Dynamique intracérébrale de l'apprentissage par renforcement chez l'humain

Maëlle Gueguen

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : PCN - Sciences cognitives, psychologie et neurocognition

Arrêté ministériel : 25 mai 2016

Présentée par

Maëlle GUEGUEN

Thèse dirigée par **Julien BASTIN (EDISCE)**, habilitation ponctuelle obtenue en 2014 pour GUEGUEN, UGA, et codirigée par **Jean-Philippe LACHAUX** INSERM

préparée au sein du **Laboratoire Grenoble Institut des Neurosciences**
dans **l'École Doctorale Ingénierie pour la santé la Cognition et l'Environnement**

Dynamique intracérébrale de l'apprentissage par renforcement chez l'humain

Intracerebral dynamics of human reinforcement learning

Thèse soutenue publiquement le **1 décembre 2017**, devant le jury composé de :

Monsieur JULIEN BASTIN
MAITRE DE CONFERENCES, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur EMMANUEL PROCYK
DIRECTEUR DE RECHERCHE, INSERM DELEGATION RHONE-ALPES, AUVERGNE, Rapporteur
Monsieur FRANCK VIDAL
PROFESSEUR, UNIVERSITE AIX-MARSEILLE, Rapporteur
Monsieur ALEXANDRE EUSEBIO
PRATICIEN HOSPITALIER, UNIVERSITE AIX-MARSEILLE, Examinateur
Monsieur MARTIAL MERMILLOD
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Président
Monsieur JEAN-PHILIPPE LACHAUX
DIRECTEUR DE RECHERCHE, INSERM DELEGATION RHONE-ALPES, AUVERGNE, Co-directeur de thèse

REMERCIEMENTS

Je n'aurais pas pu rédiger cette thèse sans l'aide et le soutien de beaucoup de personnes au cours de ces trois dernières années.

Tout d'abord, je voudrais remercier Julien Bastin, mon directeur de thèse, pour m'avoir donné l'opportunité de prendre part à ce projet de recherche. Il m'a appris à devenir une bonne chercheuse et à identifier les qualités qui sont celles d'une bonne enseignante en me permettant d'encadrer des étudiants. Il m'a montré l'intérêt de s'investir dans son travail avec passion et rigueur, ainsi qu'à communiquer cette passion au plus grand nombre. Je le remercie également énormément du temps qu'il a passé avec moi sur la dernière ligne droite pour s'assurer que le manuscrit était à la hauteur de nos attentes.

Je tiens également à remercier Jean-Philippe Lachaux, mon co-directeur de thèse, pour m'avoir guidée sur les aspects théoriques des enregistrements intracérébraux et m'avoir donné accès aux patients de Lyon. Ses livres sur l'attention m'ont également accompagnée au cours de cette thèse, y compris à l'autre bout du monde.

Merci également à Olivier David de m'avoir accueillie dans son équipe au cours de mon master et de ma thèse. Son soutien discret à mon projet professionnel a été très appréciable. Sa façon d'avoir monté une grande équipe multidisciplinaire a été un excellent moyen pour moi de prendre du recul sur mon projet et de savoir mieux communiquer dessus avec des non spécialistes.

Ce travail n'aurait pu être réalisé sans la participation de Stefano Palminteri et de Mathias Pessiglione. Merci à vous de m'avoir permis de travailler avec PBLT, pour ces échanges tout au long de la thèse quant aux résultats. Un grand merci à Stefano pour m'avoir aidée à développer le paradigme RISK, c'était très inspirant et formateur.

Un immense merci aux membres du Laboratoire de Physiopathologie de l'Epilepsie du Pr Philippe Kahane pour leur dévouement à leurs patients, la rigueur professionnelle et la bonne humeur permanente dont ils font perpétuellement preuve. Vous m'avez prouvé que de pouvoir travailler avec des patients et des cliniciens est une facette passionnante et rare du travail de chercheur. Je ne saurais vous remercier assez pour tout ce que vous m'avez apporté tant professionnellement que personnellement. Merci à Philippe de ton soutien de la première heure pour développer une nouvelle étude avec les patients et en ta foi en moi et en ce travail. Merci au Dr Lorella Minotti, je me souviendrai toujours de ta vision malicieuse

pour déclencher des crises « startle ». Merci à Patricia Boschetti et à Marie Pierre Noto pour avoir été à mes côtés depuis mon master, et m'avoir laissé la possibilité de travailler avec vos patients. Votre patience est infinie, tout comme le respect que j'ai pour vous. Votre fou rire en pleine manip n'est qu'une des raisons qui font que ce labo est génial. Merci aux internes et doctorants du labo : Sébastien Gonthier, Pauline Cuisenier, Ricardo Amorim. Merci au Dr Dominique Hoffmann pour ces discussions philosophiques qui prouvent que tous les neurochirurgiens ne sont pas dans leur tour de verre. Merci beaucoup aux secrétaires Virginie Cantale, Christelle Diaspara et Valérie Damon d'avoir supporté mes allers et venues dans leurs bureaux.

Merci beaucoup à Mathilde Petton pour avoir enregistré des patients pour moi à Lyon. Encore bravo pour ta thèse ! On entendra parler encore longtemps des « petits pois de l'attention ».

Merci à toute l'équipe des NeuroDocs pour avoir su faire de l'organisation du « European Meeting of Neuroscience for PhD students » une expérience joyeuse, de laquelle nous avons tous gagné. Je remercie tout particulièrement Julie Anne Rodier et Rebecca Powell, l'équipe NeuroHebdo, parce que c'est bien plus sympa de râler ensemble et de se changer les idées en s'amusant ! Merci aussi aux doctorants et étudiants du GIN, anciens et actuels, pour avoir tous été une part intégrante de cette grande expérience de vie. Merci également à Marie Claude Zanoni, alias MCZ, pour m'avoir sauvé la vie de nombreuses fois pour partir à Lyon en dernière minute. Discuter avec toi est toujours aussi agréable. Tu es vraiment la Chuck Norris des gestionnaires.

Enfin, je tiens à remercier tous mes proches, amis et ma famille pour avoir été là toutes ces années. Et comme rien n'est plus agréable que de vous retrouver, je vous remercierai de vive voix !

Je remercie la Région Auvergne Rhône-Alpes qui a financé cette thèse pendant trois ans dans le cadre de l'ARC2 Bien Vieillir.

Pour finir, je souhaite remercier les 115 patients et sujets sains ayant donné de leur temps pour que je puisse obtenir ces résultats. Ce travail est aussi le vôtre, j'espère qu'il vous fera honneur.

LISTE DES ÉTUDES

Etude 1 : *Rewards and punishment learning differentially modulates intracerebral brain dynamics.* (Maëlle CM Gueguen, Jean-Philippe Lachaux, Philippe Kahane, Pablo Billeke, Mathias Pessiglione, Julien Bastin)

Etude 2 : *Theta power is modulated in the human limbic thalamus during instrumental learning.* (Maëlle CM Gueguen, Jean-Philippe Lachaux, Lorella Minotti, Vincent Navarro, Philippe Kahane, Pablo Billeke, Mathias Pessiglione and Julien Bastin)

Etude 3 : *Rewards and punishment learning performance are differentially affected by risk*. (Maëlle Camille Marie Gueguen, Stefano Palminteri and Julien Bastin)

TABLE DES MATIERES

ABREVIATIONS

Anatomie :

ATV	Aire tegmentale ventrale
BA	Aire de Brodmann
CCA	Cortex cingulaire antérieur
CCM	Cortex cingulaire médian
CCP	Cortex cingulaire postérieur
CCR	Cortex cingulaire rétrospénial
dlPFC	Cortex préfrontal dorsolatéral
GABA	Acide γ-aminobutyrique (neurotransmetteur)
GPe	Globus pallidus externe
GPi	Globus pallidus interne
latOFC	Cortex orbitofrontal latéral
mPFC	Cortex préfrontal médian
NAcc	Noyau accumbens
NAT	Noyau antérieur du thalamus
NDMT	Noyau dorsomédian du thalamus
NSMT	Noyau submédian du thalamus
NST	Noyau sous-thalamique
OFC	Cortex orbitofrontal
vlOFC	Cortex orbitofrontal ventrolatéral
vlPFC	Cortex préfrontal ventrolatéral
vmPFC	Cortex préfrontal ventromédian
vPFC	Cortex préfrontal ventral
SNpc	Substance noire pars compacta
SNr	Substance noire réticulée

Techniques :

DCM	Dynamic causal modeling
DTI	Diffusion tensor imaging (tractography)
EEG	Electroencéphalographie
IRMf	Imagerie par résonnance magnétique fonctionnelle
sEEG	Electroencéphalographie stéréotaxique

Conditionnement :

RC	Réponse conditionnée
RI	Réponse inconditionnelle
RO	Réponse – Résultat (association, avec résultat = outcome)
SC	Stimulus conditionné
SI	Stimuli inconditionnel
SN	Stimulus neutre
SR	Stimulus – Réponse (association)
SRO	Stimulus – Réponse – Résultat (association, avec résultat = outcome)
SS	Stimulus – Stimulus (association)

RESUME

Chaque jour, nous prenons des décisions impliquant de choisir les options qui nous semblent les plus avantageuses, en nous basant sur nos expériences passées. Toutefois, les mécanismes et les bases neurales de l'apprentissage par renforcement restent débattus. D'une part, certains travaux suggèrent l'existence de deux systèmes opposés impliquant des aires cérébrales corticales et sous-corticales distinctes lorsque l'on apprend par la carotte ou par le bâton. D'autres part, des études ont montré une ségrégation au sein même de ces régions cérébrales ou entre des neurones traitant l'apprentissage par récompenses et celui par évitement des punitions. Le but de cette thèse était d'étudier la dynamique cérébrale de l'apprentissage par renforcement chez l'homme. Pour ce faire, nous avons utilisé des enregistrements intracérébraux réalisés chez des patients épileptiques pharmacorésistants pendant qu'ils réalisaient une tâche d'apprentissage probabiliste. Dans les deux premières études, nous avons investigué la dynamique de l'encodage des signaux de renforcement, et en particulier à celui des erreurs de prédiction des récompenses et des punitions. L'enregistrement de potentiels de champs locaux dans le cortex a mis en évidence le rôle central de l'activité à haute-fréquence gamma (50-150Hz). Les résultats suggèrent que le cortex préfrontal ventro-médian est impliqué dans l'encodage des erreurs de prédiction des récompenses alors que pour l'insula antérieure, le cortex préfrontal dorsolatéral sont impliqués dans l'encodage des erreurs de prédiction des punitions. De plus, l'activité neurale de l'insula antérieure permet de prédire la performance des patients lors de l'apprentissage. Ces résultats sont cohérents avec l'existence d'une dissociation au niveau cortical pour le traitement des renforcements appétitifs et aversifs lors de la prise de décision. La seconde étude a permis d'étudier l'implication de deux noyaux limbiques du thalamus au cours du même protocole cognitif. L'enregistrement de potentiels de champs locaux a mis en évidence le rôle des activités basse fréquence thêta dans la détection des renforcements, en particulier dans leur dimension aversive. Dans une troisième étude, nous avons testé l'influence du risque sur l'apprentissage par renforcement. Nous rapportons une aversion spécifique au risque lors de l'apprentissage par évitement des punitions ainsi qu'une diminution du temps de réaction lors de choix risqués permettant l'obtention de récompenses. Cela laisse supposer un comportement global tendant vers une aversion au risque lors de l'apprentissage par évitement des punitions et au contraire une attirance pour le risque lors de l'apprentissage par récompenses, suggérant que les mécanismes d'encodage du risque et de la valence pourraient être indépendants. L'amélioration de la compréhension des mécanismes cérébraux sous-tendant la prise de décision est importante, à la fois pour mieux comprendre les déficits motivationnels caractérisant plusieurs

pathologies neuropsychiatriques, mais aussi pour mieux comprendre les biais décisionnels que nous pouvons exhiber.

SUMMARY

We make decisions every waking day of our life. Facing our options, we tend to pick the most likely to get our expected outcome. Taking into account our past experiences and their outcome is mandatory to identify the best option. This cognitive process is called reinforcement learning. To date, the underlying neural mechanisms are debated. Despite a consensus on the role of dopaminergic neurons in reward processing, several hypotheses on the neural bases of reinforcement learning coexist: either two distinct opposite systems covering cortical and subcortical areas, or a segregation of neurons within brain regions to process reward-based and punishment-avoidance learning.

This PhD work aimed to identify the brain dynamics of human reinforcement learning. To unravel the neural mechanisms involved, we used intracerebral recordings in refractory epileptic patients during a probabilistic learning task. In the first study, we used a computational model to tackle the brain dynamics of reinforcement signal encoding, especially the encoding of reward and punishment prediction errors. Local field potentials exhibited the central role of high frequency gamma activity (50-150Hz) in these encodings. We report a role of the ventromedial prefrontal cortex in reward prediction error encoding while the anterior insula and the dorsolateral prefrontal cortex encoded punishment prediction errors. In addition, the magnitude of the neural response in the insula predicted behavioral learning and trial-to-trial behavioral adaptations. These results are consistent with the existence of two distinct opposite cortical systems processing reward and punishments during reinforcement learning. In a second study, we recorded the neural activity of the anterior and dorsomedial nuclei of the thalamus during the same cognitive task. Local field potentials recordings highlighted the role of low frequency theta activity in punishment processing, supporting an implication of these nuclei during punishment-avoidance learning. In a third behavioral study, we investigated the influence of risk on reinforcement learning. We observed a risk-aversion during punishment-avoidance, affecting the performance, as well as a risk-seeking behavior during reward-seeking, revealed by an increased reaction time towards appetitive risky choices. Taken together, these results suggest we are risk-seeking when we have something to gain and risk-averse when we have something to lose, in contrast to the prediction of the prospect theory.

Improving our common knowledge of the brain dynamics of human reinforcement learning could improve the understanding of cognitive deficits of neurological patients, but also the decision bias all human beings can exhibit.

I. CONTEXTE THEORIQUE ET SCIENTIFIQUE

A. Bases psychologiques de l'apprentissage par renforcement

L'apprentissage par renforcement est défini comme la capacité d'apprendre les associations entre stimuli, actions et conséquences de ces actions, qu'elles soient plaisantes (récompenses) ou non (punitions). De cet apprentissage découle une adaptation du comportement qui dépend directement de ces conséquences : ainsi une action récompensée aura tendance à être répétée tandis qu'une autre action qui elle sera punie se verra abandonnée. Au fur et à mesure des répétitions, l'apprentissage permet de tendre vers un comportement optimal afin d'atteindre un objectif donné.

Dans ce premier chapitre (I.A), je vais commencer par un rappel historique des travaux des pionniers de l'étude de l'apprentissage par renforcement. Cela permettra d'introduire les notions fondamentales sur lesquelles repose l'apprentissage par renforcement ainsi que le vocabulaire consacré. La terminologie utilisée dans ce manuscrit se base sur celle de la littérature en psychologie, où les termes de « conditionnement » et « d'apprentissage associatif » étaient historiquement utilisés. Plus tard, les travaux sur l'apprentissage automatique (machine learning) ont introduit le terme « d'apprentissage par renforcement ». Aujourd'hui, les termes « apprentissage par renforcement » et « conditionnement » réfèrent au même processus mais leur utilisation dépend de la formation scientifique des chercheurs (computationnelle ou psychologique). Pour plus de cohérence, j'emploierai le terme de « conditionnement » dans le cadre du rappel des travaux historiques dans le domaine, pour ensuite utiliser celui « d'apprentissage par renforcement » dans le reste du manuscrit basé sur la littérature moderne, bien que ces termes soient presque des synonymes. L'incertitude dans la prédiction des renforcements est un moteur clé de l'apprentissage par renforcement. Lorsque cette incertitude porte sur la probabilité de délivrance d'un renforcement, l'on parle de « risque ». Je m'intéresserai donc à l'influence du risque sur la prise de décision et plus particulièrement sur l'apprentissage par renforcement. Pour finir, après avoir abordé ces thèmes d'après une perspective psychologique, je développerai en quoi il est possible de compléter la compréhension de ces mécanismes de manière quantitative grâce à l'apport de la modélisation computationnelle.

1. Bref historique du conditionnement

Les prémices de la recherche sur le conditionnement remontent à la fin du XIXème siècle avec les travaux du physiologiste russe Ivan Petrovic Pavlov (1849-1936). A partir de 1889, il développa un protocole sur la digestion sur des chiens consistant à mesurer la quantité de salive produite suite à l'ingestion de nourriture. Alors qu'en début de protocole les chiens ne salivaient que lors de l'ingestion de nourriture, au fur et à mesure de ses expérimentations, Pavlov remarqua que les chiens se mettaient à saliver abondamment avant même qu'ils ne soient nourris, voire dès lors qu'ils entraient dans la pièce où ils allaient être nourris. Cette salivation avait donc lieu en réponse à une stimulation contextuelle (leur présence dans la pièce d'expérimentation) et non plus physiologique (la présence de nourriture dans leur bouche).

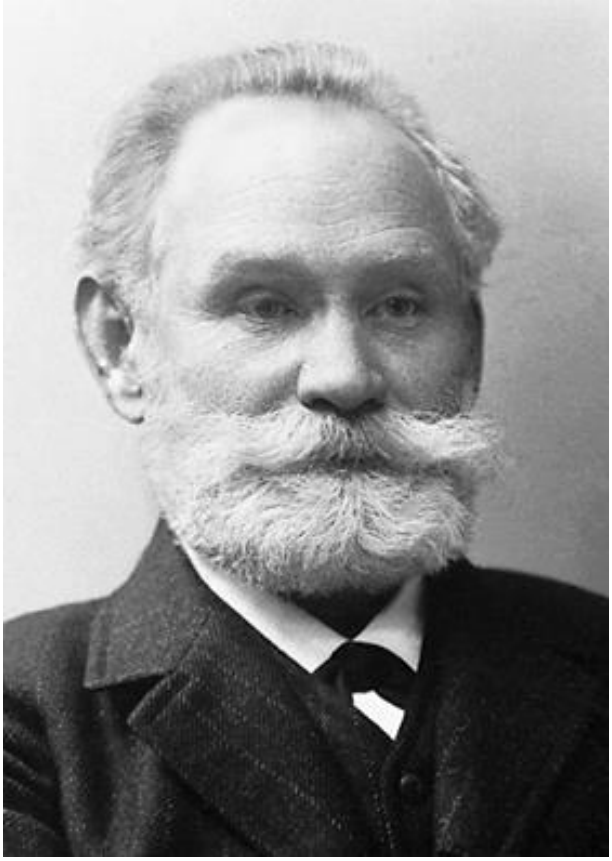


Figure 1 : Ivan Pavlov (1849-1936) lors de la remise de son prix Nobel de Physiologie en 1904

Ces observations permirent à Pavlov d'élaborer en 1927 une théorie sur ce qu'il appelait alors la « salivation psychique » et qui sera plus tard appelé « conditionnement classique » ou « conditionnement pavlovien ». L'ingestion de nourriture déclenchant systématiquement la salivation avant même le début de l'expérience, il s'agissait donc d'une réponse inconditionnelle (salivation, RI) à un stimulus inconditionnel (nourriture, SI).

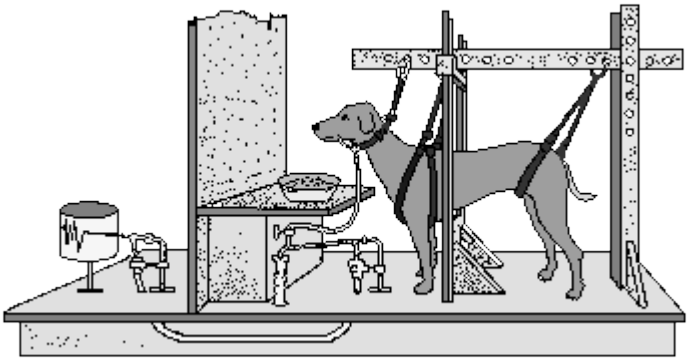


Figure 2 : Expérience de Pavlov. Un chien est placé face à un bol de nourriture pendant que sa production de salive est étudiée. La nourriture est délivrée suite au son d'une cloche. Figure issue de (Schmajuk 2008).

Au contraire, leur simple présence dans la pièce ne faisant pas saliver les chiens, elle constitue un stimulus neutre (SN) (Figure 2).

Selon Pavlov (Pavlov 1927), la réitération de l'expérience chez ses chiens leur a permis d'apprendre à associer divers stimuli neutres ayant systématiquement lieu avant l'ingestion de nourriture (l'odeur et la vue de la nourriture comme l'entrée dans la pièce) à cette ingestion de nourriture (SI). Ces stimuli auparavant neutres car n'entrainant pas de salivation en eux-mêmes, sont devenus des stimuli conditionnés (SC) puisqu'ils ont fini par déclencher la salivation après répétition de l'expérience. La salivation alors observée est devenue une réponse conditionnée (RC) en réponse à ces stimuli conditionnés.

Avant le conditionnement, un SI déclenche systématiquement une RI. Un SN seul n'induit aucune réponse. Pendant le conditionnement, la répétition de l'association SN puis SI induit tout le temps une RI. Après conditionnement, un SN devenu SC permet donc à lui seul le déclenchement d'une RC.

Figure 3 : Edward Lee Thorndike (1874-1949) en 1912 (Popular Science Monthly volume 80)

Au début du XXème siècle, à l'Université Columbia aux Etats-Unis, Edward Lee Thorndike (1874-1949) entreprit des travaux expérimentaux plus complexes afin de mieux comprendre le conditionnement. Pendant son doctorat (obtenu en 1898), il développa une boîte (« puzzle box ») de laquelle des chats devaient s'échapper en tirant sur une corde permettant l'ouverture d'une porte. Thorndike s'intéressait au temps mis par les chats pour sortir de cette boîte et atteindre la nourriture placée en évidence à l'extérieur de celle-ci. Les travaux de Thorndike diffèrent significativement de ceux de Pavlov dans le type de conditionnement induit. Là où Pavlov délivre une récompense quel que soit le comportement observé chez le chien, le protocole expérimental de Thorndike ne délivre une récompense au chat qu'à condition que celui-ci ait effectué le comportement cible.

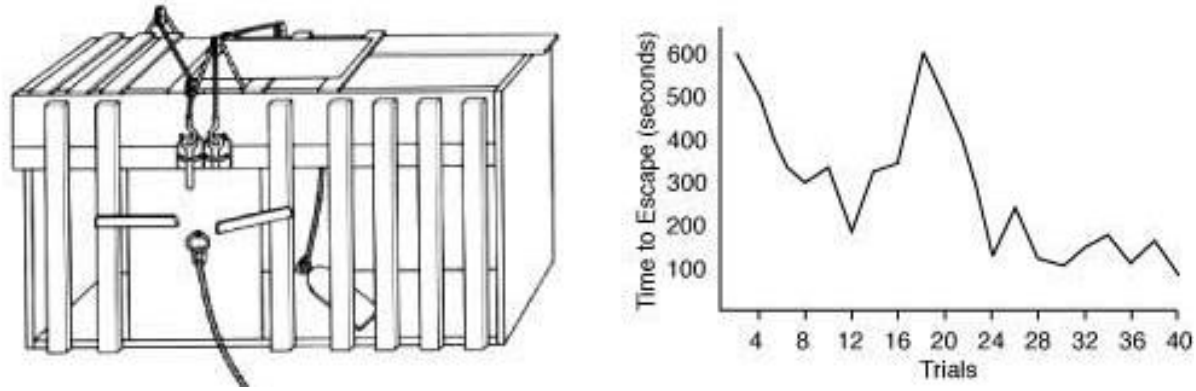


Figure 4 : Expérience de Thorndike. (Gauche) Un chat est placé dans une cage. Pour en sortir et obtenir de la nourriture, le chat doit tirer sur une ficelle pour ouvrir la porte. (Droite) Le chat sort de la boîte de plus en plus vite à force de répéter l'expérience lors de multiples essais. (Adapté de Domjan 1993 [modifié de (Thorndike 1898) (gauche) et de (Imada & Imada 1983) (droite)])

Thorndike observe qu'alors que le chat découvre le comportement cible par hasard, après plusieurs essais, celui-ci met de moins en moins de temps à sortir de la cage car il réalise le comportement cible (tirer sur la corde) plus rapidement (Figure 4). Cet apprentissage par essai erreur est permis grâce à ce que Thorndike décrira comme une association « action effet » ou « instrumentalisation ». Une action menant à l'obtention d'une récompense verra sa probabilité de réalisation renforcée. A partir de ce type de conditionnement appelé « conditionnement opérant » ou « conditionnement instrumental », Thorndike a énoncé la « Loi de l'Effet » en 1911 qui stipule qu'une action (réponse) est plus susceptible d'être reproduite si elle mène à la satisfaction de l'individu (récompense) et aura tendance à être abandonnée si elle induit une insatisfaction (punition).

Figure 5 : Burrhus Frederic Skinner (1904-1990) au département de Psychologie de Harvard (circa 1950)

Quelques années plus tard, le psychologue américain Burrhus Frederic Skinner (1904-1990) simplifia la recherche sur le conditionnement instrumental chez l'animal libre de mouvement grâce à sa « boîte de Skinner » testée sur des rats (Figure 6). Permettant d'étudier les associations entre stimulus-action-renforcement de manière

précise, ses variantes sont toujours utilisées pour étudier le comportement animal. Un grand apport de Skinner à l'étude du conditionnement opérant fût la notion de « contingence de renforcement » pour désigner l'environnement qui va induire le comportement. Trois facteurs entrent en jeu : 1) le contexte d'apparition du comportement, 2) le comportement lui-même et 3) le renforcement faisant suite au comportement. La contingence de renforcement correspond à la probabilité de la relation de cause à effet entre le comportement et le renforcement associé (Skinner 1938). Plus la contingence comportement-renforcement est forte, plus le comportement sera renforcé rapidement.

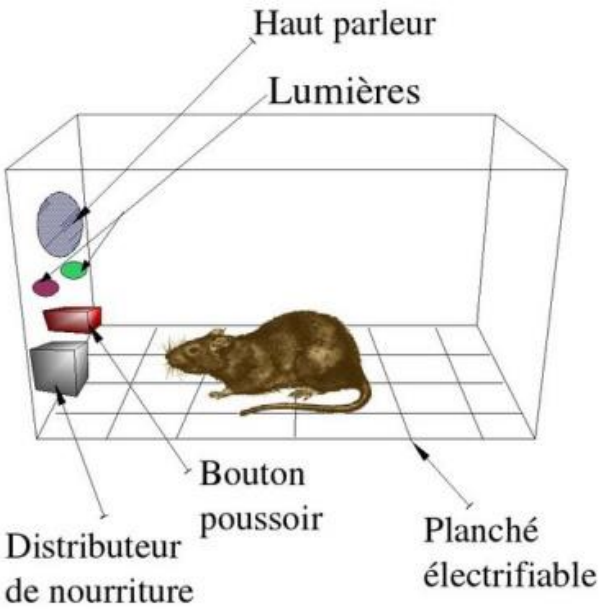


Figure 6 : Expérience de Skinner. Un rat est placé dans une boîte pour étudier le conditionnement opérant grâce à des récompenses (nourriture) et des punitions (plancher électrifié). Figure issue de (Staddon & Niv 2008).

Parallèlement, les comportementalistes Clark Hull et Kenneth Spence ont tenté de formaliser mathématiquement les lois régissant l'apprentissage par conditionnement instrumental (Hull 1943). Tout comme Thorndike avait observé que le chat sortait de plus en plus rapidement à force de répétition et Skinner pointait du doigt l'influence de la contingence du renforcement sur la vitesse d'exécution du comportement cible menant au renforcement, ils ont mis en évidence l'importance du temps pour réaliser l'action attendue pour étudier cet apprentissage. Des expériences menées sur des rats devant identifier le bras contenant de la nourriture dans un labyrinthe en T leur ont permis de montrer que la performance des animaux suit approximativement une courbe exponentielle au cours de l'apprentissage.

