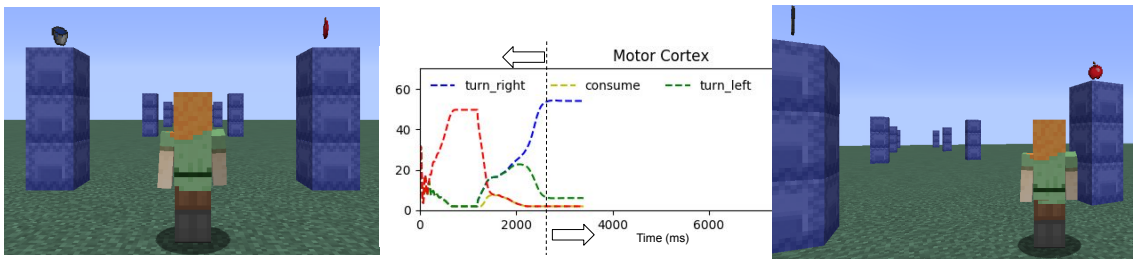


Figure 6.15: The phase diagram for the various dynamical regimens of the reduced model in (Leblois 2006) as a function of G_+ and G_-. Figure copied from (Leblois 2006)

In the model described here, the usual gain parameters for a single motor CBG loop are as described in Table 6.6. Most of the parameters in the network are fixed, mostly according to the studies done before (Topalidou et al. 2015). Throughout this work, mostly the parameters involving the cortical and striatal structures are modified according to the scenario. Fig 6.16 shows an example scenario where the agent has to make a simple motor choice between positions (between left and right, where similar stimuli are present). The algorithm that is implemented between the aspects of the framework (sensors, motors) and the aspects of the computational model are described in Alg. 1



(a) Before choice (b) Activities in Motor cortex (c) After choice `(turn_right)`

Figure 6.16: Example of simple motor action choice. (a) Before the decision, agent faces an option between two actions. (b) As the decision threshold reaches (at t=2000ms), turn_right is selected as the chosen action (blue). (c) The agent starts turning right orienting and moving towards the pillar on the right. Note that the items on the top of the pillars are not visible to the agent and since both the pillars are blue, they don't play a role in the decision.

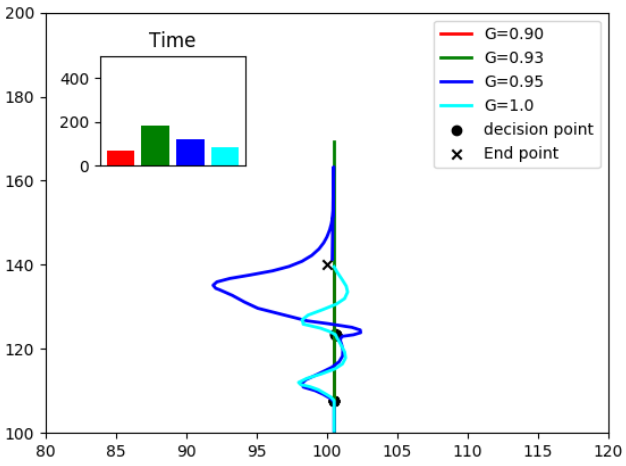


Figure 6.17: Motor choice as a result of G_{CTX_STR}. The agent, when encountered a choice situation, has to make a choice between two actions. If the agent fails to do so in a certain time, the agent moves ahead. If successful, makes a turn (action) and gets back to the initial path for more decision situations.

6.8 A computational model of parallel CBG loops for instrumental learning

Imagine the scenario in the section 6.7. But at each decision point, instead of pillars of the same color, now the agent finds a choice between two pillars, each of different

Algorithm 1 Motor decision from the model in the environment

1: $Brain = \{CTX, STR, GPI, STN, THL\}$
2: **while** True **do**
3: Brain.cycle()
4: **while** SP $>$ 0 **do** SP : SalientPositions
5: **for each :** $sp \in \mathcal{S}P$
6: $CTX[sp].Iext = CONST\_ACT$
7: **if** Delta(CTX.V) $>$ Threshold **then**
8: $P_{chosen} = ArgMax(CTX.V)$
9: $SENCTX[P_{chosen}].Desired = CONST\_ACT$
10: $Des = Any(SENCTX.Desired)$
11: **while** CurrPosition != Des **do**
12: $MOTORACTION\{Turn, Move\}$

color (see figure 6.18). The position of any colored pillar is not fixed at any time, but pseudo-randomly distributed between the positions *left* and *right* of the agent. There is a probabilistic reward, an apple, possible upon the choice. The probability that there is an apple on the top of a blue, red, brown and white pillar is $1, 2/3, 1/3$ and 0 respectively. The agent cannot 'see' what is on the top of the pillar. Only after approaching a pillar, and 'jumping', will the reward be obtained, if there was any. If there was no reward on jumping, the agent 'givesup' after a couple of attempts of jumping and it is counted as a no-reward decision. This necessity to jump around a block to obtain the reward is by design, primarily to distinguish between pavlovian scenarios (section 6.9), where there is no deliberate action for the agent to do (except moving, which is exploring). In this scenario, essentially there are two decisions for the agent to make. (i) Which color pillar to choose? (ii) Which direction to go? Intuitively, we tend to take it for granted that the decision is made in terms of which pillar to choose and then accordingly the decision of which direction to go follows. However, these aspects are distinctly represented in the brain, while being associated through the multi-modal representations in the lateral prefrontal cortex. Multi-modal representation refers to the representation that captures the relational information between modalities, for example, that the blue block is on the left and the brown block is on the right. For the decisions to be coherent in the cognitive

space (color of the pillar) and the motor space (position), the multi-modal association is crucial, so that the decision in one modality can influence the decision in the other. For example, if a decision has been made in the cognitive loop for blue pillar, the decision in the action space should subsequently be 'left', to complete the action and approach blue pillar as desired.

Note : In the traditional experimental setups of studying instrumental learning often in rodents, even in monkeys and humans, when the term 'response' or 'action' is associated to an outcome, it is a specific action that the animal performs; like pressing a particular lever (rats), touching a particular spot on the screen (monkeys) or pressing specific keys on a keyboard (humans). However, in most of the tasks studied in the context of monkeys or humans, especially n-armed bandit tasks like the one described here, 'response' per se is not a fixed action, but rather an action that is linked to choosing a particular stimulus. For example, when there is a blue pillar on the left and red pillar on the right, the rewards are not being linked to going left or right, but rather to the blue pillar that is on the left or red pillar that is on the right for this trial. Thus, it can still be seen as instrumental learning because the agent has to perform an action to get the reward, just that the action is not fixed every time as in reality the reward is linked to the pillar, not to the position.

Figure 6.19 shows a systems level model of CBG loops through OFC and PC, involving also an associative loop through lPFC and DMS, primarily to represent the binding information of the options (which object is at which position). It was shown in a single neuron recording study in monkeys to study coding of visual space in dlPFC, that lPFC has unidirectional projections to the BG that complete a loop back with lPFC through the thalamus. Thus network is effectively comprised of 3 CBG loops including the associative loop through lPFC and DMS. One CBG loop, limbic loop, is on the left, to represent the identities of the options (populations shown as circles filled in light blue in each structure). The second CBG loop, motor loop, is on the right, to represent the positions of the objects (populations shown as circles filled in pink in each structure). The associative CBG loop has two associative structures, lateral PFC (lPFC) and dorsomedial striatum (DMS). In theory, lPFC and DMS play more complex roles. lPFC is known to

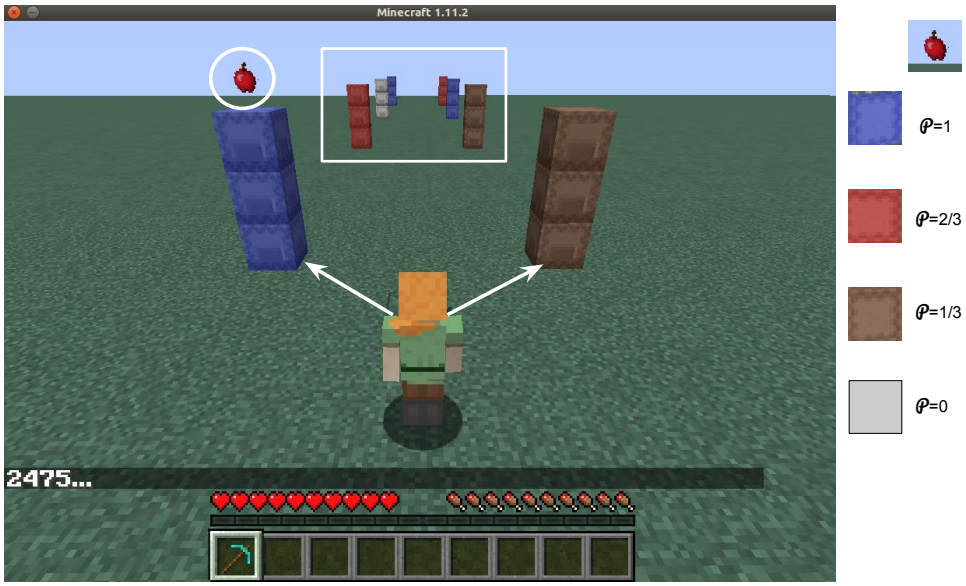


Figure 6.18: A two-armed bandit task. The agent is presented with a choice between two blocks, at every decision point. The apple (white circle) on the top of the blue block is not visible to the agent. More similar choices moving forward (white box).

be involved in high level planning, cognitive control whereas DMS is known to facilitate faster stimulus-action representation learning with OFC and the motor loops. It should be noted that given the task structure, there is no fixed stimulus-response contingencies to learn as the position of each color is never fixed. In fact, it was computationally shown previously that trying to learn by attributing rewards to the positions when the task structure doesn't do so, reduces performance of the task. Hence, both lPFC and DMS are simply used to represent the combined information of the options and also to allow the information transfer from the cognitive loop to the motor loop.

6.8.1 Network

A network of parallel and interacting CBG loops is implemented to solve this task. It will be described in later sections how different other tasks also can be solved with the similar model. Each CBG loop implemented is according to the architecture and the network connectivity similar to what was described earlier in the figures 5.3 and 6.14. There are

two CBG loops that interact through an extra associative layer. The motor loop is exactly the one described in the previous subsection. The cognitive loop is implemented through OFC as the cortical substrate of the CBG loop and the nucleus accumbens core (NAcc) as the striatal substrate. The main difference between the motor loop and the cognitive loop is that the connections between OFC and NAcc are modifiable, owing to the synaptic plasticity. That is, through the trial (or through the behavior), these weights can be modified, thanks to dopamine modulation that occurs when a reward is experienced in LH and it is signalled by VTA. In addition to the two loops, there is an associative layer involving a part of lPFC and a part of DMS. The associative layer is precisely to address the binding problem that arises when two populations in OFC and two populations in PC are activated. Because, if only OFC and PC are activated representing two options - A at position X and B at position Y - the respective positions are not encoded and it becomes equivalent to the case where A is at Y and B is at X. Hence, whenever an option is encountered, although the identity is encoded in OFC and the position (or required action) in PC, the combined information is encoded in the lPFC population, which is essentially a 2-dimensional map of identities and position. The detailed network connectivity involving the regular cognitive and motor CBG loops and the associative loop is illustrated in the figure 6.20.

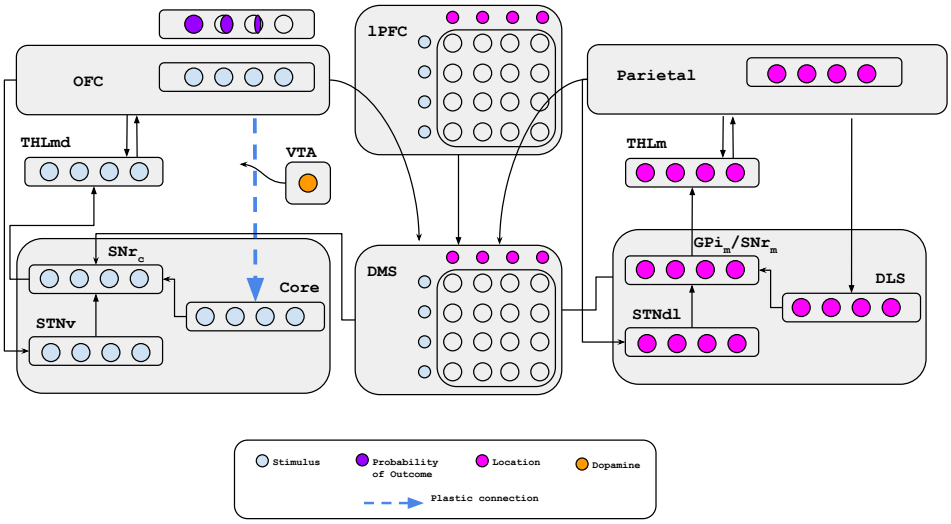
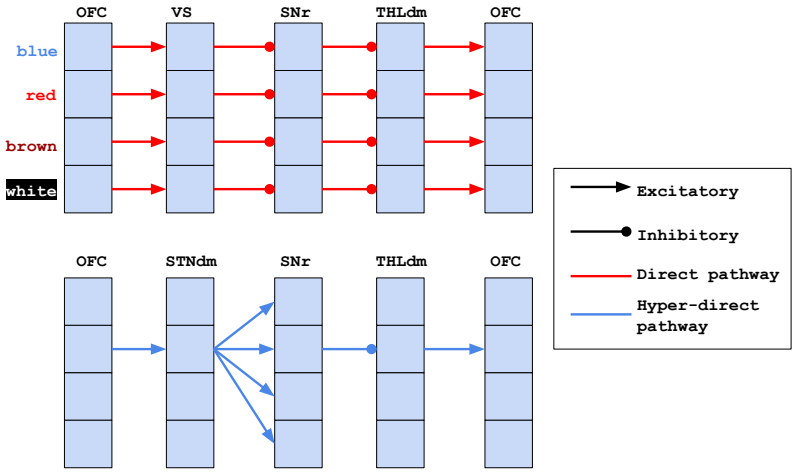
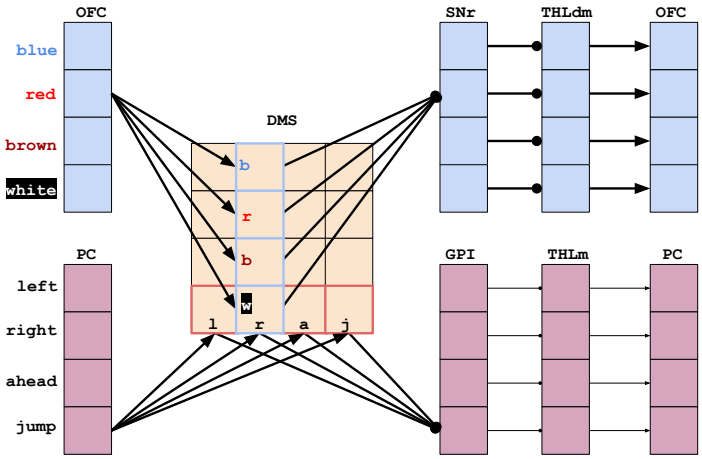


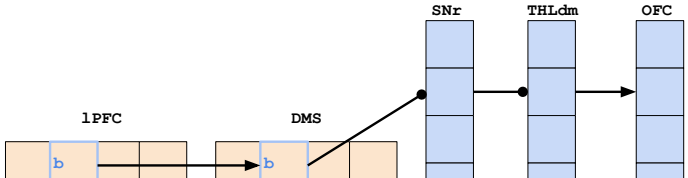
Figure 6.19: CBG loops through OFC and PC. Cognitive decision and motor decision in separate loops, with interaction via the associative loop



(a) Limbic loop connectivity



(b) Limbic to associative and associative to motor



6.8.2 Learning

As mentioned earlier, the only difference between the motor loops and the OFC loop, is that the cortico-striatal connections, between OFC and NAcc are modifiable. After every decision and verifying the outcome, the weights are updated. Although this weight update is done only to the limbic weights (OFC loop), it doesn't mean that there is no learning else where. In fact, in a different computational study (Nallapu and Rougier 2016) which will not be detailed here, we have shown that updating the cortico-striatal weights in the motor loop when the reward contingencies are not according to the positions leads the model to perform sub-optimally. Over number of trials, even if the animal tries to learn the choices according to positions they are shown at when in reality reward doesn't depend on the position, it would be realized after few trials. Hence, learning has been confined to the limbic loop. Like in the previous models, all synaptic weights are initialized to 0.5 (SD, 0.005). As described in the motor loop dynamics, gains in each pathway are used as multipliers to the weights. And upon weight changes, to make sure the weights stay within the initial bounds, every weight update is followed by a normalization of weights (equation 6.14).

$$\frac{dW_t}{dt} = \Delta W_t * (\frac{W_{max} - W_t}{W_t - W_{min}}) \tag{6.14}$$

The weight update term ΔW_t is calculated as a function of reward prediction error (RPE), which is believed to be signalled by dopamine at the level of cortico-striatal synapses. However, it was specifically found that striatal neurons involved in cortico-striatal synapses show long term potentiation (LTP) and long term depression (LTD) with respect to positive or negative prediction error, respectively (Pawlak and Kerr 2008). RPE precisely is the difference between the perceived reward value and the expected reward value. In the model, similar to a standard critic-learning RL framework, expected reward values of each stimulus population are maintained and updated.

$$\Delta W_{ij} = \alpha * \delta_t * U_j \tag{6.15}$$

$$\alpha = \begin{cases} \alpha_{LTP}, & \text{if } \delta_t > 0 \\ \alpha_{LTD}, & \text{otherwise} \end{cases} \tag{6.16}$$

The RPE, δ_t is calculated using a simple critic learning algorithm given below.

$$\delta_t = R - v_i \tag{6.17}$$

where R, the reward, is 0 or 1, depending on whether a reward was given or not on that trial. Whether a reward was given was based on the reward probability of the cue associated with the direction chosen. v_i is the value of the cue represented by neuron i in the striatal 'core' population. The value of the chosen cue is then updated by :

$$v_i \leftarrow v_i + (\delta_t * \alpha_c) \tag{6.18}$$

where α_c is the critic learning rate and is set to 0.025 and α_{LTP} and α_{LTD} are set to 0.004 and 0.002 respectively.

Figure 6.21 shows the decision dynamics in the cortical structures, OFC and Parietal (PC), in terms of the activities of the populations representing colored pillars and positions in OFC and Parietal cortices respectively. There are 4 possible positions that can be represented in the PC, two each on right and left of the agent, and one closer and farther option on each side (right_c, right_f, left_c and left_f). Closer and farther here refer to the viewing angle with respect to the agent. It should be noted that the example scenario shown in figure 6.18 has the pillars equidistant and one on each side of the agent. However, qualitatively this is no restriction for the system. The pillars could essentially be randomly positioned. As mentioned in the section 6.4.2, as the part of the cognitive architecture of the agent, when there are multiple stimuli around the agent, the agent orients towards the center of the stimuli before making a decision. Therefore, in the case of a 2-armed bandit task as described in figure 6.18, it is assumed that all the decisions taken will have positions as either right_c or left_c.

The model allows a bidirectional information flow between loops such that during

early trials, a direction can be selected randomly (Fig. 6.21 top, Trial 1), irrespective of the pillar positions. However, after repeated trials, the model is able to consistently make the cognitive decision before the motor decision in each trial. Consequently the motor decision, biased by the cognitive decision, is made towards the position of the more rewarding cue shape. This can be observed clearly in the shift of decision times by trials 60 and 120 (figure 6.21, middle and bottom).

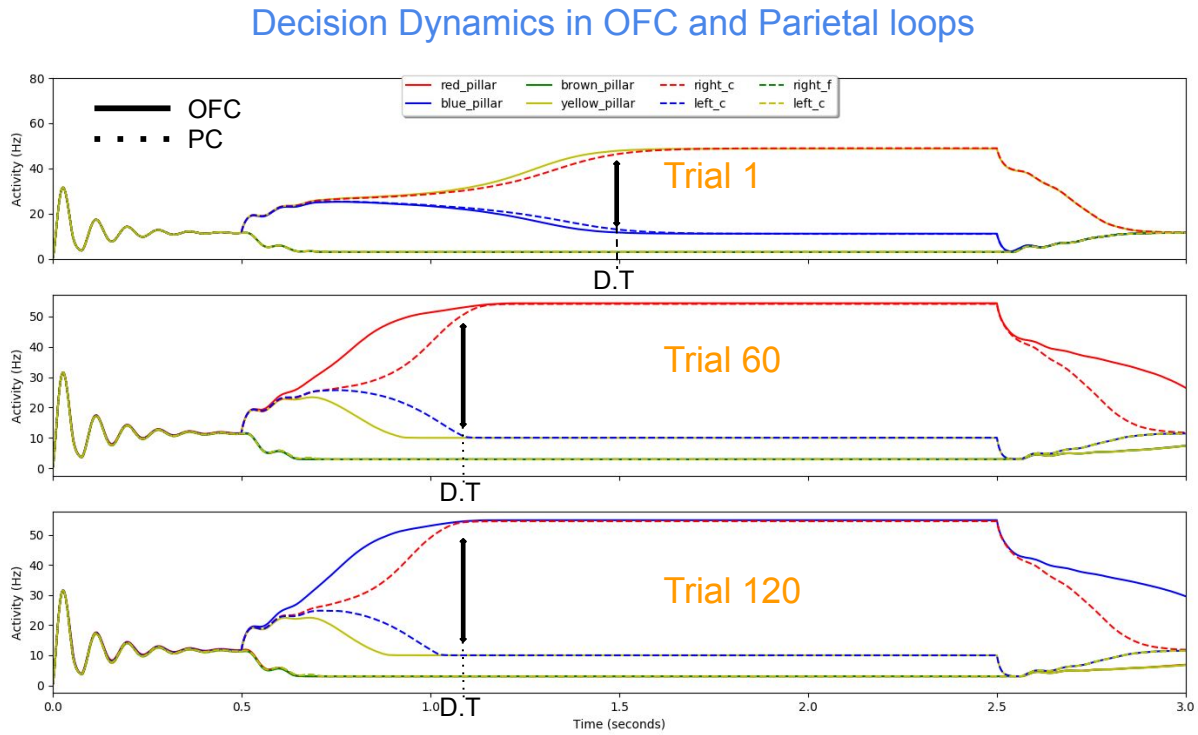


Figure 6.21: Decision in the first trial, mid-session and the last trial, in a session of 120 trials with learning. From top to bottom : Trial 1, 60 and 120. Solid lines - cognitive decisions (pillar color), Dashed lines - Position of the pillar. Among the 4 positions, right_c and right_f stand for the closer and farther position on the right respectively, with closer and farther being in terms of viewing angle. Same goes for left_c and left_f for the left side. D.T on the X-axis is decision time, when the difference in activities is greater than the threshold.

The cortical cognitive ensembles are activated with the presented cues, the motor with the presented positions, and then the combined information (which is where) is given as input to the associative ensembles. Each loop described in our model has a respective substrate of the basal ganglia involved for the local selection. This kind of architecture

has been employed to computationally model parallel feedback loops in the CBG network (Guthrie et al. 2013; Topalidou et al. 2015) and replicate primate behavior in a two-armed bandit task (Pasquereau et al. 2007).

As mentioned in the section 3.4, dopamine has a predominant effect over the CBG network after the chosen actions are performed, by signalling the reward value (more precisely, reward prediction error, RPE, often in conjunction with another neurotransmitter GABA), an outcome-dependent learning process.

6.9 A computational model of simple pavlovian conditioning in the basolateral amygdala (BLA).

Imagine a classic scenario of pavlovian conditioning (See section 2.3.1): while the animal has no action to do, a neutral stimulus, *CS*, is presented to the animal for a duration *D*. Towards the end of this duration *D*, a reward *R* is offered during a smaller duration *d*, not requiring the animal to do any action. Thus for a short time interval *d*, both CS and R coincide. If this kind of presentation is repeated sufficient number of times, CS gets associated to the reward delivery, meaning that upon the presentation of CS, right at the beginning of D, an expectation of R can be observed. Two of the most important neural correlates that will be discussed here in relation to this kind of association between a neutral stimulus CS and an unconditioned reward *R* are (i) a sub-cortical structure, basolateral nucleus of the amygdala (*BLA*) and (ii) the neurotransmitter *dopamine*.