2. Concepts de l'apprentissage par renforcement

a) Différences entre conditionnements simple et instrumental

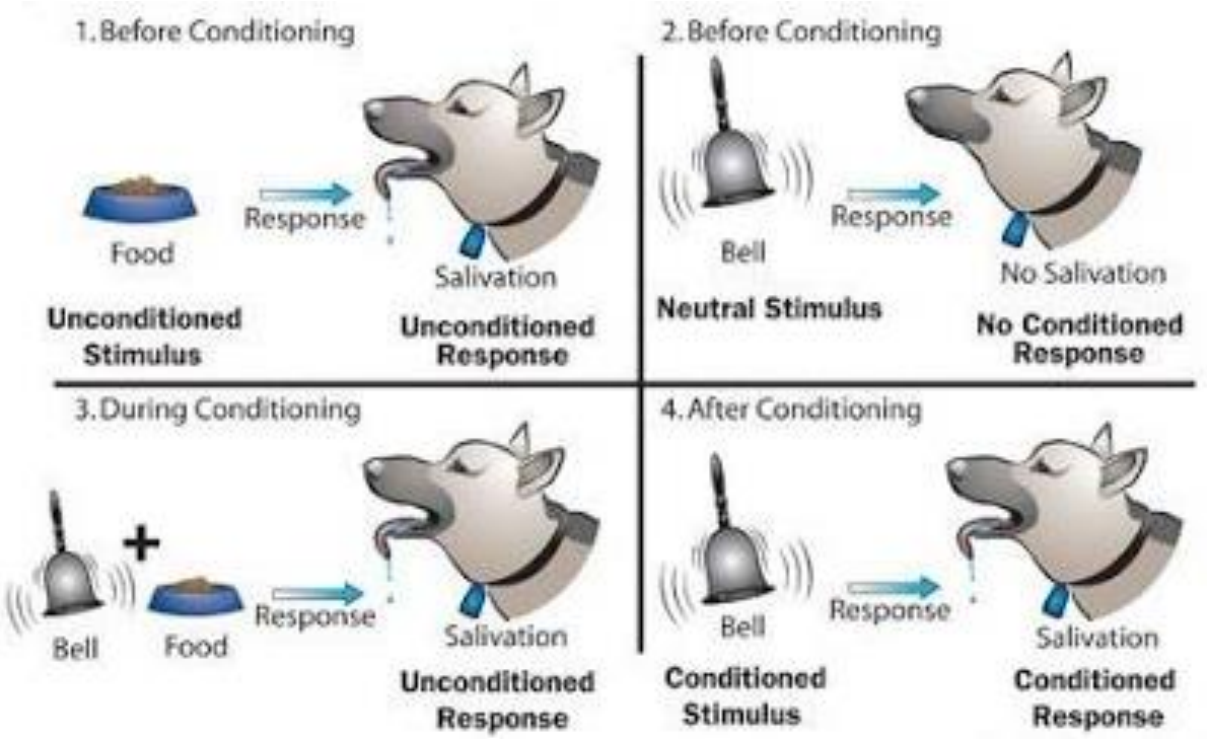
Comme évoqué précédemment, le conditionnement classique (pavlovien) consiste à conditionner un stimulus avec une réponse alors que le conditionnement instrumental (skinnérien) consiste à conditionner un comportement/une action à une réponse. Cette première différence implique qu'alors que le moteur du conditionnement simple est la répétition de l'association SN-SI, celui du conditionnement instrumental est le renforcement du comportement/de l'action cible. La délivrance du renforcement est donc directement liée au comportement du sujet en phase de conditionnement (Dickinson 1980; Staddon 1983). Le conditionnement instrumental implique que le sujet influence directement son environnement en déterminant l'occurrence du renforcement. A noter que lorsque ces occurrences sont apprises à force de répétition grâce au conditionnement instrumental, on parle également d'apprentissage par renforcement.

La seconde différence concerne la réponse en elle-même. Dans le cadre du conditionnement classique, celle-ci est innée et sa nature est définie par le renforcement utilisé (les chiens salivent face à de la nourriture, un jet d'air dans l'œil d'un lapin induit un clignement …). Au contraire, le conditionnement instrumental permet d'observer des réponses originellement non associées aux renforcements. Ainsi lors du conditionnement classique, la forme de la réponse conditionnée (RC) dépend directement du renforcement (SI) utilisé, quand seule sa fréquence (probabilité) en dépend lors du conditionnement instrumental. Ces différences ne sont cependant pas le symbole d'une séparation entre ces deux formes de conditionnement qui peuvent interagir dans ce qui est décrit comme un transfert pavlovien à instrumental. Cet effet décrit le passage de stimuli utilisés en conditionnement instrumental au statut de prédicteurs du renforcement grâce à un processus de conditionnement classique. Bien qu'ayant été largement étudié dans la littérature (Holmes et al. 2010), cet effet ne sera pas évoqué plus longuement dans cette thèse qui se concentre sur le conditionnement instrumental.

b) Qu'est-ce qu'un renforcement ?

Pavlov fut le premier à utiliser le concept de renforcement pour parler de la consolidation de l'association entre les stimuli inconditionnel et conditionnel (SI et SC). Selon lui (Pavlov 1927), la nourriture (SI) agissait comme un renforcement suite à son association répétée avec un stimulus neutre (SN). Thorndike quant à lui développa la théorie du renforcement

(Thorndike 1911) de la réponse selon laquelle les réponses suivies directement d'une satisfaction (renforcement appétitif) étaient plus fortement associées à la situation et que,



lorsque cette situation se présentait à nouveau, ces réponses étaient plus susceptibles d'apparaître. Le terme de renforcement permet alors de décrire la consolidation de l'association stimulus-réponse (SC et RC) (Figure 7).

Figure 7 : Décours temporel du protocole expérimental de Pavlov. (1) Préalablement au conditionnement, un stimulus inconditionné SI déclenche systématiquement une réponse inconditionnée RI. **(2)** Un stimulus neutre SN seul n'induit aucune réponse. **(3)** Pendant le conditionnement, la répétition de l'association SN puis SI induit tout le temps une RI. **(4)** Après conditionnement, un SN devenu SC permet donc à lui seul le déclenchement d'une réponse conditionnée RC. Figure issue de (Schmajuk 2008).

Dans le cadre de l'apprentissage par renforcement, un individu apprend, par essai-erreur, à associer à un stimulus neutre la réponse adaptée (un comportement) qu'il doit produire.

L'identification de cette réponse adaptée au détriment des réponses incorrectes est permise grâce à la délivrance d'un renforcement suite à la réponse. Le renforcement est donc la conséquence de la réponse produite. C'est un stimulus qui va modifier la probabilité de reproduction de la réponse suite à la présentation du stimulus neutre, faisant de la probabilité et de l'intensité de la reproduction de la réponse des paramètres clés de l'étude expérimentale du conditionnement instrumental (Glickman & Schiff 1967; Landauer 1969).

La délivrance de renforcements appétitifs (récompenses) a tendance à rendre plus probable la reproduction de la réponse. Au contraire, la délivrance de renforcements aversifs (punitions) rend moins probable la reproduction de la réponse. Un renforcement (récompense ou punition) peut être positif s'il induit l'ajout d'un stimulus agissant sur l'individu, ou négatif s'il entraîne le retrait d'un stimulus agissant sur l'organisme. Il existe donc quatre types de renforcements possibles dans le cadre de l'apprentissage par renforcement ou apprentissage par conditionnement instrumental :

- Renforcement appétitif positif ou récompense positive, qui **augmente** la probabilité de reproduction d'une réponse grâce à l'**ajout d'un stimulus appétitif** contingent à celle-ci.
- Renforcement appétitif négatif ou récompense négative, qui **augmente** la probabilité de reproduction d'une réponse grâce au **retrait d'un stimulus aversif** contingent à celle-ci.
- Renforcement aversif positif ou récompense positive, qui **diminue** la probabilité de reproduction d'une réponse grâce à l'**ajout d'un stimulus appétitif** contingent à celle-ci.
- Renforcement aversif négatif ou récompense négative, qui **diminue** la probabilité de reproduction d'une réponse grâce au **retrait d'un stimulus aversif** contingent à celle-ci.

Dans le cadre du travail de doctorat présenté dans ce manuscrit, les quatre types de renforcements ont été utilisés. En conséquence, en l'absence de la mention « négatif », le terme « renforcement » fera référence à un renforcement positif. En raison des protocoles expérimentaux utilisés, le terme « neutre » sera utilisé pour les renforcements négatifs (absence de récompense ou récompense neutre, absence de punition ou punition neutre) (Seymour et al. 2007).

c) Qu'est ce qui est renforcé ?

Comme évoqué plus haut, les associations formées par conditionnement au cours d'un apprentissage peuvent influencer le comportement, ce que les théories du conditionnement se sont attelées à décrire (Mowrer & Klein 2000). Ces théories ont ainsi tenté de comprendre le mécanisme associatif induisant la salivation du chien de Pavlov au son de la cloche, ou encore celui permettant au chat de Thorndike d'effectuer la bonne séquence de mouvements pour sortir de la boîte. Les explications proposées par les différents auteurs s'étant intéressés aux diverses formes de conditionnement divergent. Pavlov pensait que

l'apprentissage concernait les associations entre les représentations des stimuli (apprentissage stimulus-stimulus, SS), bien que les théories d'apprentissage SS soient inadaptées quand il s'agit d'expliquer l'ensemble des changements comportementaux observés lors du conditionnement instrumental (Pavlov 1928). L'option proposée par Thorndike suggère que le chat a appris à associer une séquence spécifique d'actions à la boîte. Ainsi, le rôle de l'objectif à atteindre (l'ouverture de la porte de la boîte) est « d'imprimer » en mémoire l'association stimulus réponse (SR) (Thorndike 1933). Une autre option est que le chat aurait appris que cette séquence d'actions permet l'ouverture de la porte, mettant en évidence une association réponse-résultat (résultat = outcome = O, RO). Skinner a réussi à concilier ces deux hypothèses grâce à la prise en compte du rôle d'un stimulus de contrôle dans ce qu'il a appelé une contingence à trois termes : le renforcement favorise une réponse en la présence de ce stimulus de contrôle ou stimulus « discriminant » (association SRO) (Skinner 1938).

La différence majeure entre ces deux hypothèses (SR : et RO) réside dans le rôle du renforcement : d'après la première, il joue un rôle dans l'apprentissage bien qu'une fois celui-ci terminé, le comportement est globalement indépendant du résultat (outcome O). La seconde hypothèse propose une représentation directe de l'outcome au sein de l'association contrôlant le comportement. Ainsi, si un chien attendait devant la boîte, l'ouverture de la porte ne serait plus désirable pour le chat qui y serait enfermé. Si seule la théorie SR s'appliquait, le chat effectuerait la séquence d'actions pour ouvrir la porte, malgré la présence du chien. Au contraire, d'après la théorie RO, le chat aurait plutôt tendance à réprimer le comportement (de Wit & Dickinson 2009; Dickinson 1985; Seymour et al. 2007). La littérature scientifique des trente dernières années a montré à plusieurs reprises que ces deux structures de contrôle existent pendant l'apprentissage instrumental (Redgrave et al. 2010; Yin & Knowlton 2006). A vrai dire, le comportement instrumental peut être subdivisé en deux catégories selon cette distinction : le comportement dirigé vers un but et le comportement habituel. Il a été suggéré qu'un comportement appris est principalement gouverné par une association RO dans sa première phase. Cette phase est dite « dirigée vers un but », puisque qu'elle apparaît afin d'atteindre l'objectif (outcome). Plus tard au cours de l'apprentissage, on parle de comportement « habituel » car il est principalement gouverné par une association SR. C'est au cours de cette phase que le comportement devient une habitude, un réflexe, qui sera exprimé automatiquement en réponse au contexte, quel que soit le résultat attendu. Il est possible de distinguer expérimentalement le comportement habituel de celui dirigé vers un but en influant sur la désirabilité du résultat (outcome). Par exemple, si un animal apprend qu'une action (appuyer sur un levier) entraine l'obtention d'un type de nourriture particulière mais qu'il est autorisé à manger à sa faim avant le début de

l'expérience, la nourriture utilisée pendant l'expérience sera dévaluée. Ainsi, si l'animal ne réalise que rarement l'action pour obtenir la nourriture dévaluée, le comportement est alors considéré comme étant dirigé vers un but. Au contraire, si l'animal continue à effectuer fréquemment le comportement (appuyer sur le levier) et ce malgré la dévaluation de la nourriture, le comportement est considéré comme étant habituel et donc sous contrôle du stimulus.

Une distinction actuelle entre les associations SR et RO se fait lors de l'utilisation d'un modèle computationnel pour prédire le comportement observé. Les associations SR observées dans le cadre de comportements habituels peuvent ainsi être modélisées par des algorithmes d'apprentissage « model-free » (sans modèle) alors que les associations RO des comportements dirigés vers un but seraient modélisées de manière plus adéquate grâce à l'utilisation de modèles d'apprentissage « model-based » (basé sur un modèle) (Daw & Abbott 2005). L'idée principale sous-tendant cette idée est que l'utilisation des algorithmes « model-free » permet le calcul de la valeur des options, à partir des erreurs de prédiction, uniquement d'après le dernier renforcement obtenu. Au contraire, un algorithme « model-based », bien qu'impliquant plus de calculs computationnels, permet de prendre en compte l'historique des renforcements déjà obtenus. Cette seconde méthode permet une vision plus complète de l'environnement, où un enchaînement d'états et de choix permettent un comportement dirigé vers un but. Bien que ces termes puissent porter à confusion, l'appellation « model-based » ou « model-free » se réfère à l'existence ou non d'un modèle interne de l'environnement chez l'individu apprenant. De récents travaux de neurobiologie ont confirmé l'existence de ces deux types de représentations, dont les localisations dans différentes régions cérébrales ont pu être mises en évidences par des études lésionnelles. Ainsi, le striatum dorsal et le cortex préfrontal prendraient en charge des représentations « model-based » et « model-free » respectivement (Balleine & O'Doherty 2010; Morris et al. 2014). Chez l'Homme, il existerait un contrôle flexible de la balance entre ces deux systèmes, situé au niveau du cortex fronto-polaire (Daw & Abbott 2005; Doll et al. 2012; Lee et al. 2014), dont le fonctionnement reste encore inexploré.

d) D'où découle l'apprentissage ?

Plusieurs facteurs ont été identifiés comme nécessaires à la création d'associations entre stimuli, actions et renforcements. Ces facteurs essentiels au conditionnement sont la contiguïté temporelle, la contingence et l'erreur de prédiction.

L'importance de la contiguïté temporelle est mise en évidence par l'influence du délai temporel séparant le SC du SI ou la réponse de l'outcome sur leur vitesse d'association. Plus deux stimuli sont contigus, plus vite leur association sera apprise (Gibbon et al. 1977). Cependant, un appariement simultané (contiguïté maximale) peut résulter en une association plus faible voire nulle, suggérant que l'apprentissage pourrait suivre une fonction de contiguïté non monotone (Plotkin & Oakley 1975). La contingence est définie comme la différence entre la probabilité de l'occurrence du SI en présence du SC et celle de l'occurrence du SI en l'absence du SC. Rescorla a montré que seules les contingences positives étaient associées à l'acquisition d'une réponse conditionnée (Rescorla 1969), et qu'une contingence nulle (pas de différence) n'induisait aucun conditionnement même si les deux stimuli étaient contigus. La contingence est donc une mesure simple de la co-occurrence statistique de deux stimuli, c'est à dire une mesure de la prédictivité du SC par rapport au SI (Rescorla 1967). La troisième condition nécessaire à l'établissement d'une association est l'erreur de prédiction. Le rôle des erreurs de prédiction peut être formulé ainsi : si l'on est capable de prédire précisément chaque évènement, l'on n'a rien à apprendre.

Historiquement, l'importance des erreurs de prédiction a été montrée grâce à deux paradigmes clés de conditionnement, nommément le blocage et l'inhibition conditionnée. A des fins explicatives pour le travail présenté dans ce manuscrit, seul le cas de l'effet de blocage (Kamin 1969a; Kamin 1969b) sera décrit en détails. Dans le paradigme de blocage de Kamin, un animal est exposé à un premier stimulus conditionné (SC1) prédisant l'occurrence d'un renforcement. Après avoir appris l'association appariant le SC1 et le stimulus inconditionné (SI), un second stimulus (SC2) est présenté en même temps que le SC1. Le décours temporel est donc le suivant : apparition simultanée de SC1 et de SC2, suivie de la délivrance du renforcement. Ainsi, le SC1 et le SC2 sont tous deux prédictifs du SI. Néanmoins, l'animal ne présente que peu voire aucune association entre le SC2 et le SI au cours des tests. L'association SC2-SI a été bloquée par la première (SC1-SI) puisque le SC1 permettait déjà de prédire complètement l'occurrence du SI, le SC2 n'apportant rien de nouveau. La délivrance du renforcement suite à l'apparition du stimulus SC2 n'induit pas d'erreur de prédiction positive (renforcement inattendu) car celui-ci était déjà entièrement prédit par l'apparition du stimulus SC1. Les erreurs de prédictions n'apparaissent donc que lorsque qu'il y a une part d'inattendu associée au renforcement : renforcement plus ou moins important que ce qui était prévu. Ces paradigmes de conditionnement ont ainsi montré que la contiguïté temporelle et la contingence (c'est à dire la simple co-occurrence temporelle et statistique) ne suffisaient pas à induire un apprentissage associatif. Afin d'expliquer ces phénomènes, Rescorla et Wagner proposèrent un modèle mathématique simple incluant le

concept d'erreur de prédiction comme facteur nécessaire au conditionnement (Rescorla & Wagner 1972).

e) Modèle de Rescorla Wagner

Le modèle de Rescorla Wagner (RW) vise à décrire les changements de force associative (V) entre un SC et un SI ultérieur à la suite d'un essai de conditionnement. Le modèle RW tente de montrer que le conditionnement a lieu non parce que deux évènements ont lieu en même temps mais parce que leur co-occurrence était inattendue d'après la force associative actuelle. Au cours d'un essai d'apprentissage où deux stimuli *A* et *X* sont suivis par un SI, le modèle RW prédit que les règles de changement de force associative de *A* et *X* sont :

(Equation 1) $\Delta V_A = [\alpha A \beta] (\lambda - V_{AX})$ et $\Delta V_X = [\alpha X \beta] (\lambda - V_{AX})$

Où $V_{AX} = V_A + V_X$.

Dans ces équations, λ est l'effet maximum que le SI peut produire, cela représente la limite de l'apprentissage. α et β, dont la valeur est comprise entre 0 et 1, sont des paramètres d'apprentissage dépendants du SC et du SI respectivement. α se rapporte à la vitesse d'apprentissage, quand β représente la propension que les évènements inattendus ont à modifier les choix (aussi appelée « température »). Ces paramètres ont des valeurs fixes basées sur les propriétés physiques du SC et du SI. A chaque essai, la force associative globale V_{AX} est comparée avec λ et la différence est considérée comme une erreur à corriger, grâce à un changement dans la force associative (ΔV) adapté. Il s'agit donc d'un modèle de correction des erreurs.

Le modèle RW a fourni un cadre pour expliquer l'effet de blocage de Kamin, selon lequel une association préexistante entre un stimulus *A* et un SI rend inefficace une association secondaire de *A* et *X* avec le SI par la génération d'une association entre *X* et le SI. Le conditionnement préalable de A implique que V_A soit proche de λ, donc lors d'un essai *AX*, puisque V_X est nulle, V_{AX} est proche de λ donc l'erreur correspond à ($\lambda - VAX$) qui est proche de zéro. Ainsi, ΔV_X est elle aussi proche de zéro donc V_X change peu.

De nombreux psychologistes se sont attelés à développer de nouveaux modèles mathématiques du conditionnement pour prendre en compte l'influence d'autres phénomènes non expliqués par le modèle RW, en intégrant les effets de la charge attentionnelle et des relations temporelles entre les stimuli conditionnés (Mackintosh 1975;

Schmajuk et al. 1998). Ces modèles sont largement étudiés dans le domaine de l'apprentissage automatique (« Machine learning »), permettant d'aller plus loin dans la compréhension du conditionnement. Malgré la grande variété de modèles mathématiques du conditionnement existant à ce jour, tous les paramètres pouvant avoir une influence sur l'apprentissage instrumental ne sont pas pris en compte par un modèle. C'est par exemple le cas du risque, dont les effets sur la prise de décision simple ont déjà été étudiés au niveau psychologique (voir paragraphe suivant et (Kahneman & Tversky 1979; Kuhnen & Knutson 2005; Mohr et al. 2010; Losecaat et al. 2014; Vermeer & Sanfey 2015) mais ne l'ont que peu été sur l'apprentissage par renforcement ni au niveau psychologique, ni au niveau computationnel (Wright et al. 2012).

3. Processus décisionnels en situation risquée

a) Effets du risque sur les choix pendant la prise de décision

Nous prenons de nombreuses décisions au cours de notre vie, et beaucoup des choix que nous faisons impliquent une part de risque. Le cas le plus simple de prise de décision économique consiste à choisir parmi plusieurs options celle qui permet d'atteindre notre objectif. Chaque décision prise, grâce au choix d'une option, a pour conséquence la délivrance d'un renforcement. De manière générale, nous avons tendance à préférer les options ayant la valeur attendue la plus grande possible, mais en pratique, des choix différents prévalent souvent. En effet, si les évènements nous sont hautement imprévisibles, un apprentissage est nécessaire pour améliorer nos prédictions futures (Niv & Schoenbaum 2008). Précédemment il avait été évoqué que la variabilité de la probabilité tout comme la force du renforcement sont des moteurs du conditionnement puisque cette variabilité entraine des erreurs de prédiction. D'après les travaux issus du domaine économique, un renforcement dont la magnitude varie induit un « risque ». Economiquement, le risque est donc défini comme la variance du renforcement (Monosov et al. 2015; Benjamin et al. 2009; Benjamin et al. 2008; McCoy & Platt 2005; O'Neill & Schultz 2010; Rothschild & Stiglitz 1970; Weber et al. 2004; Xu 2014). Il existe donc différents types d'erreurs de prédiction, comme les erreurs de prédiction du renforcement (probabilité, incertitude) et les erreurs de prédiction du risque.

Dans le cas de la prise de décision simple (sans apprentissage), chaque choix est indépendant, ce qui implique que les renforcements associés à chaque option sont connus au préalable. C'est le cas lors de protocoles expérimentaux s'intéressant à l'influence du risque sur les choix, où sont clairement explicitées les probabilités de délivrance des renforcements (incertitude) ainsi que la variance de leur magnitude (risque). Le niveau de

risque associé à une option correspond à la variance de ses renforcements potentiels (Hayden et al. 2008; McCoy & Platt 2005; Rothschild & Stiglitz 1970; Weber et al 2004; Platt & Huettel 2008), c'est à dire au carré de l'écart type de ses renforcements. Ainsi, une option risquée avec plusieurs renforcements potentiels à un niveau de risque plus élevé qu'une option sure associée à un renforcement certain. Le meilleur moyen d'étudier l'effet du risque sur la prise de décision simple est évidemment de développer un paradigme expérimental au cours duquel le risque est le seul paramètre variable, permettant ainsi d'isoler les effets du risque de tout biais possible si plusieurs paramètres entraient en jeu en même temps. Afin d'étudier l'influence du risque dans le cadre de la prise de décision simple, il est intéressant de manipuler le risque en jouant sur les probabilités d'obtenir une récompense ou une punition de variance variable, tout en conservant une valeur identique entre les options. La valeur objective *V* de chaque option correspond à la magnitude de chaque renforcement R_i multipliée par sa probabilité P_i, dans le cas où il y a *n* renforcements possibles.

(Equation 2) $$V = \sum_{i=1}^{n} R_i P_i$$

Par exemple (Figure 8), soient deux conditions possibles, l'une faisant gagner de l'argent (cas 1 = condition GAIN), l'autre faisant perdre de l'argent (cas 2 = condition PERTE). Dans chaque condition, deux options sont proposées. L'option A est dite « sure » puisque qu'elle est associée à une probabilité de 100% de renforcement R (ici R vaut 10€) qui peut être appétitif (gain de 10€) dans la condition GAIN, ou aversif (perte de 10€) dans la condition PERTE. L'option B est-elle dite « risquée » puisque la magnitude du renforcement peut être de 0€ ou de 20€ avec une probabilité de 50/50%. Dans le cas présent, les probabilités d'obtenir les différents renforcements associés à chaque option sont clairement explicitées au sujet préalablement à la prise de décision. Ces probabilités mettent en évidence que dans chaque condition (gain ou perte), les deux options ont la même valeur statistique objective (bénéfice global) de 10€ car 1*10 = 0,5*20+0,5*0. Cela veut dire que pour un nombre infini

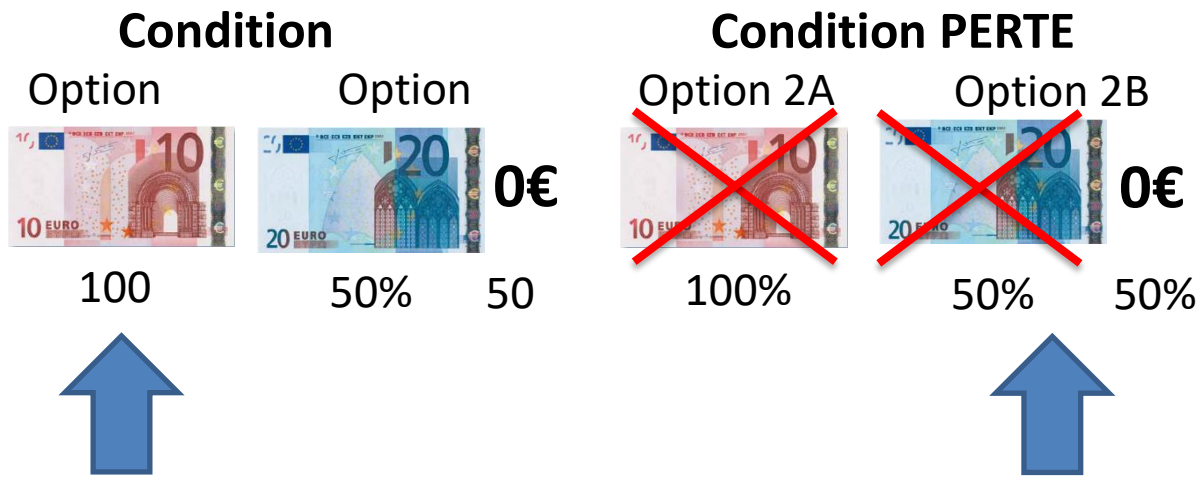


Figure 8 : Exemple d'une prise de décision simple impliquant un risque. Les probabilités des divers renforcements associés à chaque option sont explicitées sous forme de pourcentages. Les choix les plus probables d'être réalisés par le sujet dans le cadre d'une prise de décision simple sont indiqués par les flèches bleues.

d'essais, les deux options rapportent le même renforcement global.

Il est facile de comprendre que le calcul du bénéfice global des options sera d'autant plus complexe que les options seront nombreuses et associées à de multiples renforcements ayant diverses probabilités. A cause de cette complexité computationnelle, nos mécanismes de prise de décision ont évolué vers un processus plus heuristique pour se baser sur des approximations plutôt que sur des calculs computationnels difficiles (Silberberg et al. 2008; Kahneman & Frederick 2005; Kahneman et al. 1982; Kralik et al. 2012). Pour comprendre comment l'on prend des décisions, une évaluation précise du renforcement/bénéfice global n'est donc pas suffisante. Il faut également comprendre le déroulement des processus cognitifs sous-tendant ces calculs heuristiques.

Des effets asymétriques du risque ont été mis en évidence en fonction des conditions étudiées. De manière générale, nous avons tendance à présenter une aversion au risque en évitant les options pouvant résulter en une perte subjective ; qu'il s'agisse d'un gain moindre, d'un gain nul ou d'une perte effective. De plus, quand les valeurs et les probabilités des renforcements sont difficiles à intégrer, nous avons tendance à nous focaliser sur des facteurs spécifiques, comme les potentielles pertes subjectives et opportunités manquées (coûts d'opportunité), tout en ignorant d'autres facteurs, comme les probabilités. En effet dans la condition GAIN, le sujet aura plus souvent tendance à choisir l'option 1A dite « sure » afin de gagner de l'argent à tous les coups, alors même que l'option 1B a la même valeur statistique. Au contraire dans la condition PERTE, le sujet aura tendance à se tourner vers l'option 2B dite « risquée », afin d'avoir une chance de ne pas perdre, ne jouant qu'une seule fois. Cette situation où le sujet se retrouve face à chaque condition seulement une fois (ou du moins peu de fois) induit un biais dans l'attribution d'une valeur subjective (le bénéfice attendu) à chaque option. Le comportement observé indique que le sujet pense que quand il s'agit de gagner de l'argent, mieux vaut être assuré d'en gagner ne serait-ce qu'un peu, alors que quand il faut éviter d'en perdre, mieux vaut tenter un « quitte ou double » en espérant ne rien perdre. Il y a donc une aversion au risque qui apparaît dans la condition GAIN alors que l'aversion à la punition prend le dessus dans la condition PERTE. Ce que l'on appelle l'effet de réflexion est issu de la Théorie des Perspectives (Prospect Theory) (Kahneman & Tversky 1979) :

"We are risk-averse when we have something to gain, but risk-seeking when we have something to lose."

« Nous sommes réticents à prendre des risques lorsque nous avons quelque chose à gagner, mais à la recherche du risque lorsque s'il y a quelque chose à perdre ».

Certaines études suggèrent que la majorité des êtres humains n'expriment ni de préférence ni d'aversion particulière envers le risque, étant plutôt neutres (Hayden et al. 2009). Des effets contradictoires d'aversion, d'attirance et de neutralité vis-à-vis du risque, dans le domaine des récompenses comme des punitions, ont été rapportés par plusieurs études (Laury & Holt 2005; Crockett et al. 2009; Guitart-Masip et al. 2011). Ces résultats laissent entendre que la Théorie des Perspectives ne suffit pas à expliquer l'influence du risque sur les choix. Wright et ses collaborateurs montrent que la valence et le risque influencent indépendamment les choix, et que le contexte en module les effets (Wright et al. 2012).