The above described scenario is implemented in the Minecraft world, using a Pillar as **CS** and an appetitive *item* as **R**. In this case, the height of the pillar is two blocks. For instance, when a reward like apple is on the top of the block, it is automatically offered to the agent when the agent passes by a certain proximity.

The populations of basolateral amygdala (BLA) and the part of Ventral Tegmental Area (VTA) that is associated with the firing of dopamine neurons are implemented (Fig. 6.25). Other related structures like the central nucleus of amygdala (CE) and Peduncolo-

pontine nucleus (PPN) are also implemented, but in a scope that is relevant to simple pavlovian conditioning. For example, detailed roles of CE and its influence on dorsolateral striatum (DLS), and the populations of PPN that learns the precise magnitude of the reward are not taken into consideration. BLA has the representations of both CSs and the outcomes.

The BLA outcome populations (BLA_O) are implemented as the rated-coded populations similar to those described in eq. 6.10 except the synaptic input current I_s is adapted to contribute only as a phasic component, instead of the entire tonic component (all through the duration of the CS). This is done by first deriving a phasic component of the incoming I_s according to the equation 6.19 with a time constant much slower than the time constant of the membrane potential V_t. This phasic component acts as the resulting synaptic input current to calculate V_t.

$$\tau_\phi \frac{d\hat{I}}{dt} = -\hat{I} + I_s \tag{6.19}$$

$$\tau \frac{dV}{dt} = -V + (I_s - k * \hat{I}) + B \tag{6.20}$$

$$U = (V + \delta)^+ \tag{6.21}$$

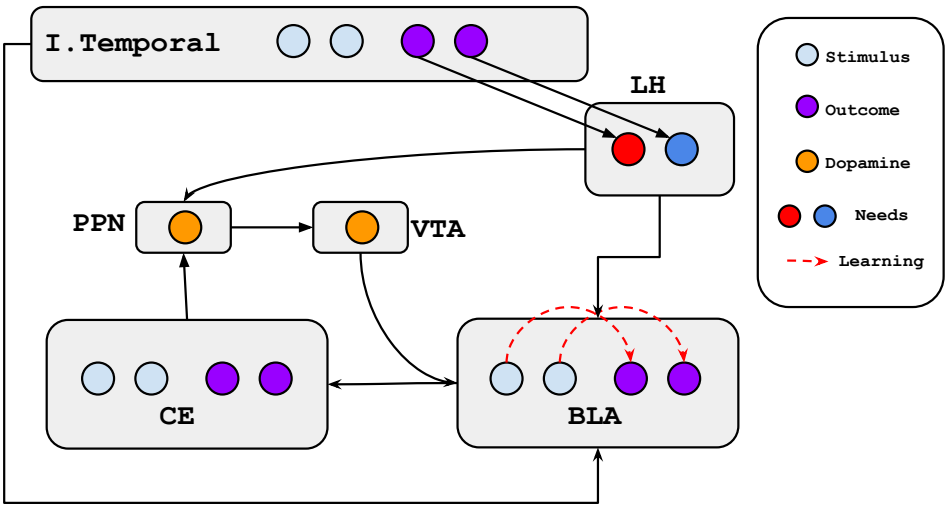


Figure 6.22: Model of learning Pavlovian Associations in Basolateral Amygdala (BLA).

In more elaborated scenarios, where finer details of reward like changing magnitude and timing are to be learned, a closer look is required at CE, the populations of PPN that maintain the magnitude of reward across trials and their connection to the dopaminergic and GABAergic systems of VTA in conjunction with ventral striatum (VS). The choice of the parameters for the structures involved the pavlovian network with amygdala are listed in the table 6.7, partly based on this consideration from previous computational accounts (Kaushik et al. 2017). The amount of tonic component retained, k, is set to 1 for all populations.

Structure	$\tau_{(ms)}$	$\tau_{\phi(ms)}$	B
BLA_O	10	10	0
CE	20	5	0
VTA_DA	5	5	0.2
PPN	5	5	0

Table 6.7: Parameters of Amygdalar Pavlovian Network

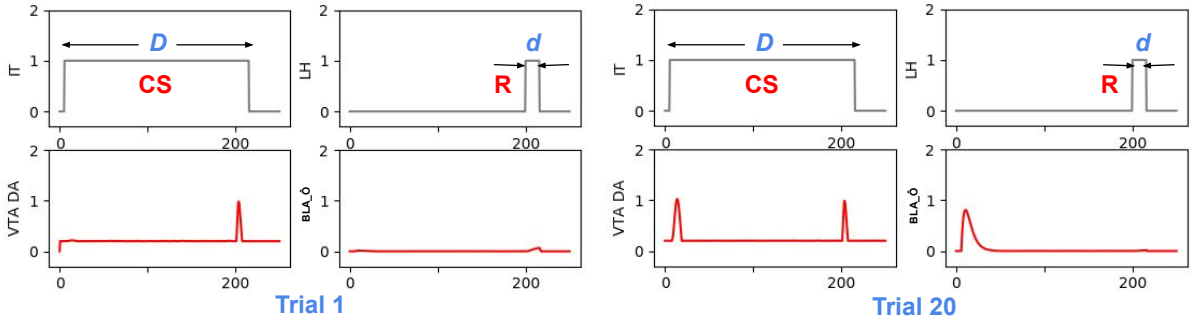
While the CS is active, BLA_CS tonically fires due to a square wave signal that is received from the Inferior Temporal (IT) cortex. When there is an outcome, the sensory aspect of outcome is transferred to BLA_O populations in the same way as BLA_CS. In

addition, the appetitive information is sent from LH, that goes both to VTA through PPN and to the corresponding outcome populations in BLA_O. BLA_CS and BLA_O are connected initially, by negligible random weights, since there is no necessary correlation to begin with. Pavlovian learning, which is a kind of *associative learning* rather than *reinforcement*, can be implemented on these weights, in a simplistic representation of the classic Hebbian rule. As the saying goes - "neurons that fire together, wire together" - the simultaneous firing of a BLA_CS population (representing the presence of a CS), and a BLA_O population (representing the presence of an outcome) is associated in by increasing the connection weights between them. Equation 6.22 is an example of a simple hebbian rule, where ΔW represents the change in weights upon learning, α_H is the learning rate, C_i is the population of BLA_CS that is present and O_i is the activity of the population of BLA_O of the outcome that is observed. Subsequently, after sufficient association, a CS population firing will cause the BLA_O population to fire.

$$\Delta W = \alpha_H * C_i * O_i \tag{6.22}$$

Since BLA_O populations also represent the presence of outcome, for the sake of clarity, the expectation in BLA_O that is learned is represented as BLA_$\hat{O}$. To saturate the learning, the input from VTA, which a fixed magnitude representation of each outcome, drives the learning between BLA_CS and BLA_$\hat{O}$. Hence, the final learning rule combines the presence of outcome as well as the magnitude and accordingly adjusts the synaptic weight update. Equation 6.23 shows the learning rule with ΔW representing the change in weights upon learning, α_H is the learning rate, C_i is the population of BLA_CS that is present and O_i is the activity of the population of BLA_O of the outcome that is observed, R is the magnitude of the outcome that is conveyed by the VTA DA every time there is an outcome and $\hat{O}_i$ is the current expectation in BLA_O population.

$$\Delta W = \alpha_H * C_i * O_i * (R - \hat{O}_i) \tag{6.23}$$



(a) Before association (b) After association

Figure 6.23: A demonstration of firing in BLA_Ô representing expectation of reward on CS onset. (a) At the beginning of experiment, VTA DA fires only when the reward is delivered (at $t = 200$). There is no firing in BLA representing the expectation of reward. (b) After learning, VTA DA fires both at the presentation of CS and the reward. BLA_Ô fires at the CS onset representing the expectation of reward.

6.10 A case for lateral and medial dissociation of OFC

As highlighted throughout chapter 4, detailed understanding of processes within OFC in the context of decision making are still unclear. However, certain implications of OFC can complement the mechanisms that have been discussed in the context of CBG loops. Especially since OFC is in a position to have more abstraction about the current state, through top-down processes, it can modulate the processes in amygdala and striatum. Although the model of CBG loops described in Fig. 6.19 could successfully solve a 2-armed bandit task with the given reward probabilities, several key attributes of the model were too simplified. Especially for the correct decision to occur, the model solely relies on the synaptic weights that are updated by learning. Although expected values of each stimulus were learned (equation 6.18) over the task session, these values were only used to derive the reward prediction error (δ_t) which was further used to calculate the change in synaptic weights. But, there was no role for these learned values in the immediate choice dynamics, nor for the history of choices. Although in principle it was sufficient to perform in the task given the reward contingencies were sufficiently dispersed ($P(r) = 1, .66, .33, 0$ respectively for each stimulus). It can be clearly found out that when the reward

contingencies are close and the synaptic weights of one stimulus population diverge to increase constantly by chance, there is no way the model would choose a different option. In this case, even if the chosen stimulus doesn't reward, if its synaptic weights have diverged sufficiently, the weight decrease doesn't ensure the choice of a different stimulus in the next trial.

In theory, in the case of reward contingencies being closer, it can be handled in the same model by adjusting the learning rate to be smaller. But, ideally this adjustment has to be intrinsic within the system, if we want the model to solve different kind of tasks. In fact, it is thought that this is one of the common roles of prefrontal cortical systems, particularly OFC, to exert a top-down control on the sub-cortical processes, in this case, learning rate at the level of cortico-striatal synapses (Kennerley and Wallis 2009). Similar argument was made in the case of OFC's influence on the Pavlovian learning in the basolateral amygdala, especially at the time of reversal of stimulus-outcome associations (Elliott 2000). The sub-cortical learning, after long training, takes longer to unlearn the changes or sudden shift in outcome contingencies, where a PFC region like OFC like identify the shift and modulate the sub-cortical learning with momentarily higher learning rates.

6.10.1 State space and Task space abstraction

It has been consistently proposed that OFC learns the abstract state space of the task. However, clear distinctions haven't been pointed out whether both lateral and medial equally contribute to learning state space of the task to guide the behavior, or either one of them has a greater role to play. To that extent, it is not even clear to what level of abstraction the state space can be learned. One of the possibilities in which one can view the structure of a task is to look at the sensory presentations and outcome contingencies separately, relating to abstracting state space information and the task space information respectively.

State space abstraction

Especially regarding the sensory representations, the abstraction can be two fold - within trial and across trials. *Within trial* state space abstraction refers to the presentation of stimuli as one distinct state and the presentation of outcome as another. And when there is an outcome in the previous trial, this can be a distinct state for the next trial - instead of stimulus alone, the distinct state represented could be stimulus-reward together. For example, when trial 1 gives a choice between A and B, this can be represented as a distinct state S1=(A,B). Following the choice, say A was made and there was a reward, in the next trial, the state could be represented as a new state S2=(A,B,A_R) and accordingly another distinct state could be represented in the following trial, if in this trial, B was chosen and rewarded, S3=(A,B,B_R). Similar state representation in OFC has been implemented in computational models that explained the possible state encoding in the OFC (Zhang et al. 2018).

Across trial state space abstraction can be described as abstracting the state of trial presentation as observed in each trial. In this case, depending on the task, there could be either a state change or not. For example, consider the two-armed bandit task described in section 6.8 which was solved by the model of CBG loops with OFC and PC described in the figure 6.19. In this case, in each trial, the stimuli presentation involved two of the four known stimuli. However, the presentation was pseudo-randomized among the pairwise combinations of the four stimuli. Therefore, at the end of each trial, the model has no basis to predict as to what would the next state be in the following trial (which pair of stimuli will be presented). If (A, B, C, D) are the four known stimuli, and each trial presents two of them, of the six (${}^4C_2$) pairwise combinations possible among the four stimuli, the animal can abstract 6 possible states (each possible pair). Whereas consider a task where the stimuli presented do not change from trial to trial, there is no new state to abstract across trials, except more statistical information like previous choice made or previous reward received.

Task space abstraction

Another aspect of the task structure that is hidden for the animal is the reward contingencies. Although the individual contingencies are learned in the form of the striatal mechanisms, modeled in the section 6.8 as the expected value in the nucleus accumbens core populations, the underlying structure in the contingency distribution is not learned and the individually learned values do not dynamically play a role in learning. It might be the case with striatal neurons that they have a different depression rate than the potentiation rate, but if the cortical counterparts have more detailed abstraction about the ongoing value structure and the estimate of not only the chosen option but also the unchosen option, they can control learning at much faster rates, especially when switching between the options is important. In fact, it was shown that vmPFC specifically encodes the value of chosen option in comparison with the unchosen option (Boorman et al. 2009; Lim et al. 2011).

6.10.2 Learning vs Choice

Although it is a premature argument to clearly dissociate, it is still safe to propose that lateral OFC might be more involved in learning whereas medial OFC could be in a better position for value comparisons. There is much evidence on how lateral OFC is crucial for proper credit assignment, to appropriately assign the outcome to the chosen stimulus. Similarly, medial OFC (or vmPFC) has strong connections from NAcc core of ventral striatum to receive the value representations. There have been propositions through computational accounts, that the value comparison in vmPFC (or medial OFC) could be explained by attractor networks with mutual and lateral inhibition. Notwithstanding the lack of evidence to point out clear dissociation between learning and choice in lateral and medial OFC respectively, one possibility is that medial OFC plays crucial role in choice at the beginning of the task and lateral OFC gradually strengthens learning and slowly takes over, as long as the predicted contingencies do not change drastically. In addition to separate roles attributed to lateral and medial OFC, a simplified memory is

implemented in each regions to reflect the general choice-based and reward-based history respectively, which is a general feature of prefrontal regions. After the reward is delivered, reward-related activities were observed in lateral OFC and action-related activities were found in medial OFC (Bouret and Richmond 2010). Since, in the paradigm used here, action is not an explicit motor action bur rather performing the action that chooses the desried option, the option-reward based history was implemented in both the regions.

6.10.3 Simplified role of ACC

ACC, has been implicated quite closely to vmPFC and OFC in general in most of the reinforcement learning scenarios. In fact, one of the striking dissociation proposed was that whereas the activity in vmPFC reflects the value difference between the options, the activity in ACC reflects the inverse value difference signal. Meaning, ACC is employed when the values of the options are too close or conflicting that vmPFC cannot arrive at a choice (Noonan et al. 2011; Rushworth et al. 2007, 2012). Several other accounts also pointed how ACC can inhibit learning in the case of no-reward choices (Kennerley and Wallis 2009). However, in the scope of this thesis, not to confound with the dissociation between the lateral and medial OFC, these aspects of ACC are not investigated. Instead, another most prominent theories about ACC, its involvement in effort-based decisions is considered in this work. ACC is appropriately connected to both valuation systems like NAcc core and the action representations from lPFC and DMS. In addition, owing to its connectivity with the anterior insula, ACC is in a position to encode action costs with respect to internal bodily situation. ACC lesions in rats resulted in lesser willingness to exert more effort to gain the high reward (Rudebeck et al. 2006; Walton et al. 2003).

6.11 Computational account of lateral and medial OFC

With the above mentioned aspects noted from the most commonly implied OFC functions, the following features have been implemented in the existing computational model of CBG

loops described in section 6.8. Later, the model will be tested on different tasks where the state space abstraction or task space abstraction can differently observed.

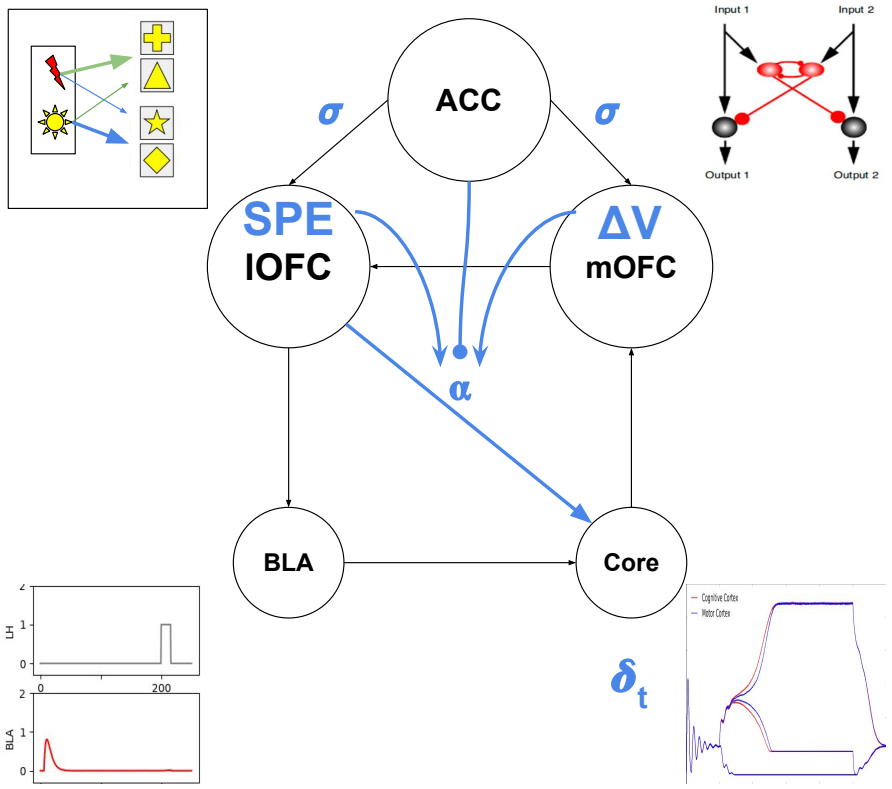


Figure 6.24: Simplified description of key model features.

6.11.1 ACC to Lateral OFC

The influence of action costs on decision by ACC has been modeled in the form of connection weights that ACC forms with the external input (eq. 6.10, I_{ext}) of the lateral OFC. These weights are learned as a function of both reward received and the cost involved. Therefore the RPE, δ_t now becomes as in equation 6.24

$$\delta_t = R * (1 - AC_i^z) - v_i \tag{6.24}$$

where AC_i^z is the fixed action cost according to the visual zone in which the stimulus represented by the population i is present.

6.11.2 Lateral OFC to Medial OFC : Long Route

Lateral OFC to BLA

While the Pavlovian learning of a stimulus-outcome association is sufficiently explained in a simple model of BLA and other amygdalar nuclei, lateral OFC is connected to BLA to modulate the learning in BLA.

BLA to Core

The Pavlovian associations between stimulus and outcome are transferred to NAcc core at the sight of a stimulus. Quite simply, this has been implemented as a bias at the population level of NAcc core, I_{ext} (recalling the population dynamics from equation 6.10).

Core to Medial OFC

With the outcome expectation signal received from BLA, the outcome expectation values in NAcc core are gated to medial OFC mapping onto the stimulus representations. Thus each stimulus is now tagged with its expected value in the medial OFC.

6.11.3 Value comparison in medial OFC

Medial OFC is believed to be performing specialized value comparisons, taking other internal factors into consideration (cf. Ch. 4). Several computational models like DDM or attentional DDMs described how lateral inhibition within the medial OFC populations can lead to comparison between similar values (Krajbich and Rangel 2011). In one of the most recent studies in rats, Malvaez et al. 2019 has shown that activating mOFC to BLA projections is sufficient for value retrieval whereas lOFC to BLA are not. Although this kind of value comparison seems much more useful in a more complex scenario where by attentional shifts, the evidence about each of the stimuli is being accumulated gradually (Gluth et al. 2018). However, in most of the simple cases described here, the same

principle of comparison could be useful when the values of stimuli learned are much closer than what a Pavlovian system like lateral OFC could distinguish.

The populations in medial OFC, representing the stimuli (CS) populations, are implemented as mutually connected laterally inhibiting each other.

$$\tau \frac{dI}{dt} = \text{-}I_d - I_l \tag{6.25}$$

$$I_d = d * I \tag{6.26}$$

$$I_l^j = \Sigma_{i \neq j} w_i * I_j \tag{6.27}$$

where I_l^j is the output of a neuron population as a result of the inhibition, $\tau = 0.01$ is the time constant, d is the decay parameter of inhibition and w_i are the inhibitory weights from rest of the populations. Typically, the decision can be chosen either by setting on the activity or the duration of inhibition.

Most importantly, the initial input to these mOFC populations, could come from both the stimuli values learned from core, and the ongoing decision process from the lateral OFC. And the winning activity is fed back as a bias into lateral OFC. Thus lateral OFC and medial OFC could affect each other during the decision. However mutual bias on each other could be controlled by specific parameters. As the decision within lateral OFC happens mostly within the first 300-400 ms, it is in this time that medial OFC could bias the input to lateral OFC.

This also explains for rapidly taking into account the values of stimuli which have been changed either by devaluation or extinction, while it might need more time for lateral OFC to modify the synaptic connections with the core.

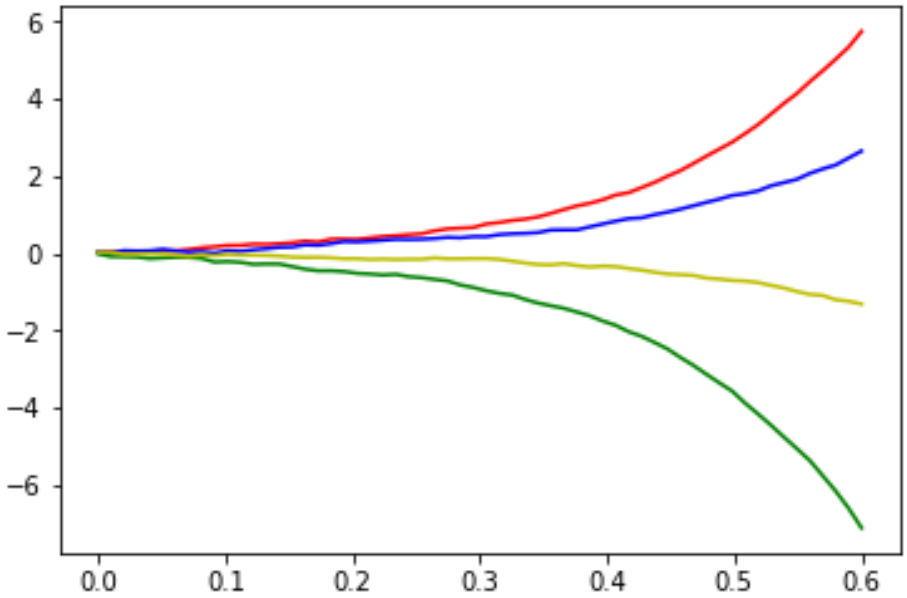


Figure 6.25: An example of lateral inhibition in medial OFC which would be provided as a bias to the lateral OFC-core loop.

6.11.4 State Prediction Errors in Lateral OFC

A model of all possible transitions from a superset of all known states is maintained. Upon every choice, irrespective of reward, only by the perceived next state, the state transitions are updated according to the equations 6.28 and 6.29.

$$SPE = 1 - P(\hat{S}_B|S_A) \tag{6.28}$$

$$P(\hat{S}_B^{t+1}|S_A) = P(\hat{S}_B^{t}|S_A) + \alpha_{MB} * SPE \tag{6.29}$$

with $P(\hat{S}_B^t$ as the probability of reaching the state S_B when in state S_A.

When SPE is non-zero, the value learning in the equation 6.18 is affected by updating the values of not only the visited state but also the other state with the available transition probabilities.

6.11.5 External bias from Medial to Lateral OFC

The decision dynamics in lateral OFC population, which is connected to the NAcc core by modifiable synaptic weights, is initially sensitive to external bias received through the populations' I_{ext} since the synaptic weights are all similar. As the weights start learning the preferences of options, depending on the stage of learning, the effect of the external bias through I_{ext} on the decision changes. It was computationally shown before that after sufficient learning, the system is robust to certain limit of bias to drive the decision according to the learned weights, not the bias (Nallapu and Rougier 2016). However, interestingly the amount bias needed to overturn a learned choice depended on the value difference of the options presented. Thus, the external bias from medial OFC will affect more the decision in lateral OFC in the beginning of the task.