A cause de tels facteurs (coûts d'opportunités et probabilités par exemple), la façon de voir le dilemme de choix affecte fortement la prise de décision (Kahneman et al. 1982; Kahneman et al. 2011). Par exemple, nos préférences sont influencées par la forme d'explicitation (orale ou écrite) des contingences, au contraire des contingences apprises par l'expérience. Les choix ont également affectés par le nombre de décisions à prendre, qu'il s'agisse d'un choix unique ou de choix multiples (Kahneman et al. 2011). Lors de choix uniques, l'aversion aux potentielles pertes subjectives apparaît comme particulièrement forte, orientant les choix vers les options sures/certaines. Lors de choix répétés, nous avons tendance à prendre plus de risques dans l'espoir que les pertes subjectives potentielles soient atténuées alors que les bénéfices réels convergent vers ceux attendus, reflétant un autre biais appelé l'illusion de Samuelson (Kahneman et al. 2011; Xu 2014).

b) Effets contextuels du risque implicite sur l'apprentissage par renforcement

La plupart de ces études se sont concentrées sur des situations dans lesquels le risque était explicite, c'est-à-dire que le sujet de l'étude connaissait, soit au préalable soit dès le début de l'expérience, les probabilités et les récompenses associés à chaque option. Dans notre vie, nous n'avons pas toujours connaissance de telles informations et nous devons généralement les apprendre par l'expérience. Alors que dans le cadre de la prise de décision simple le risque était explicite, lors de l'apprentissage par renforcement en situation de risque, le risque est implicite et doit être appris par essai et erreur. Contrairement à des choix réalisés hors apprentissage, ceux pris lors d'un apprentissage par renforcement sont moins susceptibles d'être biaisés lors de l'attribution des valeurs subjectives aux options. Ces valeurs vont être affinées par l'expérience, essai par essai. Cette situation est plus représentative de nombreux problèmes auxquels sont confrontés les êtres humains et les autres animaux dans leur état naturel, rendant son étude d'autant plus importante pour

comprendre l'influence du risque sur l'apprentissage par renforcement (Monosov et al. 2015). Une étude par Niv et ses collaborateur s'est attelée à étudier l'influence du risque sur l'apprentissage par renforcement chez l'humain, en neuroimagerie fonctionnelle (Niv et al. 2012). Grâce à l'apport d'un modèle mathématique, ils rapportent un couplage neurométrique-psychométrique entre les fluctuations de l'évaluation des options risquées (valeurs subjectives), mesurées en signal BOLD, et les fluctuations comportementales reflétant une aversion au risque. Cela suggère que la sensibilité au risque fait partie intégrante de l'apprentissage par renforcement chez l'humain, et affecte en conséquence les modèles économiques du choix, les modèles neuroscientifiques d'apprentissage affectif et les travaux sur les mécanismes neuronaux sous-jacents. A ce jour, l'apprentissage du risque au cours de l'apprentissage par renforcement reste peu étudié, ce à quoi le travail présenté dans ce manuscrit s'attelle à étudier en tentant d'identifier les effets du risque à la fois sur l'apprentissage par récompense et sur l'apprentissage par évitement des punitions.

4. Modélisation computationnelle de l'apprentissage par renforcement

a) Apprentissage automatique ou « Machine learning »

Les approches théoriques de l'apprentissage par renforcement sont de deux types : les premières visent à décrire correctement le comportement (comme les théories psychologiques descriptives), les secondes cherchent la meilleure méthode pour optimiser leur comportement dans leur environnement, en associant renforcements et actions par exemple, et permettent de comparer cette méthode optimale à celle observée dans le comportement animal (théories normatives) (Dayan & Niv 2008). La plupart des théories de ces dernières années sont des théories computationnelles. En effet, elles se basent sur une définition rigoureuse en termes d'équations décrivant l'apprentissage et la prise de décision, et permettant de prédire quantitativement les données expérimentales (Niv, Joel, et al. 2006; Niv, Daw, et al. 2006).

Ces théories computationnelles des processus d'apprentissage ont mené à l'apparition de trois catégories de modèles computationnels : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Ces catégories diffèrent selon le type d'interaction que l'apprenant/agent/sujet entretient avec son environnement (Dayan & Abbott 2002; Alpaydin 2004; Daw & Abbott 2005).

L'apprentissage supervisé regroupe des algorithmes d'apprentissage dans lesquels un superviseur donne à l'apprenant des exemples de comportement opportun. L'apprenant doit donc déterminer la règle à suivre d'après les données d'entraînement (exemples supervisés). Chaque exemple de données d'entrainement consiste en une donnée d'entrée (input) appariée à une donnée de sortie (output). L'algorithme a pour rôle de généraliser ces exemples afin d'identifier la règle (fonction mathématique) liant chaque paire de données.

L'apprentissage non supervisé regroupe les algorithmes permettant d'identifier des structures et motifs (patterns) au sein de données non labellisées. Comme ces données ne sont pas labellisées, il n'y a n'y erreur, ni signal de renforcement (récompense ou punition) pour identifier une potentielle solution. Cette distinction est observable entre l'apprentissage non supervisé d'un côté, et les apprentissages supervisés et par renforcement de l'autre.

L'apprentissage par renforcement dérive de la psychologie comportementaliste du conditionnement (voir précédemment) et fait partie de l'apprentissage automatique ou « machine learning » s'intéressant à comment un agent apprend à choisir les actions appropriées à effectuer dans son environnement de façon à maximiser/optimiser les conséquences de ses actions (renforcements appétitifs et aversifs) (Sutton & Barto 1998). Au contraire de l'apprentissage supervisé, les combinaisons correctes de données d'entrée et de sortie ne sont jamais présentées et les actions sous-optimales jamais explicitement corrigées dans le cadre de l'apprentissage par renforcement.

b) Cas de l'apprentissage par renforcement

Lorsque l'on parle d'apprentissage par renforcement, on considère un apprenant (agent/sujet) en interaction avec son environnement (tout ce qui n'est pas l'agent) par le biais d'actions. L'agent apprend donc de ses interactions avec l'environnement, et en particulier des conséquences de ses actions qu'il sélectionne en se basant sur ses expériences précédentes. Il s'agit donc d'un apprentissage par *essai et erreur*. Les conséquences des actions consistent en un renforcement, décrivant le succès ou l'échec de l'action choisie. L'agent a pour but d'apprendre à sélectionner les actions correctes afin d'accumuler les succès (renforcement appétitifs) au cours du temps. La formalisation computationnelle de l'apprentissage par renforcement se base notablement sur les apports de deux disciplines : le contrôle optimal (Bellmann 1957) et, comme mentionné plus haut, la psychologie expérimentale du conditionnement (Rescorla 1969).

Lorsque l'on considère l'environnement lors de l'apprentissage par renforcement, celui-ci est décrit comme un processus de décision Markovien (MDP) (Howard 1960). L'environnement se trouve dans un état/situation *s*, l'agent/décideur peut choisir entre diverses actions *a* possibles lors de l'état *s*. Lorsque l'agent choisit une action *a*, l'environnement réagit à cette action, mettant ainsi l'agent face à une nouvelle situation *s'* et lui délivrant un renforcement *r*. Dans le cas du chat de Thorndike, l'état *s* correspond au chat dans la boîte, l'action *a* consiste à tirer sur la corde qui mène à un état *s'* où le chat est hors de la boîte, lui permettant de manger la nourriture (renforcement *r*, appétitif en l'occurrence). Ainsi, la probabilité de passer de l'état *s* à l'état *s'* dépend de l'action *a* choisie selon la fonction de transition d'état P(*s,a,s'*). Lorsque le processus ne contient que deux états consécutifs, l'utilisation d'un MDP (où l'état futur ne dépend que de l'état présent et pas des états précédents) est suffisante car il n'y a pas de nécessité à mémoriser les conséquences de plus d'un état. Il s'agit cependant d'un cas particulier puisque la plupart du temps, de multiples états sont visités par l'agent, le menant à l'accumulation d'autant de renforcements. Ces processus où se succèdent situation-action-renforcement-situation-action sont considérés comme étant des processus continus (Sutton & Barto 1998). La valeur d'un état est définie comme la moyenne des renforcements obtenables suite aux actions disponibles. La politique/stratégie de sélection des actions peut donc évoluer. Le but de l'algorithme d'apprentissage par renforcement est donc de sélectionner les actions afin de maximiser les renforcements accumulés. En d'autres termes, maximiser les renforcements appétitifs (récompenses) et minimiser les renforcements aversifs (punitions).

En conséquence, les algorithmes d'apprentissage par renforcement doivent résoudre deux problèmes : le problème de prédiction et le problème de contrôle. Un algorithme capable d'apprendre la fonction de valeur d'un état selon la stratégie choisie permet de résoudre le problème de prédiction. En fin d'apprentissage, la fonction de valeur décrit pour chaque état visité la probabilité d'obtenir un renforcement selon l'action choisie, c'est-à-dire qu'elle calcule la valeur de chaque état. Un algorithme permettant l'obtention d'une stratégie adéquate pour maximiser l'accumulation des renforcements suite au passage d'un état à un autre permet de résoudre le problème de contrôle. Comme il est nécessaire de calculer la valeur d'un état pour choisir l'action maximisant le renforcement obtenable, un algorithme résolvant le problème de contrôle se doit également de résoudre le problème de prédiction.

Il est possible de faire des parallèles entre les paradigmes de psychologie expérimentale et la modélisation computationnelle de l'apprentissage par renforcement. Le conditionnement classique (Pavlov) ne traite ainsi que du problème de prédiction puisque la réponse de l'animal (salivation) n'influence pas son environnement, et donc n'influence pas la fonction

de transition d'état. Au contraire, le conditionnement instrumental (Thorndike) traite à la fois du problème de prédiction et du problème de contrôle puisque la réponse de l'animal (tirer sur la corde) détermine la transition d'état (ouverture de la porte permettant la sortie) et les conséquences (obtention de nourriture) et donc influence l'environnement.

(1) Formalisation computationnelle

Les éléments clés de l'apprentissage par renforcement sont :

- Plusieurs états : *s*,
- Plusieurs actions : *a*,
- Des règles de transitions entre états : *P*,
- Des règles déterminant le renforcement associé à chaque transition : *r*.

Il y a généralement deux méthodes pour déterminer la fonction de valeur et/ou la stratégie optimales. Pour rappel, la fonction de transition d'état P(s,a,s') décrit la probabilité de transition de l'état s à l'état s'suite à l'action a ; la fonction de renforcement r(s,a) détermine le renforcement obtenu à l'état s suite à l'action a. Si ces deux fonctions sont connues, les algorithmes utilisés seront dits « basés sur un modèle ». Ce type d'algorithmes a originalement été développé dans le domaine de la programmation dynamique, différant ainsi notablement des algorithmes d'apprentissage par renforcement, et ne sera donc pas discuté plus en détails dans ce manuscrit (Bellmann 1957).

Si le modèle (donc les fonctions P et r) n'est pas connu à l'avance, le processus d'apprentissage entre dans le cadre de l'apprentissage par renforcement, où un processus d'adaptation va permettre d'apprendre et d'optimiser la fonction de valeur et la stratégie de choix. L'on parle alors d'algorithme sans modèle. A ce jour, un algorithme a émergé comme étant particulièrement pertinent pour l'étude de processus de conditionnement instrumental : le modèle d'apprentissage par différence temporelle (TD learning) (Sutton & Barto 1998). L'algorithme de TD learning est utilisé pour résoudre le problème de prédiction en apprenant la fonction de valeur et sera examiné ci-après. Un autre algorithme dit de Q learning sera également abordé (Watkins & Dayan 1992). Le Q learning est une extension du TD learning, visant à résoudre à la fois le problème de prédiction et le problème de contrôle, qui a été fréquemment utilisé pour décrire les choix d'animaux pendant un conditionnement instrumental. La description qui suit des algorithmes d'apprentissage par renforcement est loin d'être exhaustive, l'apprentissage par renforcement étant un domaine particulièrement dynamique de la recherche actuelle, mais elle permet d'aborder les notions et équations clés sur lesquelles se basent actuellement les études de neuroscience expérimentale, et plus particulièrement, les travaux présentés dans ce manuscrit (Daw & Doya 2006; McClure & D'Ardenne 2009).

(2) TD learning

L'algorithme de TD learning a été majoritairement développé par Sutton et Barto (Sutton & Barto 1981; Sutton & Barto 1998). Cet algorithme dérive directement du modèle RW (Rescorla & Wagner 1972) puisque la règle centrale d'apprentissage est une règle de correction des erreurs. Le modèle TD peut ainsi être formalisé comme suit. Soit une séquence (T) d'états suivis par un renforcement (s_t, r_{t+1}, s_{t+1}, r_{t+2} … r_T, s_T), la conséquence globale (la somme des renforcements accumulés) R_t de cette séquence que l'on peut espérer dans le futur à l'état S_t suit l'équation 3 :

(Equation 3) $$R_t = r_{t+1} + \gamma_1 r_{t+2} + \ldots + \gamma_{T-t-1} * r_T$$

Où $0<\gamma<1$ est un facteur d'actualisation. Ce facteur d'actualisation γ représente l'intuition psychologique selon laquelle les renforcements distants dans le temps sont dévalués en comparaison des renforcements immédiats (Kacelnik 1997). Par exemple, si l'on a le choix entre recevoir 1€ aujourd'hui ou 1,5€ dans un mois, le renforcement distant ne sera que rarement préféré au renforcement immédiat, bien que lui étant supérieur en valeur absolue. En effet, la différence objective de valeur entre ces deux renforcements n'est pas suffisante pour contrer la dévaluation du renforcement distant, menant sa valeur subjective à être inférieure à celle du renforcement immédiat.

L'apprentissage par renforcement décrit ainsi que la valeur d'un état $V(s)$ est directement équivalente aux conséquences (renforcement) attendues :

(Equation 4) $$V(s) = R_t\,(s_t=s)$$

$V(s)$ représente la prédiction du renforcement futur dans un état donné futur, s. Ainsi, il est possible de mettre à jour la valeur de l'état s_t à chaque essai selon l'équation 5 :

(Equation 5) $$V(s_t) \rightarrow V(s_t) + \alpha[R_t - V(s_t)]$$

Où $0<\alpha<1$ est la vitesse d'apprentissage. La vitesse d'apprentissage est un paramètre permettant d'évaluer à quel point la valeur d'un état est mise à jour d'un essai à l'autre : si la vitesse d'apprentissage est de 0, il n'y a pas d'apprentissage et $V(s_t)$ reste inchangée ; si elle est de 1, $V(s_t)$ est fortement modifiée lors de la transition vers l'état suivant. Ainsi, $[R_t - V(s_t)]$ représente l'erreur de prédiction du renforcement, c'est-à-dire la différence entre le renforcement obtenu et celui attendu. Cette formulation met en évidence que si $V(s_t)$ prédisait correctement le renforcement obtenu, l'erreur de prédiction serait nulle et n'induirait pas de mise à jour donc la valeur finale de V serait d'ores et déjà connue. Bien que la formulation générale du modèle TD suppose que la mise à jour de la valeur a lieu à la fin de

la séquence d'états successifs, il est en réalité possible de suivre la mise à jour progressive de la valeur d'un état essai par essai, en utilisant l'équation selon laquelle le renforcement global correspond à la somme du renforcement obtenu lors de la transition d'état précédente et de la valeur dévaluée de l'état précédent :

(Equation 6) $R_t = r_{t+1} + \gamma V(s_{t+1})$

La mise à jour itérative (essai par essai) de $V(s_t)$ se fait donc de la manière suivante :

(Equation 7) $V(s_t) \rightarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$

Cette reformulation se base sur l'assomption que la valeur de l'état suivant $V(s_{t+1})$ est une estimation précise du renforcement attendu pour atteindre cet état s_{t+1}. En découle que l'erreur de prédiction du renforcement d'après le modèle TD ou erreur δ vaut :

(Equation 8) $\delta = [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$

Ce qui permet de simplifier le calcul de la mise à jour de la valeur de l'état en se basant uniquement sur l'état en cours :

(Equation 9) $V(s_t) \rightarrow V(s_t) + \alpha\delta_t$

Il a été prouvé que cette simplification se rapproche d'une fonction de valeur exacte, tout en proposant une solution sans-modèle au problème de prédiction. L'idée centrale du TD learning est ainsi d'ajuster les prédictions pour se rapprocher des prédictions exactes des états futurs. Sutton avait illustré ce principe grâce à l'exemple suivant :

"Suppose you wish to predict the weather for Saturday, and you have some model that predicts Saturday's weather, given the weather of each day in the week. In the standard case, you would wait until Saturday and then adjust all your models. However, when it is, for example, Friday, you should have a pretty good idea of what the weather would be on Saturday and thus be able to change, say, Monday's model before Saturday arrives."

« Admettons que l'on veuille prédire le temps qu'il fera samedi et l'on ait un modèle qui permette de prédire la météo du samedi en se basant sur la météo de chaque jour de la semaine. Dans le cas classique, il faudrait attendre samedi pour ajuster le modèle. Cependant, dès le vendredi par exemple, l'on a déjà une idée assez précise du temps qu'il fera le samedi ce qui permet de mettre le modèle à jour avant même que samedi n'arrive. » (Sutton 1988)

Ce modèle de TD learning a été largement utilisé dans la littérature neuroscientifique et psychologique puisque cette approche computationnelle de l'apprentissage par renforcement a permis de prédire précisément à la fois de données comportementales et neurales. En effet, l'activité unitaire des neurones dopaminergiques du mésencéphale suit remarquablement bien les prédictions du modèle d'apprentissage par renforcement par différence temporelle (TD learning) (Montague 1996; Schultz 1997). Comme ce modèle explique également comment apprendre efficacement, le modèle TD permet de faire un lien entre des approches neurophysiologiques et plus normatives comme la modélisation.

(3) Q learning

Le modèle de TD learning peut être utilisé tel quel, ou presque, pour prédire la valeur d'une action dans un état donné (state-action-value au lieu de state-value), c'est-à-dire le renforcement attendu si l'on effectue une action donnée dans un contexte donné (Watkins & Dayan 1992).La valeur d'une action est définie de manière similaire que le renforcement global R_t précédemment :

(Equation 10) $Q(s,a) = R_t(s_t=s|a_t=a)$

La différence est qu'il faut la calculer en supposant qu'à l'instant *t* où l'on visite un état *s* au cours duquel l'on choisit l'action *a* (ce qui n'était pas spécifié dans le modèle TD original). Il est alors possible d'évaluer la valeur approximative de Q en considérant les actions rapportant le renforcement maximal à l'état s_t :

(Equation 11) $Q(s_t,a_t) \rightarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma max_a [Q(s_{t+1},a) – Q(s_t,a_t)]$

Si l'algorithme est capable d'étudier les conséquences des différentes actions disponibles, cet algorithme dit de Q-learning permet de se rapprocher d'une fonction de valeur optimale.

Cette fonction décrivant la valeur d'une action (action-value function) est la première étape menant à la résolution du problème de contrôle, puisque la règle de décision la plus simple peut se baser uniquement sur l'estimation de la valeur des actions. Si l'on applique cette règle au cas du chat de Thorndike, les choix du choix étaient biaisés de façon à promouvoir l'action de tirer sur la cordelette, puisque la valeur de cette action était plus élevée que celles associées au fait de gratter la porte ou du miauler.

L'intérêt du Q-learning par rapport au TD-learning réside dans la capacité de son algorithme d'apprentissage par renforcement à apporter une solution au problème de contrôle (choix de

la stratégie optimale) grâce à l'estimation précise de la valeur des actions, seul support de la stratégie optimale. Néanmoins, pour que l'estimation de la valeur des actions soit la plus précise possible, il est nécessaire d'avoir l'opportunité de tester toutes les actions disponibles. En conséquence, la règle de décision (ou règle de sélection de l'action) doit prendre en compte la possibilité d'explorer ces différentes options. La règle de décision la plus simple consiste à sélectionner, à chaque instant *t*, l'action dont la valeur est maximale, *a**, pour maximiser la valeur de cette action :

(Equation 12) $Q(a_t^*) = maxQ(a_t)$

Cette méthode est dite avare car elle prend en compte l'ensemble des connaissances à l'instant *t* afin de maximiser le renforcement immédiat, tout en gagnant du temps en ignorant les actions identifiées comme inférieurs (moins favorables), y compris si cette identification était erronée. Cette stratégie de maximisation des renforcements fait ainsi gagner du temps au détriment de la mise en place d'un comportement d'exploration des choix, ce qui peut être désavantageux dans le cas de changements dynamiques, d'environnements probabilistes (comme dans le cas des protocoles utilisés au cours de ce doctorat) et mener à une estimation hautement incorrecte de la valeur des actions. Une alternative simple est donc de suivre cette stratégie avare la majorité du temps, mais de temps en temps de choisir une action de manière aléatoire en ignorant les estimations de valeurs (donc avec une faible probabilité ε). Cette règle de décision mixte presque avare est appelée une méthode ε-avare. Celle-ci permet, dans la limite d'un grand nombre d'essais disponibles, de tester toutes les actions possibles suffisamment de fois pour garantir que l'estimation de la valeur de chacune de ces actions se rapproche de leur valeur réelle (Pearl 1984).

Bien que la stratégie ε-avare soit un moyen efficace et populaire d'équilibrer l'exploration et l'exploitation au cours de l'apprentissage par renforcement, une limitation majeure vient du choix aléatoire qui implique un choix homogène entre les autres actions disponibles. Cela implique qu'il est aussi probable de choisir la pire action possible que la deuxième meilleure lors de l'essai aléatoire, ce qui rend cette exploration insatisfaisante. Une solution évidente à cette limitation est de faire varier les probabilités de choisir les actions en fonction de leur valeur estimée. La meilleure action a ainsi la plus grande probabilité d'être choisie et les autres actions sont classées selon leur valeur, afin que la pire action ait la plus faible probabilité d'être choisie. Cette stratégie de sélection de l'action est appelée une règle softmax (Bridle 1990). La règle softmax la plus connue possède une distribution de Gibbs. Soit un set d'actions possibles a_1 à a_n (n alternatives) à l'instant *t*, la probabilité de choisir l'action a_1, $P(a_1)$, est décrite par l'équation suivante :

(Equation 13) $$P(a1) = \frac{e^{Q_t(a_1)/\beta}}{\sum_{b=1}^{n} e^{Q_t(a_b)/\beta}}$$

Où β>0 est un paramètre appelé température. Quand des actions ont la même valeur estimée, leurs choix respectifs sont équiprobables. Si la température est forte, alors toutes les actions sont équiprobables (ou presque). Si la température est faible, la probabilité de choisir une action plutôt qu'une autre dépend fortement de la valeur estimée de ces actions. Cela implique que quand la fonction atteint sa limite (β→0), la règle de sélection softmax se comporte comme une réglé de sélection avare.

5. A l'interface de la psychologie expérimentale et la modélisation computationnelle

Voilà plus d'un siècle que l'étude expérimentale du conditionnement a débuté. Le conditionnement a pour processus clé la création d'associations entre un renforcement et un stimulus ou une action. Depuis le début, les paradigmes expérimentaux de conditionnement se classent en deux catégories. Le conditionnement classique se base sur la délivrance d'un renforcement quel que soit le comportement de l'apprenant. Au contraire, le conditionnement instrumental implique que la délivrance d'un renforcement soit dépendante de la réponse comportementale. Dans le cas du conditionnement instrumental, la structure associative comportementale sous-jacente comporte deux facettes : dirigée vers un but (association réponse-outcome) et habituelle (association stimulus-réponse). Plusieurs facteurs ont été mis en évidence comme nécessaires à l'élaboration d'un conditionnement : la contiguïté temporelle (l'action ou le stimulus avoir lieu peu de temps avant le renforcement pour qu'ils y soient associés), la contingence (l'action ou le stimulus doivent être des prédicteurs du renforcement) et les erreurs de prédiction (une association entre une action ou un stimulus et un renforcement n'a lieu que si le renforcement n'était pas entièrement prévisible par l'apprenant). Cette dernière condition nécessaire au conditionnement a été formalisée par Rescorla et Wagner dans un modèle mathématique de conditionnement classique, ce qui a par la suite mené à l'élaboration de nouveaux modèles plus complexes dans le cadre de l'apprentissage automatique.

L'apprentissage par renforcement est un domaine de l'apprentissage automatique visant à trouver des solutions computationnelles à des problèmes issus de processus psychologiques liés au conditionnement. L'apprenant est perçu comme navigant à travers différents états dans son environnement en sélectionnant des actions et obtenant des renforcements

associés à maximiser. L'apprenant a donc deux objectifs : prédire les renforcements obtenables au cours d'un état donné, et optimiser la stratégie de sélection des actions pour maximiser ces renforcements. Le TD-learning est une solution computationnelle sans modèle permettant d'évaluer la valeur des différents états et donc de prédire les renforcements. Le Q-learning est une évolution du TD-learning qui n'évalue pas la valeur des états mais directement celle des actions disponibles à chaque état, permettant ainsi de sélectionner de manière optimale ces actions. Ces deux méthodes computationnelles se basent sur des signaux de renforcement qui dérivent des erreurs de prédiction des renforcements décrites par le modèle RW. Grâce aux paramètres expérimentaux (contexte, réponse, fréquence de renforcement : *a*, *s*, *r*), l'utilisation d'algorithmes d'apprentissage par renforcement dans l'étude des neurosciences permet de prédire quantitativement l'évolution de données comportementales ou neurales (théories normatives), si elles reflètent des variables de l'apprentissage par modèle, comme les renforcements prédits selon les états, les actions et les erreurs de prédiction des renforcements (*V(s), Q(s,a)*). La recherche en neurosciences a ainsi tenté de rattacher des activités neurales à des variables computationnelles de ces modèles, décrivant ainsi des « bases neurales » à des « variables cachées » de l'apprentissage par renforcement. L'apprentissage par renforcement a donc permis l'étude formelle et normative des processus sous-tendant le conditionnement.

Les travaux présentés dans le cadre de cette thèse s'intéressant à la dynamique cérébrale de l'apprentissage par renforcement chez l'humain, une revue anatomique et fonctionnelle va permettre de faire le lien entre les aspects théoriques, issus de la psychologie et des sciences économiques, avec les régions cérébrales impliquées.

B. Bases neurales de l'apprentissage par renforcement

Les paradigmes expérimentaux utilisés pour l'étude de l'apprentissage par renforcement proviennent de la psychologie expérimentale avec ceux de l'étude sur le renforcement. La modélisation computationnelle a fourni un cadre formel pour mieux appréhender l'apprentissage par renforcement et générer des hypothèses quantitatives sur les réponses de certaines régions cérébrales pouvant refléter des variables spécifiques de l'apprentissage (les variables computationnelles cachées mentionnées dans le chapitre précédent (I.A).

La prise de décision et en particulier l'apprentissage par renforcement étant des processus cognitifs de haut niveau, ils impliquent de nombreuses aires cérébrales à la fois corticales (majoritairement du cortex frontal) et sous-corticales. Afin de mieux comprendre leur rôle, il est important de s'intéresser à leur anatomie au niveau cellulaire, mais aussi aux connexions qu'elles forment au sein du cerveau et aux réseaux fonctionnels dont elles font partie intégrante. Ce chapitre (I.B) aura donc pour premier rôle de présenter succinctement les acteurs de l'apprentissage par renforcement auxquels les travaux présentés dans ce manuscrit se sont intéressés. A la suite de cette première partie qualitative, une revue de la littérature actuelle présentera de manière plus quantitative les activités cérébrales observées au cours de processus de conditionnement ou décisionnels comme l'apprentissage par renforcement. Seront ainsi évoquées les contributions les plus significatives à la compréhension des bases neurales de l'apprentissage par renforcement, provenant de travaux clés en électrophysiologie chez le rongeur (rat et lapin) et le primate (humain et non-humain), ainsi qu'obtenus chez l'humain en neuroimagerie fonctionnelle, neuropsychologie et grâce à des études lésionnelles.

Dans les années 1990, Wolfram Schultz et ses collaborateurs ont été à l'origine d'un grand nombre d'expériences électrophysiologiques chez le primate non-humain, fournissant les premières preuves de l'existence d'un système cérébral permettant la représentation de plusieurs variables de l'apprentissage par renforcement, à savoir les prédictions des récompenses et les erreurs de prédiction des récompenses. A cette époque, le système dopaminergique était majoritairement associé à des affections débilitantes comme la maladie de Parkinson, le syndrome de la Tourette, la schizophrénie, les troubles de l'attention et les addictions (Kienast & Heinz 2006). C'est pour son intérêt clinique que Schultz et ses collègues ont commencé à étudier ce système neuromodulateur qu'est le système dopaminergique. Cependant, il est rapidement apparu que les variables comportementales principalement associées avec les réponses dopaminergiques étaient motivationnelles, et non motrices (Mirenowicz & Schultz 1996).

1. Neuroanatomie

Avant de nous intéresser aux découvertes expérimentales majeures concernant le rôle du système dopaminergique fronto-striatal au cours de l'apprentissage par renforcement, il nous faut s'intéresser à la neuroanatomie de base des systèmes en question. Cette revue neuroanatomique vise donc à faire un point morphologique, anatomique et fonctionnel des aires cérébrales connues à ce jour comme étant impliquées dans les processus de prise de décision. Beaucoup d'entre-elles sont directement liées au système dopaminergique et à ce que l'on appelle le « circuit de la récompense » mais d'autres n'ont qu'un lien indirect connu avec le système dopaminergique, mais sont néanmoins impliquées dans des processus essentiels de la prise de décision : l'encodage des valeurs subjectives, le traitement des erreurs ou la détection d'évènements saillants par exemple. Pour commencer, il est nécessaire de faire un point sur ce qu'est le système dopaminergique et les ganglions de la base. Le striatum et le thalamus en étant les structures de sortie et d'entrée, ils seront étudiés par la suite. Enfin, ce point sur l'anatomie s'intéressera aux régions corticales profondes, comme l'insula et le cortex cingulaire, et aux régions corticales préfrontales. Pour chaque région d'intérêt, une définition de ses limites anatomiques ainsi que de ses afférences et efférences précèdera un bref commentaire des fonctions liées à la prise de décision dans lesquelles cette région est impliquée.

a) Système dopaminergique et circuit de la récompense

Il est probable que l'idée d'un système traitant le concept de valeur au sein du cerveau ait émergé au cours des années 1950 (Olds & Milner 1954; Milner 1988). James Olds et Peter Milner voulaient prouver qu'une stimulation électrique d'une zone cérébrale particulière du mésencéphale induisait un état d'excitation chez des rats. Ils furent alors surpris de constater que les rats revenaient constamment à l'endroit où ils avaient reçu cette stimulation électrique. Olds créa alors un système permettant au rat de s'auto-stimuler grâce à une pression sur un levier, et observa que le rat actionnait alors le levier de manière répétée, en abandonnant toute activité autre (Olds & Milner 1954). Les scientifiques découvrirent par la suite que l'électrode n'était pas localisée dans le mésencéphale mais dans l'hypothalamus latéral, juste à côté de l'aire septale. Ces sites de stimulation furent rapidement connus sous le nom de « centres du plaisir » (Olds 1956), en raison du lien entre leur activité neurale et le renforcement du comportement (pression sur le levier). Bien qu'une association stimulus-réponse ait été suffisante pour expliquer un tel comportement, l'existence même d'une région cérébrale dont l'activité pourrait être responsable de l'induction d'une telle préférence

comportementale fut à l'origine de la notion selon laquelle il existerait un système cérébral spécialisé permettant l'encodage du plaisir, ou du moins des valeurs (Wise 2002; Corrado et al. 2005).

La méthode de stimulation cérébrale profonde développée par Olds et Milner chez le rat fut plus tard étendue à d'autres espèces (lapins, singes et humains) et permit de définir un circuit cérébral (Olds 1956) incluant le striatum ventral (VS), le noyau accumbens (NAcc) et les aires ventrales du cortex préfrontal. Le rôle de ces régions dans des comportements impliquant des renforcements appétitifs fut démontré grâce à des expériences d'autostimulation (Rolls et al. 1980; Mora et al. 1980). Des expériences pharmacologiques (Falck & Hillarp 1959; Falck et al. 1959; Shizgal 1997; McBride et al. 1999) ont mis en évidence l'innervation de plusieurs régions par des projections en provenance des neurones dopaminergiques localisés dans le mésencéphale, à savoir l'aire tegmentale ventrale (ATV) et la substance noire pars compacta (SNc).

Depuis la description de la dopamine (DA) comme étant un neurotransmetteur essentiel au sein du système nerveux central (Carlsson 1959), son implication dans la motivation et le contrôle du mouvement a souvent été mise en exergue.

Tout comme pour les autres systèmes neuromodulateurs catécholaminergiques dans le cerveau, les corps cellulaires des neurones dopaminergiques sont situés au niveau du mésencéphale (Arias-Carrión et al. 2010). Plus précisément, la substance noire *pars compacta* (SNpc) et l'aire tegmentale ventrale (ATV) sont deux régions du mésencéphale ventral contenant les neurones dopaminergiques. Des fibres dopaminergiques ascendantes en provenance de ces deux régions innervent un grand nombre de structures corticales et sous-corticales (Ashby et al. 2015). Les structures des ganglions de la base, en particulier le striatum, reçoivent les connexions dopaminergiques les plus denses en provenance de la SNpc, et de loin. Au sein du cortex, il est largement accepté que les aires sensorielles primaires (comme le cortex visuel primaire) sont relativement dépouillées de connexions dopaminergiques ascendantes, alors que le cortex frontal est bien plus densément innervé. Parmi les aires frontales étant fortement innervées par le système dopaminergique l'on trouve les aires préfrontales, grâce à la voie mésocorticale (Figure 9). Ces projections corticales ont principalement pour origine l'ATV. Une autre voie dopaminergique part de l'ATV et se projette sur le système limbique au niveau cortical (hippocampe et certaines parties du cortex préfrontal) et sous-cortical (amygdale, striatum ventral) : il s'agit de la voie mésolimbique.

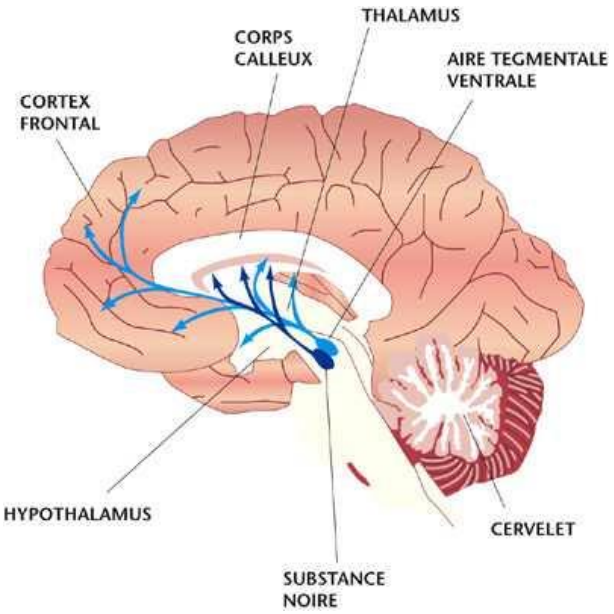


Figure 9 : Système dopaminergique. Les neurones dopaminergiques sont localisés dans deux régions du mésencéphale : l'aire tegmentale ventrale ATV et la substance noire compacte SNc. Les voies mésocorticale (bleu clair) et mésolimbique (bleu foncé) forment des connexions excitatrices partant respectivement de l'ATV et de la SNc vers le cortex préfrontal et le striatum. Figure issue de (Anon 2007).

A ce jour, tous les récepteurs dopaminergiques ayant été identifiés sont des récepteurs couplés à des protéines G, c'est-à-dire que ce sont de récepteurs métabotropiques lents qui modulent fonctionnellement d'autres récepteurs et/ou canaux ioniques. Par conséquent, à part quelques exceptions, l'activation des récepteurs dopaminergiques en soi au niveau du prosencéphale ne peut pas résulter en de forts courants postsynaptiques (Vallone et al. 2000). L'activation de ces récepteurs module plutôt, par le biais de cascades de signalisation intracellulaire, un grand nombre de propriétés biophysiques des synapses et des neurones post-synaptiques, ce qui change leurs capacités de traitement de l'information. Donc, par exemple dans le striatum, les projections et récepteurs dopaminergiques sont situés au niveau des synapses cortico-striatales (Smith & Bolam 1990). En l'occurrence, les neurones dopaminergiques sont sollicités afin de moduler positivement ou négativement (en fonction du récepteur dopaminergique D1 ou D2 exprimé) la force des synapses cortico-striatales. Cette propriété des neurones dopaminergiques permet de créer des phénomènes de potentialisation et de dépression à long terme, ce qui est cohérent avec l'implication de ce neuromodulateur (la dopamine) dans les processus d'apprentissage (Tremblay & Schultz 1999).

b) Ganglions de la base

Les neurones dopaminergiques de la SNpc et de l'ATV fournissent des signaux modulateurs majeurs aux noyaux des ganglions de la base, et principalement au striatum (Figure 10). Les ganglions de la base comportent deux structures d'entrée principales, le striatum et le noyau sous-thalamique, et deux structures de sortie principales, la substance noire *pars réticulée* (SNpr) et le globus pallidus interne (GPi).

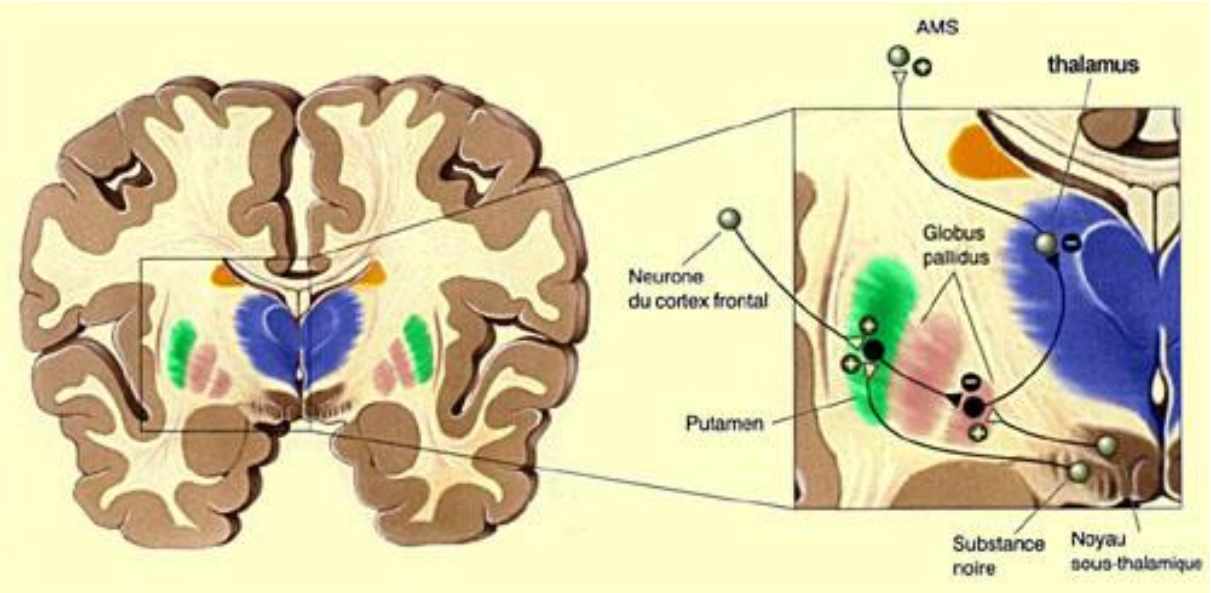


Figure 10 : Localisation anatomique des ganglions de la base au sein du mésencéphale humain. Source : Université McGill, Montréal, Canada.

Le striatum est le plus grand noyau des ganglions de la base. Chez les primates, le striatum comprend le noyau caudé et le putamen, et chez les mammifères, le striatum ventral (ou noyau accumbens). Il reçoit des afférences directes en provenance de la plupart des régions corticales, mais également des structures limbiques comme l'amygdale et l'hippocampe.

Le noyau sous-thalamique (NST) est considéré comme un important relai au sein de la voie indirecte liant le striatum au globus pallidus externe (GPe). En plus de cette fonction, il a été plus récemment mis en évidence le rôle du NST en tant que seconde structure d'entrée des ganglions de la base (avec le striatum), créant ainsi ce que l'on appelle la voie hyper directe (du cortex aux ganglions de la base sans passer par le striatum).

Le GPe est principalement une structure intrinsèque car il reçoit la plupart de ses afférences et projetant ses efférences à d'autres noyaux au sein des ganglions de la base (Graybiel 2000; Nambu 2004; Yelnik 2002). Ainsi le GPe reçoit des afférences inhibitrices du striatum (GABAergiques) et excitatrices du NST (glutamate) et projette des efférences GABAergiques inhibitrices à tous les noyaux d'entrée et de sortie des ganglions de la base, ainsi qu'à la SNpc.

Le GPi est l'une des deux structures de sortie des ganglions de la base, faisant le lien entre les noyaux des ganglions de la base d'un côté et le thalamus et le tronc cérébral de l'autre. Il reçoit des afférences inhibitrices GABAergiques du striatum et du GPe, et excitatrices glutamatergiques du NST. Les neurones du GPi sont GABAergiques et exercent une inhibition puissante sur leurs cibles au niveau thalamique.

La SNpc est la seconde structure de sortie principale. Elle reçoit également des afférences depuis les autres noyaux des ganglions de la base et ses efférences se projettent sur le thalamus. Ses afférences sont inhibitrices (GABAergiques) et proviennent du striatum et du GPe, ses efférences sont également GABAergiques.

Il existe 3 voies reliant les structures d'entrée (striatum et NST) aux structures de sortie (GPi/SNr) des ganglions de la base (zone grisée sur la Figure 11).

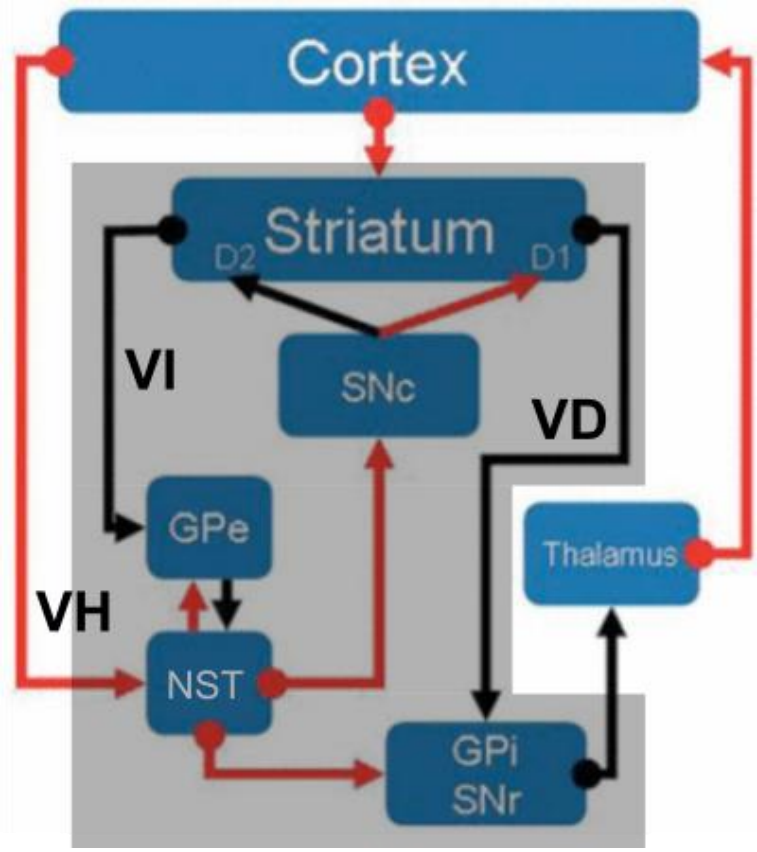


Figure 11 : Schéma structurel et fonctionnel du circuit des ganglions de la base. Les connexions entre les régions peuvent être excitatrices (rouge) ou inhibitrices (noir). Figure adaptée de (Anon 2013).

La voie directe (VD) a le striatum pour structure d'entrée dans les ganglions de la base. Celui-ci est connecté de façon monosynaptique à la structure de sortie, permettant une levée de l'inhibition du thalamus et donc une activation du cortex. La voie indirecte (VI) a pour différence le passage de l'information du striatum aux structures de sortie via deux noyaux, le GPe puis le NST. L'activation de la VI résulte en une inhibition du thalamus et donc une inhibition du cortex. Pour finir, la voie hyper directe relie le cortex directement au NST (Nambu et al. 2002; Nambu 2004; Aron & Poldrack 2006), permettant ainsi une inhibition très rapide du thalamus pouvant contrer les effets de la voie directe (Isoda & Hikosaka 2008).

Une théorie influente de l'organisation intrinsèque des ganglions de la base date de la fin du XXème siècle (Alexander et al. 1990). Dans cette représentation, les signaux provenant du cortex cérébral sont distribués en deux populations de neurones épineux moyens GABAergiques au niveau du striatum. Ces neurones inhibiteurs représentent environ 95% de la population totale des neurones striataux (Yager et al. 2015) et se divisent en deux catégories selon leur phénotype, basé sur les récepteurs dopaminergiques majoritaires qu'ils possèdent (Yager et al. 2015; Ferré et al. 2010; Nishi et al. 2011). Les neurones contenant majoritairement des récepteurs dopaminergiques de type D1 créent des projections directes avec la structure de sortie des ganglions de la base : c'est la voie directe. En parallèle, les neurones striataux exprimant majoritairement des récepteurs dopaminergiques de type D2 créent des projections indirectes avec les structures de sortie des ganglions de la base via le GPe et le NST : c'est la voie indirecte. Les efférences provenant des ganglions de la base ont été supposées refléter la balance entre les projections de la voie directe et de la voie indirecte. Les afférences parvenant au striatum de toutes les sources majeures (cortex frontal et structure limbiques en particulier) sont organisées topographiquement. Similairement, les efférences des ganglions de la base atteignent le thalamus puis se projettent directement sur les régions du cortex à l'origine des afférences en premier lieu, conservant l'organisation topographique : on observe donc des boucles fermées (Yelnik et al. 2007). Les connexions entre le cortex et les ganglions de la base peuvent être considérées comme une série de projections parallèles, largement ségréguées, sous la forme de boucles cortico-striato-nigro-thalamo-corticales.

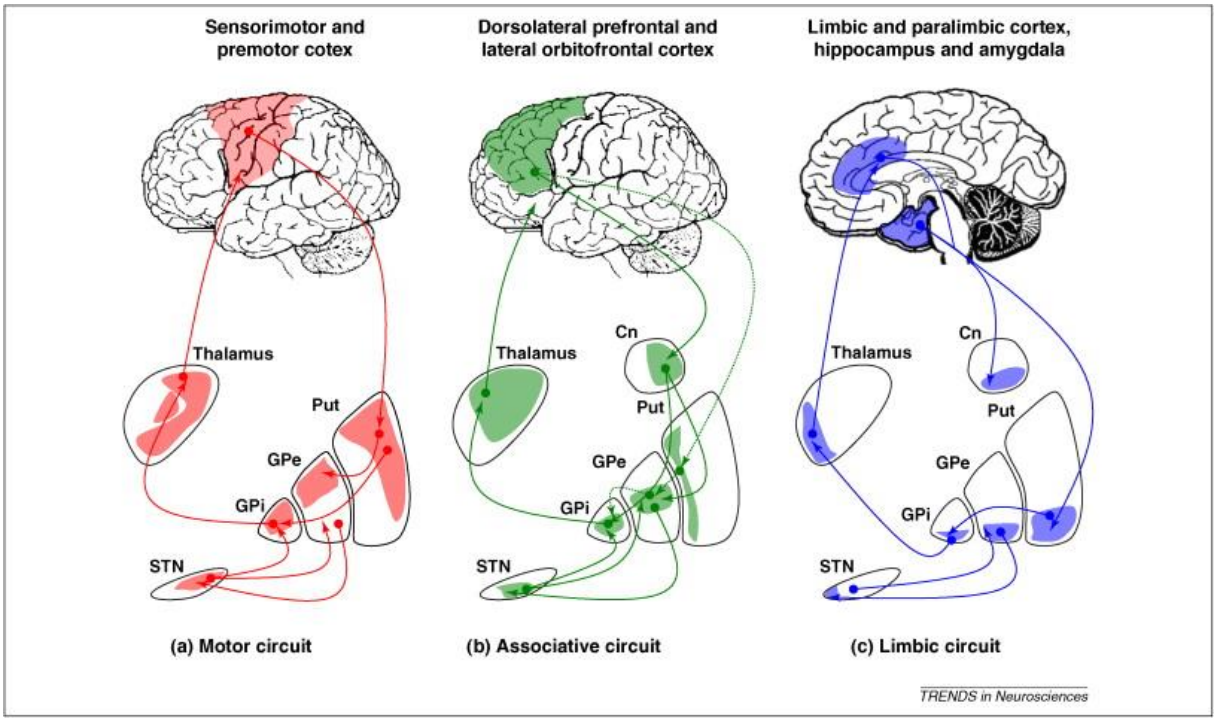


Figure 12 : Structure anatomofonctionnelle en boucles fermées cortico-striato-nigro-thalamo-corticales (Krack et al. 2010).

Ainsi, un nombre important des projections vers des territoires fonctionnels du cortex cérébral se font vers des territoires recevant exclusivement des projections provenant des ganglions de la base, créant des boucles isolées du reste du cerveau. Parmi ces boucles atomiques, trois circuits fonctionnels principaux ont été identifiés (Figure 12) : la boucle limbique la boucle associative et la boucle sensorimotrice, qui possèdent chacune des composants corticaux et sous-corticaux spécifiques. Ces trois boucles sont impliquées dans le traitement d'informations dans le cadre de fonctions différentes : le circuit limbique (ventral) est impliqué dans le traitement des informations émotionnelles et motivationnelles, le circuit associatif (préfrontal dorsal) est responsable de la planification des stratégies comportementales et cognitives, le circuit sensorimoteur est essentiel à l'exécution motrice. Ainsi, au sein des différentes régions du cortex frontal et des ganglions de la base, l'ensemble des informations motivationnelles, cognitives et motrices nécessaires pour créer un comportement adapté trouvent leur place (Brown & Pluck 2000). La description de ces circuits anatomofonctionnels ségrégués provient majoritairement d'études sur des primates non-humains. Plus récemment, des preuves anatomiques d'une organisation similaire chez l'être humain ont été apportées grâce à l'utilisation de la technique d'imagerie DTI (diffusion tensor imaging) (Draganski et al. 2008; Lehéricy et al. 2004). Des preuves fonctionnelles de cette organisation en circuits parallèles ont également été fournies par l'observation clinique de déficits comportementaux causés par des lésions au sein de ces différents circuits, que

ce soit au niveau cortical ou sous-cortical (Arbuthnott & Garcia-Munoz 2009; Levy & Dubois 2006). Cependant, ce modèle de circuits séparés ne permet pas d'apporter les bases anatomiques du trajet pris par l'information au sein de ces circuits, permettant la génération d'un comportement dirigé vers un but, couvrant des aspects motivationnels, cognitifs et moteurs. Des pistes sur l'anatomie du trajet pris par l'information pour aller du cortex aux régions sous-corticales ont été apportées pour les différents circuits (Haber 2003; Haber et al. 2006). Les bases neurales impliquées seraient le réseau striato-nigro-strié et le réseau thalamo-cortico-thalamique. Au sein de ces groupes de structures connectées, il existe à la fois des connexions réciproques liant des régions associées à des fonctions similaires, et des connexions non-réciproques liant des régions associées à des circuits corticaux-basaux distincts. Au sein de chaque sous-territoire striatal, du plus antéro-ventral (limbique) au plus postéro-dorsal (sensorimoteur) il existe des connections permettant l'envoi d'informations antérogrades (feedforward) et rétrogrades (feedback), régissant ainsi un flux directionnel de l'information (Figure 12 et Figure 13). Le striatum est une structure clé des ganglions de la base qu'il est essentiel de considérer lors de l'étude de processus cognitifs comme la prise de décision. De plus, de par sa localisation anatomofonctionnelle, le thalamus a donc un rôle essentiel à jouer à l'interface entre le cortex et les ganglions de la base lors de ces même processus cognitifs.

c) Le striatum

Comme décrit brièvement précédemment, le striatum est un noyau sous-cortical faisant partie intégrante des ganglions de la base. Il est essentiel au système moteur ainsi qu'au circuit de la récompense en raison de ses nombreux afférences glutamatergiques et dopaminergiques et de ses efférences servant de voie d'entrée majeure au réseau des ganglions de la base.

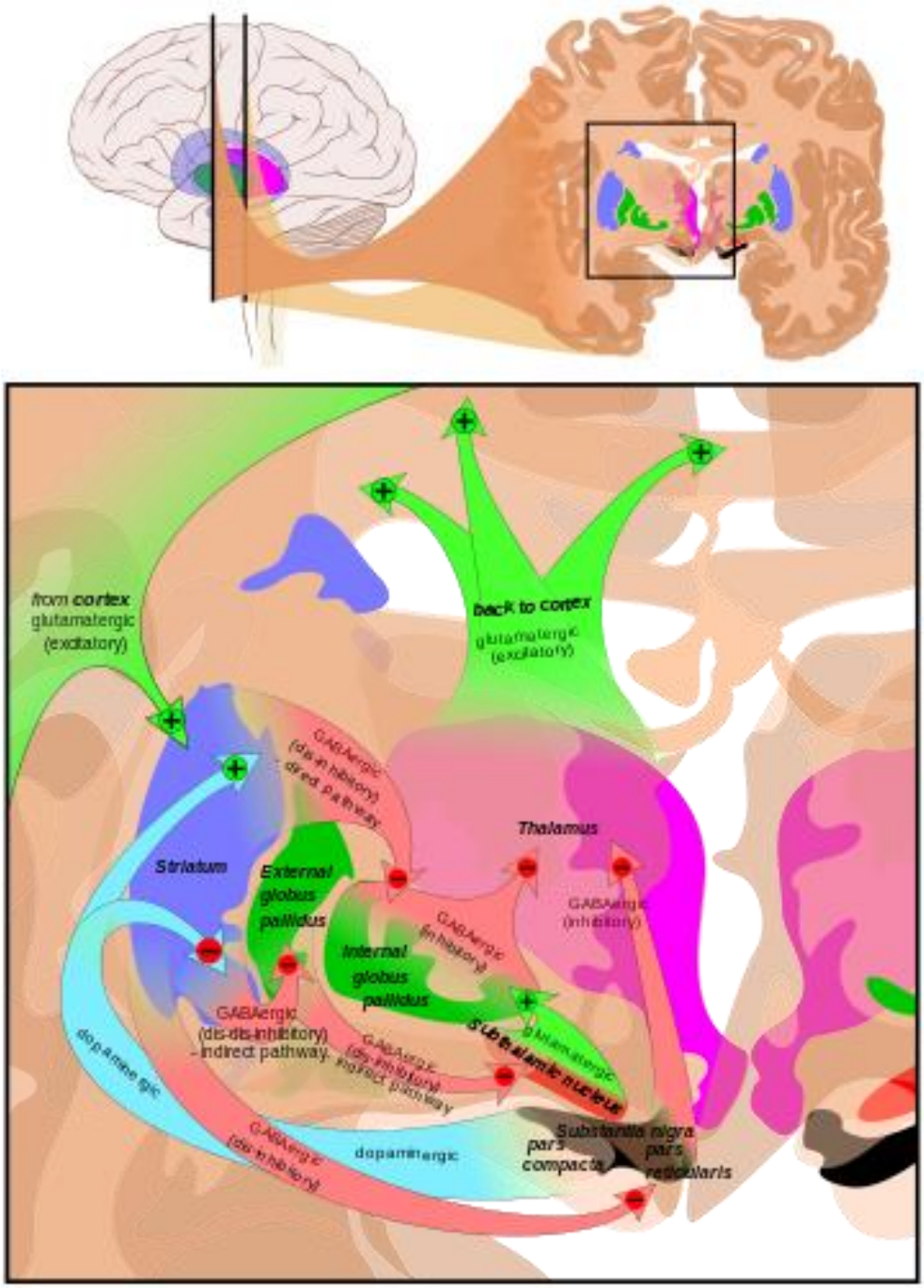


Figure 13 : Anatomie des circuits principaux au sein des ganglions de la base.

Le striatum est indiqué en bleu. Plusieurs coupes (en haut) ont été superposées pour observer toutes les structures impliquées au niveau des ganglions de la base (en bas). Les signes + et – des flèches indiquent si la voie est excitatrice ou inhibitrice. Les flèches vertes représentent les connexions excitatrices glutamatergiques, les rouges les inhibitrices GABAergiques et en turquoise sont indiquées les connexions dopaminergiques excitatrices de la voie directe et inhibitrices de la voie indirecte. Figure issue de (Häggström 2017).

Le striatum est composé de 95% de neurones épineux moyens, GABAergiques, créant des afférences vers la voir directe s'ils possèdent un phénotype D1 (majoritairement des récepteurs dopaminergiques activateurs de type D1 au niveau présynaptique) ou vers la voie indirecte s'ils possèdent un phénotype de type D2 (majoritairement des récepteurs dopaminergiques inhibiteurs de type D2 au niveau présynaptique). Le fonctionnement des voies directe et indirecte est décrit dans le paragraphe précédent (I.B.1.b). Les 5% restants représentent les interneurones cholinergiques sécrétant de l'acétylcholine, répondant aux stimuli environnementaux saillants grâce à l'établissement de réponses stéréotypées, de façon contigüe avec les neurones dopaminergiques de la substance noire (Goldberg & Reynolds 2011; Morris et al. 2004). Ces interneurones sont eux-mêmes sensibles à la dopamine puisqu'ils possèdent des récepteurs dopaminergiques de type D5 (Bergson et al. 1995). Sont également inclus dans ces 5% des interneurones GABAergiques présents sous des types variés (pour plus d'informations sur ces interneurones, voir la revue de Tepper et ses collaborateurs (Ibáñez-Sandoval et al. 2015).

Le striatum ventral reçoit des afférences directes de multiples régions du cortex et de structures limbiques comme l'amygdale, le thalamus ou encore l'hippocampe. La voie mésolimbique en provenance de l'ATV se projette directement sur le noyau accumbens au sein du striatum ventral.

Le striatum possède aussi des afférences bien connues venant de la SNpc par la voie nigrostriée. Contrairement aux neurones corticaux formant des synapses au niveau des têtes des épines dendritiques des neurones striataux épineux, les neurones nigraux forment des synapses principalement sur les troncs de ces mêmes épines. Chez les primates, l'afférence striatale provenant du thalamus provient des noyaux du groupe médian du thalamus et est glutamatergique. Enfin, le striatum reçoit des afférences d'autres noyaux des ganglions de la base comme des afférences glutamatergiques du NST et GABAergiques du globus pallidus.

Les neurones du striatum se projettent majoritairement sur le pallidum ventral, et le noyau dorso-médian du thalamus (NDMT) dans le cadre du circuit fronto-striatal. D'autres efférences venant du striatum ventral existent vers le globus pallidus et la SNr, entre-autres. Ces projections sont majoritairement celles des neurones épineux, le long du faisceau striato-pallidonigral visant les ganglions de la base. Les neurones de ce faisceau sont soumis à une inhibition par des synapses GABAergiques venant du striatum dorsal. D'autres projections existent vers l'extérieur de ce système, à savoir vers le colliculus supérieur ou le thalamus. Deux voies séparées vont vers le thalamus : la première traverse le globus

pallidus pour atteindre des noyaux ventraux du thalamus pour ensuite rejoindre l'aire motrice supplémentaire du cortex frontal, la deuxième passe par la substance noire puis les noyaux antérieurs du thalamus avant d'atteindre les cortex frontal et oculomoteur.