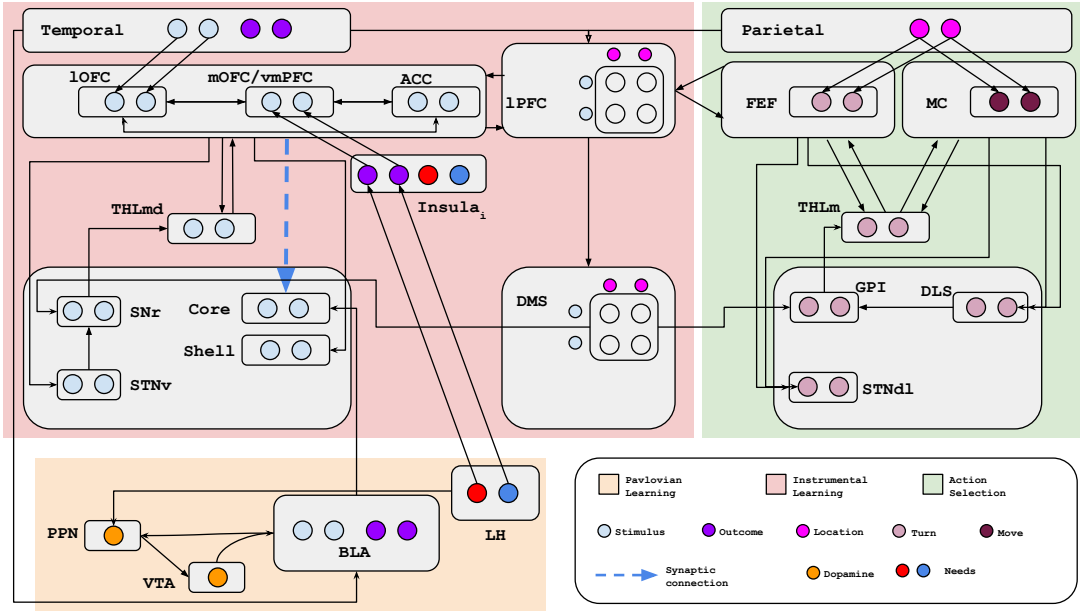


Figure 6.26: A schematic of the model of CBG loops, dissociating lateral and medial OFC

The primary goal has been to investigate the dissociation between lateral and medial OFC. However, after implementing several mechanisms that were specifically implicated for lateral or medial OFC individually, it appears that it does not necessarily have to be only dissociation between both these sub-regions, there could very well be interacting processes between them, for the amount of shared functional connectivity there is between

the lateral and medial OFC particularly considering ventral striatum as a crucial nexus. Some standard decision-making tasks have been demonstrated so far in this chapter using different aspects of the model. Since it involves much more closer and detailed look to analyze the dynamics in all the different pathways involved in the comprehensive model, it is assessed in terms of overall behavioral performance and detailed analyses will be of future tasks on the topic. Several experiments which highlight different roles of lateral and medial OFC are chosen and the model was made to solve the tasks in the experiments, analyzing either the dissociation or interaction between both the sub-regions.

Chapter 7

Experimentation

Sommaire

As mentioned in the previous chapter, what looked like an apparent dissociation could very well also be an interaction between both the sub-regions - lateral and medial OFC. Also, given the number of sub-systems implied together with both the sub-regions, the analysis of the comprehensive model is restricted to behavioral performance to rather than in-depth analyses of the entire network. Moreover, most of the sub-systems used within the comprehensive model were the models which already accounted for some in-depth analyses in their respective contexts. To investigate the behavioral performance of the model, the experiments were chosen such that the role of ACC can be investigated at a preliminary level and then move to tasks that are more centered on the distinct roles of

lateral and medial OFC. Involving ACC, a task similar to standard 2-arm bandit task but with added action costs, was chosen. As an investigation into a possible dissociative role of lateral and medial OFC, a 3-arm bandit task is used, that was earlier studied in a lesion study in monkeys (Noonan et al. 2010). The setup of the task is such that it evokes several observations raised in the theoretical review of OFC in chapter 4. Finally, a task where, not dissociation but a combined role of lateral and medial OFC is assessed. A 2-stage Markov task, which the current model, though not explicitly designed for, with minor adaptations is tested with the model. The resulting behavior is compared to relate to one of the prominent questions in the computational theory of reinforcement learning (RL) - model-based vs model-free RL (Dayan and Niv 2008).

7.1 Learning vs Value comparison

First the performance of an existing model of decision-making and learning on a 2-arm bandit task with probabilistic reward is described (like the scenario described in the previous chapter 6.8). The advantage of generic nature of the task highlights the fundamental dynamics of the model. Then it is shown that the model presented here with the distinct description of lateral and medial OFC replicates the results of basic model, robustly and in more realistic timescales. Further complementary findings of separate lesions (simulated) of the lateral and medial OFC components in the model are presented. The effect of these findings on the performance in different task contingencies is discussed, replicating a neuroscientific evidence found in monkeys with lesions to different subregions of OFC.

7.1.1 2-Arm Bandit Task and Probabilistic Reward Learning

Multi-arm bandit task is a classic reinforcement learning problem that has been used in the study of decision-making in experimental (Noonan et al. 2010; Pasquereau et al. 2007; Walton et al. 2010) and computational neuroscience (Garenne et al. 2011; Guthrie et al. 2013; Topalidou et al. 2015). Typically, in an N-arm bandit task, there are N

possible cues (bandits) each carrying a different probability of reward and requiring a particular action to do, in order to select the cue. Fig. 7.1A shows an example trial of a 2-arm bandit task that has been used to study the computational models of probabilistic reward-based learning involving the basal ganglia (BG) (**Nallapu2016**; Topalidou et al. 2015). In this case, cue is one of the four possible shapes. The reinforcement in the model during the task is driven by the probabilistic reward offered at the end of each trial, with a different probability for each cue. It has been shown that monkeys learn to perform the task (Pasquereau et al. 2007), learning the reward contingencies over time and choosing always the best rewarding option after learning.

The basic model (referred hereafter as OFC model) is a set of inter-connected CBG loops and an associative network (ASC), each network processing different information and contributing for a decision within the network (Fig 7.1C). In each trial, the CBG_{cue} labeled 'limbic' takes as the input, the activation for the shapes that are presented in the trial. This activation represents a constant visual salience component, that in the simplest case, is same for every stimulus (shape). Similarly the other CBG position loops (CBG_{pos}) takes as the input, the activation of the positions where the shapes are presented. Since the positions are chosen randomly and carry no significance in obtaining reward, there is no value-learning in this CBG_{pos} loop. Hence the activation of a position represents just the presence of a cue at that position. Finally, the ASC network takes as the input, the combined information of binding specific shape to a specific position. The ASC network represent the associative loop through lateral PFC and the dorsomedial striatum (DMS) which is believed to represent a multi-modal information of stimulus-vs-position mapping (Funahashi et al. 1989). This is implemented in the form a 2 dimensional mapping for each shape against all possible position and each position against all possible shapes (Fig 7.1B, blue squares). The networks are inter-connected in such a way that while each of the CBG loops independently processes the information that it is activated with, it also affects the activities in the other through the ASC network. The network architecture within each CBG loop that guarantees the resolution of competition between the options is based on classical BG pathways that have been previously explained with computational

accounts (Guthrie et al. 2013; Leblois 2006; Topalidou et al. 2015).

In each trial of the task, the model is presented with pseudo-randomized pairwise presentation of the four possible shapes in any two of the four possible positions ('Cue presentation' phase in Fig 7.1A and first 6 panels in Fig 7.1D). Although the performance of the model is assessed in terms of the shape it chooses for optimal reward probability, the choice is confirmed only if the corresponding position of the shape is chosen as the 'motor' decision (Fig 7.1A, black + sign under 'Decision' phase, dashed lines under 'CBG' in Fig 7.1D). Thus, after the 'Decision' phase of the trial, the shape at the chosen position is considered as the choice of 'cue' and the reward is delivered according to the predetermined probability associated to that cue.

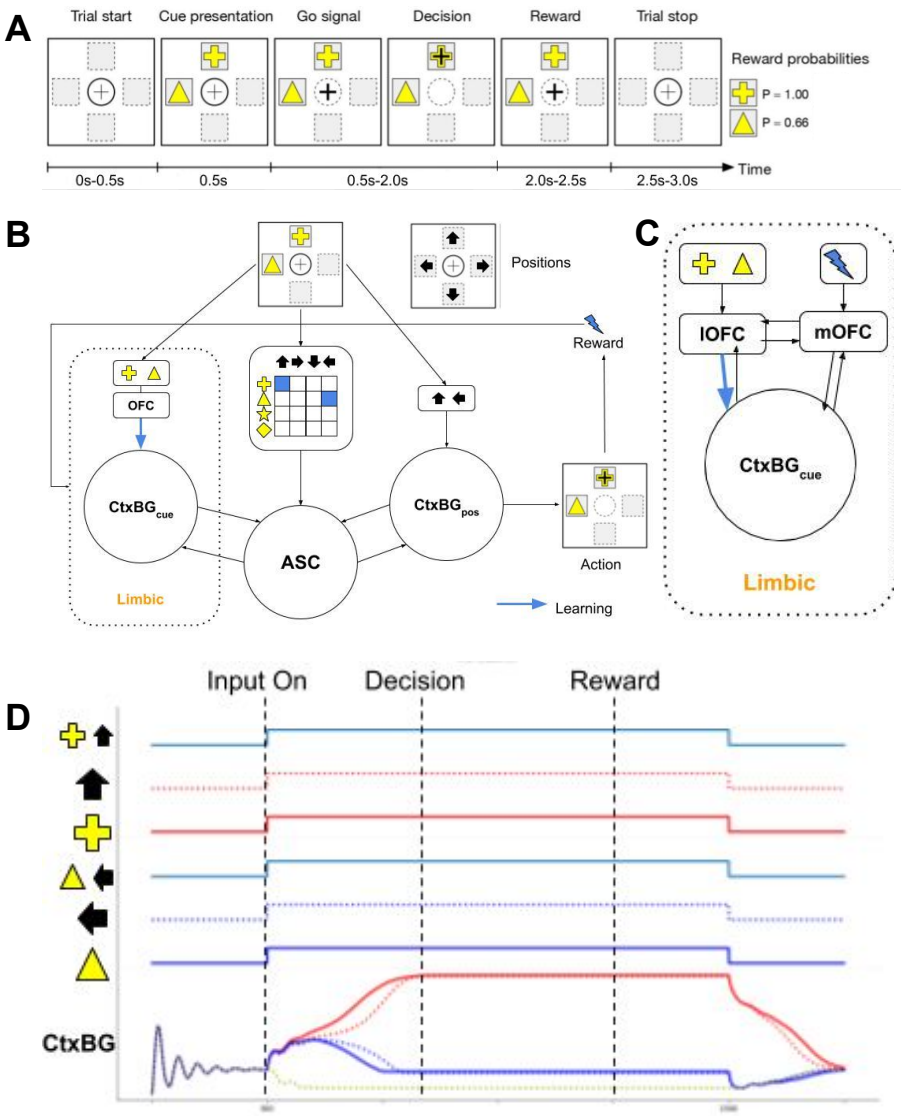


Figure 7.1: **2-arm bandit task. A.** Sample trial from a 2-arm bandit task. Out of four possible shapes (cues) are shown at two random positions (of the four cardinal positions). The position that is chosen implies the choice of the shape made. **B.** Basic model involving two CBG loops and an associative loop (ASC), one CBG loop leading to a choice between the two cues and the other between the two positions. The final output that is considered from the model within a trial is that of the decision of CBG position, the cue shown at the *chosen position* is considered as the *chosen cue*. Note the CBG Cue is labelled *limbic*, as it will be developed more into components representing sub-regions of the OFC. Blue arrow represents the connection that can be modified by learning. **C.** The proposed change in the original model which will be described in detail in the following section. **D.** Activation of each cue that is shown in a choice, its position and the combined information. Also, the evolution of activity in a CBG loop - solid lines for cue, dashed lines for positions.

The performance of the model is demonstrated under two conditions : EASY and DIFFICULT. EASY is the condition where the reward probabilities related to each shape are fairly separated and DIFFICULT is the condition where the reward probabilities are either lower or closer, thus making the reinforcement difficult (Fig 7.2A). The effect of learning in the model after each trial can be observed in terms of the decision times over the duration of the task. A decrease in decision times of both cue and position is observed (Fig 7.2B, left). A running average over the choice of 10 trials is considered for the performance over 120 trials. The performance of the model under the EASY condition replicates animals' behavior (Pasquereau et al. 2007) (Fig 7.2D, blue). In the DIFFICULT condition (Fig. 7.2A, right), the reward probabilities of both the shapes are lower or closer. This should result in lower rate of reinforcement and thereby make it difficult to make a correct choice. Animals however, with considerable amount of training, were shown to identify the option with more chance of reward and thus make correct choices (Noonan et al. 2010; Walton et al. 2010). The same model is tested in this case as in the previous EASY case (Fig 7.2A, left), but the model couldn't learn the appropriate contingencies well. The Decision Times (DTs) were longer compared to the previous case (Fig 7.2C) and the overall performance was sub-optimal (Fig 7.2D, red).

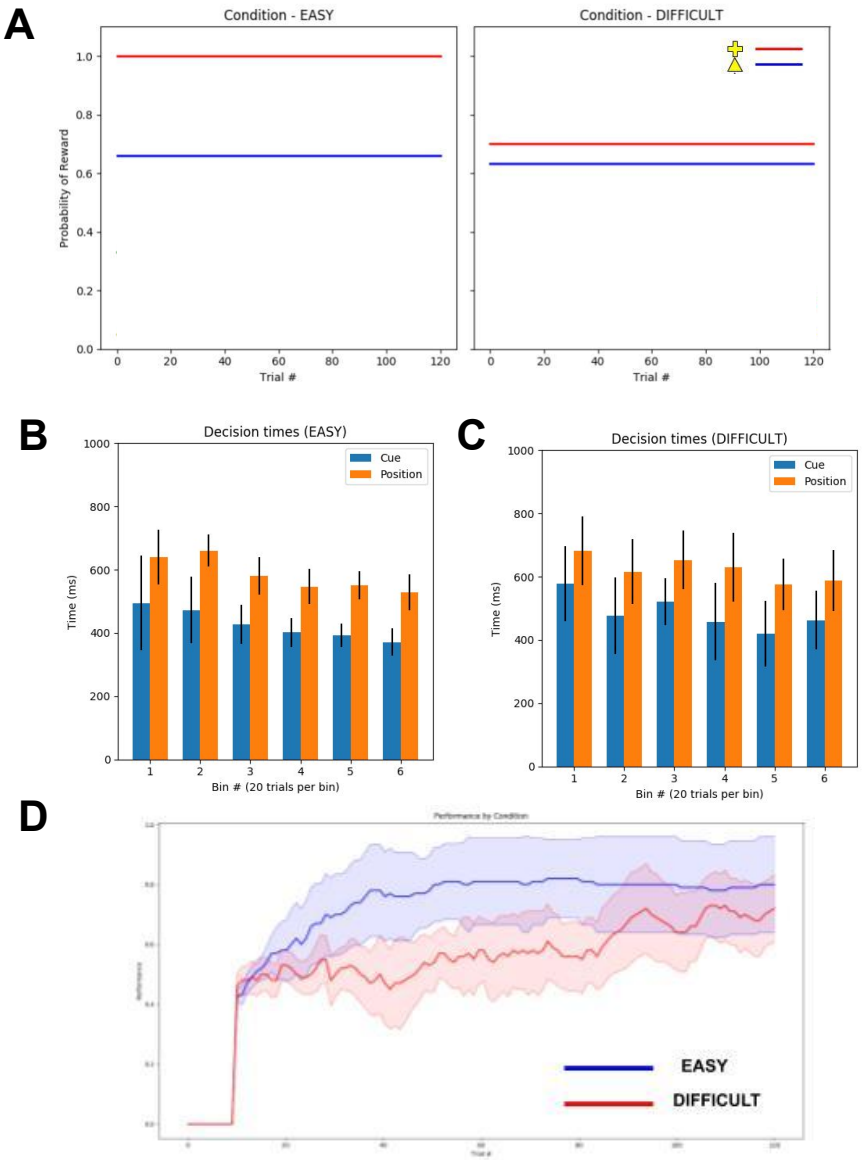


Figure 7.2: **N-arm bandit task.** The task described in Fig 7.1**A** has two possible cues (shapes) each with a predetermined probability of reward upon choice. **A.** Left and right figures show two different reward probability schemes, in EASY and DIFFICULT task scenarios. Each color represents a particular shape, as in the legend of right sub-figure. **B, C.** Decision Times (DTs) in the model after cue presentation. 120 trials are divided into 6 bins, with 20 trials per bin, and the DTs of both cue decisions and the position decisions are averaged per bin. **B** shows the DTs in EASY condition of the task and **C** in DIFFICULT condition. **D.** Performance of the model. Running average of number of correct choices across 10 trials, averaged over 10 sessions. Correct choice means the shape that rewards the most according to the predetermined probabilities. Lighter color filling represents the standard deviation.

7.1.2 Precise Value Comparison

Then the 'limbic' CBG loop is extended to individually describe two separate CBG loops - one representing the lateral OFC and the other representing the medial OFC. Here after this version of the model will be referred as *lmOFC* model. The CBG loop involving lateral OFC builds on the top of the single limbic loop from the basic model (described in Fig 7.1B). In addition to the activation (I_{ext}) to the network, a Current Subjective Value (CSV) for each shape is also added to the input. CSV represents the subjective value of a shape at any moment taking the externally learned reward contingencies and internal bodily desire for the reward that the shape leads to (see *Materials*, CSV). Another key aspect of lOFC is that it properly assigns the obtained reward to the appropriate choice made in that trial (referred as *credit assignment*). There has been evidence that neurons in lateral OFC are particularly active after the reward delivery in a choice (Bouret and Richmond 2010) and also the fact that medium spiny neurons neurons which are extensively involved in decision-making are consistently active for a while after reward delivery (Bissonette et al. 2013). These evidences support the possibility that cortico-striatal synaptic plasticity is a plausible phenomenon in the context of obtaining reward. Similar arguments were made by other experimental findings (Walton et al. 2010).

The CBG loop with medial OFC receives input from the CSV layer. Medial OFC has a separate value comparison mechanism implemented as a simple 'recurrent excitation lateral inhibition' model, activated by the CSVs received. It was shown that the activity in medial OFC correlated to the value difference between the options (FitzGerald et al. 2009). Supporting the view that the relative difference of the presented options is represented in vmPFC, multiple value comparison mechanisms have been proposed. This value difference signal further allows vmPFC to perform a value comparison to facilitate the choice through principles of recurrent excitation and lateral inhibition (Grabenhorst and Rolls 2011; Rolls et al. 2010; Strait et al. 2014; Wang 2008). The output activities of mOFC are fed into its CBG loop. It has been shown that one of the general function of populations in the PFC is to maintain history of decision events such as previous action,

previous reward etc (Tsutsui et al. 2016). Accordingly, a simple history of rewards in mOFC, without cue-specific information is implemented. As the lOFC maintains the current choice until the reward delivery and later (Bouret and Richmond 2010), possibly a history of choices is maintained in lOFC. It was shown that lesions to lOFC affect the appropriate consolidation of the reward history with the choice history (Walton et al. 2010). Hence, for the sake of simplicity, both the histories in lOFC and mOFC are combined within the lOFC to provide a combined choice-reward history up to one previous choice and reward, and it is fed into the CBG loop of lOFC along with the activation of the cue. In addition, a synaptic connection is added to the ASC layer outside the limbic network, from each cue population in lOFC to all the possible position populations in the 2-D mapping of ASC network. However the learning in these connections would be less influential on the decision, compared to that in the lOFC network, since the learning happens with respect to all four possible positions corresponding to the cue in the 2-D mapping of ASC (Fig 7.1B, input to ASC). Although it has been shown that mOFC / vmPFC encodes action-outcome associations in several task settings (Daw et al. 2006; Hampton et al. 2006; Kim et al. 2006; Tanaka et al. 2016), the design of the n-arm bandit task setting does not allow much of learning action-values. The reason is that the task randomizes the positions where the cues are present and hence the action required to chose a cue.

Then the lmOFC model is tested on the DIFFICULT condition as in the previous task. The model performed considerably well compared to the previous OFC model, with much faster DTs. Both the models have an estimated value difference for the ongoing task, across all the trials. Interestingly, the precise value comparison in mOFC estimates the value difference across all the trials better than that estimated by the OFC model under DIFFICULT condition (Fig 7.3D).

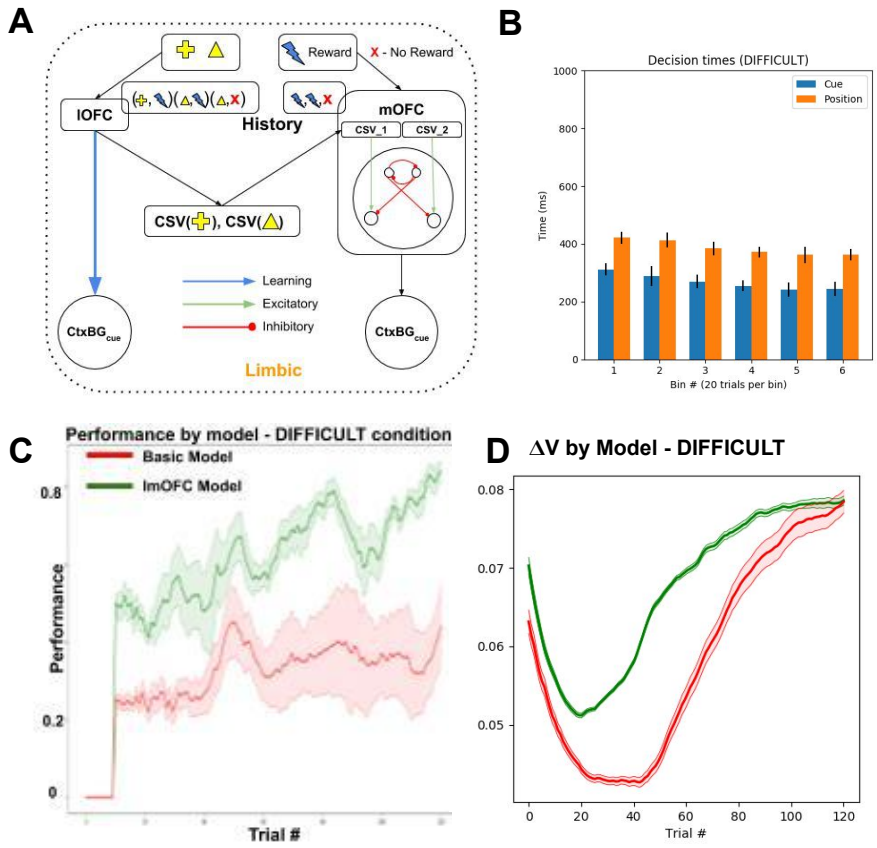


Figure 7.3: **lmOFC Model : CBG loops with lateral and medial OFC. A.** lmOFC Model. Changes in the 'limbic' CBG loop, compared to the basic model. Lateral OFC (lOFC) has access to cue identity (shape), hence drives learning the connections to its CBG loop. lOFC also activates the Current Subjective Value (CSV) for each of the presented cues from elsewhere. Medial OFC (mOFC) has a value comparison mechanism to compare the CSVs of the presented cues it receives. mOFC further drives its CBG loop with the ongoing value comparison outputs. Both lOFC and mOFC also maintain general history of chosen cue-reward association and reward respectively. This input is also used in the activation to their respective loops. **B.** The average DTs of decisions choosing cue and position, across 120 trials binned every 20 trials. **C.** The performance of the lmOFC model (green) in comparison with the performance of the basic model in DIFFICULT condition (red). **D.** Average value difference of the presented options estimated in lmOFC model (green) and basic model under DIFFICULT condition (red).

7.1.3 Proximity of Values and Decision Making

The lmOFC model is then tested on a 3-arm bandit task (Fig 7.4). Each of the three cues that are shown in every trial has a reward probability upon its choice. As shown in Fig 7.4, V1, V2 and V3 are the reward probabilities associated to the cues *plus*, *delta* and *star* respectively in a given experimental session. The task is carried out under three different reward schedules (Fig 7.5A-C). In all the sessions, V1 and V3 are fixed to be .7 and 0.05. V2 value is changed across three types of sessions : V2_HIGH, V2_MID and V2_LOW where V2 is set to 0.6, 0.3 and 0.1. Similar task schedule was used on animals to test the effects of lesions of lateral and medial OFC separately (Noonan et al. 2010).