D'un point de vue fonctionnel, le striatum permet la coordination d'aspects multiples de la cognition, allant de la planification motrice et des actions à la prise de décision via un rôle dans la motivation, les efforts (Schmidt et al. 2012), le renforcement et la perception des récompenses (Yager et al. 2015; Taylor et al. 2013; Ferré et al. 2010). Chez les primates, des implications différentes ont été rapportées entre le striatum ventral (noyau accumbens et tubercule olfactifs (Ferré et al. 2010) et le striatum dorsal (noyau caudé et putamen). Le striatum ventral et le noyau accumbens en particulier, grâce à ses associations avec le système limbique, a surtout été mis en évidence comme jouant un rôle vital dans les circuits de la prise de décision, la motivation et des comportements liés aux récompenses. Le striatum dorsal quant à lui est principalement impliqué dans les fonctions cognitives motrices, exécutives (contrôle inhibiteur par exemple) et l'apprentissage stimulus-réponse (Yager et al. 2015; Roesch et al. 2006). Il semblerait aussi qu'il ait un rôle dans le traitement des punitions (Pessiglione et al. 2006; Palminteri et al. 2012) mais a été bien moins étudié que son équivalent ventral. Certaines fonctions sont partiellement partagées entre le striatum ventral et le striatum dorsal puisque la partie dorsale fait aussi partie du système de récompense via la médiation de l'encodage de nouveaux programmes moteurs associés à l'obtention de nouvelles récompenses (comme dans le cas des réponses motrices conditionnées à des stimuli appétitifs) (Taylor et al. 2013). La présence de récepteurs dopaminergiques métabotropiques (à protéine G) sur les neurones striataux épineux implique que leur activation déclenche des cascades cellulaires pouvant moduler à court ou long terme les fonctions pré- et post-synaptiques (Greengard 2001; Cachope & Cheer 2014). Chez l'humain, le striatum est ainsi activé par des stimuli associés à des récompenses mais aussi par des stimuli aversifs (Pessiglione et al. 2006), inattendus ou encore intenses, ainsi que par les symboles associés à de tels évènements (Volman et al. 2013). Des résultats de neuroimagerie fonctionnelle suggèrent que la propriété commune de tous ces stimuli, à laquelle le striatum réagit, est la saillance des stimuli (Luna & Sweeney 2004). Un grand nombre d'autres régions et circuits cérébraux sont associés aux récompenses, comme le cortex frontal. Des cartes fonctionnelles du striatum révèlent des interactions largement distribuées avec diverses aires corticales importantes pour une grande variété de fonctions (Choi et al. 2012). Toutes ses fonctions peuvent être altérées lorsque le striatum est affecté par une pathologie comme la maladie de Parkinson (réduction des afférences dopaminergiques du striatum dorsal (Drui et al. 2014; Carnicella et al. 2014; Favier et al. 2017) ou la maladie de Huntington (atrophie progressive du striatum dorsal puis ventral)

(Walker 2007). De par son implication dans le circuit de la récompense et la motivation, on retrouve un fonctionnement anormal du striatum dans nombre de pathologies psychiatriques (addictions (Nestler 2013; Olsen 2011), troubles bipolaires (McDonald et al. 2012), troubles obsessionnels compulsifs (Pena-Garijo et al. 2011).

d) Thalamus

Le thalamus est une structure sous-corticale diencéphalique (présente dans chaque hémisphère) située de part et d'autre du troisième ventricule. De par sa localisation intermédiaire entre le cortex et le tronc cérébral, le thalamus a un rôle de relais et d'intégration des afférences sensitives et sensorielles (en provenance du tronc cérébral) et des efférences motrices (vers le tronc cérébral), ainsi que dans la régulation de la conscience, de la vigilance et du sommeil. Le thalamus est une structure importante de projection des ganglions de la base vers le cortex, majoritairement à l'origine de projections efférentes GABAergiques vers le cortex, déjà mises en évidences au sein du réseau des ganglions de la base (Figure 11).

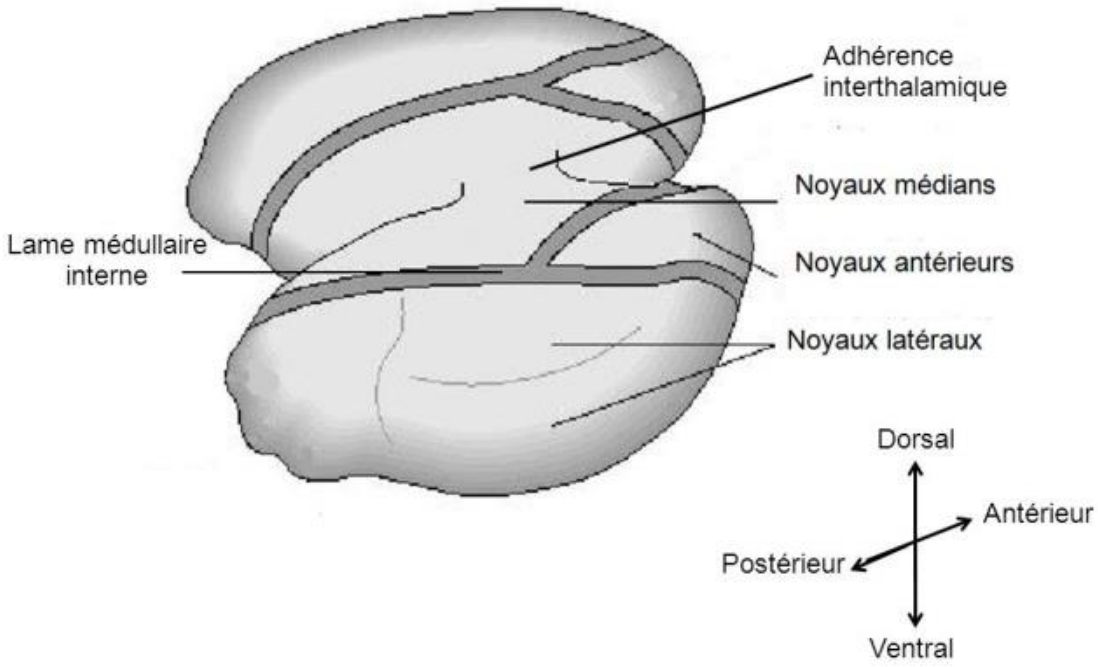


Figure 14 : Anatomie externe du thalamus et de ses noyaux principaux. Figure issue de (Landmann & Langlais 2014).

Au sein et autour des deux thalamus se rencontrent des cloisons fibreuses formées de fibres myélinisées, appelées « lames médullaires ». À l'intérieur de chaque thalamus se trouve la lame médullaire interne formant une fourche aux pôles antérieur et postérieur, Dans chaque hémisphère, la lame médullaire sépare différents groupes de noyaux au sein du thalamus : noyaux antérieurs, latéro-postérieurs et médians (Figure 14). Chacun de ces groupes de noyaux thalamiques est subdivisé en différents noyaux en fonction de leur localisation et de leurs afférences/efférences notamment corticales, striatales et cérébelleuses. Les connexions corticales des noyaux latéro-postérieurs se font majoritairement vers les cortex temporaux, pariétaux et occipitaux. Au contraire, les noyaux antérieurs et médians présentent nombre de connexions vers les cortex frontaux, préfrontaux, orbitaires, cingulaires et temporaux polaires, ainsi que provenant de la substance noire. Les afférences et efférences des noyaux antérieurs et médians en font donc des régions légitimement impliquées dans des processus cognitifs de haut niveau et en particulier motivationnels comme la prise de décision. Pour cela, je me limiterai à l'étude des principaux noyaux des groupes antérieurs et médians du thalamus.

(1) Le noyau antérieur du thalamus et le circuit de Papez

Le noyau antérieur du thalamus (NAT) se divise ainsi en trois sous-noyaux (antérodorsal, antéroventral, antéromédial). Le noyau antérieur reçoit des afférences en provenance des corps mamillaires et forme des efférences se projetant majoritairement vers le cortex cingulaire. La boucle fonctionnelle liant les corps mamillaires au cortex cingulaire en passant par le NAT est connue sous le nom de circuit de Papez (Figure 15). Le NAT est donc un relai au sein du circuit de Papez, surtout connu pour sa fonction limbique.

Le circuit de Papez est un ensemble de structures nerveuses impliquées dans le contrôle des émotions. Il a reçu le nom du neuroanatomiste américain James Papez (1883-1958) qui théorisa le rôle de ce circuit dans l'expérience émotionnelle (Papez 1995).

Ce circuit relie différentes structures du système limbique comprenant les cortex temporal et cingulaire, le thalamus, l'hypothalamus et certaines de leurs interconnexions, en passant par l'hippocampe, donnant ainsi son autre nom au circuit : le circuit hippocampo-mamillo-thalamique. L'hippocampe et le circuit de Papez jouent un rôle essentiel dans la mémorisation et la formation de souvenirs durables. En effet, une fois une information captée par les différents cortex sensoriels, c'est l'hippocampe, par le biais du circuit de Papez qui va répéter cette information (les activations). Celle-ci va donc transiter de l'hippocampe vers les corps mamillaires (les noyaux de l'hypothalamus) puis vers le thalamus (le relais de

l'information dans le cerveau), mais elle ne va pas s'arrêter là. Bien au contraire, elle va ensuite voyager vers le cortex cingulaire (lié aux émotions) pour enfin retourner vers l'hippocampe par le biais du cortex entorhinal (un autre relais de l'information). Quand cette boucle est finie, elle recommence. L'information va donc emprunter ce circuit de très nombreuses fois et c'est grâce à celui-ci que les différentes traces de l'expérience que je suis en train de vivre vont être associées. Au bout d'un long moment, cette association va se stabiliser et devenir indépendante de l'hippocampe. C'est ainsi que cette association des différentes traces mnésiques de mon expérience va donner naissance à un souvenir. Chacune de ces traces va ensuite retourner vers le cortex d'où elle vient, où elle pourra être stockée de façon quasi permanente.

La mémorisation de l'association de différentes informations liées entre-elles est essentielle dans le cadre de l'apprentissage par renforcement. Les associations entre stimuli, actions et renforcements doivent être mémorisées grâce à l'action du circuit de Papez pour permettre la production d'un comportement optimal. Le noyau antérieur du thalamus est donc une cible sous-corticale de choix pour étudier les bases neurales de l'apprentissage par renforcement au niveau sous-cortical.

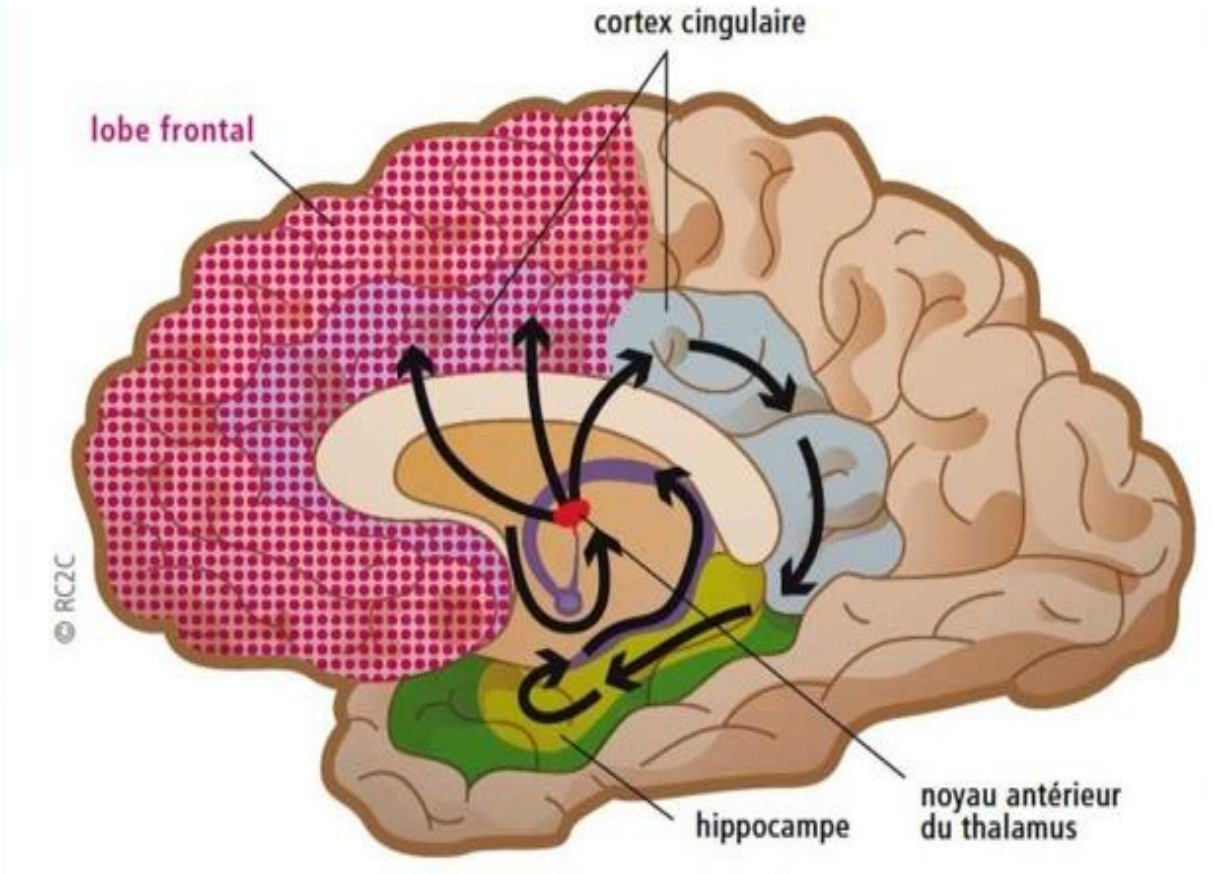


Figure 15 : Le circuit de Papez ou circuit hippocampo-mamillo-thalamique joue un rôle essentiel dans la formation de souvenirs durables. (Source Université de McGill)

(2) Le noyau dorso-médian du thalamus

Juste à côté du NAT se trouve le noyau dorso-médian du thalamus (NDMT), aux fonctions proches bien que différentes à celles du NAT. Le NDMT se compose également de trois sous-noyaux appelés magnocellulaire (médial), parvocellulaire (latéral) et paralaminaire. La partie parvocellulaire reçoit entre-autres des afférences venant de la SNc et forme des projections vers le cortex préfrontal. La partie magnocellulaire quant à elle reçoit des informations venant entre-autres de l'amygdale et de l'hypothalamus et forme des connexions efférentes vers les cortex orbitofrontal, préfrontal médian et temporal, grâce au faisceau thalamo-cortical. De par ses connexions, le NDMT a principalement une fonction limbique.

(3) Le noyau thalamique sous-médian ou noyau ventro-médian du thalamus

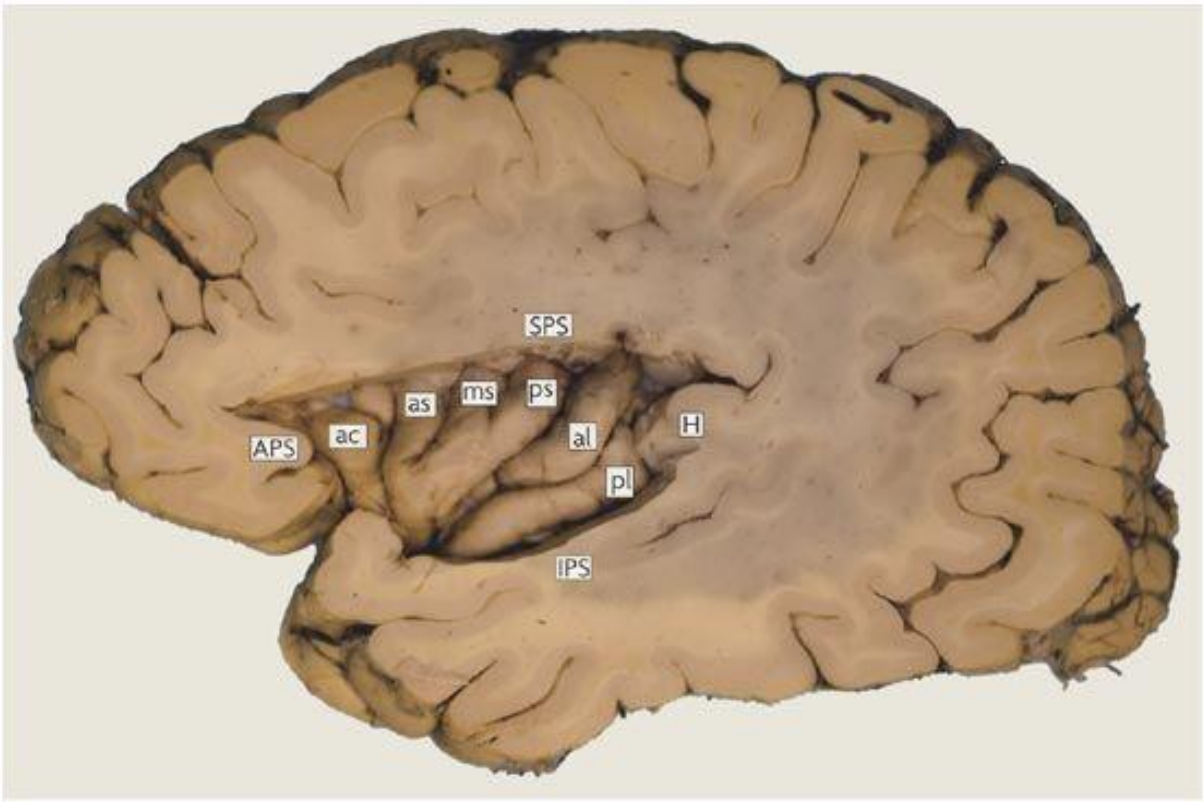
Le noyau sous-médian du thalamus (NSMT) reçoit des afférences GABAergiques en provenance des noyaux réticulaires du thalamus (Wang et al. 2005). Le NSMT possède des connexions efférentes allant vers le cortex insulaire ainsi que vers de cortex orbitofrontal ventro-latéral (vlPFC). Les neurones du NSMT ainsi que le lien direct observé avec le vlOFC ont été mis en évidence comme étant impliqués dans la modulation de l'information nociceptive (Fu et al. 2002; Tang et al. 2009). De plus, des connexions ont été rapportées entre le groupe des noyaux thalamiques médians et le cortex insulaire (Craig et al. 2000), déjà connu pour être impliqué dans le traitement de stimuli nociceptifs (Kurth et al. 2010; Meyniel et al. 2013; Craig 2003; Hayes et al. 2014), entre-autres grâce à ses afférences venant de l'amygdale identifiées dans les années 1980 chez le singe rhésus (Mufson et al. 1981).

e) Cortex insulaire

Cachée sous les replis du cortex frontal, du cortex orbitaire et du cortex operculo-temporal, au niveau du sillon sylvien, se trouve le cortex insulaire humain (originalement appelée « l'Ile

de Reil »). L'insula est une structure bilatérale de forme trapézoïdale, séparée des lobes corticaux voisins par le gyrus péri-insulaire (Augustine 1985; Augustine 1996; Türe et al. 1999; Afif & Mertens 2010; Afif et al. 2013; Craig 2009b). L'origine de l'insula est très probablement ancienne au cours de l'évolution, étant en effet considérée comme un des premiers lobes corticaux à être apparu (Craig 2003; Craig 2010), proposant ainsi une explication à la multitude de fonctions prises en charge ou impliquant l'insula (entéroception (Craig 2003), émotions comme la douleur, la peur ou le dégoût (Caruana et al. 2011), saillance (Ham et al. 2013; Menon & Uddin 2010). De par son organisation cellulaire très similaire à celle du cortex orbitofrontal latéral (Simon et al. 2006), elle possède des liens structuraux et fonctionnels forts avec le cortex préfrontal.

Anatomiquement, l'insula est divisée en deux zones par le sillon insulaire central, considéré comme étant une « continuation convexe » du sillon central séparant les lobes pariétaux et frontaux, délimitant une limite entre les portions antéro-ventrale et dorso-postérieure de l'insula. Alors que l'insula antérieure est principalement composée de trois gyri courts (antérieur, médian et postérieur), et plus rarement du gyrus accessoire, l'insula postérieure est composée de deux gyri longs (antérieur et postérieur) étant parfois séparés de manière



incomplète ((Türe et al. 1999) et Figure 16).

Figure 16 : Gyri et sillons de l'insula. L'insula possède trois gyri courts (antérieur as, médian ms et postérieur ps), deux gyri longs (antérieur al et postérieur pl) et un gyrus accessoire (ac). Figure issue de (Craig 2009b).

Ces gyri ne sont pas parfaitement parallèles, mais convergent plutôt vers deux pôles ventraux et relativement rostraux (antérieur et postérieur) (Afif & Mertens 2010) localisés de chaque côté des bordures de l'insula.

Certaines des propriétés de l'organisation cellulaire du cortex insulaire furent mises à jour grâce à des études cytoarchitectoniques qui permettent souvent de caractériser la « granularité » du tissu nerveux (c'est-à-dire le nombre de couches cellulaires, la présence et les caractéristiques des couches « granulaires » internes) comme un critère de base pour la parcellisation du cortex en sous-divisions. Des études princeps chez le primate (Mesulam & Mufson 1982) ont décrit la présence d'un gradient « concentrique » de granularité croissance allant de l'insula antérieure à l'insula postérieure, commençant par un cortex agranulaire (absence de la couche 4 de corps cellulaires) au niveau des limites de l'insula antérieure, ensuite entouré par un cortex dysgranulaire, lui-même encerclé par un cortex granulaire au niveau de l'insula postérieure dorsale. Des travaux plus récents chez le macaque rhésus combinant des études cytoarchitectoniques et immunohistochimiques (Gallay et al. 2012) confirment ce gradient progressif de granularité sur l'axe antéro-ventral vers dorso-postérieur, et suggèrent l'inclusion du cortex operculaire voisin (aire VS possiblement) dans les limites anatomo-morphologiques de l'insula. Cette organisation générale semble se confirmer chez l'humain. En effet, trois sous-divisions distinctes du cortex insulaire postérieur caudal furent retrouvées chez dix cerveaux humains post-mortem (Kurth et al. 2010), les deux divisions les plus dorsales furent labellisées comme étant granulaires, et du cortex dysgranulaire fut retrouvé plus ventralement. L'existence d'une relation entre la cytoarchitecture de ces différentes sous-divisions du cortex insulaire et des rôles fonctionnels distincts reste à être explorée. Il a d'ailleurs été souvent rapporté que ces aires définies selon la cytoarchitectonie ne sont pas parfaitement cohérentes avec des limites anatomiques (Nieuwenhuys 2012). Cependant, un schéma global semble ressortir à travers les différentes délimitations du cortex insulaire proposées, qu'elles soient anatomiques, morphologiques ou encore fonctionnelles (Klein et al. 2013).

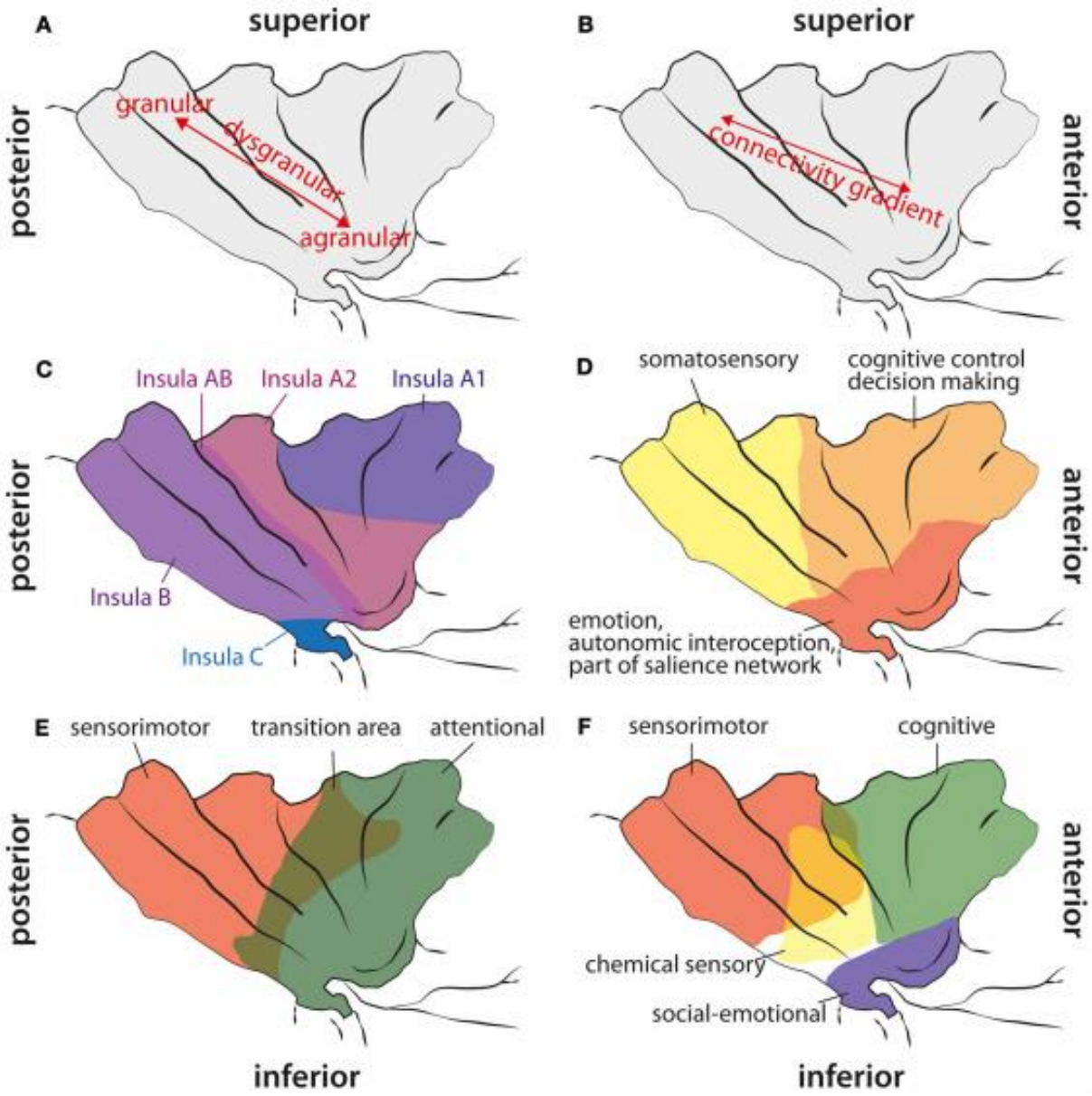


Figure 17 : Cartes de cytoarchitectonique, de connectivité structurelle et fonctionnelle de l'insula. (A) Gradient cytoarchitectonique allant de l'insula antérieure inférieure (ventrale) à l'insula postérieure, passant d'un cortex agranulaire à un cortex granulaire, avec une transition en cortex dysgranulaire. **(B)** Gradient de connectivité structurale dans l'insula (Cerliani et al. 2012). Contrairement à d'autres régions cérébrales comme le cortex prémoteur, aucune limite claire de l'insula ne fut identifiée d'après le profil de connectivité structurelle. A la place, il fut rapporté l'existence d'un changement graduel du profil de connectivité allant des aires antérieures aux aires postérieures de l'insula. **(C)** Carte de cytoarchitectonique adaptée de Von Economo et Koskinas (Triarhou 2013). Aucune région agranulaire ne fut trouvée au sein de l'insula (hormis une zone fronto-insulaire antérieure à ce qui est montré ici), mais une aire moins granulaire appelée « Insula A1) et des aires plus granulaires (Insula B) au niveau postérieur ont été observées. A noter que la présence d'une

aire de transition (Insula AB) fut rapportée entre les régions antérieure et postérieure. **(D)** Aires fonctionnelles de l'insula (Deen et al. 2011). **(E)** Différenciation fonctionnelle de l'insula (Cauda et al. 2011). A noter la présence d'une aire de transition entre les parties antérieure et postérieure. **(F)** Aires fonctionnelles de l'insula (Kurth et al. 2010). Quatre aires différentes ont été identifiées, ainsi qu'un recoupement clair de ces aires fonctionnelles au centre de l'insula. Figure adaptée de (Klein et al. 2013).

L'insula est bien située pour intégrer l'information relative à l'état du corps et rendre cette information disponible pour des processus cognitifs et émotionnels d'ordre supérieur. Elle reçoit des afférences sensorielles homéostatiques par l'entremise du thalamus, et elle envoie ses projections à plusieurs structures liées au système limbique comme l'amygdale, le striatum ventral et le cortex orbitofrontal. L'insula est aussi déjà bien associée aux processus de douleur ainsi qu'à plusieurs émotions de base comme la colère, la peur, le dégoût, la joie ou la tristesse. Sa portion la plus antérieure est considérée comme faisant partie du système limbique. L'insula serait aussi grandement impliquée dans les désirs conscients, comme la recherche active de nourriture ou de drogue. Ce qu'il y a de commun dans tous ces états, c'est qu'ils affectent le corps entier en profondeur. Un constat qui tend à renforcer son rôle probable dans la représentation que nous nous faisons de notre propre corps ainsi que dans l'aspect subjectif de l'expérience émotionnelle. Enfin, l'insula humaine, et à un moindre degré celle des grands singes, aurait deux innovations évolutives qui lui permettrait de porter la lecture de notre état corporel à un niveau inégalé chez les autres mammifères. D'abord la partie antérieure de l'insula, et plus particulièrement de l'insula de l'hémisphère droit, serait davantage développée chez les humains et les grands singes que chez les autres espèces animales. Ceci permettrait un décodage plus précis de nos états viscéraux, et donc par exemple à une simple mauvaise odeur de devenir un sentiment de dégoût (Caruana et al. 2011), ou encore au toucher d'une personne aimée de se transformer en sentiment de délice (Craig et al. 2000) grâce à des afférences des noyaux ventromédians du thalamus qui sont hautement spécialisées pour transmettre les informations homéostatiques (douleur, la température, les démangeaisons, le niveau d'oxygène et le toucher sensuel).

L'autre modification majeure à notre insula est la présence d'un type de neurones que l'on retrouve seulement chez les grands singes et l'humain. Il s'agit de grandes cellules nerveuses allongées en forme de cigare appelées neurones de Van Economo (Triarhou 2013). De plus, on ne retrouve ce type de neurones que dans l'insula et le cortex cingulaire antérieur. Ces neurones font des connexions avec diverses parties du cerveau, ce qui serait un atout essentiel pour les fonctions supérieures qu'on attribue à ces deux structures cérébrales. Une étude de neuroimagerie humaine utilisant la méthode DTI (diffusion tensor

imaging) (Jakab et al. 2012) a révélé que l'insula antérieure est interconnectée à des régions des lobes temporaux et occipitaux, au cortex operculaire et orbitofrontaux ainsi qu'au gyrus frontal inférieur. Cette même étude a révélé des différences anatomiques de profils de connections entre les hémisphères gauche et droit.en

Ainsi l'insula antérieure serait plutôt peu granulaire voire agranulaire (Figure 17 A et C) et associée à des fonctions cognitives et attentionnelles comme le contrôle cognitif et la prise de décision (Figure 17 D, E et F), justifiant son étude dans ce manuscrit dans le cadre de la prise de décision et plus particulièrement de l'apprentissage par renforcement.

f) Cortex préfrontal

La partie antérieure du cortex frontal est classiquement désignée sous le terme de cortex préfrontal (aires de Brodmann 8 à 14, 24 et 25, 32 et 44 à 47) (Murray et al. 2016). De manière générale, il est dit que c'est dans le cortex préfrontal que réside notre personnalité et ce qui fait de nous des êtres humains (DeYoung et al. 2010). Cette région cérébrale est en effet impliquée dans la planification de comportements cognitifs complexes, l'expression de notre personnalité, la prise de décision, le contrôle du comportement social (Yang & Raine 2009) et les fonctions exécutives en général (Shimamura 2000). L'activité de base du cortex préfrontal est considérée comme étant l'orchestration de nos actions et pensées en accord avec nos objectifs internes (Miller et al. 2002). Les fonctions exécutives se réfèrent à la capacité de différencier des pensées conflictuelles, de déterminer le bien du mal, le mieux du meilleur, l'identique du différent, de prédire les conséquences futures d'activités présentes, d'engager un comportement dirigé vers un but, d'exercer un contrôle des interactions sociales (supprimer des impulsions pouvant avoir des conséquences sociales délétères (Damasio 1994; Macmillan 2000; Macmillan 2008). De plus, le cortex préfrontal est le support de l'apprentissage suivant des règles (voir partie I) à un haut niveau d'abstraction (Badre et al. 2010).

La définition des limites anatomiques de cette région n'est pas universelle. Parmi les définitions les plus formelles et abouties, il y a les suivantes : 1) la partie du cortex frontal qui, lorsqu'elle est stimulée électriquement, n'élicite pas de mouvements (définition fonctionnelle) ; 2) la partie du cortex frontal recevant des projections provenant du noyau dorso-médian du thalamus (NDMT) (définition anatomique) ; 3) la partie du cortex frontal présentant une couche cellulaire granulaire (définition histologique). Le cortex préfrontal, en plus du NDMT, reçoit également des afférences directes depuis l'hypothalamus, l'amygdale, le cortex limbique et d'autres aires corticales associatives sensorielles. Comme mentionné

précédemment, le cortex préfrontal est une aire corticale recevant la majorité des projections dopaminergiques en provenance de l'ATV.

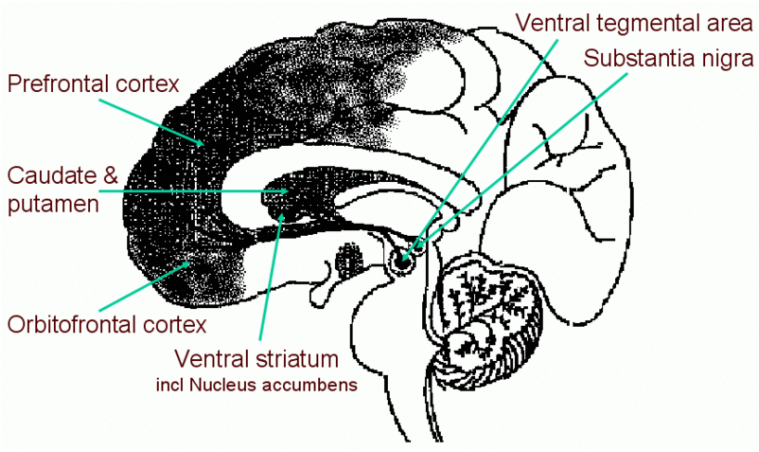


Figure 18 : Le système dopaminergique forme de multiples connexions avec le cortex préfrontal. Les zones innervées par le système dopaminergique sont indiquées en nuances de gris. Figure adaptée de (Schultz 2007).

En plus de la dopamine (Figure 18), le cortex préfrontal reçoit aussi des projections d'autres systèmes de neurotransmission utilisant pour neurotransmetteurs : les systèmes noradrénergique (depuis le locus coeruleus), sérotoninergique (depuis les noyaux du raphé) et acétylcholinergique (depuis les noyaux septaux). Pratiquement toutes les connexions préfrontales sont réciproques : les structures projetant leurs efférences sur le cortex préfrontal reçoivent également ses afférences. La seule exception concerne les ganglions de la base, auxquels le cortex préfrontal envoie des projections afférentes directes mais non-réciproques au niveau du striatum (Fuster 1997). Au cours de l'évolution, le cortex préfrontal a subi une expansion plus massive que le reste du cortex. Sa taille relative (par rapport à la taille du cerveau ou au poids corporel) atteint son maximum avec l'espèce humaine, où le cortex préfrontal représente presque un tiers de l'ensemble du néocortex. C'est pour cette raison qu'il a souvent été associé au développement de fonctions cognitives supérieures chez l'humain (Semendeferi et al. 2002).

Au sein du cortex préfrontal, une sous-division anatomo-fonctionnelle a été faite, impliquant au moins trois principaux sous-territoires : le cortex préfrontal latéral (latPFC, subdivisé en parties dorsolatérale : BA 8 9lat 46, et ventrolatérale : BA 12 44 45 47) impliqué dans des actions complexes de planification de fonctions exécutives (Koechlin & Summerfield 2007), le cortex préfrontal médian (mPFC, incluant le cortex cingulaire et le cortex préfrontal dorsomédian, BA 9med 10med 24 25 32) impliqué dans le contrôle de la performance (Ridderinkhof 2014) et pour finir le cortex préfrontal ventral (vPFC, recouvrant les cortex orbitofrontaux, préfrontaux ventrolatéral et ventromédian, BA 10 11 13 14) jouant un rôle clé dans les processus affectifs et émotionnels (Kringelbach & Rolls 2004).

(1) Le cortex préfrontal médian

Le cortex préfrontal médian (mPFC) est composé d'aires corticales granulaires (parties médianes de BA 9 et 10) et agranulaires (BA 24, 25 et 32) et comprend le cortex cingulaire antérieur (BA 24), le cortex infralimbique (BA 25) et le cortex prélimbique (BA 32). La neuranatomie du cortex cingulaire est décrite plus en détails plus loin. Il a été rapporté l'implication du mPFC dans la génération du sommeil à ondes lentes ainsi que dans la consolidation de la mémoire (Mander et al. 2013).

(2) Le cortex préfrontal ventral

Une région particulièrement intéressante innervée par des cellules dopaminergiques est la région la plus ventrale du cortex préfrontal (vPFC). Cette région est fortement interconnectée avec des régions cérébrales impliquées dans les émotions (Price 1999) et reçoit des afférences des systèmes d'arousal (excitation) du tronc cérébral, rendant sa fonction hautement dépendante de l'environnement neurochimique (Robbins & Arnsten 2009), ce qui permet une coordination directe entre l'état d'arousal et l'état mental (Arnsten et al. 2010). Le cortex préfrontal ventral se compose de trois régions cytoarchitectoniques majeures : les aires de Brodmann 11 (granulaire), 13 et 14 (à la fois granulaires et agranulaires).

En se basant sur les données de cytoarchitectonie et de connectivité, le vPFC semble ségrégué en deux circuits différenciables : les aires 11 et 13, classiquement appelées cortex orbitofrontal (OFC), et des régions plus médianes regroupées dans une aire fonctionnellement définie comme le cortex préfrontal ventromédian (vmPFC) (Ongür & Price 2000; Bechara et al. 1998; Wallis 2011).

D'un point de vue fonctionnel, le cortex orbitofrontal est impliqué dans des processus cognitifs de prise de décision. L'OFC est considéré comme un synonyme du vmPFC sur des bases anatomiques (MacPherson et al. 2002). La distinction doit donc se faire d'après les connexions neurales et les fonctions mises en jeu (Barbas et al. 2003). L'OFC est défini comme étant la partie du cortex préfrontal recevant des projections des noyaux médians du thalamus (NDMT magnocellulaire) et est considéré comme représentant les émotions et les récompenses au cours de la prise de décision (Fuster 1997). L'OFC possède des connexions réciproques avec le cortex insulaire (aires granulaires, agranulaires et

dysgranulaires) ainsi qu'avec le parahippocampe et l'hippocampe (Cavada et al. 2000). L'OFC possède aussi de nombreuses connexions réciproques avec l'amygdale. Ces connexions avec l'hippocampe et l'amygdale contribuent à la modulation des processus d'apprentissage associatif et de régulation émotionnelle (Barbas & Zikopoulos 2007). D'autres projections sous-corticales proviennent du striatum et en particulier des aires ventrales associées aux récompenses (Kringelbach 2005; Hollerman et al. 2000). Des boucles parallèles corticostriatales semblent impliquées dans les processus de comportement dirigé vers un but et des comportements habituels, alors que les boucles corticolimbiques semblent plus impliquées dans la sélection de l'action, en collaboration avec l'amygdale, et l'intégration d'informations pour adapter le comportement (Balleine & O'Doherty 2010). Des études invasives de tractographie DTI ont été réalisées chez le macaque rhésus (Lehéricy et al. 2004) pour cartographier la connectivité du cortex orbitofrontal avec d'autres structures corticales et sous-corticales. Les connexions de l'OFC humain conservent un même schéma que celles observées chez le singe rhésus, bien que ce schéma diffère au niveau du striatum.

La partie médiale du cortex orbitofrontal (mOFC) est aussi connue sous le nom de cortex préfrontal ventro-médian (vmPFC). Bien que le vmPFC ait été originellement identifié comme se référant à l'aire 10, des études ont mis en évidence chez le singe et l'humain qu'il se référait en réalité à l'aire 14 ainsi qu'à la partie ventrale de l'aire 10m (Wallis 2011; Wallis & Kennerley 2011). Des lésions réalisées sur des macaques (Mackey & Petrides 2010) prêchent en faveur d'une correspondance entre le vmPFC chez le macaque et chez l'être humain. La couche granulaire 4 du vmPFC fait de cette région une adaptation probablement spécifique aux primates (Mackey & Petrides 2010; Tsujimoto et al. 2012). Il a été suggéré que le mOFC est impliqué dans la création d'associations stimulus-récompense et dans le renforcement du comportement.

Au contraire, la partie latérale de l'OFC (latOFC) est impliquée dans les associations stimulus-conséquence (outcome) et dans l'évaluation et possiblement l'inversion du comportement (Walton et al. 2010). Le latOFC est activé lors de l'encodage de nouvelles attentes de punitions et de représailles sociales (Campbell-Meiklejohn et al. 2012). Le latOFC permet aussi la suppression d'émotions négatives, en particulier dans les situations d'approche et d'évitement (Astolfi et al. 2010). Le latOFC joue un rôle important dans la résolution de conflits et des lésions de cette aire entrainent des manifestations inappropriées de colère ainsi que des réponses tout aussi inappropriées à la colère des autres (Meyers et al. 1992).

Malgré tout, l'OFC humain compte parmi les aires cérébrales les moins bien comprises, en particulier en raison de sa localisation profonde compliquant fortement son étude de façon non invasive. Il a été proposé que l'OFC soit impliqué dans l'intégration sensorielle, la représentation des valeurs affectives de renforcement, dans la prise de décision et l'attente des conséquences d'actions (Kringelbach 2005). En particulier, l'OFC semble important dans la signalisation de récompenses et de punitions attendus suite à une action (Schoenbaum et al. 2011).

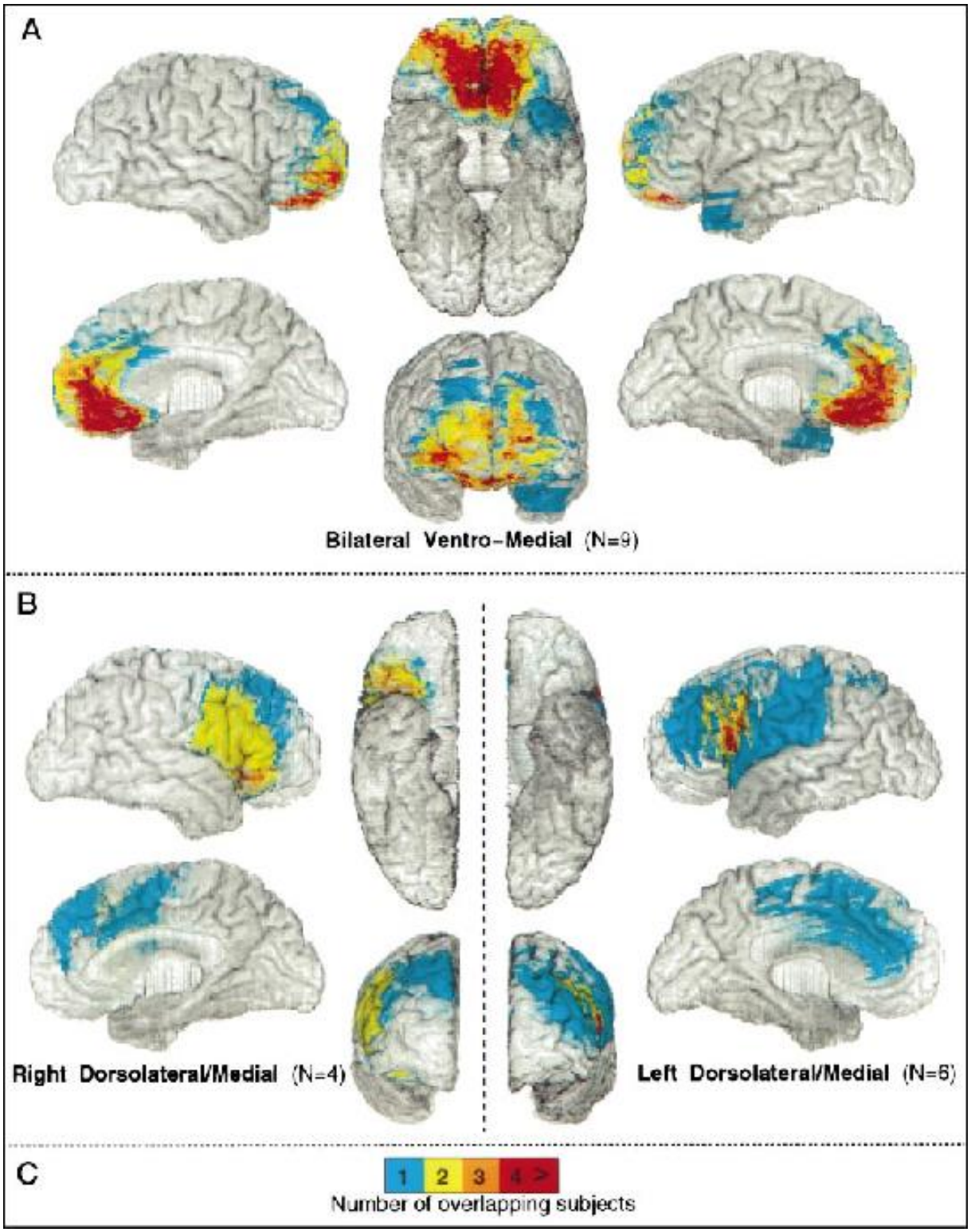


Figure 19 : Localisation du cortex ventromédian préfrontal (en rouge en A) et du cortex préfrontal dorsal (en jaune en B). Figure issue de (Bechara et al. 1998). Le recoupement des lésions observées chez des patients (C) indique le nombre de patients correspondant à ces recoupements.

Grâce à cela, le cerveau est capable de comparer les récompenses et punitions attendues à celles réellement obtenues, permettant ainsi l'adaptation du comportement. Ces rôles sont

supportés par des travaux chez l'humain, les primates non-humains et les rongeurs. La recherche humaine s'est focalisée sur des travaux de neuroimagerie (Volz et al. 2008) chez des sujets sains et de neuropsychologie (Bechara et al. 1998) chez des patients présentant des lésions discrètes au niveau du cortex orbitofrontal (Figure 19). L'implication de l'OFC dans des processus de jugement de valeur, de motivation et ses connexions dopaminergiques expliquent ses liens avec l'addiction (Leong et al. 2017).

(3) Le cortex préfrontal latéral

Le cortex préfrontal latéral se découpe en deux sous-divisions : le cortex préfrontal dorsolatéral (dlPFC, BA 9lat et 46) et le cortex préfrontal ventrolatéral (vlPFC, BA 12, 44, 45 et 47).

Le cortex préfrontal dorsolatéral est une région du cortex préfrontal parmi les plus récentes de l'évolution, apparue chez les primates et très développée chez l'être humain. Sa maturation se termine à l'âge adulte (Nelson & Collins 2008). Il s'agit plus d'une délimitation fonctionnelle qu'anatomique. Le dlPFC est fortement interconnecté avec des régions cérébrales impliquées dans les processus attentionnels, cognitifs et liés à l'action (Goldman-Rakic 1988) comme le cortex orbitofrontal, le thalamus, les ganglions de la base (surtout le noyau caudé du striatum) et l'hippocampe (Moss & Simmon 2013). Il a également été montré que des lésions au niveau du dlPFC sont responsables de déficits dans la mémoire à court-terme (Jacobsen 1935), soutenant un rôle du dlPFC dans le stockage de la mémoire à court terme (Funahashi et al.1993). Ses fonctions majeures couvrent donc les fonctions exécutives comme la mémoire de travail, la flexibilité cognitive (Kaplan et al. 2016), la planification, l'inhibition et le raisonnement abstrait (Miller & Cummings 2007). Ses liens avec des régions limbiques et son rôle exécutif expliquent son implication dans la prise de décision risquée et morale (Greene et al. 2001).

Il a été rapporté des distinctions fonctionnelles entre les cortex préfrontaux ventrolatéraux gauche et droit (Levy & Wagner 2011). Le vlPFC droit serait ainsi une région clé dans le contrôle cognitif et l'inhibition motrice (Aron et al. 2004). Une autre hypothèse suggère l'existence de deux circuits pariéto-frontaux distincts (Badre & Wagner 2007) régissant l'attention spatiale, avec le vlPFC droit permettant la réorientation de l'attention en gouvernant le réseau ventral de l'attention. Le vlPFC est en effet la partie terminale de la voie ventrale apportant des informations sur les caractéristiques des stimuli (Lee et al. 2013). Le vlPFC droit est actif au cours de l'inhibition motrice, avec une activation critique pour stopper l'action du cortex moteur. C'est aussi le lieu de la mise à jour des plans d'action et

cette région est sensible à l'incertitude au cours de la prise de décision (Levy & Wagner 2011). Cette latéralisation et prédominance du vlPFC droit a été rapportée chez des individus droitiers. Il est possible que cette prédominance soit plus tenue voire inversée chez des individus ambidextres et gauchers.

g) Le cortex cingulaire

Le gyrus cingulaire est un gyrus du lobe limbique du cortex cérébral. Il est situé sur la face médiale des hémisphères, au-dessus du corps calleux. Au-dessus, il est séparé du gyrus frontal supérieur par le sillon cingulaire et du précuneus par le sillon sous-pariétal. En dessous, sa limite est le sillon du corps calleux. Au niveau du splénium du corps calleux, le gyrus cingulaire se rétrécit dans l'isthme qui se poursuit par le gyrus parahippocampique (Houde et al. 2002). On peut diviser le gyrus cingulaire en quatre grandes parties, chacune d'elles accomplissant des tâches spécifiques. Le cortex cingulaire antérieur (CCA) occupe ainsi un rôle dans les états affectifs alors que le cortex cingulaire moyen (CCM) intervient dans le choix des réponses. Le cortex cingulaire postérieur (CCP), quant à lui, tient une place dans la fonction mémorielle. Enfin, le cortex cingulaire rétrospénial (CCR), participe au traitement des informations visuo-spatiales. L'ancien "cortex cingulaire antérieur" est maintenant divisé en cortex cingulaire antérieur et moyen et l'ancien "cortex cingulaire postérieur", maintenant nommé gyrus cingulaire postérieur est divisé en cortex cingulaire postérieur et rétrosplénial.

Le cortex cingulaire reçoit des afférences du thalamus et du néocortex et se projette sur le cortex entorhinal et le cingulum. C'est une partie intégrante du système limbique, qui est impliqué dans la formation et le traitement des émotions (Hadland et al. 2003b), l'apprentissage (Bussey et al. 1996), la sélection d'actions apportant une récompense (Hadland et al. 2003a) et la mémoire (Kozlovskiy et al. 2012; Rushworth et al. 2003). La combinaison de ces trois fonctions fait du cortex cingulaire un pilier influent dans l'association des conséquences des actions à la motivation (par exemple si une action induit une réponse émotionnelle positive, permettant un apprentissage) (Hayden & Platt 2010). Ce rôle implique que le cortex cingulaire est particulièrement important dans des troubles de type dépressifs (Drevets et al. 2008) ou schizophréniques (Adams & David 2007). Il joue aussi un rôle dans les fonctions exécutives et le contrôle respiratoire.

(1) Cortex cingulaire antérieur

Nous avons mentionné un lien direct entre le cortex cingulaire antérieur (CCA) avec le cortex insulaire et le circuit de Papez. Comme l'indiquent ces connexions, le cortex cingulaire antérieur (aires de Brodmann 24, 32 et 33 (Hellwig 1993; Triarhou 2013; Bailey & Von Bonin 1957) voir Figure 20) joue lui aussi un rôle d'interface important entre l'émotion et la cognition, plus précisément dans la transformation de nos sentiments en intentions et en actions. Il est impliqué dans des fonctions supérieures comme le contrôle de soi sur ses émotions, la concentration sur la résolution d'un problème, la reconnaissance de nos erreurs, la promotion de réponses adaptatives en réponse à des conditions changeantes. Des fonctions qui toutes impliquent un lien étroit avec nos émotions. Les primates, et donc les humains, sont des créatures hautement sociales. Connaître les intentions des autres (empathie) a de tout temps été crucial pour notre survie (Proverbio et al. 2010).

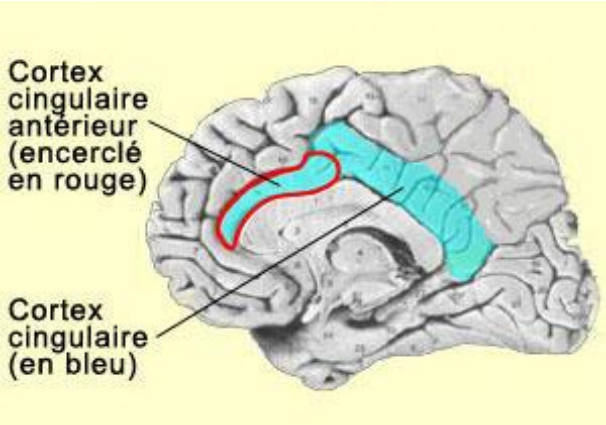


Figure 20 : Division du cortex cingulaire en deux parties antérieure et postérieure. La partie bleue non cerclée de rouge correspond au cortex cingulaire postérieur. Figure adaptée de (Wheatley et al 2007).

Le CCA reçoit principalement des afférences venant des noyaux intralaminaires et médians du thalamus. Le noyau antérieur du thalamus reçoit des afférences mamillo-thalamiques au sein du circuit de Papez (voir ci-dessus) et projette des connexions vers le CCA. En plus de son implication dans la détection d'éléments émotionnellement chargés nous concernant ou pas (cas de la douleur empathique via les neurones miroirs) (Hutchison et al. 1999), le CCA est impliqué dans des processus de détection des erreurs et des conflits.

(2) Cortex cingulaire postérieur

Le cortex cingulaire postérieur (CCP) correspond aux aires 23 et 31 de Brodmann. Sa structure est granulaire. Le CCP reçoit une grande partie de ses afférences du noyau supérieur du thalamus, lui-même innervé par le subiculum de l'hippocampe. Dans une certaine mesure, ces connexions répliquent le circuit de Papez. Le CCP (BA23) se projette sur le cortex préfrontal dorsolatéral (BA 9), le cortex préfrontal antérieur (BA 10), le cortex orbitofrontal (BA 11), le gyrus parahippocampique et plusieurs régions du lobe temporal

(Pandya et al. 1981). Le cortex cingulaire postérieur a été identifié pour son implication dans certains processus cognitifs et dans certaines pathologies de par son lien avec les processus de mémorisation. Il existe en effet un hypométabolisme du CCP au cours de la maladie d'Alzheimer (Leech & Sharp 2014).

2. Corrélats neuronaux de l'apprentissage par renforcement et du risque

Nous venons de voir que le système dopaminergique est depuis de nombreuses années reconnu pour son implication dans les processus motivationnels et ce que l'on appelle le « système de la récompense ». Par ailleurs, de nombreuses aires cérébrales lui étant directement ou indirectement connectées semblent fonctionnellement impliquées dans divers processus cognitifs liés à la prise de décision.

Pour étudier les implications fonctionnelles des activations cérébrales liées à la prise de décision, il est important de considérer à la fois les déficits comportementaux observables suite à des dommages neurologiques (approche neurologique des neurosciences cognitives) ainsi que les mesures directes de l'activité cérébrale pouvant être corrélés avec les stimuli issus du monde extérieur et les comportements produits (approche psychologique des neurosciences cognitives) (Glimcher & Fehr 2014). Comme longuement expliqué au cours du premier chapitre (I.A), l'utilisation de modèles computationnels est un apport majeur permettant de lier entre les déficits, activations et comportements observés. Cette association de la neurobiologie et de modèles computationnels pour l'étude des bases neurales de la prise de décision liées aux valeurs des actions est connue sous le nom de Neuroéconomie, et est au cœur des travaux présentés par la suite (Rangel et al. 2008). La première démonstration claire de la corrélation entre une activité neuronale et un choix observé remonte à 1989 (Newsome et al. 1989) avec la mise en évidence d'un lien psychométrique-neurométrique (choix-fréquence de décharge).

Cette seconde partie vise donc à faire une revue aussi exhaustive que possible des réponses fonctionnelles rapportées à ce jour au niveau de ces aires cérébrales au cours de la prise de décision. Ces résultats peuvent venir à la fois de travaux chez l'humain que chez l'animal et être issus d'études électrophysiologiques (Fischer & Ullsperger 2013), lésionnelles (Damasio et al. 1994; Bechara et al. 2000; Palminteri et al. 2012; Fleming et al. 2014) ou de neuroimagerie (Ursu & Carter 2005; Bartra, McGuire & Kable 2013; Rouault 2015). En raison de l'axe de recherche suivi par les travaux présentés dans ce manuscrit, à savoir la dynamique cérébrale de l'apprentissage par renforcement, cette partie sur les corrélats neuronaux s'intéressera séparément aux résultats rapportés au cours de

l'apprentissage par récompense (et du conditionnement en condition appétitive hors apprentissage) et ceux rapportés au cours de l'apprentissage par évitement des punitions (et du conditionnement en condition aversive hors apprentissage). Comme mentionné à plusieurs reprises dans le premier chapitre (I.A) et dans la revue d'anatomie (I.B.1), un lien entre apprentissage par renforcement et risque existe et mérite d'être abordé. Pour cela, les réponses fonctionnelles au risque connues seront rapportées séparément au sein d'une troisième section.

L'intérêt de cette organisation vise à mettre en évidence l'existence de deux théories coexistant actuellement sur les bases neurales de l'apprentissage par renforcement au niveau cortical, l'une supposant un encodage des récompenses et des punitions au sein d'un système unique et selon un axe fonctionnel continu (d'appétitif à aversif en passant par neutre), quand l'autre théorie propose l'existence de deux systèmes distincts traitant séparément les récompenses et les punitions (Pessiglione et al. 2006; Palminteri et al. 2012; Palminteri et al. 2015; O'Doherty et al. 2001; Jung et al. 2010; Jung et al. 2011). Au niveau sous-cortical, un consensus n'a pas non plus été trouvé en raison de résultats divers suggérant soit l'existence de neurones encodant tout le spectre de valence des renforcements grâce à une augmentation ou à une dépression dans leur fréquence de décharge (Mirenowicz & Schultz 1996; Schultz 1997; Monosov et al. 2015), soit la ségrégation au sein des noyaux eux-mêmes de neurones traitant les récompenses ou les punitions (Namburi et al. 2015), soit encore deux systèmes sous-corticaux se basant sur des noyaux distincts.