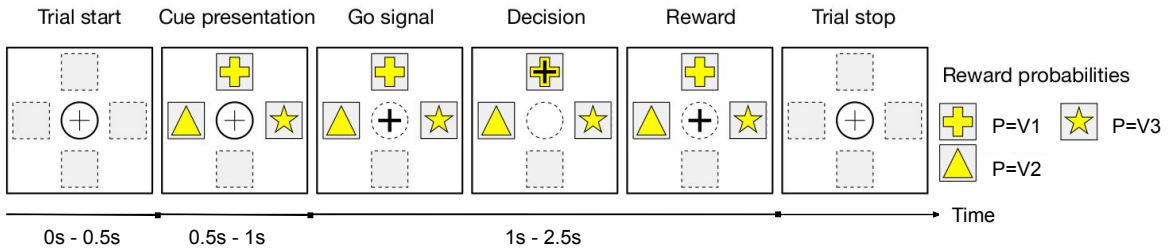


Figure 7.4: **3-arm bandit task** A sample trial from the 3-arm bandit task. Three possible shapes (cues) are shown in three random positions (of the four cardinal positions). The position that is chosen implies the choice of the shape made. Upon selection of a shape, a reward is delivered with a probability (p), which is different for each of the shapes (V1, V2 and V3).

The lmOFC model is used without any changes. At the time of presentation in every trial, 3 cues are activated simultaneously (along with their positions in the CBG_{pos} loop and in the ASC network). In terms of the model parameters, just the input activation which represents the cue salience had to be increased as compared to when the choice was between 2 options (as in the previous tasks). In this task, a correct choice or a good choice is a V1 choice. The model reached optimal performance (more than 80% V1 choices) in less than 150 trials in each session, in all three reward schedules (V2_HIGH, V2_MID and V2_LOW). This is referred as the 'Control' condition, green in Fig 7.4E-J.

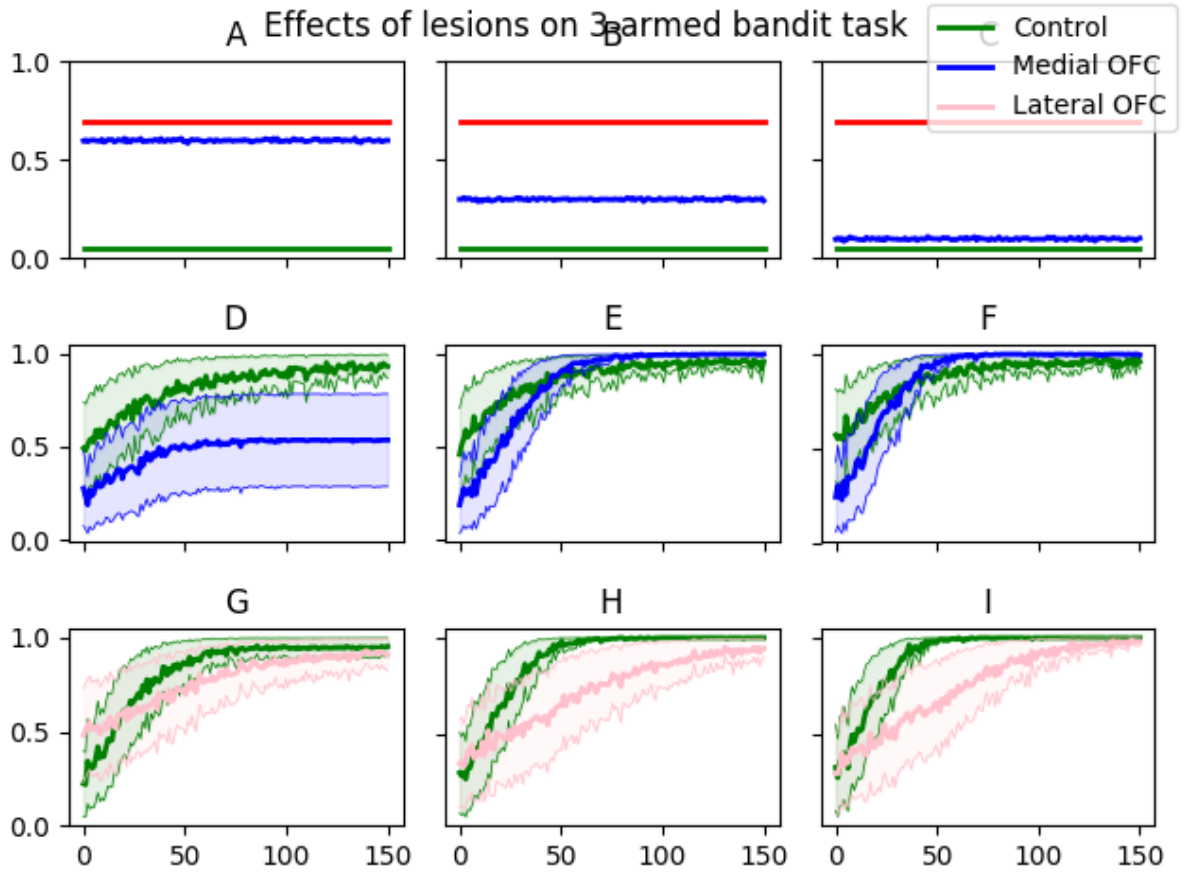


Figure 7.5: **Effects of lateral and medial OFC lesions in the model. A-C.** 3 conditions of the task (each column). Of the reward probabilities V1, V2 and V3 described in Fig 7.4, V1 and V3 are fixed in all 3 task conditions (A-C, red and green respectively). The 3 conditions depend on the value V2 (A-C, blue) : V2_HIGH, V2_MID and V2_LOW. **D-F.** Lesion of mOFC under each task condition. The average performance in each condition with a lesion to mOFC (blue) is compared to the control performance (green). **G-I.** Lesion of lOFC under each task condition. The average performance in each condition with a lesion to lOFC (pink) is compared to the control performance (green).

Furthermore lesions of lateral and medial OFC are simulated in the model. Since the model generates a decision through at least one of the 'limbic' CBG loops and the other ASC and CBG_{pos} loop, even in case of a lesion to lOFC or mOFC, a valid decision should be made. First the changes in the model with respect to each of the lesions and the corresponding results are described .

Medial OFC Lesion

A lesion of mOFC to the lmOFC model shown in Fig 7.3A makes the model slightly similar to the basic OFC model described in Fig 7.1B. The credit assignment still works by lOFC during the period after reward (Fig 7.1D, CBG, after 'Reward' phase), because the identity of the chosen cue maintained in lOFC is available for learning after the reward. However, in the absence of mOFC, the input to lOFC from the ongoing precise value comparison in mOFC is absent.

In all the control experiments, the model reached optimal performance within 150 trials. In the case of medial OFC lesions however, the performance was significantly impaired in the case of V2_HIGH scenarios, when V1 and V2 values were proximate. In the case of V2_MID and V2_LOW, the performance was observed to be similar to that of controls, except for a slight delay in reaching better performance. Such a normal performance in the case of V2_MID and V2_LOW can be attributed to the appropriate *credit assignment* by lateral OFC happening during learning. When the value difference is sufficiently large, and the credit correctly assigned to the correct choice, as the V2 anyway does not reward as much as V1, it is easily learned between the lOFC-CBG$_{cue}$ synaptic connections and they can drive the decision without a precise comparison (Fig. 7.4 D-F).

Lateral OFC Lesion

One of the major changes in case of the lateral lesion is the credit assignment. In the control condition, when there is a reward delivered, the activation of the chosen cue in lOFC is active (Fig 7.1D, CBG, after 'Reward' until 2500ms). When there is no lOFC in the network, the association of current reward to only current choice can no longer be done. In this case, we still consider that the CSV for each cue is sent as an input to mOFC, because mOFC/vmPFC has been shown to receive projections from the ventral striatum (Carmichael and Price 1995a; Carmichael and Price 1995b; Haber and Knutson 2010), which is a crucial component of the CSV layer.

Striking of the observations, in the case when the difference between V1 and V2 was

close, monkeys with mOFC lesions showed impaired performance while the controls and animals with lOFC lesions fairly performed well as V1 and V2 were distinct. Such an impairment argued for the role of mOFC to be more sensitive to the value difference between the options. Conversely, when the difference between the values was more, quite surprisingly animals with lOFC lesions were impaired whereas the controls and animals with mOFC lesions could steadily perform optimal choices. While it was an interesting observation to see how mOFC could not compensate even when the difference between the values is high (meaning it is an easy choice), it highlighted the role of lOFC in appropriate credit assignment, i.e. assigning the reward to the appropriate choice made in the current trial rather than to the previous or even the succeeding choice or even to the choice that rewarded the most historically.

In the case of lOFC lesions, the performance was affected in rather contrasting manner. Although eventually the performances reached near-optimal in all three cases of V2_HIGH, V2_MID and V2_LOW, the performance was sub-optimal for most of the earlier part of the sessions especially in the cases where the value difference was larger. This highlights the importance of lOFC in appropriately assigning the credit of reward to the correct option. Impairment of performance in the absence of lateral OFC in the case of V2_MID and V2_LOW was observed for the initial part of the session. This may be due to partial learning in the form of reward-based history maintained in medial OFC (as it was maintained when lateral OFC in intact) (Fig. 7.4 H-I).

7.1.4 Discussion

The OFC is positioned on top of classical sub-cortical decision-making systems, with the descriptions of experimentally observed roles of its individual sub-regions. The seemingly dissociate yet more complicated effects of the sub-regions of the OFC on the task performance depending on the task structure (value difference between the options) are described. The OFC is clearly a crucial prefrontal region with heterogeneous representations and dynamics that result in complex behavior. Therefore clearly it is not a feasible

idea to attempt a simplistic representation that relies on a unique way of information processing within the OFC, without implying several other brain regions that closely interact with the OFC during the behavior. Instead, the positioning of the OFC in the grand picture of several prefrontal and sub-cortical brain regions is acknowledged, as well as the heterogeneity within itself. Before attempting to model the possible mechanisms within the sub-regions of OFC in detail, it is crucial to build a framework that embeds a representation of environment in an embodied manner (with bodily needs and relevant behavior protocols for testing). Hence this work points towards the interest for modeling the dynamics of OFC as a part of a larger framework of related brain systems that the OFC interacts with, and the valuation systems it employs to guide decisions and learning. One of the well-accounted frameworks of decision-making and reinforcement learning involving the BG (CBG loops) and complemented it with the specialized representations of the sub-regions of the OFC was chosen. The simplistic model as in Fig 7.1B is shown to be sufficient for simple tasks. Further it is demonstrated that a more informed decomposed model (*lmOFC* model, Fig 7.3), while performing equally well on the simple tasks, also allows to study performance on more complex tasks in which the basic model cannot perform well.

Choice

Although the primary role of lateral OFC has been implied to be appropriate credit assignment, here we use the fact that lateral OFC still plays an important role in driving the activities in the downstream BG loops with the dynamic subjective values added to the visual salience. Since the synaptic weights that are changed through learning are the ones connecting lOFC to the CBG_{cue} loop, the dynamics would passively favor a choice whose connection weights have been sufficiently learned. For instance, in the case of a lesion to mOFC (Fig 7.5), the initial decisions before any learning may be guided by lOFC through the CBG loop, randomly with the help of intrinsic noise. However, as in the case of Fig 7.5E,F where only cue rewards significantly more, the learning can rapidly

increase the synaptic weights in the network corresponding to that cue and thus guide the subsequent decisions to that cue. This could be one reason why the performance slowly picked up towards the latter end of the trials, in the cases of V2_MID and V2_LOW in case of medial OFC lesions. Whereas in the case of mOFC lesion under V2_HIGH condition, since V1 and V2 almost similarly reward, even if appropriate credit assignment is done between the cues corresponding to V1 and V2, the network may not be able to definitively guide the decision necessarily towards V1.

Learning

Learning in the system occurs at the level of both CSV (for expected values) and cortico-striatal synapses. The learning that occurs at the level of cortico-striatal synapses indirectly represents the reward contingencies of stimuli in terms of their probability. One of the possible motivations behind multiple learning mechanisms in the system is the feasibility of a shift of control from the value-comparison based processes in the mOFC/vmPFC at the beginning of the trials to a faster, network strength based decision through the lOFC-BG loops driven by the learned connection weights, in the trials after substantial learning. Such a distinction was reported where activities in vmPFC were more remarkably distinct between more deliberative situations with slower reaction times as opposed to trials towards the end of the experiment or even no-brainer trials (highly probable high reward versus the opposite). Moreover, the involvement of value-difference signal in vmPFC consistently decreased towards the later trials of the task (Hunt et al. 2012). However, it is important to note that, in a different formal description, it has been highlighted that ventrolateral PFC (vlPFC) encodes the *Availability* (probability) of rewards whereas the OFC was shown to encode the *Desirability* (palatability) of rewards (Rudebeck et al. 2017). However it was shown activity in medial and lateral orbitofrontal cortex, extending into vmPFC, was correlated with the probability assigned to the action actually chosen on a given trial (Daw et al. 2006).

Lateral and Medial in Learning and Choice : Dissociation or Interaction ?

What is important to note here is that the dissociate effect observed does not necessarily imply that both the sub-regions have dissociate roles in decision-making and learning, as recently suggested (Miller 2018). The dissociation might be observed in terms of the anatomical connectivity in the sense that lateral OFC predominantly receives inputs from sensory regions about external environment whereas medial OFC has more inputs from internal bodily states and visceral responses. But we argue that seeming dissociation in terms of internal processes within these sub-regions is a possible network effect as a result of their temporal dynamics. Because, as we have shown here, both the sub-regions are involved in the circuitry that is capable of both guiding decisions and learning from outcomes. Albeit, the dissociation might be apparent because of the fact that each of them might have access to different information about the state of the environment and exert control at a different stage of the behavior.

7.2 Better rewarding v/s closer choice

Imagine the scenario of probabilistic reward based choice task same as the one described in section 6.8, except that the pillars at each choice point are not equidistant. Pseudo-randomly, one of the pillars will be in the *See* zone and the other in *Appear* zone. In the framework described in this work, to be able to estimate action costs to reach a particular pillar, information is needed from different structures. The identity of the pillar from sensory cortex, the binding information of the pillar from lPFC - the signal the position of the pillar and the strength of the signal from sensory cortex to decide which zone the pillar is in - *See* or *Appear*. ACC is at the center of these regions and receive connections from sensory areas and lPFC. And most importantly, ACC also is involved in reinforcement of actions as closely as OFC in general.

The network is activated exactly as in the case of probabilistic reward task (section 6.8) and in addition activation in ACC pairing the pillar with the pseudo-randomized

distance (*See* or *Appear*). The reward given probabilistically was always the same for any choice except for the probability with which it would be awarded. The action cost introduced is assumed to be of the same dimension as reward (reward was an apple related to the need hunger, hence the action cost would a slight increase in hunger for the effort made).

Results

After learning for 120 trials with the action costs chosen high as $AC_i^z = .5$ for z=*Appear* and 0 for z=*See*, as can be seen in the figure 7.6, almost all the choices are the closest offered ones, except for the exceptions highlighted in black columns. It can observed that the exceptional cases are when the rewarding probability of the farther option is so high that it will be chosen even if it is far (p=1 vs p=0 and p=1 vs p=1/3).

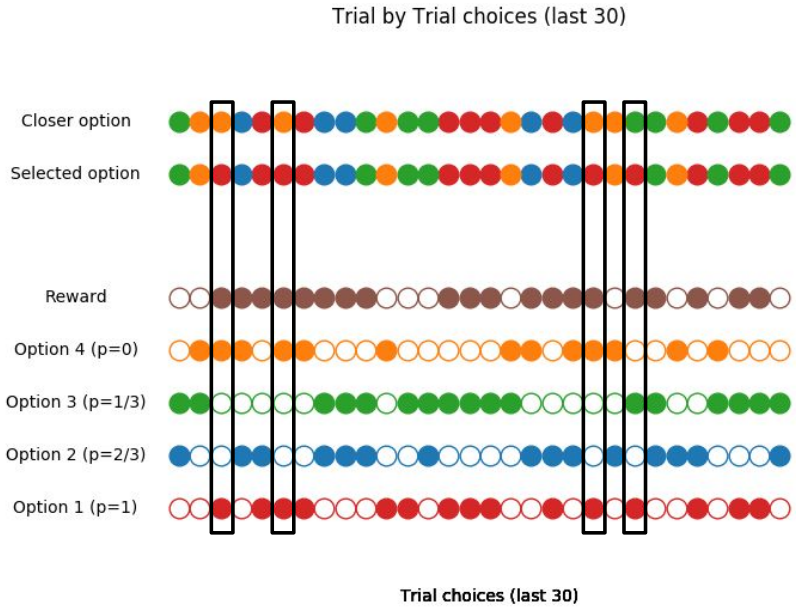


Figure 7.6: Closer choice vs Rewarding choice. Last 30 trials out of a 120-trial simulation where the reward probabilities are learned with each option (A, B, C and D). Of the two options presented each trial, pseudo-randomly one of them is closer (lesser action cost) and the other farther (more action cost). The choice made by the agent is compared to that of closer option.

7.3 2-stage markov task

2-stage Markov task is a variant of 2-arm bandit task, where at any given decision point, the choice is between 2 options. In the computational theory of reinforcement learning (RL), this task has been used widely in relation to understanding the model-based and model-free variants of RL, comparing against behavioral performances of animals (Daw et al. 2011; Dezfouli and Balleine 2019; Gläscher et al. 2010). Each trial has two stages. In the first stage, the choice leads to another 2-option choice without any outcome. In the second stage, the choice from both the options leads to a probabilistic reward. In the first stage, the choice is always between two cues (referred as S1, Fig. 7.7, shown as lightning and sun). In the second stage the choice could be between one of the two possible pairs. Stage 1 could lead to either S2 (Fig. 7.7, a triangle and a "plus") or S3 (Fig. 7.7, a star and a diamond). The transitions to S2 or S3 depending on the choice at Stage 1 are probabilistically controlled, in this case 0.7 ("common" transition) and 0.3 ("rare" transition) for S2 and S3 respectively upon choosing "lightning" and vice-versa upon choosing "sun" (Fig. 7.7, thick green/blue arrows). Therefore each choice at stage 1 has a respective "common" and "rare" transition to the choice in stage 2. And eventually, upon each stage 2 choice, a probabilistic reward is delivered according to a probability associated with that option. These reward probabilities were distributed between 0.25 and 0.75 and were diffused by adding independent Gaussian noise (mean 0, SD 0.025) at each trial (similar to the task in Daw et al. 2011).

This task was not performed in the virtual environment. The model was directly tested by presenting numerical simulated trial presentations. To be consistent with the previous aspects of the model, although the task does not require, the positions of presentation were pseudo-randomized across all possible four positions.

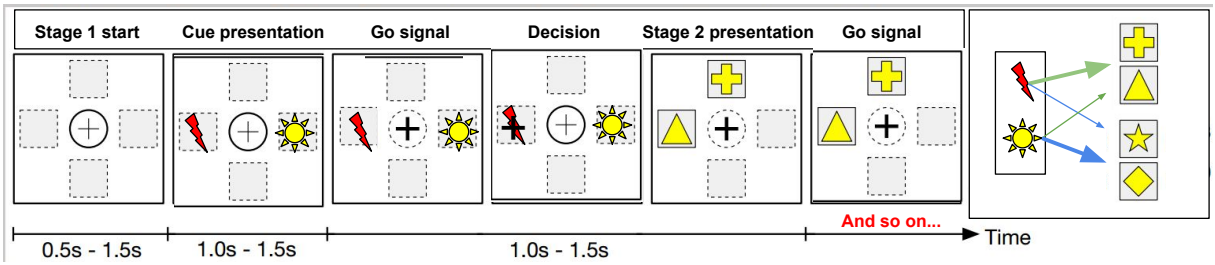


Figure 7.7: A 2 stage Markov task.

The model in its original form in 6.8 was not designed for a 2-stage task. But a minor adaptation in the presentation phase to retain the activity of stage 1 choice, allowed the model to perform the task, as in essence it is not too far from a standard 2-arm bandit task. The key change is in the way the model should learn. Since the first stage has no reward, and only the reward from second stage has to contribute to the learning in the first stage. The implementation of SPE in lateral OFC managed to learn the transition probabilities of the resulting states and further these probabilities are used in updating the values of first stage options. Although this is the case in any other 2-arm bandit task, it is not prominent because either the state change probabilities are the same (equally random) like in section 6.8 or they do not change at all, as in the case of the 3-arm bandit task in section 7.1.

Results

In each trial, the model leads to a stable choice as in the previous tasks (Fig. 7.8, first, middle and the last trials of the session of 240 trials). As there is not obvious optimal choice, the performance of the model is not assessed as was done in the previous tasks. Instead, since the key aspect to understand is the differential contribution of attributes of lateral OFC such as state space learning and the effect of recent choices, the performance was evaluated in terms of whether the model "stays" with the previous choice of stage 1 for the following trial if the second stage of the current trial rewarded. The performance of the model in terms of the fraction of "stay" is reported, for each of the cases - two cases of rewarded and unrewarded each for a common transition and a rare transition. The results

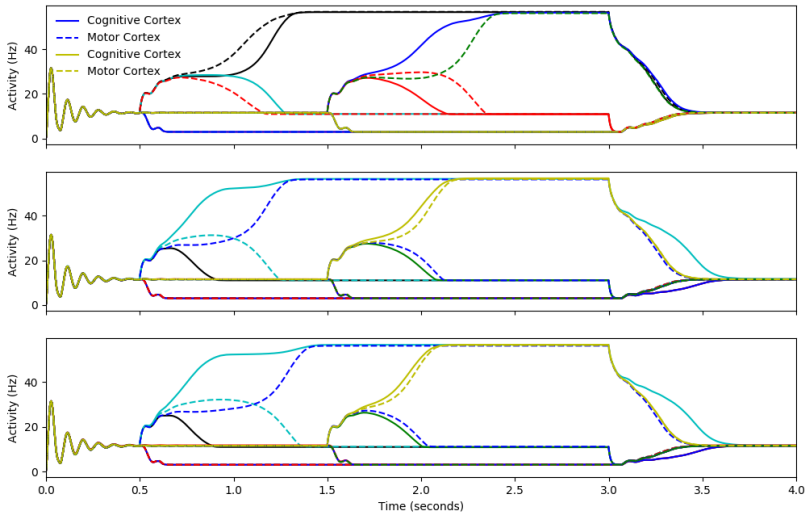


Figure 7.8: Decision activities in the model from the first, middle and the last trials of the total 240 trials. The first stage choice is always, as represented in colors, cyan and black. The second stage choice is some times blue vs red (first trial) or yellow vs green. The positions are shown as dashed lines.

are compared to the performances of humans performing the similar task (Daw et al. 2011) and also in rats (Dezfouli and Balleine 2019), shown in the figure 7.9. Theoretically, if the reward after stage 2 choice is directly attributed to the stage 1 choice in each trial, without taking into account the transition probabilities, then the model should be more likely to ”stay” with the decision in all the rewarded cases and less likely to ”stay” following unrewarded trials. On the other hand, if the transition probabilities between the stages are strictly taken into account, the likeliness to ”stay” should be higher when a ”common” transition was rewarded and a ”rare” transition was unrewarding. Likewise, the likeliness to ”stay” should be lower when a ”common” transition was unrewarding and a ”rare” transition was rewarding. It can be seen that the performance matches neither of these theoretical expectations but qualitatively similar to those of behavioral studies in humans and rats.

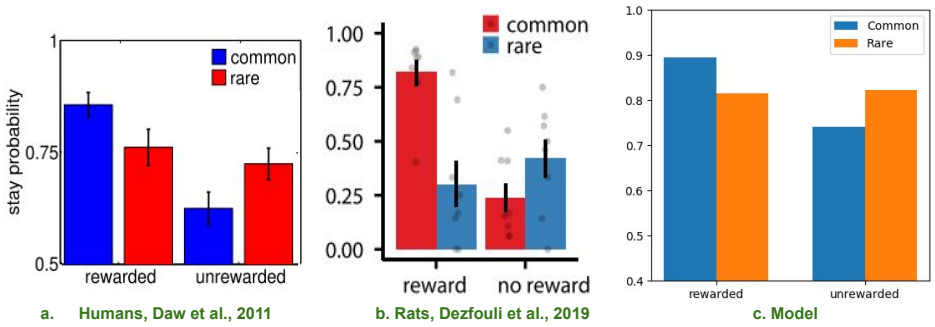


Figure 7.9: Model performance in terms of stay probability, the likeliness to "stay" with the stage 1 choice of the previous trial. The performance is different from theoretical expectations, qualitatively similar to that of experimental studies. **a** and **b** adapted from Daw et al. 2011 and Dezfouli and Balleine 2019 respectively. **c.** results from the model.

Conclusion and Perspectives

Sommaire

I . Discussion

This work primarily aimed at understanding the generic principles that contribute to flexible (naturally intelligent) behaviors in high-functioning animals that can be applied for the development of intelligent agents. The work has been carried out across two major axes : (i) Closed-loop experimental framework of voluntary behavior (ii) Representations and processes of value-based decision-making in OFC. As divided into several precise goals in the chapter 5, now these goals will be verified as to what extent they were achieved.

II . Closed-loop experimental framework of voluntary behavior

Goal I : Describe an experimental framework to study voluntary behavior

eventually a generic framework to study artificial agents was described. The key attributes besides the agent and the environment itself, that are crucial for a minimalistic representation of voluntary behavior are pointed out. Not only the elements of the environment, but the crucial processes within the agent itself, interactions between the agent and the environment, in conjunction with the other elements are essential to elicit voluntary behavior. It has been established that the closest reference to voluntary behavior in changing environment is the animals trying to survive in an open world. Thus these building elements and the processes involved in the agent's behavior have been grounded against those of a voluntarily behaving animal. Most importantly, agent's emotions and motivations are highlighted as the fundamental driving factor of behavior. This is not the first time that modeling emotions and motivations forms the basis of understanding intelligence in artificial agents (Canamero 1997; Sequeira et al. 2011).

Goal II : Adapt a virtual environment according to the behavioral framework

Further, a video game environment, Minecraft has been identified as a suitable experimental testbed that would provide scenarios similar to naturalistic settings that animals survive in. Malmo (Johnson et al. 2016), a software platform provided by Microsoft, conviniently allows to communicate with the video game Minecraft, thus allowing the behavioral framework to interact with the environment in the form of agent and the bodily characteristics the agent takes. Since the characteristics of the agent were desired to be animal-like, certain adaptations were made to the otherwise purely software platform Malmo. These adaptations provide a great flexibility to apply biological principles to

drive the behavior of the agent close to that of an animal.

Goal III : Summarize global organization in Brain underlying flexible behavior

As proposed in Alexander 1986 and compiled well in Alexandre 2016, the brain of a high-functioning animal is known to be organized well into parallel feedback loops, at least in the regions that drive the animal's voluntary, and even flexible behaviors. A great part of the description as shown in figure 1.1 in chapter 1 has inspired the cognitive architecture that has been implemented in this work, representing the generic parallel loops that segregate the information about the environment into four fundamental questions - *What?, Why?, Where? and How?*. Most importantly, such a segregation of information in the architecture allows to study them individually but still in a combined setting to account for the effects of individual loop on the rest of the network.

Goal IV : Integrate a biologically-inspired model of parallel loops in the virtual environment

This framework of parallel loops in the brain was projected onto a cognitive framework, and implemented in an algorithmic fashion. The artificial agent depends on sensors and actions to interact with the environment. By identifying the corresponding elements in the biological framework, a seem-less integration has been done between the biologically inspired cognitive framework of behavior and the experimental platform. The agent could successfully exhibit fundamental behaviors as a result of the interacting parallel feedback loops, closing a closed-loop of behavioral interaction between the agent and the environment.

III . Representations and processes of value-based decision-making in OFC

Goal V : Implement known computational neuronal models in the framework of parallel loops.

The animal behavior is viewed as an ensemble of fundamentally different individual behavioral paradigms. Pavlovian learning, instrumental action learning and Pavlovian instrumental transfer were identified as the miminal set of behavioral paradigms that would give rise to an observable voluntary behavior in the animals. Since these paradigms are extensively studied in the field of neuroscience, several computational accounts explaining these paradigms were studied to understand the underlying the cognitive and modeling principles. Within the cognitive framework of algorithmic parallel loops that was built with the experimentation platform, the computational models of Pavlovian learning, probabilistic reward learning and the transfer of emotional value to motivational value have been implemented.

Goal VI : Replicate previous numerical simulations as virtual experiments

Although same mathematical principles were used as those in the previously accounted models, expected behaviors were replicated in the interacting agent. Most importantly, previous models with just numerical simulations do not account for basic processes of action execution and goal sustenance after explaining the decision mechanism. This was one of the key contributions by the cognitive framework. With biologically inspired mechanism of sustained activation in sensory modules to maintain the goal in a desired state, the agent is able to execute actions until the desired goals are reached.

Goal VII : Implement ACC to interact with both lateral and medial OFC

As mentioned earlier, the cognitive framework for parallel loops was built basing on biological inspiration from the parallel loops found in the brain and these loops involve prefrontal cortex (PFC), basal ganglia (BG) and sensory cortices. There are three different groups of loops (Krack et al. 2010) - limbic (representing emotion and motivation), sensori-motor (representing action in the environment) and associative loops (multi-modal representations). The prefrontal cortex (PFC) has been quite generally implicated in many high-level behaviors, often termed as *executive behaviors*. In addition, many of the roles that PFC has been implied involve memory structures like hippocampus which contribute to functions like short-term memory, working memory and high-level planning. Hence it was important to first dissociate PFC in assessing the possible roles it plays uniquely contributing to the emotional and motivational behavior of the animal. Thus the focus has been set on the *limbic* loops and commonly briefly a specific PFC region related to motivational control, anterior cingulate cortex (ACC) was studied and then the interest was shifted to Orbitofrontal cortex (OFC).

Goal VIII : Implement credit assignment in lateral OFC and core

Goal IX : Implement value-comparison system between core and medial OFC

Goal X : Implement minimalistic reward history in both lateral and medial OFC

Animal behavior, which is basically a series of responses to the sensory states based on decisions, can be essentially expressed as the consequences of value-based decision making, for most part it. And OFC was found to play wide range of different roles related to value-based decision making. A detailed survey into what OFC was implied for, and how the information is organized and processed within OFC, has been done. This survey

highlighted a strong dissociation between two of its sub-regions - lateral and medial OFC, observed to be playing very different roles, both in value-based decision-making. The computational models implemented in the framework, that were described earlier, were implemented according to this PFC-BG loops architecture. Therefore, conviniently a more detailed model of OFC is conjured, separately implementing lateral and medial OFC and complemented by ACC, with rest of the BG structures involved. As pointed by several studies, first distinct roles of each of lateral and medial OFC were considered to implement respectively. A consistent finding about the role lateral OFC, credit assignment (Walton et al. 2010) was implemented combined with the nucleus accumbens core (ventral striatum) in the model. Similarly, as it was consistently pointed that medial OFC/ vmPFC plays a key role in retrieving values from BLA/ core (Malvaez et al. 2019; O'Doherty 2011). In addition, following another recent theory that OFC might represent the abstraction of state space of the task, different attributes of the task space are implemented in each of lateral and medial OFC.

Experimental results

The effect of bias between the cortical systems was shown as a part of effort-based experiment. The same two-arm bandit task that has been reproduced using the initial model of instrumental learning using OFC and lPFC loops, has been modified to introduce different action costs to the options. The role of ACC in allowing to learn the action costs in conjunction with the emotional value, whereas the lateral OFC system could learn the emotional value of stimuli irrespective of action cost is demonstrated. The action cost was linked to the distance of the object, as the framework allows such a configuration by the virtue of its sensory zones. The results of model's choice behavior showed a combination of these two effects, choosing closer option when the options are closer in value and choosing the better option when the value difference is large even if it is farther.

In a more directed experiment to dissociate the roles of lateral and medial OFC, the model has been adapted to solve a 3-arm bandit task with some minor parameter

adjustments to the previous model which performed the 2-arm bandit task. This task was earlier studied in monkeys with different lesions - in lateral and medial OFC, performing such a task (Noonan et al. 2010). The task involved different scenarios of reward schedules of reinforcement, with the schedule fixed in each scenario. The goal of the task was to see the effect of value difference between the options on decision making. Fixing the reward probabilities of the first (V1, highest) and the third (V3, lowest) options, the reward probability of the second option was changed across scenarios. The second option was referred in each scenario as V2_HIGH, V2_MID or V2_LOW depending on whether it is close to V1, between V1 and V3, or close to V3 respectively. The results highlighted the importance of medial OFC when the values are closer and difficult to compare. Since value comparison was implemented in the medial OFC, lesions of medial OFC in the model resulted in sub-optimal performance when V1 and V2 were closer. On the contrary, lesions of lateral OFC did not affect the performance in the case where the V1, V2 values were closer (V2_HIGH). Rather, the lesions of lateral OFC affected the performance in the other two cases - V2_MID and V2_LOW. Although these two cases are apparently easier to make a decision in general, the lack of appropriate credit assignment confounds with the value learning in core. That is, any reward that was offered for the choice V1, would be wrongly assigned to any other option and eventually the value learned for each of the options tend to be near the average value of all the options. Essentially the comparison process in medial OFC is not compromised, but the values provided to medial OFC by the core are. Whereas in the case V2_HIGH, since the problem with credit assignment also causes the rewards obtained for other option choices to be assigned to V1 and thus the estimated value of V1 starts to increase, thanks to the reward-based history implemented in medial OFC. Considerably similar results were found in the study performed in monkeys (Noonan et al. 2010). In figure 7.4 H,I, it can be observed how the impairment in performance is prominent in the first half of the trials.

There is an extensive ongoing interest in the field of reinforcement learning about two of its variants - model-based and model-free learning (Daw et al. 2005). The topic is studied usually in more detail with respect to navigation strategies (Khamassi and Humphries

2012) but also more recently in the cognitive aspects Nogueira et al. 2017; Stalnaker et al. 2016, implying OFC as a candidate for learning state representations of the task. The model described in this work especially with the learning of state space and task-space in lateral and medial OFC relates to similar aspects in what is described as model-based and the usual cortico-striatal learning and decision driven through lateral OFC and core can be compared to model-free. In fact, more analogies have been drawn further considering the PIT related pathways involving lateral OFC, shell, BLA and medial OFC, core and CeN. To test how relevant the behavior of the model could be to these formal paradigms, a 2-stage markov task was performed, which was tested in humans and rats before (Daw et al. 2011; Dezfouli and Balleine 2019; Gläscher et al. 2010). Although it is still not clear if any of the lateral and medial counterparts would directly correlate to these paradigms, nor how these different kinds of learning are implemented in the brain while the general theory of reinforcement involves specific neural correlates, it could be observed from the model behavior, that neither of model-free nor model-based was uniquely observed (7.9c). Although the implementation and the results shown here are preliminary, the kind of functional connectivity used in this model with more detail involving different sub-structures of amygdala and ventral striatum can help study this important theoretical dissociation in more detail. More generally, it was pointed out that there could be mechanisms in PFC as a whole may be a suitable candidate circuitry to implement learning mechanisms similar to these theoretical paradigms, owing to its representations of the expected values of actions, objects and states, recent history of actions and rewards (Wang et al. 2018).

Value-based decision-making and learning mechanisms have been placed at the forefront of understanding voluntary animal behavior. The behavior referred in this work is naturally intelligent behavior - that is a sophistication beyond being purely reflexive or just adaptive to the changes in the environment, but actively modifying existing behaviors with flexibility. Studying this question contributes to the understanding of the general principles of intelligent behavior, simultaneously from the perspective of both the fields - artificial intelligence (for software or hardware agents with human-like intelligence and behaviors) and cognitive science & neuroscience together (for a closer look at underlying

brain processes in healthy as well as psychiatric disorders). Given that it involves studying the natural processes of animal behaviors, the field of neuroscience and the advances that have been made so far in understanding the neural correlates of decision-making offer a great starting point to understand behavior.

Experimentation in virtual environments that mimic naturalistic settings opens up a wider discussion of taking many relevant attributes such as internal body, motivations and external body into account and study their role in decision-making. And by bridging biologically informed cognitive architectures into the simulation and the visualization of an artificial agent, this work places itself at the junction of advancing the grounding principles of general intelligence in artificial agents and the verifying the understanding of the existing biological models. In this perspective, Malmo offers a striking advantage to describe our model and its behavior to neuroscientists, who are more accustomed to life than algorithms and equations. Thanks to the inherent software abstraction followed in this work, it simultaneously allows exploring more complex scenarios of behavior with theoretical implementations as well as testing more detailed biologically informed computational models in the context of known behaviors.

Computational modeling of neuronal circuits have been instrumental in leveraging the understanding from neuroscientific and psychological studies and create a structure which would help further the understanding. However, in the case of this thesis topic, which is understanding high-level voluntary animal behavior as a function of decision-making and learning, even before venturing into underpinning the involved cognitive aspects and their neural correlates in brain, it is clear that such a study requires a tool that is more sophisticated than traditional numerical simulations and symbolic localized representations used in such models. Primarily this is due to the fact that the field of neuroscience of decision making has already highlighted intertwined roles of several brain regions working in parallel as well as interacting with each other. Even from a computational point of view, different kind of computational processes are thought to be taking place in each of these subsystems. The only common substrate that links all of the sub-processes is the emergent behavior and it is the key to understanding individual processes involved.

Extensive review into the literature of OFC gave some key insights related to decision-making and learning in general. OFC is proposed to represent an abstract state space of the environment, that can contribute to behavior in the absence of obvious state information from the environment which on the other hand sub-cortical structures cannot (Schuck et al. 2018; Wilson et al. 2014). This forms an important aspect of cognition, since the agent has to acess the information about the environment that is not obviously apparent. Furthermore it was also proposed that the very dynamics of learning in the sub-cortical structures like amygdala and VS might be modulated by OFC through top-down processes, to quickly adapt to changing environments (Elliott 2000; Kennerley and Wallis 2009). Understandably, the computational accounts of Pavlovian learning and instrumental learning chosen, like any of other similar models (Gurney et al. 2001b; Guthrie et al. 2013; Vitay and Hamker 2014) are from the point of view of sub-cortical mechanisms. It was identified that to account for the findings with respect to the possible roles of OFC, the structure of learning in the models has to be modified. Learning rates play a huge role in the performance and it is not something that has been explored much in studies. There is some dynamic factor within the task that should control and change the learning rates, for example changing value differences. But essentially what stands out from this observation is that for any agent, it is beneficial to have a system which responds fast when most of the environment is predictable and another system that can flag the unpredicted or novel situation to switch back to careful evaluation.

IV . Summary of contributions

As the primary goal of this thesis is to contribute to the understanding of the neuronal processes, most of the contributions are directed towards the field computational neuroscience. In addition, following these contributions are several hypotheses that can be explored in the fields of both experimental neuroscience and reinforcement learning.

Contributions to Neuroscience

- A bio-inspired cognitive architecture
  - multiple behaviours without a central executive
  - a testbed for neuro-computational models of decision-making and learning
  - a closed-loop framework to study the effect of prefrontal regions on behaviour
- OFC as a part of the prefrontal and the sub-cortical systems
  - the limbic sub-system of the prefrontal cortex including ACC
  - the emotional and motivational valuation
  - the basolateral amygdala (BLA) and ventral striatum (VS)
- A system-level account of the interacting roles of lateral and medial subregions of the OFC

Hypotheses for Reinforcement Learning

1. Episodic-meta RL : In its current state, episodic-meta RL stores abstractions and visited states for later, which could be static representations in Lateral OFC-Hippocampal interactions. Further, more dynamic representations could be possible in the form of interacting pathways of the limbic and sub-cortical circuits.

2. Fast and Slow RL : Multiple choice and learning processes at different timescales could be described by understanding interacting computations between lateral and medial OFC, parallel decision processes through the BG and value representations in BLA.

3. V(s) or Q(s, a) : Current RL frameworks most commonly employ either of the two value functions. However, current neurofunctional understanding is that the OFC circuits encode aspects comparable to V(s) and ACC circuits encode aspects

comparable to Q(s,a) in RL. Possibly combined representations of V(s) and Q(s,a) are present in the brain with differential influence on the selection processes and behaviour.

4. Dynamic learning rates : The learning rate decay is currently used in certain RL frameworks and machine learning in general, which simply reduces the learning rates carefully so as to stabilize learning. In the prefrontal circuits involving OFC and ACC, more relevant task-related information was found to influence the downstream sub-cortical structures to adapt to faster or slower learning rates. Such state abstraction in RL could facilitate such adaptation of learning rates

V . Limitations

Reinforcement

Notwithstanding the detailed neural architectures and basic neuronal descriptions used in certain parts of this work, the neural mechanisms of all the behavioral paradigms were discussed at a very simplistic level. Throughout the work, only appetitive behavior has been described, whereas most of the processes described in this work are also known to account for aversive behaviors like avoiding punishments. For example, Pavlovian systems have been very well described to explain context-based fear reversal and extinction (Carrere and Alexandre 2015; Vlachos et al. 2011). Although several processes were argued to be respond in a similar way for both positive and negative reinforcements Walton et al. 2010, it is not necessary that it holds true for all the cases.

In addition, the role of dopamine as the neurotransmitter facilitating learning has been extremely simplified. As mentioned in the initial chapters (section 3.4), the dopaminergic system is a wide complicated network that reaches different parts of the brain with different dynamics (tonic and phasic firing of dopamine neurons). Furthermore, with multiple systems of reinforcement learning involved in the framework, it demands for a detailed role of how dopamine could have a differential effect on these systems. Several

observations on dopamine firing even within a single system of Pavlovian learning is an emerging interest of various studies (Kaushik et al. 2017; Vitay and Hamker 2014). More complicated learning scenarios - timing and magnitude differences in reward - that are not discussed in this work, were explained using the systems that are already mentioned in this work, describing the roles of other neurotransmitters in conjunction with dopamine (Chuhma et al. 2011; Xia et al. 2011).

Memory

One of the most important elements of behavior that is not accounted for in the framework is memory, as can be seen as the difference of missing 'Hippocampus' in the figure 1.1 that clearly laid out the behavioral framework in brain and figure 1.2 that is directly referred in this work. The involvement of processes of working memory or episodic memory in behavior is very well certain and several works have been accounting for the involvement of PFC and hippocampus. In fact by complementing the framework with an existing computational account of a minimal working memory model (Strock et al. 2019), the mechanisms of sustained activities to maintain goals until achieving, aspects like giving up if the goal hasn't been reached for a long time etc, can be explored.

Adding an explicit memory to store minimum spatial and episodic information would allow the framework to explain more flexible behaviors like pure goal-directed or opportunistic behaviors. When the agent is hungry or thirst with no stimuli in sight, they would serve well in controlling the direction of agent's exploration by recollecting the direction where historically, finding the stimuli related to a certain need is more likely. Even when there are no stimuli around, if there is a need arising, this converts into 'desired' in Insula, subsequently the afferent representations are looked up in the VS[Core]. Further by modulating the OFC loop, the sensory representations that previously led to the outcome that satisfies this need, and the representations of the reinforcers specific to that outcome are activated to be desired in the sensory cortex. Using the episodic memory, the agent remembers the location where the desired outcome was abundantly found in previous

experience, and this position is activated to be desired in the motor loops. However, that would require much sophisticated implementations of motor loops where a desired position can be navigated. Also, when the agent is engaged in a goal-driven behavior and suddenly perceives a stimulus corresponding to the currently irrelevant need but with a strong preference, agent could choose it for several reasons. This could be particularly the case if the stimulus with the strong preference is rare.

Functional connectivity

Some of the crucial brain regions referred in this work, Orbitofrontal cortex (OFC), Nucleus accumbens (NAcc) and Amygdala, have been extensively explored in terms of their sub-divisions. Although this work focussed on the lateral and medial subregions of OFC, there are very distinct sub-structures within NAcc - shell and core, and amygdala - basolateral (BLA) and central (CeN) nuclei. What is more to it that most likely, all these sub-structures seem to be in a dissociate connectivity suggesting a possible deeper relation to the dissociation of OFC investigated in this work. Lateral and medial parts of OFC may be functionally more closely involved with the basolateral (BLA) and the central nucleus (CeA) of amygdala respectively. Anatomically, although it is not an outright clear distinction, several reviews that argued for functional dissociation of BLA and CeA pointed out that CeA makes direct projections to vmPFC (particularly medial OFC), implicated in the motivational control of habitual actions (Balleine and Killcross 2006; Killcross and Coutureau 2003; Yin et al. 2004). This is consistent with the proposition that shell is more involved in Pavlovian consummatory responses whereas core is clearly more involved in instrumental preparatory responses. Especially the studies into specific and general Pavlovian instrumental transfer highlighted possible dissociations between BLA interacting with the shell and CeN interacting with the core. Essentially, ventral striatum is described as the place where associations between the outcomes and their motivational value in the shell from BLA and between actions and the outcomes in the core from CeA (Mannella et al. 2013).

VI . Perspectives

Understanding the basis of flexible behavior is expressed in terms of fundamental animal behaviors that have been long detailed in psychology and cognitive science. The *limbic* part of the high-level architecture, which is more concerned with the emotional and cognitive aspects of behavior, is expressed as an interplay between the more fundamental processes of Pavlovian and instrumental behaviors. And by revisiting the detailed neural descriptions of these underlying processes, it allowed to place the individual neural accounts in their respective place in the high-level framework. In the process of exploring the neuroscientific evidence of these fundamental systems, an extensive review into the orbitofrontal cortex (OFC) pointed out a distinct possibility of dissociate contribution of two of its subregions - lateral and medial OFC. This possibility is studied in a closer detail, by exploiting the available computational models of the underlying processes and their integration into the high-level framework that has been built already. While retaining the individual accounts of each of these models, the combined interaction of these models in used to study the role of OFC, and several experimental studies have been used to compare the results of the hypothesis of possible dissociation within OFC to contribute to flexible decision-making.

In the wake of the field of Artificial General Intelligence that advocates intelligent agents should be able to solve general problems, it is crucial to first sketch the organization of fundamental building blocks of the best known intelligent agents - humans and high-level animals. The fields of cognitive science and neuroscience, and computational neuroscience have a lot to offer in that front by contributing the understanding from the human and animal studies and the models that explain them.

On the other hand, the AI that will be used to study these artificial intelligent agents enables neuroscientists to obtain further insights into how computation works in the brain, and particularly in a global view of the brain than individual sub-regions. Especially in the cases of psychiatric disorders such as depression, anxiety, addictions where the role of motivation drastically affects behaviors and the underlying dysfunctions are difficult to

relate to, a global representation of underlying systems can provide a useful test-bed for the existing knowledge. Virtual experimentation has already been proven to help in the case of physical therapy of stroke recovery (Singla et al. 2017). If the attributes involved in the motivational behavior can be well represented in the form an interactive survival scenario, there is a potential that virtual environment could offer a possibility, if not for a non-invasive treatments to these disorders, but at least an assessment of the patient's disposition of these underlying motivational factors.

In the field of robotics, there have been accounts of motivation driven robotic systems and a few design frameworks Konidaris and Barto 2006; Lewis and Cañamero 2016; Strannegård et al. 2017. Most of them address the task of making a choice based on motivation. This work complements them by pushing for a more involved biological knowledge. While it is by no means a detailed model of how cognition works in brain nor a detailed explanation of the models used in the work itself, it is reminded that building such frameworks in a bio-inspired manner requires further more understanding of higher level neural mechanisms involved in animal behavior Constantino and Daw 2015; Kolling et al. 2012. It is an attempt to place the key components of intelligent behavior in a comprehensive systems level view by leveraging the existing biological understanding. It is the execution of the most unique capability bestowed upon us - *reasoning* - to *reason* why we don't just *react* but rather *think* (well, at least often, if not always) before *responding*.