Nous avons vu précédemment que les erreurs de prédictions servent probablement de moteur à l'apprentissage par renforcement. Ces erreurs de prédiction représentent la différence entre le renforcement obtenu et celui attendu (valeur subjective). Pour cela, cette revue des corrélats neuronaux de l'apprentissage par renforcement s'intéressera à la bibliographie s'étant intéressée aux réponses aux renforcements, aux valeurs subjectives et aux erreurs de prédiction au sein des régions cérébrales préalablement identifiées dans la revue anatomique (I.B.1).

a) Apprentissage par récompense

(1) Réponses aux stimuli appétitifs

RONGEUR ET PRIMATE NON-HUMAIN

Les premières études sur le traitement cérébral des récompenses chez le primate ont mis en évidence une activité de l'OFC lorsque les animaux recevaient une récompense sous la forme de jus à boire (Rosenkilde et al. 1981). Ces études ont montré que bien que la satiété ne modifiait pas l'activité neurale de l'aire gustative primaire de l'insula et de l'opercule frontal (Yaxley et al. 1988), elle réduisait la réponse de l'aire gustative secondaire située dans l'OFC caudal et dans une partie plus médiale de l'OFC (Rolls et al. 1996). De manière concomitante, d'autres auteurs ont montré que certains neurones de l'OFC répondent à une récompense de jus mais pas à une solution saline (punition), et pourraient ainsi potentiellement apporter l'information qu'une récompense vient d'être reçue (Thorpe et al. 1983). L'ensemble de ces résultats suggère que l'activité de l'OFC reflète la magnitude des renforcements reçus, plutôt que leurs propriétés sensorielles.

Dans des études ultérieures, un encodage des renforcements a également été trouvé dans diverses régions du cortex frontal, avec la magnitude et le type des renforcements modulant l'activité des neurones du cortex cingulaire antérieur (ACC), de l'OFC et du cortex préfrontal latéral (latPFC) (Hikosaka & Watanabe 2000; Roesch et al. 2006; Schoenbaum et al. 2006).

En enregistrant des neurones du striatum ventral et dorsal pendant un protocole de conditionnement Go/No-Go, Apicella et ses collaborateurs (Apicella et al. 1991) ont trouvé qu'une proportion significative des neurones enregistrés présentait une augmentation de leur activité en réponse à la délivrance de jus (récompense liquide) dans les essais Go comme No-Go. Ils ont aussi noté que ces réponses aux récompenses étaient prédominantes et similaires à celles déjà rapportées au niveau de l'OFC. Ils identifièrent des activations phasiques signalant une récompense dès le stimulus de prédiction (valeur attendue), des activations phasiques encodant les valeurs obtenues au moment de la délivrance des récompenses, et des réponses toniques graduelles signalant l'occurrence prochaine d'une récompense (Figure 21 milieu).

HUMAIN

Dans le cadre d'une étude comparative des réponses cérébrales aux stimuli appétitifs et aversifs chez l'humain et d'autres mammifères, Hayes et ses collaborateurs (Hayes et al. 2014) rapportent des activations au sein de régions cérébrales sélectivement suite aux récompenses (vmPFC et ATV par exemple), sélectivement suite aux punitions (comme le cortex cingulaire et le noyau périaqueducal gris). Ces résultats laissent penser que deux systèmes existeraient en parallèle pour traiter les récompenses séparément des punitions. Cependant, ils ont aussi mis en évidence des régions répondant à la fois aux récompenses

et aux punitions mais de façon ségréguée (insula antérieure et amygdale), ainsi que des régions répondant asymétriquement à ces deux types de renforcements (le cortex orbitofrontal et le striatum ventral par exemple). Ces résultats vont dans le sens d'une éventuelle ségrégation fonctionnelle au sein même de régions cérébrales au cours de l'apprentissage par récompense et de l'apprentissage par évitement des punitions. L'hypothèse avancée par ces chercheurs est celle d'une dynamique spatio-temporelle différente entre récompenses et punitions, pouvant expliquer à la fois les similarités et les différences d'activités liées aux stimuli appétitifs et aversifs. Des travaux de neuroimagerie chez l'humain indiquent que les réponses du vmPFC suite au gain ou à la perte d'argent suggèrent qu'éviter un stimulus aversif peut être considéré comme une récompense relative (Kim et al. 2006).

Au cours d'un protocole de Go/No-Go, Thut et ses collaborateurs ont mis en évidence le rôle du thalamus dans la détection des stimuli appétitifs (Thut et al. 1997). Ils rapportent une augmentation du signal BOLD au sein du thalamus suite à la délivrance de récompenses monétaires. Ils n'écartent cependant pas une réponse en réalité à la saillance du stimulus plutôt qu'à la valence (appétitive en l'occurrence). Cette hypothèse est en lien avec des travaux plus récents où un protocole cognitif « oddball » a été utilisé pour tester les réponses à la saillance (Brázdil et al. 2007). Le thalamus avait là aussi été identifié comme étant activé suite à un évènement saillant, et cette réponse n'était pas liée à la valence. La question du rôle du thalamus dans la détection de la valence ou de la saillance reste donc ouverte.

(2) Réponses aux valeurs subjectives

RONGEUR ET PRIMATE NON-HUMAIN

Le vmPFC est fortement interconnecté avec les structures limbiques et du système nerveux autonome, alors que l'OFC est fortement interconnecté avec les aires sensorielles (Ongür & Price 2000). Des études précédentes suggèrent que la ségrégation anatomique est accompagnée de différences fonctionnelles (Balleine & Dickinson 1998; Gottfried et al. 2003; Kringelbach et al. 2003; Kable & Glimcher 2007; Ostlund & Balleine 2007; Rushworth & Behrens 2008; Gläscher et al. 2009; Hare et al. 2009), mais la plupart des études sur la prise de décision basée sur la valeur des options réalisées chez le singe se sont intéressées à l'OFC. Ceci est en partie dû à la localisation anatomique de l'OFC, le rendant plus accessible au cours d'études électrophysiologiques en raison de sa localisation proche de la surface du cerveau, et suffisamment latérale pour éviter tout contact accidentel avec le sinus central. Les lésions préfrontales ont été mises en évidence dans l'induction de changement de

schémas d'appétit et d'alimentation : des singes avec des ablations au niveau de l'amygdale et du vmPFC présentaient des préférences inhabituelles pour la viande crue (Ursin et al. 1969) ; les singes avec une ablation de l'OFC exhibaient plus de tendances à produire des comportements oraux, des réponses instrumentales persistantes envers des objets non-comestibles (Butter et al. 1969) et avaient des choix inhabituels de nourriture. Plus important encore, les singes avec de telles lésions présentaient une diminution dans leur classification des aliments selon s'ils les appréciaient ou pas, perdant ainsi la dissociation entre récompense et punition dans une certaine mesure (Baylis & Gaffan 1991). Les singes présentaient donc un déficit de l'encodage des valeurs subjectives des stimuli. Des déficits similaires furent rapportés chez des singes avec des lésions bilatérales de l'OFC (Izquierdo et al. 2004). Malgré ces changements, les animaux n'avaient pas de difficulté au cours des tâches de discrimination visuelle à la suite de ces lésions. Le déficit lié à une lésion de l'OFC n'affectait donc pas la capacité à distinguer des aliments différents mais plus à leur attribuer des valeurs subjectives d'appétence (appétitivité ou aversivité gustative).

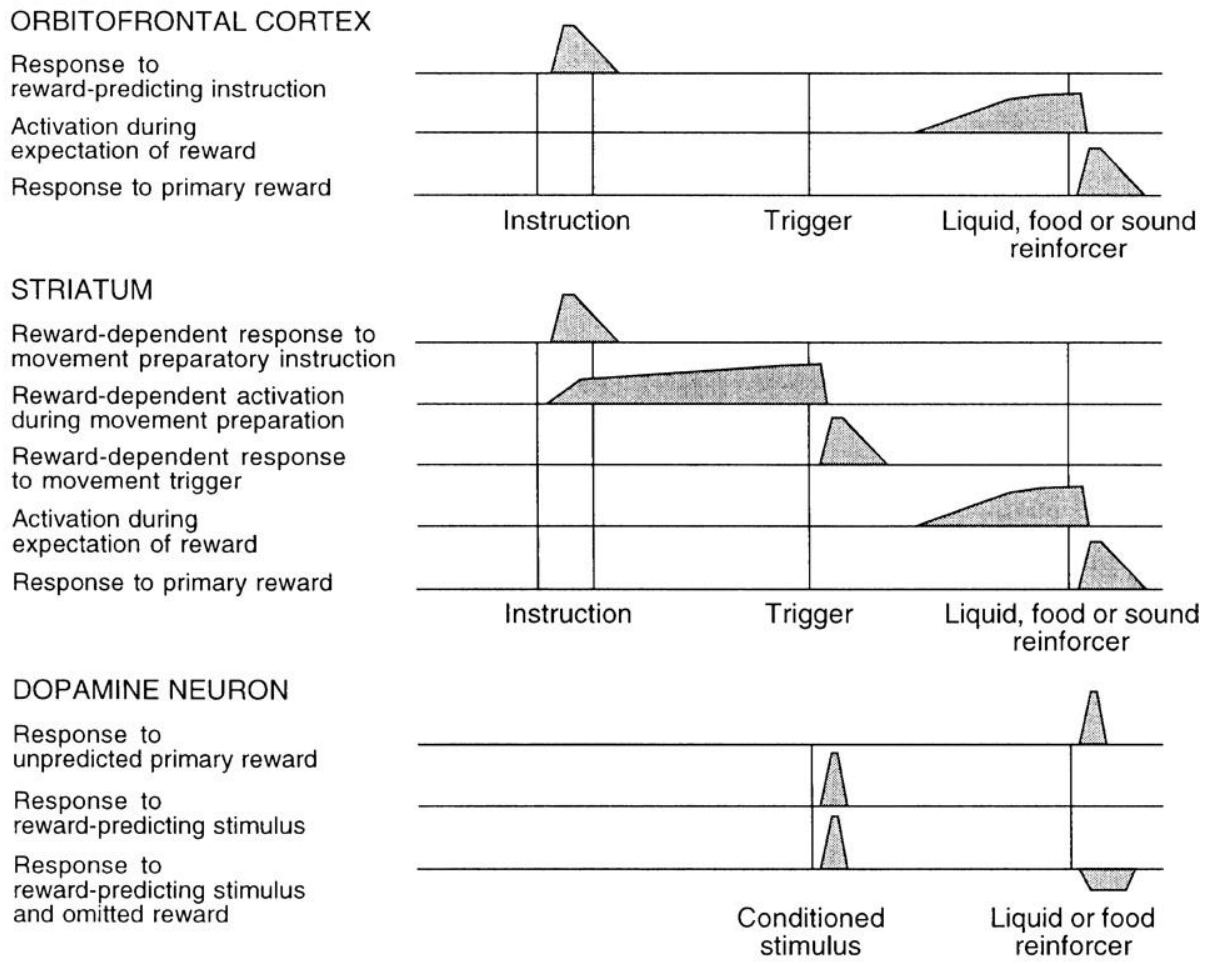


Figure 21 : Vue schématique du traitement des récompenses par les systèmes fronto-striataux et dopaminergiques. Principales réponses aux récompenses dans l'OFC (haut) et le striatum (milieu). Les neurones dopaminergiques sont activés par des récompenses inattendues et par les stimuli conditionnés prédisant des récompenses mais présentent une dépression dans leur fréquence de décharge en réponse à des omissions de récompenses attendues. Figure reproduite de (Tremblay & Schultz 2000).

D'autres études lésionnelles chez le singe confirment ce rôle du cortex orbitofrontal dans l'encodage des valeurs. Ces lésions pouvaient empêcher la mise à jour des représentations des valeurs des options associées à des récompenses (Buckley et al. 2009) ainsi que rendre plus difficile l'identification des stimuli associés à la plus grande récompense lors d'apprentissage par renversement (Walton et al. 2010).

Les paradigmes de conditionnement se sont révélés être des outils efficaces pour étudier les valeurs attendues ainsi que les valeurs obtenues : le stimulus conditionné n'a aucune valeur pour l'animal en début de conditionnement, mais prédit graduellement la récompense attendue, tout en restant non-récompensant en soi. Cela a permis de dissocier la valeur du renforcement de ses autres caractéristiques. De telles études sur le renforcement mirent en

évidence une grande variété d'activités neuronales encodant les différents aspects des évènements appétitifs, très bien décrits par Tremblay et Schultz (Tremblay & Schultz 2000). Dans l'OFC, ils distinguèrent trois types d'activité liés à la valeur : 1) des activations en réponse au stimulus prédisant la valeur du renforcement, encodant la valeur attendue ou la valeur de décision, 2) des activations pendant la période d'attente juste avant la délivrance de la récompense, encodant la valeur attendue, et enfin 3) des activations en réponse au renforcement, encodant la valeur obtenue (Figure 21 en haut). Il a également été montré que des neurones du striatum sont activés en lien avec la valeur attendue et obtenue (Figure 21 au milieu). D'autres neurones striataux ont une activité liée à la préparation, l'initiation et l'exécution de mouvements, dont beaucoup dépendent de la valeur du renforcement dont il est question.

Finalement, il a été mis en évidence que les neurones du vmPFC encodent la valeur attendue des stimuli annonçant une récompense au cours d'un protocole de conditionnement opérant (Bouret & Richmond 2010). L'enregistrement de l'activité unitaire de neurones de l'OFC et du vmPFC tout en manipulant des facteurs externes (quantité d'eau reçue) et internes (soif du singe) a montré une modification de la valeur des stimuli. Des neurones encodant la valeur attendue ont été identifiés dans les deux aires cérébrales. Cependant, alors que les neurones de l'OFC étaient plus sensibles à la magnitude des récompenses (modulations externes de la valeur), la réponse des neurones du vmPFC semblait être plus fortement modulée par la satiété (modulation interne de la valeur). Des enregistrements unitaires ont été réalisés chez le macaque afin de comparer l'activité unitaire des neurones de vmPFC et du cortex orbitofrontal avec les activations liées à l'encodage des valeurs subjectives retrouvées en IRMf chez l'humain (Abitbol et al. 2015). Ils mirent en évidence un encodage au sein du vmPFC de la magnitude de la récompense attendue, c'est-à-dire de la valeur subjective du stimulus associé à cette récompense.

Parce que le concept de valeur économique a été principalement défini comme déterminant les choix, il est important d'investiguer l'encodage de la valeur au cours de protocole de choix. L'une des études ayant eu le plus d'influence à propos de l'encodage des valeurs économiques chez le primate non-humain fut réalisée par Padoa-Schioppa et Assad (Padoa-Schioppa & Assad 2006). Des singes faisant face à des choix binaires entre différentes quantités de deux jus différents et devaient faire une saccade oculaire vers leur option préférée. Au niveau comportemental, les auteurs ont rapporté un biais entre le type de jus et la quantité de jus (Figure 22).

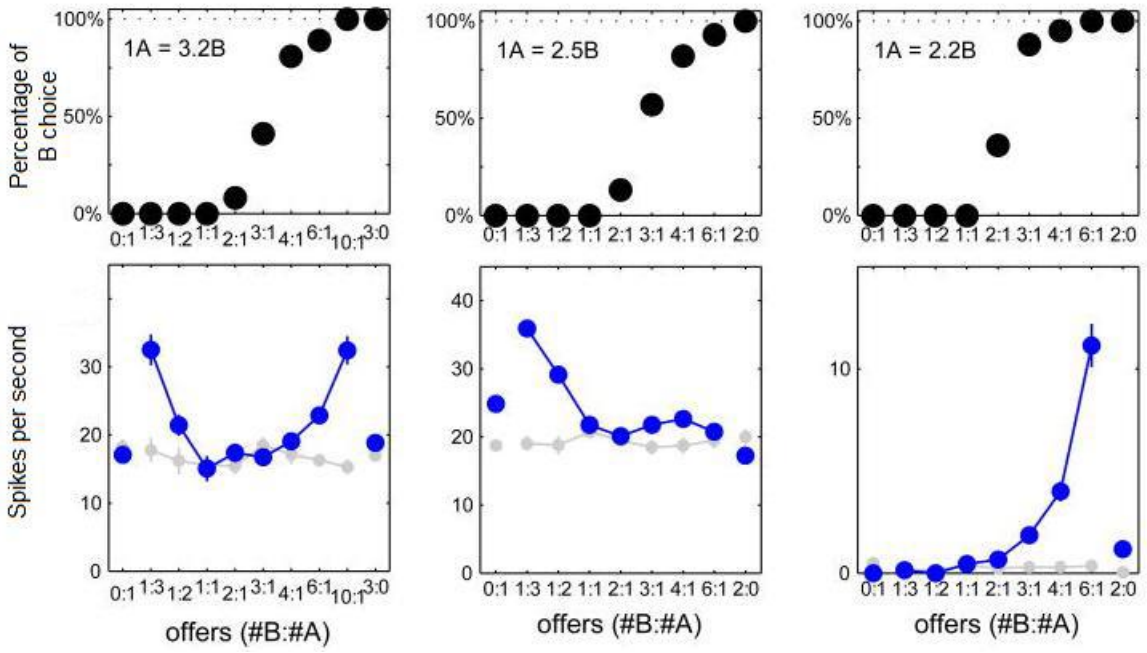


Figure 22 : Activations neuronales liées à la décision dans l'OFC de singes. Haut : profil de choix observés dans les sessions spécifiques. En bas : Activité neuronale. A gauche : Encodage de la valeur choisie. Milieu : Encodage de la valeur du jus A offert. A droite : Encodage de la valeur du jus B offert. Figure issue de (Padoa-Schioppa & Assad 2006).

Immédiatement après qu'une offre de jus ait été faite, les auteurs ont enregistré des neurones de l'OFC encodant la valeur de l'offre A ainsi que d'autres neurones encodant la valeur de l'offre B, qu'elle que soit la configuration spatiale des deux stimuli visuels ou la direction de la saccade oculaire à produire pour indiquer le choix. Quelques millisecondes plus tard, ils ont identifié des neurones encodant la valeur de l'option choisie, suivis par des neurones encodant l'identité du jus choisi. Ces résultats sont importants pour plusieurs raisons. Tout d'abord, ils suggèrent que les décisions pourraient être basées sur une étape d'évaluation des options (valeurs subjectives) distincte de l'étape de sélection de l'action, comme suggéré par le modèle de prise de décision basé sur la valeur de type Q-learning (I.A). Deuxièmes, les valeurs choisies sont particulièrement intéressantes parce qu'en plus d'être indépendantes des contingences visuo-motrices du protocole expérimental, elles sont indépendantes des caractéristiques du renforcement (type et quantité du jus), et sont donc peut-être responsables de l'encodage non-spécifique des valeurs économiques. Leur activité est fondamentalement dissociable de la sélection de l'action, et reflète probablement la réelle valeur motivationnelle. Ces résultats furent ultérieurement répliqués dans les études complémentaires (Padoa-Schioppa& Assad 2008; Padoa-Schioppa 2009) par cette même équipe, dont une revue exhaustive et une discussion du travail a été faite (Padoa-Schioppa 2007; Padoa-Schioppa & Cai 2011). Certains de leurs travaux sur des enregistrements unitaires chez le singe (Cai & Padoa-Schioppa 2014) suggèrent l'existence de populations

fonctionnellement ségréguées de neurones du cortex orbitofrontal permettant l'intégration de ces informations liées à la valeur subjective des stimuli.

Au niveau du striatum, trois régions ont été identifiées depuis longtemps : le striatum ventral, dorsomédian et dorsolatéral. Alors qu'il était jusqu'à présent supposé que ces trois régions remplissent chacune un rôle distinct, des travaux récents montrent qu'elles travaillent en réalité de manière hiérarchique et coordonnée (Ito & Doya 2015a; Ito & Doya 2015b). Des enregistrements unitaires ont été effectués au sein de ces trois zones pendant un protocole de choix binaire entre deux orifices permettant d'obtenir une récompense sucrée. Au cours des essais, le taux de probabilité de récompense pour un orifice donné évolue, forçant le rat à adapter son comportement afin de toujours choisir celui associé à la plus grande probabilité de récompense.

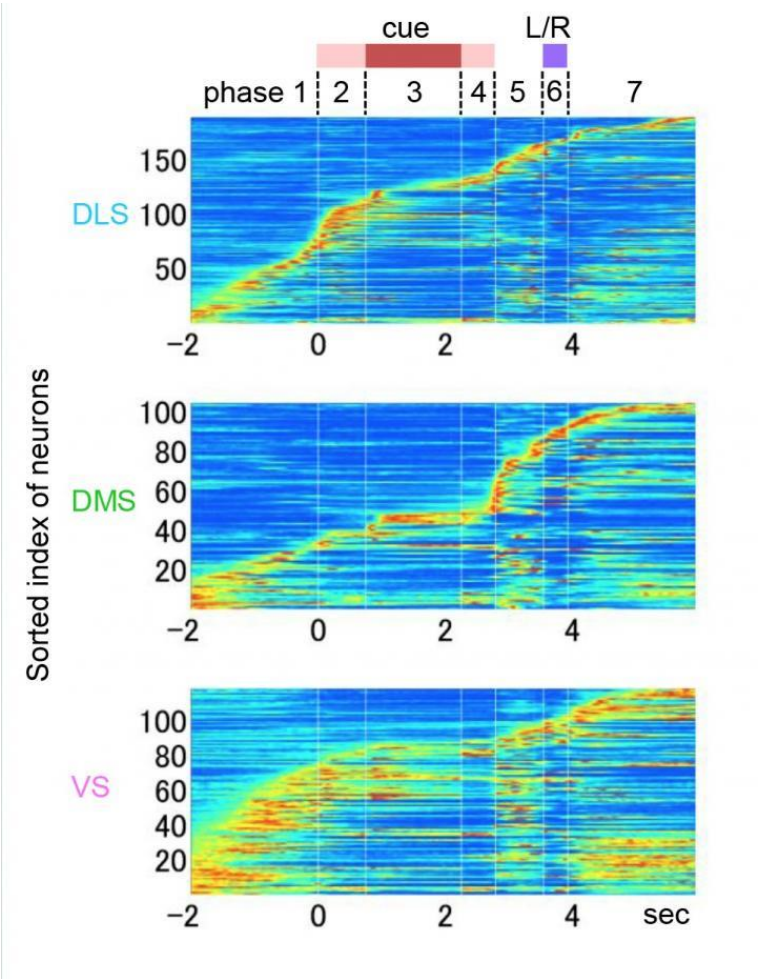


Figure 23: Organisation hiérarchique du processus décisionnel lors de la recherche de récompense au sein du striatum.

L'axe des ordonnées représente le nombre de neurones et l'activité de chacun est indiquée par les couleurs rouge et jaune. Figure issue de (Ito & Doya 2015a).

Les chercheurs ont ainsi constaté que les trois régions du striatum travaillent ensemble dans les différentes phases du processus de décision : le striatum ventral (VS) était le plus actif au début, lorsque le rat a dû choisir s'il participerait à l'activité proposée (ou non). Le striatum dorso-médian (DMS) s'est occupé de la décision suivante, à savoir s'il fallait se diriger vers

l'orifice situé à gauche ou celui de droite. Le striatum dorso-latéral (DLS) est plusieurs fois intervenu au cours de l'ensemble de l'exercice (Figure 23). Ainsi, le striatum ventral (VS) déciderait d'effectuer une nouvelle tâche ou non, le striatum dorsomédian (DMS) évaluerait l'orientation à prendre et le striatum dorsolatéral (DLS) réaliserait la tâche. Cette découverte contredit l'idée communément admise selon laquelle les décisions habituelles et les décisions nouvelles seraient gérées par des zones différentes : le DLS pour les habituelles et le DLM pour les nouvelles. De manière étonnante, les chercheurs n'ont constaté aucune différence significative d'activité au sein du DLS et du DLM pendant les choix habituels et les choix nouveaux d'actions effectuées pour obtenir les récompenses, ce qui suggère fortement que les rats analysaient à chaque essai les avantages potentiels de choisir l'orifice de gauche comme l'orifice de droite pour être récompensés.

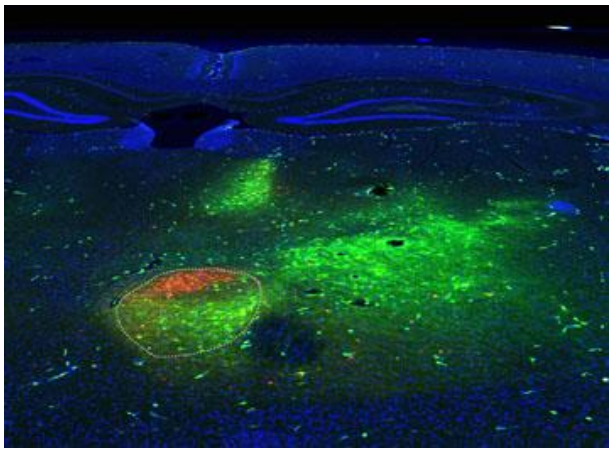


Figure 24 : Localisation du noyau submédian du thalamus chez le rat. Pour visualiser les voies nerveuses, deux marqueurs (rouge et vert) sont appliqués dans deux régions du cortex orbitofrontal. Ces molécules migrent ensuite pour s'accumuler dans les neurones thalamiques. Un marquage dense des deux traceurs est visible au niveau du thalamus submédian (délimité par des pointillés). Figure issue de (Alcaraz et al. 2015).

Bien que le cortex et le système dopaminergique jouent un rôle essentiel pour prendre des décisions adaptées, d'autres régions cérébrales sous-corticales sont également impliquées dans la prise de décision. Une autre zone du cerveau a récemment été mise en évidence pour son implication capitale dans la prise de décision chez le rat : le noyau sous-médian du thalamus ou NSMT (Alcaraz et al. 2015). Par une technique de marquage (Figure 24), les scientifiques ont mis en évidence une zone cérébrale fortement connectée au cortex préfrontal et au rôle jusqu'à présent inconnu : le thalamus submédian. Ils ont testé le rôle de cette structure et du cortex préfrontal dans la prise de décision et l'adaptation à l'environnement chez trois groupes de rats : le premier présentant des lésions du cortex préfrontal, le deuxième au niveau du thalamus submédian, et le troisième regroupant des rats témoins sans lésion. L'expérience, en deux étapes, consistait à tester leur capacité à établir un lien entre un son et l'obtention d'une récompense alimentaire. D'abord, la phase d'apprentissage, qui a permis aux animaux d'apprendre à identifier deux sons différents, annonçant chacun la survenue d'une récompense alimentaire spécifique. Les trois groupes d'animaux visitent donc la mangeoire dès qu'un signal auditif est perçu. Les lésions

n'empêchent pas les animaux d'apprendre qu'un stimulus auditif prédit l'obtention de la récompense. Lors de la deuxième étape (la phase de "dégradation"), la procédure reste inchangée pour le premier son, mais pour le deuxième, les chercheurs ont distribué des récompenses alimentaires durant et surtout en dehors des périodes sonores. Ce son perd donc sa valeur prédictive et un animal sans lésion en vient à négliger ce deuxième stimulus auditif pour ne venir à la mangeoire que lorsqu'il entend le son 1. En revanche, les animaux présentant une lésion que ce soit au niveau du cortex préfrontal ou du thalamus submédian se sont montrés incapables de faire une telle distinction, et donc, de s'adapter. Cette expérience confirme ainsi l'existence d'un circuit entre le thalamus et le cortex préfrontal primordial dans la prise de décision adaptée à l'environnement. Dans cette zone, dite "thalamocorticale", d'autres circuits neuronaux pourraient eux aussi être impliqués dans la prise de décision mais aussi *dans la survenue de pathologies, comme la schizophrénie ou l'addiction.*

HUMAIN

De nombreux déficits ont été rapportés chez les patients présentant des lésions du cortex orbitofrontal, et en particulier du vmPFC : ils ont des difficultés à adapter leurs choix aux changements de valeur des renforcements, et échouent à mettre à jour la valeur des renforcements lorsqu'elle change comme au cours de protocoles dits d'inversion d'apprentissage (reversal learning) (Fellows & Farah 2003). Plus récemment, il a été montré que les patients ayant des lésions au lobe frontal ventro-médian étaient plus susceptibles de faire des choix irrationnels ne suivant pas leurs préférences ou leurs objectifs (Camille et al. 2011). Ces résultats indiquent qu'un circuit comprenant l'OFC et le vmPFC est essentiel pour calculer des valeurs cohérentes de valeurs attendue et obtenues, faisant de ces aires cérébrales un probable lieu majeur dans la formation des préférences.

Des études chez l'humain ont souvent rapporté un encodage des valeurs au niveau du vmPFC. Similairement aux études électrophysiologiques chez le singes, les études en neuroimagerie chez l'Homme ont tout d'abord rapporté un encodage des erreurs de prédiction des récompenses, à la fois dans le striatum et dans le vmPFC au cours de protocoles de conditionnement classique (Berns et al. 2001) et de conditionnement pavlovien appétitif (Peterson 2005; Gottfried et al. 2003; Knutson & Greer 2008). Cet encodage des erreurs de prédiction des récompenses a également été rapporté au niveau du mésencéphale (D'Ardenne et al. 2008).