Bibliography

Abler, Birgit, Henrik Walter, and Susanne Erk (2005). "Neural correlates of frustration". In: *Neuroreport* 16.7, pp. 669–672. ISSN: 09594965. DOI: 10.1097/00001756-200505120-00003.

Adams, Geoffrey K. et al. (2012). *Neuroethology of decision-making*. DOI: 10.1016/j.conb.2012.07.009.

Alexander, G. (1986). "Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex". In: *Annu. Rev. Neurosci.* ISSN: 0147006X. DOI: 10.1146/annurev.neuro.9.1.357.

Alexander, William H. and Joshua W. Brown (2011). "Medial prefrontal cortex as an action-outcome predictor". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/nn.2921.

Alexandre, Frédéric (2016). "A behavioral framework for a systemic view of brain modeling". In: *Comput. Model. Brain Behav.* URL: https://hal.inria.fr/hal-01246653.

Anderson, M. C. and C. Weaver (2009). "Inhibitory Control over Action and Memory". In: *Encycl. Neurosci.* ISBN: 9780080450469. DOI: 10.1016/B978-008045046-9.00421-6.

Arana, F. Sergio et al. (2003). "Dissociable Contributions of the Human Amygdala and Orbitofrontal Cortex to Incentive Motivation and Goal Selection". In: *J. Neurosci.* ISSN: 02706474.

Aston-Jones, G., J. Rajkowski, and P. Kubiak (1997). "Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task". In: *Neuroscience*. ISSN: 03064522. DOI: 10.1016/S0306-4522(97)00060-2.

Baddeley, A. et al. (1986). "Dementia and Working Memory". In: *Q. J. Exp. Psychol. Sect. A*. ISSN: 14640740. DOI: 10.1080/14640748608401616.

Badre, David (2008). *Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes*. DOI: 10.1016/j.tics.2008.02.004.

Balasubramani, Pragathi P. et al. (2014). "An extended Reinforcement Learning model of basal ganglia to understand the contributions of serotonin and dopamine in risk-based decision making, reward prediction, and punishment learning". In: *Front. Comput. Neurosci.* 8.1 APR, pp. 1–12. ISSN: 16625188. DOI: 10.3389/fncom.2014.00047.

Balleine, Bernard W and Anthony Dickinson (1998). "Balleine and Dickinson 1998". In: 37, pp. 1–13. URL: papers2://publication/uuid/2F2B9DC2-C3EF-4FA9-992F-54B876380E47.

Balleine, Bernard W. and Simon Killcross (2006). "Parallel incentive processing: an integrated view of amygdala function". In: *Trends Neurosci.* 29.5, pp. 272–279. ISSN: 01662236. DOI: 10.1016/j.tins.2006.03.002.

Bandura, A. (1997). "The anatomy of stages of change." In: *Am. J. Health Promot.* ISSN: 08901171.

Bault, Nadège et al. (2019). "Dissociation between private and social counterfactual value signals following ventromedial prefrontal cortex damage". In: *J. Cogn. Neurosci.* ISSN: 15308898. DOI: 10.1162/jocn_a_01372.

Baxter, M G and E A Murray (2000). "Reinterpreting the behavioural effects of amygdala lesions in non-human primates". In: *amygdala, a Funct. Anal.*

Bechara, Antoine and Antonio R. Damasio (2005). "The somatic marker hypothesis: A neural theory of economic decision". In: *Games Econ. Behav.* ISSN: 10902473. DOI: 10.1016/j.geb.2004.06.010.

Bechara, Antoine et al. (1997). "Deciding advantageously before knowing the advantageous strategy". In: *Science (80-.).* 275.5304, pp. 1293–1295. ISSN: 00368075. DOI: 10.1126/science.275.5304.1293.

Bechara, Antoine et al. (1999). "Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making". In: *J. Neurosci.* ISSN: 02706474.

Behrens, Timothy E.J. et al. (2007). "Learning the value of information in an uncertain world". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/nn1954.

Berridge, Kent C. and Terry E. Robinson (1995). "The Mind of an Addicted Brain: Neural Sensitization of Wanting Versus Liking". In: *Curr. Dir. Psychol. Sci.* ISSN: 14678721. DOI: 10.1111/1467-8721.ep10772316.

Bissonette, Gregory B. et al. (2013). "Separate Populations of Neurons in Ventral Striatum Encode Value and Motivation". In: *PLoS One* 8.5, pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0064673.

Blair, Karina et al. (2006). "Choosing the lesser of two evils, the better of two goods: Specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1640-06.2006.

Bogacz, Rafal et al. (2006). "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks". In: *Psychol. Rev.* 113.4, pp. 700–765. ISSN: 0033295X. DOI: 10.1037/0033-295X.113.4.700.

Boorman, Erie D. et al. (2009). "How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action". In: *Neuron* 62.5, pp. 733–743. ISSN: 08966273. DOI: 10.1016/j.neuron.2009.05.014. URL: http://dx.doi.org/10.1016/j.neuron.2009.05.014.

Boorman, Erie D. et al. (2016). "Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus". In: *Neuron*. ISSN: 10974199. DOI: 10.1016/j.neuron.2016.02.014.

Boraud, Thomas, Arthur Leblois, and Nicolas P. Rougier (2018). "A natural history of skills". In: *Prog. Neurobiol.* ISSN: 18735118. DOI: 10.1016/j.pneurobio.2018.08.003.

Bouret, S. and B. J. Richmond (2010). "Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys". In: *J. Neurosci.* 30.25, pp. 8591–8601. ISSN: 0270-6474. DOI: 10.1523/jneurosci.0049-10.2010. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0049-10.2010.

Bouret, Sebastien and Susan J. Sara (2005). "Network reset: A simplified overarching theory of locus coeruleus noradrenaline function". In: *Trends Neurosci.* ISSN: 01662236. DOI: 10.1016/j.tins.2005.09.002.

Bradfield, Laura A. et al. (2015). "Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations". In: *Neuron* 88.6, pp. 1268–1280. ISSN: 08966273. DOI: 10.1016/j.neuron.2015.10.044. URL: http://dx.doi.org/10.1016/j.neuron.2015.10.044https://linkinghub.elsevier.com/retrieve/pii/S0896627315009381.

Brodmann, Korbinian (2007). *Brodmann's: Localisation in the cerebral cortex*. Springer Science & Business Media.

Bryden, Daniel W. et al. (2011). "Attention for learning signals in anterior cingulate cortex". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.4715-11.2011.

Cabanac, Michel (1992). "Pleasure: the common currency". In: *J. Theor. Biol.* ISSN: 10958541. DOI: 10.1016/S0022-5193(05)80594-6.

Calandreau, Ludovic et al. (2006). "Extracellular hippocampal acetylcholine level controls amygdala function and promotes adaptive conditioned emotional response". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.3713-06.2006.

Canamero, Dolores (1997). "Modeling motivations and emotions as a basis for intelligent behavior". In: *Proc. First Inrl. Can)'. Auton. Agents*, pp. 148–155.

Carmichael, S. T. and J. L. Price (1994). "Architectonic subdivision of the orbital and medial prefrontal cortex in the macaque monkey". In: *J. Comp. Neurol.* 346.3, pp. 366–402. ISSN: 0021-9967. DOI: 10.1002/cne.903460305. URL: http://doi.wiley.com/10.1002/cne.903460305.

— (1995a). "Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys". In: *J. Comp. Neurol.* ISSN: 10969861. DOI: 10.1002/cne.903630408.

— (1996). "Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys". In: *J. Comp. Neurol.* ISSN: 00219967. DOI: 10.1002/(SICI)1096-9861(19960722)371:2<179::AID-CNE1>3.0.CO;2-\#.

Carmichael, S. T. and Joseph L. Price (1995b). "Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys". In: *J. Comp. Neurol.* 363.4, pp. 642–664. ISSN: 10969861. DOI: 10.1002/cne.903630409.

Carrere, Maxime and Frédéric Alexandre (2015). "A pavlovian model of the amygdala and its influence within the medial temporal lobe". In: *Front. Syst. Neurosci.* 9.March. ISSN: 1662-5137. DOI: 10.3389/fnsys.2015.00041. URL: http://www.frontiersin.org/Systems{_}Neuroscience/10.3389/fnsys.2015.00041/abstract.

Cartoni, Emilio, Stefano Puglisi-Allegra, and Gianluca Baldassarre (2013). "The three principles of action: A Pavlovian-instrumental transfer hypothesis". In: *Front. Behav. Neurosci.* ISSN: 16625153. DOI: 10.3389/fnbeh.2013.00153.

Cassey, Thomas C. et al. (2013). "Adaptive Sampling of Information in Perceptual Decision-Making". In: *PLoS One* 8.11. Ed. by Daniel Osorio, e78993. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0078993. URL: https://dx.plos.org/10.1371/journal.pone.0078993.

Cavada, C. (2000a). "The Anatomical Connections of the Macaque Monkey Orbitofrontal Cortex. A Review". In: *Cereb. Cortex*. DOI: 10.1093/cercor/10.3.220.

— (2000b). "The Mysterious Orbitofrontal Cortex. Foreword". In: *Cereb. Cortex*. DOI: 10.1093/cercor/10.3.205.

Chakravarthy, V. S., Denny Joseph, and Raju S. Bapi (2010). *What do the basal ganglia do? A modeling perspective*. DOI: 10.1007/s00422-010-0401-y.

Chan, Stephanie C.Y., Yael Niv, and Kenneth A. Norman (2016). "A probability distribution over latent causes, in the orbitofrontal cortex". In: *J. Neurosci.* ISSN: 15292401. DOI: 10.1523/JNEUROSCI.0659-16.2016.

Choi, Jung Seok et al. (2004). "Left anterior subregion of orbitofrontal cortex volume reduction and impaired organizational strategies in obsessive-compulsive disorder". In: *J. Psychiatr. Res.* 38.2, pp. 193–199. ISSN: 00223956. DOI: 10.1016/j.jpsychires.2003.08.001.

Christopoulos, George I. et al. (2009). "Neural correlates of value, risk, and risk aversion contributing to decision making under risk". In: *J. Neurosci.* 29.40, pp. 12574–12583. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.2614-09.2009.

Chudasama, Y. and Trevor W. Robbins (2003). "Dissociable contributions of the orbitofrontal and infralimbic cortex to pavlovian autoshaping and discrimination reversal learning: Further evidence for the functional heterogeneity of the rodent frontal cortex". In: *J. Neurosci.* 23.25, pp. 8771–8780. ISSN: 02706474.

Chuhma, Nao et al. (2011). "Functional connectome of the striatal medium spiny neuron". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.3833-10.2011.

Cisek, Paul (2011). "Cortical mechanisms of action selection: the affordance competition hypothesis". In: *Model. Nat. Action Sel.* April, pp. 208–238. DOI: 10.1017/CBO9780511731525.015.

Clarke, Hannah F., Trevor W. Robbins, and Angela C. Roberts (2008). "Lesions of the medial striatum in monkeys produce perseverative impairments during reversal learning similar to those produced by lesions of the orbitofrontal cortex". In: *J. Neurosci.* 28.43, pp. 10972–10982. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1521-08.2008.

Cohen, Jeremiah Y. et al. (2012). *Neuron-type-specific signals for reward and punishment in the ventral tegmental area*. DOI: 10.1038/nature10754.

Conen, Katherine E. and Camillo Padoa-Schioppa (2019). "Partial Adaptation to the Value Range in the Macaque Orbitofrontal Cortex". In: *J. Neurosci.* Pp. 2279–18. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2279-18.2019. URL: http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2279-18.2019.

Constantino, Sara M. and Nathaniel D. Daw (2015). "Learning the opportunity cost of time in a patch-foraging task". In: *Cogn. Affect. Behav. Neurosci.* 15.4, pp. 837–853. ISSN: 15307026. DOI: 10.3758/s13415-015-0350-y.

Cools, Roshan, Kae Nakamura, and Nathaniel D. Daw (2011). "Serotonin and dopamine: Unifying affective, activational, and decision functions". In: *Neuropsychopharmacology* 36.1, pp. 98–113. ISSN: 0893133X. DOI: 10.1038/npp.2010.121. URL: http://dx.doi.org/10.1038/npp.2010.121.

Corbit, Laura H., Sarah C. Fischbach, and Patricia H. Janak (2016). "Nucleus accumbens core and shell are differentially involved in general and outcome-specific forms of Pavlovian-instrumental transfer with alcohol and sucrose rewards". In: *Eur. J. Neurosci.* 43.9, pp. 1229–1236. ISSN: 14609568. DOI: 10.1111/ejn.13235.

Coricelli, Giorgio et al. (2005). "Regret and its avoidance: a neuroimaging study of choice behavior". In: *Nat. Neurosci.* 8.9, pp. 1255–1262. ISSN: 1097-6256. DOI: 10.1038/nn1514. URL: http://www.nature.com/articles/nn1514.

Cos, Ignasi, Lola Cañamero, and Gillian M. Hayes (2010). "Learning affordances of consummatory behaviors: Motivation-driven adaptive perception". In: *Adapt. Behav.* 18.3, pp. 285–314. ISSN: 10597123. DOI: 10.1177/1059712310375471.

Croxson, Paula L. et al. (2005). "Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1311-05.2005.

Daw, Nathaniel D., Sham Kakade, and Peter Dayan (2002). "Opponent interactions between serotonin and dopamine". In: *Neural Networks* 15.4-6, pp. 603–616. ISSN: 08936080. DOI: 10.1016/S0893-6080(02)00052-7.

Daw, Nathaniel D., Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nat. Neurosci.* 8.12, pp. 1704–1711. ISSN: 1097-6256. DOI: 10.1038/nn1560. URL: http://www.nature.com/articles/nn1560.

Daw, Nathaniel D. et al. (2006). "Cortical substrates for exploratory decisions in humans". In: *Nature*. ISSN: 14764687. DOI: 10.1038/nature04766.

Daw, Nathaniel D. et al. (2011). "Model-Based Influences on Humans' Choices and Striatal Prediction Errors". In: *Neuron* 69.6, pp. 1204–1215. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.02.027. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627311001255.

Dayan, Peter (2009). "Goal-directed control and its antipodes". In: *Neural Networks*. ISSN: 08936080. DOI: 10.1016/j.neunet.2009.03.004.

— (2012). “Twenty-Five Lessons from Computational Neuromodulation”. In: *Neuron* 76.1, pp. 240–256. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.09.027. URL: http://dx.doi.org/10.1016/j.neuron.2012.09.027.

Dayan, Peter and Yael Niv (2008). “Reinforcement learning: The Good, The Bad and The Ugly”. In: *Curr. Opin. Neurobiol.* 18.2, pp. 185–196. ISSN: 09594388. DOI: 10.1016/j.conb.2008.08.003.

Denoyelle, Nicolas et al. (2016). “From biological to numerical experiments in systemic neuroscience: A simulation platform”. In: *Biosyst. Biorobotics* 12, pp. 1–17. ISSN: 21953570. DOI: 10.1007/978-3-319-26242-0_1.

Depue, Brendan E., Tim Curran, and Marie T. Banich (2007). “Prefrontal regions orchestrate suppression of emotional memories via a two-phase process”. In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.1139560.

Dezfouli, Amir and Bernard W. Balleine (2019). “Learning the structure of the world: The adaptive nature of state-space and action representations in multi-stage decision-making”. In: *PLOS Comput. Biol.* 15.9, e1007334. DOI: 10.1371/journal.pcbi.1007334. URL: http://dx.doi.org/10.1371/journal.pcbi.1007334.

Dolan, Ray J and Peter Dayan (2013). “Goals and habits in the brain.” In: *Neuron* 80.2, pp. 312–25. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2013.09.007. URL: http://www.ncbi.nlm.nih.gov/pubmed/24139036{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3807793.

Domínguez D, Juan F. et al. (2018). “Lateral orbitofrontal cortex activity is modulated by group membership in situations of justified and unjustified violence”. In: *Soc. Neurosci.* ISSN: 17470927. DOI: 10.1080/17470919.2017.1392342.

Doya, K. (1999). “What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?” In: *Neural Networks*. ISSN: 08936080. DOI: 10.1016/S0893-6080(99)00046-5.

Doya, Kenji et al. (2002). “Multiple model-based reinforcement learning”. In: *Neural Comput.* 14.6, pp. 1347–1369. ISSN: 08997667. DOI: 10.1162/089976602753712972.

Dum, Richard P and Peter L Strick (1993). "Cingulate motor areas." In: *Neurobiol. cingulate cortex limbic thalamus A Compr. handbook.* Cambridge, MA, US: Birkhäuser, pp. 415–441. ISBN: 0-8176-3568-8 (Hardcover); 3-7643-3568-8 (Hardcover). DOI: 10.1007/978-1-4899-6704-6_15.

Elliott, R. (2000). "Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies". In: *Cereb. Cortex*. DOI: 10.1093/cercor/10.3.308.

Elliott, R., Z. Agnew, and J. F.W. Deakin (2008). "Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans". In: *Eur. J. Neurosci.* 27.9, pp. 2213–2218. ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2008.06202.x.

Eshel, Neir et al. (2015). "Arithmetic and local circuitry underlying dopamine prediction errors". In: *Nature*. ISSN: 14764687. DOI: 10.1038/nature14855.

Fearing, Franklin, I. P. Pavlov, and G. V. Anrep (1929). "Conditioned Reflexes. An Investigation of the Physiological Activity of the Cerebral Cortex". In: *J. Am. Inst. Crim. Law Criminol.* ISSN: 08854173. DOI: 10.2307/1134737.

Fellows, Lesley K. (2003). "Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm". In: *Brain* 126.8, pp. 1830–1837. ISSN: 1460-2156. DOI: 10.1093/brain/awg180. URL: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awg180.

— (2011). "Orbitofrontal contributions to value-based decision making: Evidence from humans with frontal lobe damage". In: *Ann. N. Y. Acad. Sci.* 1239.1, pp. 51–58. ISSN: 17496632. DOI: 10.1111/j.1749-6632.2011.06229.x.

Fellows, Lesley K. and Martha J. Farah (2005). "Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans". In: *Cereb. Cortex*. ISSN: 10473211. DOI: 10.1093/cercor/bhh108.

— (2007). "The role of ventromedial prefrontal cortex in decision making: Judgment under uncertainty or judgment per se?" In: *Cereb. Cortex* 17.11, pp. 2669–2674. ISSN: 10473211. DOI: 10.1093/cercor/bhl176.

Fiorillo, Christopher D., Philippe N. Tobler, and Wolfram Schultz (2003). "Discrete coding of reward probability and uncertainty by dopamine neurons". In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.1077349.

FitzGerald, T. H. B., B. Seymour, and R. J. Dolan (2009). "The Role of Human Orbitofrontal Cortex in Value Comparison for Incommensurable Objects". In: *J. Neurosci.* 29.26, pp. 8388–8395. ISSN: 0270-6474. DOI: 10.1523/jneurosci.0717-09.2009.

Floresco, Stan B. (2015). "The Nucleus Accumbens: An Interface Between Cognition, Emotion, and Action". In: *Annu. Rev. Psychol.* ISSN: 0066-4308. DOI: 10.1146/annurev-psych-010213-115159.

Frith, Chris and Raymond J. Dolan (1997). *Brain mechanisms associated with top-down processes in perception.* DOI: 10.1098/rstb.1997.0104.

Fu, Bo, Natalya F. Noy, and Margaret Anne Storey (2017). "Eye tracking the user experience - An evaluation of ontology visualization techniques". In: *Semant. Web.* ISSN: 22104968. DOI: 10.3233/SW-140163.

Funahashi, S., C. J. Bruce, and P. S. Goldman-Rakic (1989). "Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex". In: *J. Neurophysiol.* 61.2, pp. 331–349. ISSN: 00223077. DOI: 10.1152/jn.1989.61.2.331.

Gaffan, David (1996). "Memory, action and the corpus striatum: Current developments in the memory-habit distinction". In: *Semin. Neurosci.* ISSN: 10445765. DOI: 10.1006/smns.1996.0005.

Gardner, Matthew P.H. et al. (2017). "Lateral Orbitofrontal Inactivation Dissociates Devaluation-Sensitive Behavior and Economic Choice". In: *Neuron* 96.5, 1192–1203.e4. ISSN: 10974199. DOI: 10.1016/j.neuron.2017.10.026. URL: https://doi.org/10.1016/j.neuron.2017.10.026.

Garenne, André et al. (2011). "Basal Ganglia Preferentially Encode Context Dependent Choice in a Two-Armed Bandit Task". In: *Front. Syst. Neurosci.* 5.May, pp. 1–9. ISSN: 1662-5137. DOI: 10.3389/fnsys.2011.00023. URL: http://journal.frontiersin.org/article/10.3389/fnsys.2011.00023/abstract.

Gläscher, Jan, Alan N. Hampton, and John P. O'Doherty (2009). "Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making". In: *Cereb. Cortex*. ISSN: 14602199. DOI: 10.1093/cercor/bhn098.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning." In: *Neuron* 66.4, pp. 585–595. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2010.04.016.States. URL: http://www.ncbi.nlm.nih.gov/pubmed/20510862{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2895323.

Glasser, Matthew F. et al. (2016). "A multi-modal parcellation of human cerebral cortex". In: *Nature*. ISSN: 14764687. DOI: 10.1038/nature18933.

Gluth, Sebastian, Mikhail S Spektor, and Jörg Rieskamp (2018). "Value-based attentional capture affects multi-alternative decision making". In: *Elife* 7, pp. 1–36. DOI: 10.7554/elife.39659.

Goldman-Rakic, Patricia S. (2011). "Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory". In: *Compr. Physiol.* DOI: 10.1002/cphy.cp010509.

GOLDMAN-RAKIC, PATRICIA S. (1995). "Architecture of the Prefrontal Cortex and the Central Executive". In: *Ann. N. Y. Acad. Sci.* ISSN: 17496632. DOI: 10.1111/j.1749-6632.1995.tb38132.x.

Gottfried, Jay A, J P O'Doherty, and R J Dolan (2003). "Value in Human Amygdala and Orbitofrontal Cortex". In: *Science.* 301.5636, pp. 1104–1108. ISSN: 10959203. DOI: 10.1126/science.1087919.

Grabenhorst, Fabian and Edmund T. Rolls (2009). "Different representations of relative and absolute subjective value in the human brain". In: *Neuroimage* 48.1, pp. 258–268. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.06.045. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811909006818.

— (2011). "Value, pleasure and choice in the ventral prefrontal cortex". In: *Trends Cogn. Sci.* 15.2, pp. 56–67. ISSN: 13646613. DOI: 10.1016/j.tics.2010.12.004. URL: http://dx.doi.org/10.1016/j.tics.2010.12.004.

Grabenhorst, Fabian, Edmund T. Rolls, and Benjamin A. Parris (2008). "From affective value to decision-making in the prefrontal cortex". In: *Eur. J. Neurosci.* ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2008.06489.x.

Grabenhorst, Fabian et al. (2010). "A common neural scale for the subjective pleasantness of different primary rewards". In: *Neuroimage* 51.3, pp. 1265–1274. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2010.03.043. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811910003307.

Graybiel, Ann M. (1995). "Building action repertoires: memory and learning functions of the basal ganglia". In: *Curr. Opin. Neurobiol.* ISSN: 09594388. DOI: 10.1016/0959-4388(95)80100-6.

Groman, Stephanie M. et al. (2019). "Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes". In: *Neuron*, pp. 1–13. ISSN: 08966273. DOI: 10.1016/j.neuron.2019.05.042. URL: https://doi.org/10.1016/j.neuron.2019.05.042.

Gurney, K., T. J. Prescott, and P. Redgrave (2001a). "A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour". In: *Biol. Cybern.* 84.6, pp. 411–423. ISSN: 03401200. DOI: 10.1007/PL00007985.

Gurney, K, T J Prescott, and P Redgrave (2001b). "Gurney Et Al. 2001a". In: 410, pp. 1–10. URL: papers://82ac23f7-2eaf-4339-a5e1-4600c19d7f01/Paper/p2316.

Guthrie, M. et al. (2013). "Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study". In: *J. Neurophysiol.* 109.12, pp. 3025–3040. ISSN: 0022-3077. DOI: 10.1152/jn.00026.2013.

Ha, David and Jürgen Schmidhuber (2018). "Recurrent World Models Facilitate Policy Evolution". In: C. arXiv: 1809.01999. URL: http://arxiv.org/abs/1809.01999.

Haber, Suzanne N. (2003). "The primate basal ganglia: Parallel and integrative networks". In: *J. Chem. Neuroanat.* DOI: 10.1016/j.jchemneu.2003.10.003.

Haber, Suzanne N. and Brian Knutson (2010). *The reward circuit: Linking primate anatomy and human imaging*. DOI: 10.1038/npp.2009.129.

Hampton, Alan N., Peter Bossaerts, and John P. O'Doherty (2006). "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1010-06.2006.

Hatfield, Tammy et al. (1996). "Neurotoxic lesions of basolateral, but not central, amygdala interfere with pavlovian second-order conditioning and reinforcer devaluation effects". In: *J. Neurosci.* ISSN: 02706474.

Hebscher, Melissa et al. (2016). "Memory, decision-making, and the ventromedial prefrontal cortex (vmpfc): The roles of subcallosal and posterior orbitofrontal cortices in monitoring and control processes". In: *Cereb. Cortex* 26.12, pp. 4590–4601. ISSN: 14602199. DOI: 10.1093/cercor/bhv220.

Hornak, J. et al. (2004). "Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans". In: *J. Cogn. Neurosci.* 16.3, pp. 463–478. ISSN: 0898929X. DOI: 10.1162/089892904322926791.

Howard, James D. and Thorsten Kahnt (2017). "Identity-specific reward representations in orbitofrontal cortex are modulated by selective devaluation". In: *J. Neurosci.* ISSN: 15292401. DOI: 10.1523/JNEUROSCI.3473-16.2017.

Hulme, Oliver J., Tobias Morville, and Boris Gutkin (2019). *Neurocomputational theories of homeostatic control*. DOI: 10.1016/j.plrev.2019.07.005.

Humphries, Mark D., Mehdi Khamassi, and Kevin Gurney (2012). "Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia". In: *Front. Neurosci.* ISSN: 16624548. DOI: 10.3389/fnins.2012.00009.

Hunt, Laurence T. and Benjamin Y. Hayden (2017). "A distributed, hierarchical and recurrent framework for reward-based choice". In: *Nat. Rev. Neurosci.* 18.3, pp. 172–182. ISSN: 14710048. DOI: 10.1038/nrn.2017.7. URL: http://dx.doi.org/10.1038/nrn.2017.7.

Hunt, Laurence T. et al. (2012). "Mechanisms underlying cortical activity during value-guided choice". In: *Nat. Neurosci.* 15.3, pp. 470–476. ISSN: 10976256. DOI: 10.1038/

nn.3017. URL: http://www.nature.com/articles/nn.3017http://dx.doi.org/10.1038/nn.3017.

Iversen, Susan D. and Mortimer Mishkin (1970). "Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity". In: *Exp. Brain Res.* ISSN: 00144819. DOI: 10.1007/BF00237911.

Izquierdo, Alicia and J. David Jentsch (2012). *Reversal learning as a measure of impulsive and compulsive behavior in addictions*. DOI: 10.1007/s00213-011-2579-7.

Izquierdo, Alicia, Robin K. Suda, and Elisabeth A. Murray (2004). "Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency". In: *J. Neurosci.* 24.34, pp. 7540–7548. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1921-04.2004.

Joel, Daphna, Yael Niv, and Eytan Ruppin (2002). "Actor-critic models of the basal ganglia: new anatomical and computational perspectives". In: *Neural Networks* 15.4-6, pp. 535–547. ISSN: 08936080. DOI: 10.1016/S0893-6080(02)00047-3.

Johnson, Matthew et al. (2016). "The malmo platform for artificial intelligence experimentation". In: *IJCAI Int. Jt. Conf. Artif. Intell.*

Jones, B. and M. Mishkin (1972). "Limbic lesions and the problem of stimulus-Reinforcement associations". In: *Exp. Neurol.* ISSN: 10902430. DOI: 10.1016/0014-4886(72)90030-1.

Jones, J. L. et al. (2012). "Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values". In: *Science (80-.).* 338.6109, pp. 953–956. ISSN: 0036-8075. DOI: 10.1126/science.1227489. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1227489.

Kable, Joseph W and Paul W Glimcher (2007). "The neural correlates of subjective value during intertemporal choice." In: *Nat. Neurosci.* 10.12, pp. 1625–33. ISSN: 1097-6256. DOI: 10.1038/nn2007. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2845395{\&}tool=pmcentrez{\&}rendertype=abstract.

Kable, Joseph W. and Paul W. Glimcher (2009). "The Neurobiology of Decision: Consensus and Controversy". In: *Neuron* 63.6, pp. 733–745. ISSN: 08966273. DOI: 10.1016/j.neuron.2009.09.003. URL: http://dx.doi.org/10.1016/j.neuron.2009.09.003.

Kahneman, Daniel and Amos Tversky (1984). "Choices, values, and frames". In: *Am. Psychol.* ISSN: 0003066X. DOI: 10.1037/0003-066X.39.4.341.

Kahnt, Thorsten et al. (2012). "Connectivity-based parcellation of the human orbitofrontal cortex". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0257-12.2012.

Kaushik, Pramod S. et al. (2017). "A biologically inspired neuronal model of reward prediction error computation". In: *Proc. Int. Jt. Conf. Neural Networks*. ISBN: 9781509061815. DOI: 10.1109/IJCNN.2017.7966306.

Kawagoe, Reiko, Yoriko Takikawa, and Okihide Hikosaka (1998). "Expectation of reward modulates cognitive signals in the basal ganglia". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/1625.

Kazama, Andy and Jocelyne Bachevalier (2009). "Selective aspiration or neurotoxic lesions of orbital frontal areas 11 and 13 spared monkeys' performance on the object discrimination reversal task". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.4655-08.2009.

Kennerley, Steven W and Jonathan D Wallis (2009). "Evaluating Choices By Single Neurons in the Frontal Lobe". In: *Neuroscience* 29.10, pp. 2061–2073. DOI: 10.1111/j.1460-9568.2009.06743.x.Evaluating.

Kennerley, Steven W., Timothy E.J. Behrens, and Jonathan D. Wallis (2011). "Double dissociation of value computations in orbitofrontal and anterior cingulate neurons". In: *Nat. Neurosci.* 14.12, pp. 1581–1589. ISSN: 10976256. DOI: 10.1038/nn.2961.

Keramati, Mehdi and Boris Gutkin (2011). "A reinforcement learning theory for homeostatic regulation". In: *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pp. 1–9.

— (2014). "Homeostatic reinforcement learning for integrating reward collection and physiological stability". In: *Elife* 3, pp. 1–26. ISSN: 2050084X. DOI: 10.7554/eLife.04811.

Khamassi, Mehdi and Mark D. Humphries (2012). *Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies*. DOI: 10.3389/fnbeh.2012.00079.

Killcross, Simon and Etienne Coutureau (2003). "Coordination of actions and habits in the medial prefrontal cortex of rats". In: *Cereb. Cortex* 13.4, pp. 400–408. ISSN: 10473211. DOI: 10.1093/cercor/13.4.400.

Kim, Jenna and Michael E. Ragozzino (2005). "The involvement of the orbitofrontal cortex in learning under changing task contingencies". In: *Neurobiol. Learn. Mem.* ISSN: 10747427. DOI: 10.1016/j.nlm.2004.10.003.

Kim, Mina et al. (2006). "Anatomical correlates of the functional organization in the human occipitotemporal cortex". In: *Magn. Reson. Imaging*. ISSN: 0730725X. DOI: 10.1016/j.mri.2005.12.005.

Knutson, Brian et al. (2005). "Distributed neural representation of expected value". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0642-05.2005.

Kobayashi, S., O. Pinto de Carvalho, and W. Schultz (2010). "Adaptation of Reward Sensitivity in Orbitofrontal Neurons". In: *J. Neurosci.* 30.2, pp. 534–544. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4009-09.2010. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4009-09.2010.

Koechlin, Etienne, Chrystèle Ody, and Frédérique Kouneiher (2003). "The Architecture of Cognitive Control in the Human Prefrontal Cortex". In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.1088545.

Kolb, B. (2007). "Do all mammals have a prefrontal cortex?" In: *Evol. Nerv. Syst.* ISBN: 9780123708786. DOI: 10.1016/B0-12-370878-8/00081-1.

Kolling, Nils et al. (2012). "Neural Mechanisms of Foraging". In: *Science (80-.).* 336.6077, pp. 95–98. ISSN: 0036-8075. DOI: 10.1126/science.1216930.

Kondo, Yumiko et al. (2005). "Changes in brain activation associated with use of a memory strategy: A functional MRI study". In: *Neuroimage*. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2004.10.033.

Konidaris, George and Andrew Barto (2006). "An Adaptive Robot Motivational System". In: pp. 346–356. DOI: 10.1007/11840541_29.

Krack, Paul et al. (2010). "Deep brain stimulation: From neurology to psychiatry?" In: *Trends Neurosci.* 33.10, pp. 474–484. ISSN: 01662236. DOI: 10.1016/j.tins.2010.07.002. URL: http://dx.doi.org/10.1016/j.tins.2010.07.002.

Krajbich, Ian and Antonio Rangel (2011). "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions." In: *Proc. Natl. Acad. Sci. U. S. A.* 108.33, pp. 13852–7. ISSN: 1091-6490. DOI: 10.1073/pnas.1101328108. URL: http://www.ncbi.nlm.nih.gov/pubmed/21808009{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3158210.

Krettek, J. E. and J. L. Price (1977). "The cortical projections of the mediodorsal nucleus and adjacent thalamic nuclei in the rat". In: *J. Comp. Neurol.* 171.2, pp. 157–191. ISSN: 0021-9967. DOI: 10.1002/cne.901710204. URL: http://doi.wiley.com/10.1002/cne.901710204.

Kringelbach, M. L. et al. (2003). "Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness". In: *Cereb. Cortex*. ISSN: 10473211. DOI: 10.1093/cercor/13.10.1064.

Kringelbach, Morten L. (2005). "The human orbitofrontal cortex: linking reward to hedonic experience". In: *Nat. Rev. Neurosci.* 6.9, pp. 691–702. ISSN: 1471-003X. DOI: 10.1038/nrn1747. URL: http://www.nature.com/articles/nrn1747.

Kringelbach, Morten L. and Edmund T. Rolls (2004). *The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology*. DOI: 10.1016/j.pneurobio.2004.03.006.

Leblois, A. (2006). "Competition between Feedback Loops Underlies Normal and Pathological Dynamics in the Basal Ganglia". In: *J. Neurosci.* 26.13, pp. 3567–3583. ISSN:

0270-6474. DOI: 10.1523/JNEUROSCI.5050-05.2006. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5050-05.2006.

Lebreton, Maël et al. (2009). "An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2009.09.040.

Lee, Sang Wan, Shinsuke Shimojo, and John P. O'Doherty (2014). "Neural Computations Underlying Arbitration between Model-Based and Model-free Learning". In: *Neuron* 81.3, pp. 687–699. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.11.028. URL: http://dx.doi.org/10.1016/j.neuron.2013.11.028https://linkinghub.elsevier.com/retrieve/pii/S0896627313011252.

Lewis, Matthew and Lola Cañamero (2016). "Hedonic quality or reward? A study of basic pleasure in homeostasis and decision making of a motivated autonomous robot". In: *Adapt. Behav.* 24.5, pp. 267–291. ISSN: 17412633. DOI: 10.1177/1059712316666331.

Li, Yansong et al. (2015). "Local Morphology Predicts Functional Organization of Experienced Value Signals in the Human Orbitofrontal Cortex". In: *J. Neurosci.* ISSN: 15292401. DOI: 10.1523/JNEUROSCI.3058-14.2015.

Li, Yansong et al. (2016). "The neural dynamics of reward value and risk coding in the human orbitofrontal cortex". In: *Brain* 139.4, pp. 1295–1309. ISSN: 14602156. DOI: 10.1093/brain/awv409.

Lim, S.-L., J. P. O'Doherty, and A. Rangel (2011). "The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention". In: *J. Neurosci.* 31.37, pp. 13214–13223. ISSN: 0270-6474. DOI: 10.1523/jneurosci.1246-11.2011.

Luk, Chung Hay and Jonathan D. Wallis (2009). "Dynamic encoding of responses and outcomes by neurons in medial prefrontal cortex". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0386-09.2009.

Luria, Aleksandr Romanovich (2012). *Higher cortical functions in man*. Springer Science & Business Media.

Mackey, Scott and Michael Petrides (2010). "Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains". In: *Eur. J. Neurosci.* ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2010.07465.x.

Málková, Ludiše, David Gaffan, and Elisabeth A. Murray (1997). "Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys". In: *J. Neurosci.* ISSN: 02706474.

Malvaez, Melissa et al. (2019). "Distinct cortical–amygdala projections drive reward value encoding and retrieval". In: *Nat. Neurosci.* ISSN: 15461726. DOI: 10.1038/s41593-019-0374-7.

Mannella, Francesco, Kevin Gurney, and Gianluca Baldassarre (2013). "The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis". In: *Front. Behav. Neurosci.* 7.October, pp. 1–29. DOI: 10.3389/fnbeh.2013.00135.

Marsden, C. D. (1982). "The mysterious motor function of the basal ganglia: The Robert Wartenberg Lecture". In: *Neurology.* ISSN: 1526632X.

Marsden, C. D. and J. A. Obeso (1994). *The functions of the basal ganglia and the paradox of stereotaxic surgery in parkinson's disease.* DOI: 10.1093/brain/117.4.877.

Matiisen, Tambet et al. (2017). "Teacher-Student Curriculum Learning". In: arXiv: 1707.00183. URL: http://arxiv.org/abs/1707.00183.

Matsumoto, Masayuki and Okihide Hikosaka (2007). "Lateral habenula as a source of negative reward signals in dopamine neurons". In: *Nature* 447.7148, pp. 1111–1115. ISSN: 14764687. DOI: 10.1038/nature05860.

Maturana, Humberto and Francisco Varela (1991). *Autopoiesis and Cognition : The Realization of the Living (Boston Studies in the Philosophy of Science).* ISBN: 9027710163.

McDannald, Michael A. et al. (2011). "Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning". In: *J. Neurosci.* 31.7, pp. 2700–2705. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.5499-

10.2011. URL: http://www.ncbi.nlm.nih.gov/pubmed/21325538{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3079289http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5499-10.2011.

Meunier, Martine, Jocelyne Bachevalier, and Mortimer Mishkin (1997). "Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys". In: *Neuropsychologia* 35.7, pp. 999–1015. ISSN: 00283932. DOI: 10.1016/S0028-3932(97)00027-4.

Miller, Earl K. et al. (2003). *Neural correlates of categories and concepts*. DOI: 10.1016/S0959-4388(03)00037-0.

Miller, Kevin J (2018). "Value Representations in Orbitofrontal Cortex Drive Learning , but not Choice". In: *bioRXiv Prepr.* Pp. 1–25. DOI: 10.1101/245720. URL: https://www.biorxiv.org/content/biorxiv/early/2018/01/10/245720.1.full.pdf.

Milosavljevic, Milica et al. (2010). "The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure". In: *Judgm. Decis. Mak.* 5.6, pp. 437–449. ISSN: 19302975. DOI: 10.2139/ssrn.1901533.

Mink, Jonathan W. (1996). "The basal ganglia: Focused selection and inhibition of competing motor programs". In: *Prog. Neurobiol.* 50.4, pp. 381–425. ISSN: 03010082. DOI: 10.1016/S0301-0082(96)00042-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0301008296000421http://www.sciencedirect.com/science/article/pii/S0301008296000421.

Minsky, Marvin Lee (1954). *Theory of neural-analog reinforcement systems and its application to the brain model problem*. Princeton University.

Montague, P. Read and Gregory S. Berns (2002). *Neural economics and the biological substrates of valuation*. DOI: 10.1016/S0896-6273(02)00974-1.

Montague, P. Read, Peter Dayan, and Terrence J. Sejnowski (1996). "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". In: *J. Neurosci.* ISSN: 02706474.

Morecraft, Robert J. et al. (2007). "Amygdala interconnections with the cingulate motor cortex in the rhesus monkey". In: *J. Comp. Neurol.* ISSN: 00219967. DOI: 10.1002/cne.21165.

Morrison, Sara E. et al. (2011). "Different Time Courses for Learning-Related Changes in Amygdala and Orbitofrontal Cortex". In: *Neuron* 71.6, pp. 1127–1140. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.07.016.

Murray, Elisabeth A. and Alicia Izquierdo (2007). "Orbitofrontal cortex and amygdala contributions to affect and action in primates". In: *Ann. N. Y. Acad. Sci.* 1121, pp. 273–296. ISSN: 17496632. DOI: 10.1196/annals.1401.021.

Murray, Elisabeth A. and Peter H. Rudebeck (2018). "Specializations for reward-guided decision-making in the primate ventral prefrontal cortex". In: *Nat. Rev. Neurosci.* 19.7, pp. 404–417. ISSN: 14710048. DOI: 10.1038/s41583-018-0013-4. URL: http://dx.doi.org/10.1038/s41583-018-0013-4.

Nallapu, B.T. Bhargav Teja and N.P. Nicolas P. Rougier (2016). "Dynamics of reward based decision making: A computational study". In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 9886 LNCS, pp. 322–329. ISSN: 16113349. DOI: 10.1007/978-3-319-44778-0_38.

Nisenbaum, Eric S. and Charles J. Wilson (1995). "Potassium currents responsible for inward and outward rectification in rat neostriatal spiny projection neurons". In: *J. Neurosci.* ISSN: 02706474.

Niv, Yael et al. (2007). "Tonic dopamine: Opportunity costs and the control of response vigor". In: *Psychopharmacology (Berl).* ISSN: 00333158. DOI: 10.1007/s00213-006-0502-4.

Nogueira, Ramon et al. (2017). "Lateral orbitofrontal cortex anticipates choices and integrates prior with current information". In: *Nat. Commun.* 8. ISSN: 20411723. DOI: 10.1038/ncomms14823.

Noonan, M. P. et al. (2010). "Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex". In: *Proc. Natl. Acad. Sci.* 107.47, pp. 20547–20552. ISSN: 0027-8424. DOI: 10.1073/pnas.1012246107.

Noonan, M. P., R. B. Mars, and M. F. S. Rushworth (2011). "Distinct Roles of Three Frontal Cortical Areas in Reward-Guided Behavior". In: *J. Neurosci.* 31.40, pp. 14399–14412. ISSN: 0270-6474. DOI: 10.1523/jneurosci.6456-10.2011.

Noonan, M. P. et al. (2012). "Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement". In: *Eur. J. Neurosci.* 35.7, pp. 997–1010. ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2012.08023.x.

O'Doherty, J. et al. (2001). "Abstract reward and punishment representations in the human orbitofrontal cortex". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/82959.

O'Doherty, J. et al. (2003). "Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness". In: *Neuropsychologia* 41.2, pp. 147–155. ISSN: 00283932. DOI: 10.1016/S0028-3932(02)00145-8.

O'Doherty, John P. (2011). "Contributions of the ventromedial prefrontal cortex to goal-directed action selection". In: *Ann. N. Y. Acad. Sci.* 1239.1, pp. 118–129. ISSN: 17496632. DOI: 10.1111/j.1749-6632.2011.06290.x.

O'Doherty, John P. et al. (2006). "Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2005.11.014.

Olds, James and Peter Milner (1954). "POSITIVE REINFORCEMENT PRODUCED BY ELECTRICAL STIMULATION OF SEPTAL AREA AND OTHER REGIONS OF RAT BRAIN". In: *J. Comp. Physiol. Psychol.* ISSN: 00219940. DOI: 10.1037/h0058775.

O'Neill, Martin and Wolfram Schultz (2010). "Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2010.09.031.

Öngür, Dost, Amon T. Ferry, and Joseph L. Price (2003). "Architectonic subdivision of the human orbital and medial prefrontal cortex". In: *J. Comp. Neurol.* 460.3, pp. 425–449. ISSN: 00219967. DOI: 10.1002/cne.10609. URL: http://doi.wiley.com/10.1002/cne.10609.

O'Reilly, Randall C., Seth A. Herd, and Wolfgang M. Pauli (2010). "Computational models of cognitive control". In: *Curr. Opin. Neurobiol.* 20.2, pp. 257–261. ISSN: 09594388. DOI: 10.1016/j.conb.2010.01.008. URL: http://dx.doi.org/10.1016/j.conb.2010.01.008.

Padoa-Schioppa, C. (2009). "Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex". In: *J. Neurosci.* 29.44, pp. 14004–14014. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3751-09.2009. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3751-09.2009.

Padoa-Schioppa, Camillo and John A. Assad (2006). "Neurons in the orbitofrontal cortex encode economic value". In: *Nature* 441.7090, pp. 223–226. ISSN: 14764687. DOI: 10.1038/nature04676. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/16633341http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2630027http://www.nature.com/articles/nature04676.

— (2008). "The representation of economic value in the orbitofrontal cortex is invariant for changes of menu". In: *Nat. Neurosci.* 11.1, pp. 95–102. ISSN: 10976256. DOI: 10.1038/nn2020.

Padoa-Schioppa, Camillo and Katherine E. Conen (2017). "Orbitofrontal Cortex: A Neural Circuit for Economic Decisions". In: *Neuron* 96.4, pp. 736–754. ISSN: 10974199. DOI: 10.1016/j.neuron.2017.09.031. URL: https://doi.org/10.1016/j.neuron.2017.09.031.

Padoa-Schioppa, Camillo, Lucia Jandolo, and Elisabetta Visalberghi (2006). "Multi-stage mental process for economic choice in capuchins". In: *Cognition* 99.1, B1–B13. ISSN: 00100277. DOI: 10.1016/j.cognition.2005.04.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027705001009.

Parkinson, John A. et al. (2001). "The role of the primate amygdala in conditioned reinforcement". In: *J. Neurosci.* ISSN: 02706474.

Pasquereau, B. et al. (2007). "Shaping of Motor Responses by Incentive Values through the Basal Ganglia". In: *J. Neurosci.* 27.5, pp. 1176–1183. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3745-06.2007.

Passingham, R E (1993). *The frontal lobes and voluntary action.* Oxford psychology series, No. 21. New York, NY, US: Oxford University Press, pp. xxii, 299–xxii, 299. ISBN: 0-19-852185-5 (Hardcover).

Paton, Joseph J. et al. (2006). "The primate amygdala represents the positive and negative value of visual stimuli during learning". In: *Nature* 439.7078, pp. 865–870. ISSN: 14764687. DOI: 10.1038/nature04490.

Pawlak, Verena and Jason N.D. Kerr (2008). "Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity". In: *J. Neurosci.* 28.10, pp. 2435–2446. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.4402-07.2008.

Pears, Andrew et al. (2003). "Lesions of the Orbitofrontal but not Medial Prefrontal Cortex Disrupt Conditioned Reinforcement in Primates". In: *J. Neurosci.* ISSN: 02706474.

Petrides, Michael (1996). "Specialized systems for the processing of mnemonic information within the primate frontal cortex". In: *Philos. Trans. R. Soc. B Biol. Sci.* ISSN: 09628436. DOI: 10.1098/rstb.1996.0130.

Petrides, Michael and D. N. Pandya (2002). "Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey". In: *Eur. J. Neurosci.* 16.2, pp. 291–310. ISSN: 0953816X. DOI: 10.1046/j.1460-9568.2001.02090.x. URL: http://doi.wiley.com/10.1046/j.1460-9568.2001.02090.x.

Pezzulo, Giovanni et al. (2011). "The mechanics of embodiment: A dialog on embodiment and computational modeling". In: *Front. Psychol.* 2.JAN, pp. 1–21. ISSN: 16641078. DOI: 10.3389/fpsyg.2011.00005.

Pickens, Charles L. et al. (2003). "Different Roles for Orbitofrontal Cortex and Basolateral Amygdala in a Reinforcer Devaluation Task". In: *J. Neurosci.* ISSN: 02706474.

Platt, Michael L. and Paul W. Glimcher (1999). "Neural correlates of decision variables in parietal cortex". In: *Nature*. ISSN: 00280836. DOI: 10.1038/22268.

Porrino, L. J., A. M. Crane, and P. S. Goldman-Rakic (1981). "Direct and indirect pathways from the amygdala to the frontal lobe in rhesus monkeys". In: *J. Comp. Neurol.* 198.1, pp. 121–136. ISSN: 10969861. DOI: 10.1002/cne.901980111.

Preuss, Todd M (1995). "Do rats have prefrontal cortex? The Rose-Woolsey-Akert program reconsidered". In: *Journal of cognitive neuroscience* 7.1, pp. 1–24.

Price, Joseph L. (2007). "Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions". In: *Ann. N. Y. Acad. Sci.* ISBN: 9781573316835. DOI: 10.1196/annals.1401.008.

Quilodran, René, Marie Rothé, and Emmanuel Procyk (2008). "Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.11.031.

Rangel, Antonio, Colin Camerer, and P. Read Montague (2008). *A framework for studying the neurobiology of value-based decision making*. DOI: 10.1038/nrn2357.

Ravlin, Elizabeth C. and Bruce M. Meglino (1987). "Effect of values on perception and decision making: A study of alternative work values measures." In: *J. Appl. Psychol.* 72.4, pp. 666–673. ISSN: 1939-1854(Electronic),0021-9010(Print). DOI: 10.1037/0021-9010.72.4.666.

Redgrave, P., T. J. Prescott, and K. Gurney (1999). "The basal ganglia: A vertebrate solution to the selection problem?" In: *Neuroscience* 89.4, pp. 1009–1023. ISSN: 03064522. DOI: 10.1016/S0306-4522(98)00319-4.

Rescorla, Robert A and Allan R Wagner (1972). "A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement BT - Clasical conditioning II: current research and theory". In: *Clasical Cond. II Curr. Res. theory*. Ed. by A H Black and W F Prokasy. New York: Appleton-Century-Crofts, pp. 64–99. URL: http://jshd.pubs.asha.org/Article.aspx?articleid=1775379papers3://publication/uuid/1A852E2C-BD69-44DE-BAE6-3DFAFA705330.

Riceberg, Justin S. and Matthew L. Shapiro (2012). "Reward stability determines the contribution of orbitofrontal cortex to adaptive behavior". In: *J. Neurosci.* 32.46, pp. 16402–16409. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0776-12.2012.

Rich, Erin L., Frederic M. Stoll, and Peter H. Rudebeck (2018). "Linking dynamic patterns of neural activity in orbitofrontal cortex with decision making". In: *Curr. Opin.*

Neurobiol. 49, pp. 24–32. ISSN: 18736882. DOI: 10.1016/j.conb.2017.11.002. URL: http://dx.doi.org/10.1016/j.conb.2017.11.002.

Robbins, Trevor W. and Verity J. Brown (1990). "The Role of the Striatum in the Mental Chronometry of Action: A Theoretical Review". In: *Rev. Neurosci.* ISSN: 21910200. DOI: 10.1515/REVNEURO.1990.2.4.181.

Roberts, A. C. (2006). "Primate orbitofrontal cortex and adaptive behaviour". In: *Trends Cogn. Sci.* 10.2, pp. 83–90. ISSN: 13646613. DOI: 10.1016/j.tics.2005.12.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661305003347.

Rolls, E T (2009). "The Anterior and Midcingulate Cortices and Reward". In: *Cingulate Neurobiol. Dis.*

Rolls, E. T. et al. (1994). "Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage". In: *J. Neurol. Neurosurg. Psychiatry* 57.12, pp. 1518–1524. ISSN: 00223050. DOI: 10.1136/jnnp.57.12.1518.

Rolls, Edmund T. and Fabian Grabenhorst (2008). *The orbitofrontal cortex and beyond: From affect to decision-making*. DOI: 10.1016/j.pneurobio.2008.09.001.

Rolls, Edmund T., Ciara McCabe, and Jerome Redoute (2008). "Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task". In: *Cereb. Cortex*. ISSN: 10473211. DOI: 10.1093/cercor/bhm097.

Rolls, Edmund T., Fabian Grabenhorst, and Gustavo Deco (2010). "Choice, difficulty, and confidence in the brain". In: *Neuroimage*. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2010.06.073.

ROSE, J. E. and C. N. WOOLSEY (1948). "The orbitofrontal cortex and its connections with the mediodorsal nucleus in rabbit, sheep and cat". In: *Res. Publ. Assoc. Res. Nerv. Ment. Dis.* ISSN: 00917443.

Rudebeck, P. H. and E. A. Murray (2011). "Dissociable Effects of Subtotal Lesions within the Macaque Orbital Prefrontal Cortex on Reward-Guided Behavior". In: *J. Neurosci.* 31.29, pp. 10569–10578. ISSN: 0270-6474. DOI: 10.1523/jneurosci.0091-11.2011. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0091-11.2011.

Rudebeck, P. H. et al. (2008). "Frontal Cortex Subregions Play Distinct Roles in Choices between Actions and Stimuli". In: *J. Neurosci.* 28.51, pp. 13775–13785. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3541-08.2008.

Rudebeck, Peter H. and Elisabeth A. Murray (2014). *The orbitofrontal oracle: Cortical mechanisms for the prediction and evaluation of specific behavioral outcomes*. DOI: 10.1016/j.neuron.2014.10.049.

Rudebeck, Peter H. et al. (2006). "Separate neural pathways process different decision costs". In: *Nat. Neurosci.* 9.9, pp. 1161–1168. ISSN: 10976256. DOI: 10.1038/nn1756.

Rudebeck, Peter H. et al. (2013). "Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating". In: *Nat. Neurosci.* 16.8, pp. 1140–1145. ISSN: 10976256. DOI: 10.1038/nn.3440. URL: http://dx.doi.org/10.1038/nn.3440.

Rudebeck, Peter H. et al. (2017). "Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes". In: *Neuron* 95.5, 1208–1220.e5. ISSN: 10974199. DOI: 10.1016/j.neuron.2017.07.042. URL: http://dx.doi.org/10.1016/j.neuron.2017.07.042.

Rushworth, Matthew FS et al. (2007). *Functional organization of the medial frontal cortex*. DOI: 10.1016/j.conb.2007.03.001.

Rushworth, Matthew FS et al. (2012). "Valuation and decision-making in frontal cortex: one or many serial or parallel systems?" In: *Curr. Opin. Neurobiol.* 22.6, pp. 946–955. ISSN: 09594388. DOI: 10.1016/j.conb.2012.04.011. URL: http://dx.doi.org/10.1016/j.conb.2012.04.011https://linkinghub.elsevier.com/retrieve/pii/S0959438812000694.

Samuelson, P. A. (1938). "A Note on the Pure Theory of Consumer's Behaviour". In: *Economica*. ISSN: 00130427. DOI: 10.2307/2548836.

Sandstrom, Michael I. and George V. Rebec (2003). "Characterization of striatal activity in conscious rats: Contribution of NMDA and AMPA/kainate receptors to both spontaneous and glutamate-driven firing". In: *Synapse*. ISSN: 08874476. DOI: 10.1002/syn.10142.

Schoenbaum, Geoffrey, Andrea A. Chiba, and Michela Gallagher (1998). "Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/407.

— (1999). "Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning". In: *J. Neurosci.* 19.5, pp. 1876–1884. ISSN: 02706474.

Schoenbaum, Geoffrey et al. (2002). "Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations". In: *Neuroreport* 13.6, pp. 885–890. ISSN: 09594965. DOI: 10.1097/00001756-200205070-00030.

Schoenbaum, Geoffrey et al. (2003). "Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala". In: *Neuron* 39.5, pp. 855–867. ISSN: 08966273. DOI: 10.1016/S0896-6273(03)00474-4.

Schoenbaum, Geoffrey, Matthew R. Roesch, and Thomas A. Stalnaker (2006). *Orbitofrontal cortex, decision-making and drug addiction*. DOI: 10.1016/j.tins.2005.12.006.

Schrodt, Fabian et al. (2017). "Mario Becomes Cognitive". In: *Top. Cogn. Sci.* 9.2, pp. 343–373. ISSN: 17568765. DOI: 10.1111/tops.12252.

Schuck, Nicolas W. et al. (2016). "Human Orbitofrontal Cortex Represents a Cognitive Map of State Space". In: *Neuron*. ISSN: 10974199. DOI: 10.1016/j.neuron.2016.08.019.

Schuck, Nicolas W., Robert Wilson, and Yael Niv (2018). "A State Representation for Reinforcement Learning and Decision-Making in the Orbitofrontal Cortex". In: *Goal-Directed Decis. Mak.* Elsevier, pp. 259–278. DOI: 10.1016/B978-0-12-812098-9.00012-7. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128120989000127.

Schultz, W., P. Dayan, and P. R. Montague (1997). "A neural substrate of prediction and reward". In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.275.5306.1593.

Schultz, Wolfram (2015). "Neuronal reward and decision signals: From theories to data". In: *Physiol. Rev.* 95.3, pp. 853–951. ISSN: 15221210. DOI: 10.1152/physrev.00023.2014.

Seamans, Jeremy K., Christopher C. Lapish, and Daniel Durstewitz (2008). "Comparing the prefrontal cortex of rats and primates: Insights from electrophysiology". In: *Neurotox. Res.* ISSN: 10298428. DOI: 10.1007/BF03033814.

Seo, Hyojung and Daeyeol Lee (2007). "Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.2369-07.2007.

Sequeira, Pedro, Francisco S. Melo, and Ana Paiva (2011). "Emotion-based intrinsic motivation for reinforcement learning agents". In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. ISBN: 9783642245992. DOI: 10.1007/978-3-642-24600-5_36.

Serences, John T. (2008). "Value-Based Modulations in Human Visual Cortex". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2008.10.051.

Sescousse, Guillaume et al. (2013). "Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies". In: *Neurosci. Biobehav. Rev.* 37.4, pp. 681–696. ISSN: 01497634. DOI: 10.1016/j.neubiorev.2013.02.002. URL: http://dx.doi.org/10.1016/j.neubiorev.2013.02.002.

Shallice, T. (1982). "Specific impairments of planning." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* ISSN: 09628436. DOI: 10.1098/rstb.1982.0082.

Shallice, Tim (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Shidara, Munetaka and Barry J. Richmond (2002). "Anterior cingulate: Single neuronal signals related to degree of reward expectancy". In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.1069504.

Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489. DOI: 10.1038/nature16961.

Singla, Rajat, Ravi Raja Ganta, and Kavita Vemuri (2017). "An Exergame Themed on the Power of Religious Belief for Stroke/Motor Rehabilitation". In: *Hci 2018*, pp. 1–6. DOI: 10.14236/ewic/hci2018.155.

Skinner, B. F. (1938). "The behavior of organisms: an experimental analysis. Appleton-Century". In: *New York*.

Skinner, Burrhus Frederic (1965). *Science and human behavior*. 92904. Simon and Schuster.

Sommer, Marc A. and Robert H. Wurtz (2004). "What the Brain Stem Tells the Frontal Cortex. I. Oculomotor Signals Sent from Superior Colliculus to Frontal Eye Field Via Mediodorsal Thalamus". In: *J. Neurophysiol.* ISSN: 00223077. DOI: 10.1152/jn.00738.2003.

Stalnaker, Thomas A. et al. (2007). "Basolateral Amygdala Lesions Abolish Orbitofrontal-Dependent Reversal Impairments". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.02.014.

Stalnaker, Thomas A. et al. (2010). "Neural correlates of stimulus-response and response-outcome associations in dorsolateral versus dorsomedial striatum". In: *Front. Integr. Neurosci.* ISSN: 16625145. DOI: 10.3389/fnint.2010.00012.

Stalnaker, Thomas A., Nisha K. Cooch, and Geoffrey Schoenbaum (2015). "What the orbitofrontal cortex does not do". In: *Nat. Neurosci.* 18.5, pp. 620–627. ISSN: 1097-6256. DOI: 10.1038/nn.3982. URL: http://www.nature.com/articles/nn.3982.

Stalnaker, Thomas A. et al. (2016). "Cholinergic interneurons use orbitofrontal input to track beliefs about current state". In: *J. Neurosci.* ISSN: 15292401. DOI: 10.1523/JNEUROSCI.0157-16.2016.

Steiner, Adam P. and A. David Redish (2014). "Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task". In: *Nat. Neurosci.* 17.7, pp. 995–1002. ISSN: 1097-6256. DOI: 10.1038/nn.3740. URL: http://dx.doi.org/10.1038/nn.3740http://www.nature.com/articles/nn.3740.

Strait, Caleb E., Tommy C. Blanchard, and Benjamin Y. Hayden (2014). "Reward value comparison via mutual inhibition in ventromedial prefrontal cortex". In: *Neuron*. ISSN: 10974199. DOI: 10.1016/j.neuron.2014.04.032.

Strannegård, Claes et al. (2017). "Generic animats". In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. ISBN: 9783319637020. DOI: 10.1007/978-3-319-63703-7_3.

Strannegård, Claes et al. (2018). "Learning and decision-making in artificial animals". In: *J. Artif. Gen. Intell.* 9.1, pp. 55–82. DOI: 10.2478/jagi-2018-0002.

Strock, Anthony, Nicolas Rougier, and Xavier Hinaut (2019). "Using Conceptors to Transfer Between Long-Term and Short-Term Memory". In: *Int. Conf. Artif. Neural Networks*. Springer, pp. 19–23.

Sugden, Robert et al. (1996). "Economic Choice Theory: An Experimental Analysis of Animal Behaviour." In: *Econ. J.* ISSN: 00130133. DOI: 10.2307/2235234.

Sugrue, Leo P., Greg S. Corrado, and William T. Newsome (2004). "Matching behavior and the representation of value in the parietal cortex". In: *Science (80-.).* 304.5678, pp. 1782–1787. ISSN: 00368075. DOI: 10.1126/science.1094765.

Sutton, Richard S and Andrew G Barto (1998). *Reinforcement Learning : An Introduction*. ISBN: 0262193981.

Takahashi, Yuji K. et al. (2011). "Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex". In: *Nat. Neurosci.* ISSN: 10976256. DOI: 10.1038/nn.2957.

Tanaka, Saori C., Bernard W. Balleine, and John P. O'Doherty (2008). "Calculating consequences: Brain systems that encode the causal effects of actions". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1808-08.2008.

Tanaka, Saori C. et al. (2016). "Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops". In: *Behav. Econ. Prefer. Choices, Happiness*, pp. 593–616. ISBN: 9784431554028. DOI: 10.1007/978-4-431-55402-8_22.

Thorndike, Edward L (1898). "Animal intelligence: An experimental study of the associative processes in animals." In: *Psychol. Rev. Monogr. Suppl.* 2.4, pp. i–109. ISSN: 0096-9753(Print). DOI: 10.1037/h0092987.

Tom, Sabrina M. et al. (2007). "The neural basis of loss aversion in decision-making under risk". In: *Science (80-.).* ISSN: 00368075. DOI: 10.1126/science.1134239.

Topalidou, Meropi et al. (2015). "[Re] Interaction between cognitive and motor cortico-basal ganglia loops during decision making : a computational study To cite this version : HAL Id : hal-01201790 Re Science [Re] Interaction between cognitive and motor cortico-basal ganglia loops du". In: pp. 0–6.

Topalidou, Meropi et al. (2018). "A computational model of dual competition between the basal ganglia and the cortex". In: *eNeuro* 5.6, ENEURO.0339–17.2018. ISSN: 23732822. DOI: 10.1523/ENEURO.0339-17.2018. URL: http://eneuro.org/lookup/doi/10.1523/ENEURO.0339-17.2018.

Tsuchida, Ami, Bradley B. Doll, and Lesley K. Fellows (2010). "Beyond reversal: A critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback". In: *J. Neurosci.* 30.50, pp. 16868–16875. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1958-10.2010.

Tsutsui, Ken Ichiro et al. (2016). "A dynamic code for economic object valuation in prefrontal cortex neurons". In: *Nat. Commun.* 7. ISSN: 20411723. DOI: 10.1038/ncomms12554.

Uylings, Harry BM, Henk J Groenewegen, and Bryan Kolb (2003). "Do rats have a prefrontal cortex?" In: *Behavioural brain research* 146.1-2, pp. 3–17.

Valentin, V. V., A. Dickinson, and J. P. O'Doherty (2007). "Determining the Neural Substrates of Goal-Directed Learning in the Human Brain". In: *J. Neurosci.* 27.15, pp. 4019–4026. ISSN: 0270-6474. DOI: 10.1523/jneurosci.0564-07.2007.

Van Hoesen, Gary W, Robert J Morecraft, and Brent A Vogt (1993). "Connections of the monkey cingulate cortex". In: *Neurobiol. cingulate cortex limbic thalamus.* Springer, pp. 249–284.

Vanni-Mercier, Giovanna et al. (2009). "The hippocampus codes the uncertainty of cue-outcome associations: An intracranial Electro physiological study in humans". In: *J. Neurosci.* 29.16, pp. 5287–5294. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.5298-08.2009.

Varela, F (1992). "Autopoiesis and a biology of intentionality". In: *Autopoiesis Percept. A Work. with ESPRIT …*

Varela, Francisco J. (1991). "Organism: A Meshwork of Selfless Selves". In: DOI: 10.1007/978-94-011-3406-4_5.

Verplanken, Bas and Rob W. Holland (2002). *Motivated decision making: Effects of activation and self-centrality of values on choices and behavior.* Verplanken, Bas: U Tromsø, Dept of Psychology, Tromsø, Norway, N-9037, verplanken@psyk.uit.no. DOI: 10.1037/0022-3514.82.3.434.

Verschure, Paul F.M.J. (2012). "Distributed Adaptive Control: A theory of the Mind, Brain, Body Nexus". In: *Biol. Inspired Cogn. Archit.* 1, pp. 55–72. ISSN: 2212683X. DOI: 10.1016/j.bica.2012.04.005.

Vitay, Julien and Fred H. Hamker (2014). "Timing and expectation of reward: A neurocomputational model of the afferents to the ventral tegmental area". In: *Front. Neurorobot.* ISSN: 16625218. DOI: 10.3389/fnbot.2014.00004.

Vlachos, Ioannis et al. (2011). "Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala". In: *PLoS Comput. Biol.* ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1001104.

Von Neumann, J and O Morgenstern (1944). *Theory of games and economic behavior.* Princeton, NJ, US: Princeton University Press, pp. xviii, 625–xviii, 625.

Voorn, Pieter et al. (2004). "Putting a spin on the dorsal-ventral divide of the striatum". In: *Trends Neurosci.* 27.8, pp. 468–474. ISSN: 01662236. DOI: 10.1016/j.tins.2004.06.006.

Walker, A. Earl (1940). "A cytoarchitectural study of the prefrontal area of the macaque monkey". In: *J. Comp. Neurol.* ISSN: 10969861. DOI: 10.1002/cne.900730106.

Wallis, Jonathan D. (2012). "Cross-species studies of orbitofrontal cortex and value-based decision-making". In: *Nat. Neurosci.* 15.1, pp. 13–19. ISSN: 10976256. DOI: 10.1038/nn.2956. URL: http://dx.doi.org/10.1038/nn.2956.

Wallis, Jonathan D., Kathleen C. Anderson, and Earl K. Miller (2001). "Single neurons in prefrontal cortex encode abstract roles". In: *Nature*. ISSN: 00280836. DOI: 10.1038/35082081.

Walton, Mark E. et al. (2003). "Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions". In: *J. Neurosci.* ISSN: 02706474.

Walton, Mark E. et al. (2010). "Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning". In: *Neuron* 65.6, pp. 927–939. ISSN: 08966273. DOI: 10.1016/j.neuron.2010.02.027. URL: http://dx.doi.org/10.1016/j.neuron.2010.02.027https://linkinghub.elsevier.com/retrieve/pii/S0896627310001443.

Walton, Mark E. et al. (2011). "Giving credit where credit is due: Orbitofrontal cortex and valuation in an uncertain world". In: *Ann. N. Y. Acad. Sci.* 1239.1, pp. 14–24. ISSN: 17496632. DOI: 10.1111/j.1749-6632.2011.06257.x.

Wang, Jane X et al. (2018). "Prefrontal cortex as a meta-reinforcement learning system". In: *Nat. Neurosci.* 21.6, pp. 860–868. ISSN: 1097-6256. DOI: 10.1038/s41593-018-0147-8. URL: https://www.biorxiv.org/content/early/2018/04/13/295964{\%}0Ahttp://www.nature.com/articles/s41593-018-0147-8http://www.nature.com/articles/s41593-018-0147-8.

Wang, Tao, Jie Deng, and Bin He (2004). "Classifying EEG-based motor imagery tasks by means of time-frequency synthesized spatial patterns". In: *Clin. Neurophysiol.* ISSN: 13882457. DOI: 10.1016/j.clinph.2004.06.022.

Wang, Xiao Jing (2008). "Decision Making in Recurrent Neuronal Circuits". In: *Neuron* 60.2, pp. 215–234. ISSN: 08966273. DOI: 10.1016/j.neuron.2008.09.034. URL: http://dx.doi.org/10.1016/j.neuron.2008.09.034.

Wickens, Jeffery R. et al. (2007). "Dopaminergic mechanisms in actions and habits". In: *J. Neurosci.* 27.31, pp. 8181–8183. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1671-07.2007.

Williams, S. Mark and Patricia S. Goldman-Rakic (1993). "Characterization of the dopaminergic innervation of the primate frontal cortex using a dopamine-specific antibody". In: *Cereb. Cortex*. ISSN: 14602199. DOI: 10.1093/cercor/3.3.199.

Wilson, Charles J. and Philip M. Groves (1981). "Spontaneous firing patterns of identified spiny neurons in the rat neostriatum". In: *Brain Res.* ISSN: 00068993. DOI: 10.1016/0006-8993(81)90211-0.

Wilson, Robert C. et al. (2014). "Orbitofrontal cortex as a cognitive map of task space". In: *Neuron* 81.2, pp. 267–279. ISSN: 10974199. DOI: 10.1016/j.neuron.2013.11.005. URL: http://dx.doi.org/10.1016/j.neuron.2013.11.005.

Xia, Yanfang et al. (2011). "Nucleus accumbens medium spiny neurons target non-dopaminergic neurons in the ventral tegmental area". In: *J. Neurosci.* ISSN: 02706474. DOI: 10.1523/JNEUROSCI.1504-11.2011.

Yang, Yaling and Adrian Raine (2009). *Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: A meta-analysis*. DOI: 10.1016/j.pscychresns.2009.03.012.

Yin, Henry H., Barbara J. Knowlton, and Bernard W. Balleine (2004). "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning". In: *Eur. J. Neurosci.* ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2004.03095.x.

Yu, Angela J. and Peter Dayan (2005). "Uncertainty, neuromodulation, and attention". In: *Neuron*. ISSN: 08966273. DOI: 10.1016/j.neuron.2005.04.026.

Zald, David H. et al. (2014). "Meta-Analytic Connectivity Modeling Reveals Differential Functional Connectivity of the Medial and Lateral Orbitofrontal Cortex". In: *Cereb. Cortex* 24.1, pp. 232–248. ISSN: 1460-2199. DOI: 10.1093/cercor/bhs308. URL: https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhs308.

Zhang, Zhewei et al. (2018). “A neural network model for the orbitofrontal cortex and task space acquisition during reinforcement learning”. In: *PLoS Comput. Biol.* 14.1, pp. 1–24. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005925.

Annexes

Sommaire

A . Definitions

Définition 1. Credit assignment The ability to learn that a particular outcome (in experiments, this is typically food or fluid) was produced by a particular choice.

Définition 2. Value- based decision- making The ability to make informed choices that optimize subjective value.

Définition 3. menu invariance Values that are assigned to different options on the ”menu” do not vary depending on what else is on the ”menu”.

Définition 4. Cognitive map A neural representation of stimuli, actions and other sensory features that occur in association with outcomes in a multidimensional array. The cognitive map has been theorized to guide value- based decision-making.

Définition 5. Aspiration lesion A technique for removing grey matter (that is, neurons) that is based on subpial aspiration of tissue. Lesions are typically carried out with the aid of an operating microscope.

Définition 6. Excitotoxic lesions Lesions created using a technique for selectively removing grey matter (that is, neurons) and sparing white matter (that is, axons) that is based on

the injection of neurotoxins. injections are often carried out via a stereotaxic approach based on coordinates obtained from magnetic resonance images of the brain.