Tout comme chez le singe, la simulation gustative a été utilisée chez l'humain pour étudier la valeur d'un renforcement obtenu puisqu'elle permet de dissocier les caractéristiques objectives d'un renforcement de son évaluation subjective à force que la satiété progresse et que le renforcement auparavant fortement récompensant est dévalué de plus en plus. L'activité du vmPFC observée est ainsi liée au caractère plaisant, appétitif du renforcement, et diminue spécifiquement avec la satiété progressive envers un jus donné (O'Doherty et al. 2000; Kringelbach et al. 2003). Ces résultats ont été répliqués en utilisant la tomographie par émission de positons (TEP) (Small et al. 2001). En remplaçant la nourriture par des récompenses liquides (de l'eau) chez des sujets assoiffés, les mêmes effets d'appétence et de satiété ont été observés dans le vmPFC (de Araujo et al. 2003).

Au contraire des singes, les êtres humains sont habitués à manipuler des récompenses secondaires et des concepts tels que l'argent. Les réponses cérébrales à l'argent furent d'abord étudiées grâce à des protocoles très proches de ceux utilisés avec les singes (Thut et al. 1997; O'Doherty et al. 2001) et ont depuis été étudiées avec de nombreux paradigmes (Elliott et al. 2000; Knutson et al. 2000; Breiter et al. 2001; Knutson et al. 2001). Les expériences ont eu des résultats très similaires à ceux des études de stimulation sensorielle, entrainant (entre-autres) des activations du vmPFC ainsi que du mésencéphale et du thalamus. A noter qu'un gradient a été observé entre l'encodage des renforcements concrets (nourriture) et plus abstraits (argent) au sein du vmPFC avec un niveau d'abstraction grandissant dans les aires cérébrales les plus antérieures du vmPFC, au contraire de la partie plus postérieure du vmPFC encodant plus les renforcements concrets (Kringelbach & Rolls 2004).

Une étude en IRMf s'est intéressée aux réponses élicitées par des renforcements appétitifs de deux types : une récompense monétaire et un encouragement verbal (Kirsch et al. 2003). Les résultats ont révélé une activation significative de la substance noire, du thalamus, du striatum et du cortex orbitofrontal, ainsi qu'au niveau de l'insula et du cortex cingulaire antérieur pendant la présentation de stimuli conditionnés prédisant la délivrance d'un renforcement appétitif. L'anticipation de la délivrance d'une récompense est le signe d'un encodage des valeurs subjectives appétitives au sein de ces régions. De manière intéressante, l'anticipation d'une délivrance monétaire a produit une activation de ces régions plus forte que l'anticipation d'un encouragement verbal, justifiant l'utilisation d'argent (virtuel ou non) dans les études sur l'apprentissage par renforcement.

Pour synthétiser les résultats de plus de 200 expériences en IRMf utilisant des choix ou des notations pour étudier la valeur, Bartra et ses collaborateurs (Bartra et al. 2013) ont identifié un premier ensemble de régions dont l'activité neurale est modulée de façon cohérente avec

la valeur de décision, la valeur attendue et la valeur obtenue de renforcement (Figure 25), indépendamment de la valence. Ces régions comprennent l'insula antérieure, le dlPFC, le striatum dorsal postérieur et le thalamus. Cet effet ambivalent sous forme d'effets positifs et négatifs de la valeur subjective sur le signal BOLD peut être le signe d'un encodage de la saillance. Dans un second groupe de régions, on observe une activation spécifique lors de l'encodage des valeurs subjectives. Ces régions activées incluent le vmPFC et le striatum ventral.

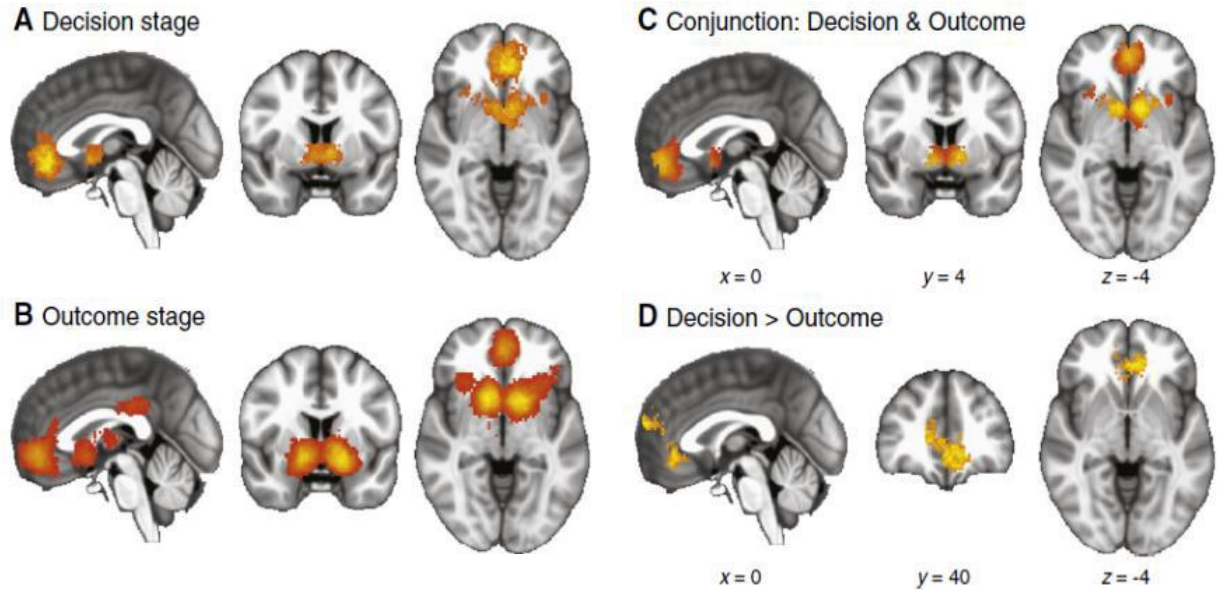


Figure 25 : Les différents types de valeurs subjectives encodés dans le cerveau humain. Les cartes indiquent la significativité statistique de la corrélation entre le signal BOLD et les valeurs subjectives. (A) Etapes 2 ou 3 du modèle syncrétique utilisé. (B) Etape 4 du modèle syncrétique. (C) Carte de conjonction montrant la corrélation positive pendant la prise de décision et la délivrance du renforcement. (D) Régions présentant significativement plus d'effets positifs pendant la prise de décision que pendant la délivrance du renforcement. Figure issue de Bartra et al. 2013).

(3) Réponses aux erreurs de prédiction des récompenses

RONGEUR ET PRIMATE NON-HUMAIN

Initialement, il était supposé que la réponse des neurones dopaminergiques aux renforcements est assez homogène et encode ce que l'on appelle les erreurs de prédictions des récompenses (Figure 21 en bas) : les neurones sont activés quand un stimulus

conditionné annonce la délivrance d'un renforcement appétitif dans un futur proche (Figure 26 en haut), ainsi que lors de la délivrance d'une récompense inattendue (Figure 26 au milieu). Néanmoins, il n'y a pas de modification de l'activité de ces neurones dopaminergiques lorsqu'une récompense entièrement prédite est délivrée (Figure 26 en bas).

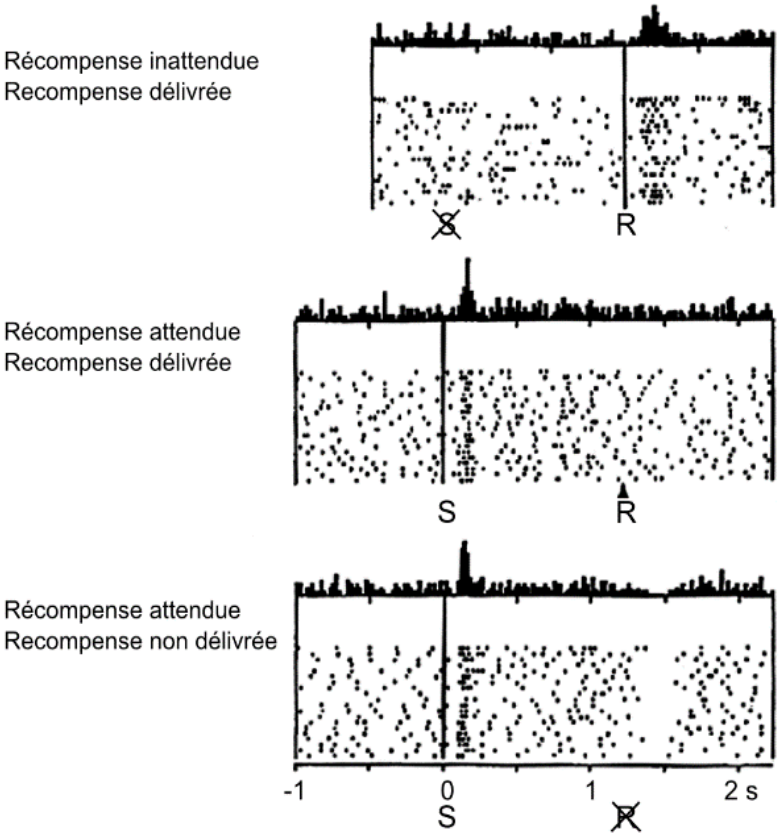


Figure 26 : Les neurones dopaminergiques du mésencéphale encodent les erreurs de prédiction des récompenses. Haut : réponses à une récompense inattendue. Milieu : réponse à une récompense attendue et au stimulus la prédisant. Bas : réponses à l'omission d'une récompense, pourtant prédite par un stimulus, et à ce stimulus. Figure adaptée de Schultz et al 1997.

Une autre spécificité des neurones dopaminergiques est que les mêmes neurones dopaminergiques présentent les trois types de réponses selon si la récompense est délivrée ou omise, ce qui veut dire que ces neurones dopaminergiques du mésencéphale encodent les erreurs de prédiction des récompenses (Figure 21 et Figure 26). Au contraire de l'OFC et du striatum qui possèdent différentes populations de neurones pour différentes réponses à la valeur (récompense ou neutre). Ce profil spécifique d'activité a mené les auteurs à proposer un rôle des neurones dopaminergiques du mésencéphale dans l'encodage des erreurs de prédiction des récompenses, c'est-à-dire la différence entre la récompense obtenue et celle attendue. De plus, la réponse à une récompense attendue s'est décalée temporellement de la délivrance de la récompense (R) à la présentation du stimulus conditionné la précisant (CS) au cours de l'apprentissage.

L'étude des propriétés des neurones dopaminergiques mésencéphaliques a commencé avec les mouvements auto-initiés chez des singes libres de se déplacer. Les premières études ont rapporté que les neurones dopaminergiques répondaient quand les animaux touchaient un morceau de nourriture caché pendant des mouvements exploratoires, et que cette activité disparaissait lorsqu'une récompense attendue était omise (Romo & Schultz 1990). Dans les protocoles de conditionnement et d'apprentissage, les neurones dopaminergiques sont activés de manière phasique par la délivrance d'une récompense pendant l'apprentissage, mais cette activité disparait après l'apprentissage quand la récompense est totalement prévisible ((Schultz et al. 1993; Schultz, Apicella & Ljungberg 1993) et Figure 26). Similairement, l'activité des neurones dopaminergiques répond à la délivrance d'une récompense liquide quand les singes devaient réaliser une tâche pour l'obtenir, mais pas quand la récompense était juste délivrée à des intervalles réguliers sans qu'ils n'aient à faire quoi que ce soit (Ljungberg et al. 1992). Des études extensives de ce phénomène ont démontré que l'imprédictibilité d'une récompense était importante pour l'encodage des valeurs obtenues par les neurones dopaminergiques mésencéphaliques (Mirenowicz & Schultz 1994; Ljungberg et al. 1992; Romo & Schultz 1990). De manière importante, ces études montrant des activations en réponses à la délivrance inattendue d'une récompense, ont également rapporté des activations liées aux récompenses au moment des stimuli les prédisant, suite à un conditionnement (Mirenowicz & Schultz 1994; Ljungberg et al. 1992; Schultz et al. 1993), suggérant un encodage des valeurs attendues par les neurones dopaminergiques mésencéphaliques (Figure 26).

Cette observation, en parallèle des résultats indiquant que les neurones dopaminergiques mésencéphaliques incorporent un timing précis de la délivrance de la récompense lorsque les délais entre R et CS variaient (Hollerman & Schultz 1998), apporte un argument supplémentaire en faveur de l'hypothèse selon laquelle ces neurones encodent l'erreur de prédiction théorique, correspondant à celle utilisée dans les modèles computationnels formels d'apprentissage par renforcement (voir I.A). Le rôle des neurones dopaminergiques mésencéphaliques dans l'encodage des valeurs implique de nombreuses facettes des stimuli à encoder (Schultz & Dickinson 2000; Schultz et al. 2011). En effet, des travaux suggèrent un encodage plus complexe des erreurs de prédictions au niveau des neurones dopaminergiques (Bromberg-Martin et al. 2010; Morrens 2014), certains encodant la magnitude de la récompense quand d'autres signalent plutôt la saillance de l'évènement et les erreurs de prédiction positives des récompenses. Des enregistrements unitaires chez le singe au cours de l'apprentissage par récompense ont mis en évidence un encodage quantitatif du signal d'erreur de prédiction des récompenses, mais uniquement lorsque que ces erreurs de prédiction des récompenses étaient positives (Bayer & Glimcher 2005).

Une autre lignée de travaux investiguant la valeur grâce à un protocole de choix a aidé à affiner le rôle des neurones de l'OFC dans l'encodage des valeurs de décision (Kennerley et al. 2011; Wallis & Kennerley 2011). A l'aide d'un paradigme de choix binaires, il a été rapporté que beaucoup de neurones de l'OFC et de l'ACC encode les valeurs choisies, mais que ces deux régions utilisent des méthodes d'encodage opposées : les neurones de l'OFC, mais pas ceux de l'ACC, encodent la valeur choisie d'après l'historique récent des valeurs de choix ; au contraire, une population spécifique de neurones de l'ACC, mais pas de l'OFC, encodent les erreurs de prédiction des récompenses. Cela suggère l'existence de processus complémentaires d'évaluation des options au sein de ces deux régions, avec les neurones de l'OFC évaluant dynamiquement les choix en cours en fonction des valeurs des choix récemment effectués, et les neurones de l'ACC encodant les prédictions de choix et les erreurs de prédiction qui reflètent l'intégration de paramètres de décision multiples. Des enregistrements unitaires chez le singe ont montré que le cortex cingulaire antérieur est à la fois capable de signaler les récompenses réelles mais aussi les récompenses fictives, ainsi que leurs erreurs de prédiction (Hayden et al. 2009).

Plusieurs études ont rapporté un rôle du thalamus, et en particulier du noyau antérieur du thalamus, au cours de l'apprentissage par récompense. Des lésions du NAT chez des rats induit un déficit d'apprentissage par récompense (Wright et al. 2015), de même que des lésions similaires chez le lapin (Smith et al. 2002). L'étude de Smith un rôle du NAT dans L'identification des stimuli corrects et incorrects, affectant à la fois les phénomènes d'approche et d'évitement au cours du conditionnement instrumental. Pour finir, une étude chez le macaque a mis en évidence le rôle critique du noyau dorsomédian du thalamus au cours de choix dirigés vers les récompenses (Chakraborty et al. 2016). Des lésions excitotoxiques du NDMT ont induit un déficit sévère dans la mise à jour des choix suite à l'inversement des probabilités des récompenses, où face à plusieurs choix associés à des récompenses. Ces singes lésés étaient dans l'incapacité de promouvoir la répétition de choix corrects, suggérant un rôle du NDMT dans la prise en compte de récompenses et dans la mise à jour des valeurs subjectives des stimuli pendant l'apprentissage par récompenses. Par ailleurs, Parnaudeau et ses collaborateurs ont également mis en évidence le rôle du NDMT dans l'apprentissage par récompenses chez la souris (Parnaudeau et al. 2015). Ils ont utilisé une approche pharmacologique afin d'induire une hypofonction du NDMT, qui a causé un déficit d'apprentissage par renversement. L'hypoactivité du NDMT a diminué la capacité des souris à s'adapter aux changements de contingences entre actions et renforcements, et donc une incapacité à adapter le comportement suite aux récompenses.

HUMAIN

Des travaux ayant mis en perspective des études de neuroimagerie chez l'humain grâce à l'utilisation de la technique d'EEG a été réalisée par Krigolson et ses collaborateurs en 2014. Ils ont ainsi identifié une réponse aux erreurs sous la forme d'un potentiel lié à un évènement (ERP). Ils rapportent des informations sur la dynamique de la réponse cérébrale aux récompenses. Le signal encodant l'erreur de prédiction d'une récompense apparait rapidement lors de la délivrance de la récompense, et disparait au cours de l'apprentissage lors d'une tâche d'apprentissage par récompense. Que ce signal diminue avec l'apprentissage, c'est-à-dire quand les récompenses sont de plus prévisibles, est un indice fort qu'il s'agit d'un signal d'erreur de prédiction des récompenses (Krigolson et al. 2014). Une autre étude en EEG a identifié un encodage des erreurs de prédictions des récompenses sous la forme d'une onde négative suite à celles-ci. La reconstruction de source indique une origine au sein du cortex frontal médian (Chase et al. 2011). Dans leur étude en EEG sur l'apprentissage par récompenses de 2010, Philiastides et ses collaborateurs ont trouvé la même onde 300ms (P300) en réponse aux récompenses inattendues qu'Ullsperger et Fischer quelques années plus tard, avec une localisation centrale au sommet du scalp (Philiastides et al. 2010; Fischer & Ullsperger 2013).

Une revue des travaux de neuroimagerie sur l'apprentissage par renforcement a été réalisée par Garrison et ses collaborateurs (Garrison et al. 2013). Les résultats proviennent à la fois d'étude sur le conditionnement pavlovien que sur le conditionnement instrumental, c'est-à-dire l'apprentissage par renforcement. Ainsi, le cortex préfrontal médian et le cortex cingulaire antérieur ont un rôle dans la détection des erreurs de prédiction. Plus précisément, concernant les erreurs de prédiction des récompenses au cours de l'apprentissage par récompenses, les chercheurs mettent en évidence un rôle clé du striatum.

Dans le cadre d'une étude pharmacologique chez des sujets sains, Jocham et ses collaborateurs ont étudié les effets d'un agoniste des récepteurs dopaminergiques D2 sur l'apprentissage par récompense, en IRMf (Jocham et al. 2014). Bien que cet agoniste D2 n'influence pas le comportement lors de la phase initiale d'apprentissage par renforcement, les sujets traités étaient beaucoup plus efficaces lors de la phase suivante quand il s'agissait de faire des choix corrects afin d'accumuler des récompenses. Ils ont identifié un rôle du striatum dans la première phase d'apprentissage par renforcement, et celui du vmPFC lors de la seconde phase dite de transfert. Bien qu'il n'y ait plus de récompenses délivrées lors de la phase de transfert, le rôle du vmPFC est celui de l'encodage des valeurs subjectives, déjà détaillé précédemment (I.B.2.a)(1)).

b) Apprentissage par évitement de punition

(1) Réponses aux stimuli aversifs

RONGEUR ET PRIMATE NON-HUMAIN

Au niveau sous-cortical, de rares travaux suggèrent un rôle du noyau antérieur (NAT) et du noyau dorsomédian (NDMT) du thalamus dans l'apprentissage par évitement des punitions (Corbit et al. 2003; Gabriel et al. 1977; Sparenborg & Gabriel 1992; Freeman et al. 1996; Smith et al. 2002; Bradfield et al. 2013). Chez des lapins, différentes lésions ont été effectuées au niveau du thalamus touchant 1) uniquement le noyau dorsomédian du thalamus, ainsi que 2) des lésions partielles du NDMT ou du NAT et 3) des lésions combinées totales du NDMT et de NAT. Par rapport aux animaux contrôles, les lapins présentant des lésions simples (1 ou 2) mettaient plus de temps pour apprendre à éviter les punitions (+50%), et les animaux présentant des lésions combinées (type 3) triplaient le temps nécessaire pour apprendre à éviter les punitions en comparaison des animaux contrôles sans lésion. Cela prouve une implication fonctionnelle du NAT et du NDMT dans l'apprentissage par renforcement, et en particulier dans l'évitement des punitions. Cependant, les mécanismes neuronaux impliqués ne sont pas connus à ce jour.

L'insula est une structure cérébrale affective impliquée dans la conscience et les états affectifs (Blackford et al. 2010; Wu et al. 2014; Brass & Haggard 2010; Kuhnen & Knutson 2005), dans l'intégration de processus cognitifs et d'émotions (Craig 2003; Harlé et al. 2012; Naqvi & Bechara 2010) et semble avoir un rôle important dans les mécanismes d'attribution de valeur à des options dans le cadre de la prise de décision (Harlé et al. 2012; Paulus et al. 2003; Naqvi & Bechara 2010; Palminteri et al. 2012).

Le cortex insulaire antérieur reçoit des afférences dopaminergiques en provenance de l'aire tegmentale ventrale et il est bien documenté qu'il possède des récepteurs dopaminergiques (Gaspar et al. 1995; Richfield et al. 1989; Richfield et al. 1989). Des études chez le singe rhésus ont montré que ces neurones dopaminergiques mésolimbiques du mésencéphale encodent la valeur relative des récompenses (Tobler et al. 2005) et mesurent la probabilité de récompense (Fiorillo et al. 2003). Des études anatomiques, toujours chez le macaque rhésus, ont révélé l'existence de projections efférentes partant de l'insula antérieure vers le striatum ventral (Chikama et al. 1997; Fudge et al. 2005) où des sous-régions du striatum ventral ont été identifiées comme encodant la valeur et la probabilité des récompenses (Yacubian 2006).

L'insula antérieure apparaît aussi comme signalant les représentations indésirables dans notre environnement, comme l'anticipation du risque et le dégoût (Calder et al. 2000; Naqvi & Bechara 2009; Preuschoff et al. 2008). La micro stimulation de cette région cérébrale, de façon à en augmenter l'activité neurale, a produit des réactions de dégoût chez des macaques rhésus vis-à-vis de nourriture auparavant appétitive (Caruana et al. 2011) et Figure 27). De manière intéressante, cette étude chez le singe a mis en évidence des réponses de l'insula à des évènements aversifs de plusieurs natures : le dégoût envers un aliment ou un mal-être suite à une privation de contact social. Cela indique que l'insula

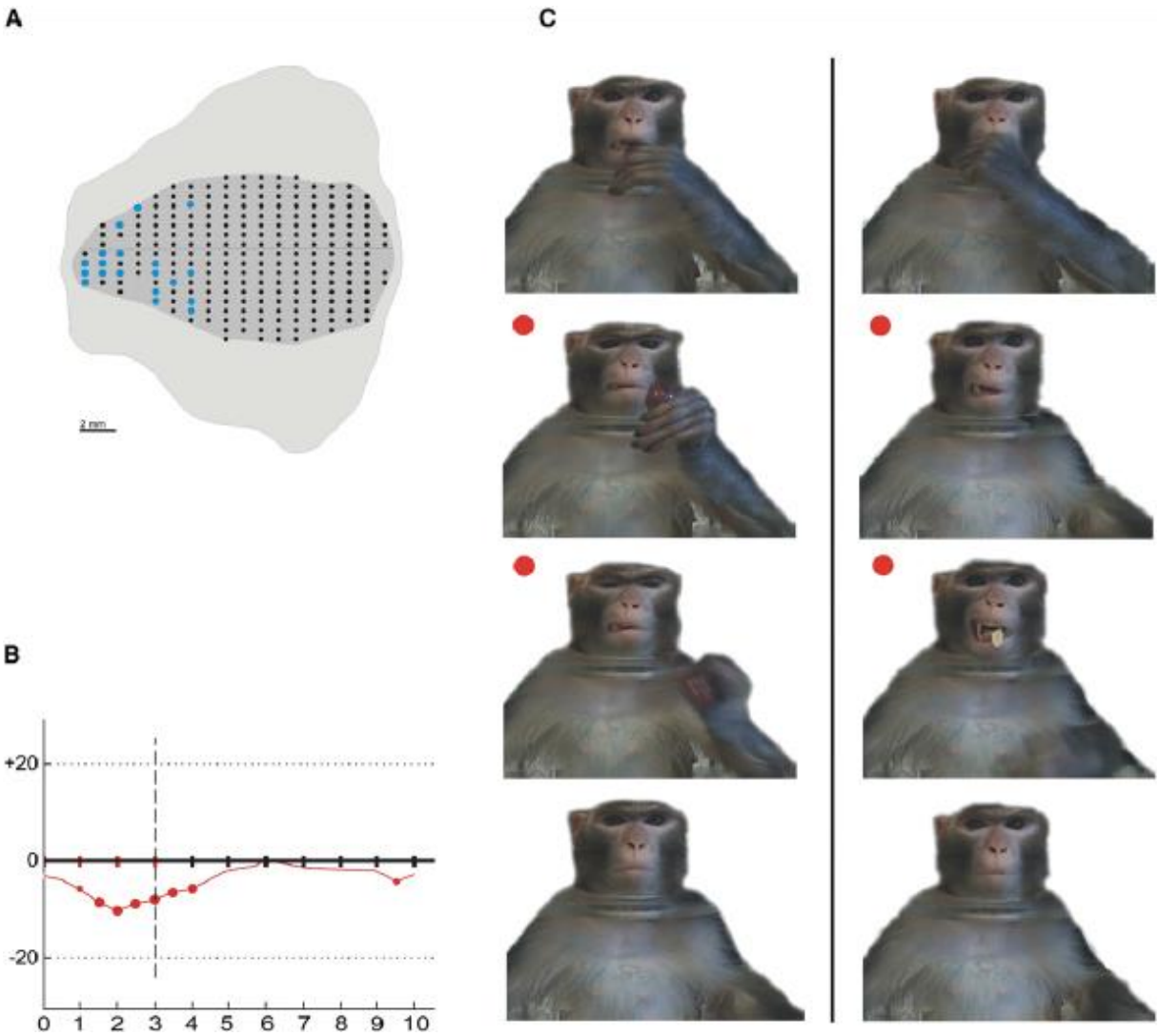


Figure 27 : Induction d'une sensation de dégoût chez un singe rhésus suite à la microstimulation électrique directe de l'insula antérieure.
(A) Les sites de stimulations utilisés (bleu) sont indiqués sur une représentation dépliée de l'insula (en gris). **(B)** Fréquence cardiaque instantanée moyenne suite aux stimulations (ligne pointillée) au cours du temps (abcisse). **(C)** Comportements apparentés à du dégoût observés. Si le singe avait de la nourriture dans la main (à gauche) lorsque la stimulation (points rouges) commençait, l'animal jettait la nourriture. S'il la stimulation commençait alors que le singe avait déjà mis la nourriture dans la bouche, il la crachait immédiatement. Figure adaptée de (Caruana et al. 2011).

encode différents types d'aversivité des stimuli.

Au niveau sous-cortical, l'amygdale, connue pour son rôle dans la détection d'évènements saillants et ses connexions au sein du circuit de Papez, possède deux populations distinctes de neurones au niveau du complexe amygdalien basolatéral : 1) des neurones se projetant sur le noyau accumbens, impliqués dans l'apprentissage par récompense, et 2) des neurones se projetant sur l'amygdale contramédiale impliqués dans l'apprentissage par évitement des punitions. Ces résultats (Namburi et al. 2015) permettent une explication mécanistique de la représentation d'associations appétitives et aversives au sein de l'amygdale. Les travaux de Namburi sont à mettre en lien avec ceux de Esber (Esber & Holland 2014) qui a mis en évidence causalement le rôle des neurones de l'amygdale basolatérale dans l'encodage des erreurs de prédiction au cours de l'apprentissage par associations. Des lésions localisées au niveau de la partie basolatérale de l'amygdale de rats a induit un déficit dans le traitement des stimuli suite à la délivrance d'un renforcement, tout en n'affectant pas l'apprentissage. En temps normal, la délivrance d'un renforcement induit une erreur de prédiction qui va permettre le traitement du stimulus choisi pour adapter la stratégie de choix. Cela indique que les neurones de l'amygdale basolatérale encodent les erreurs de prédiction positives (renforcement obtenu plus intéressant que celui attendu). Fait intéressant, ces mêmes neurones semblent également encoder les erreurs de prédiction négatives (renforcement obtenu moins intéressant que celui attendu), suggérant la possibilité que l'amygdale basolatérale puisse utiliser ces erreurs de prédiction pour améliorer le processus décisionnel.

HUMAIN

Pour rappel (voir I.B.2.a)(1)), Hayes et ses collaborateurs ont rapporté des activations au sein de régions cérébrales sélectivement suite aux punitions dans le cortex cingulaire et le noyau périaqueducal gris. Les activations de ces régions étaient indépendantes des récompenses, suggérant un système distinct traitant les punitions au sein du cerveau humain. Des résultats complémentaires indiquant des réponses fonctionnellement ségréguées aux récompenses et aux punitions dans l'insula antérieure et l'amygdale suggèrent l'existence de sous-régions pour le traitement des renforcements (Hayes et al. 2014). L'hypothèse avancée par ces chercheurs est celle d'une dynamique spatio-temporelle différente entre récompenses et punitions, pouvant expliquer à la fois les similarités et les différences d'activités liées aux stimuli appétitifs et aversifs. Une étude en EEG de surface a mis en évidence un aspect dynamique clé de la détection des renforcements aversifs pour

l'adaptation du comportement. Blank et ses collaborateurs (Blank et al. 2013) ont identifié une réponse aux punitions, au niveau central sur le scalp, ayant lieu pendant et après la délivrance de la punition. La durée de cette réponse au-delà de la délivrance simple de la punition suggère un rôle dans des processus décisionnels d'adaptation du comportement suite à la punition. Cette adaptation du comportement est d'autant plus importante suite à une punition qu'à une récompense, lorsqu'il s'agit d'éviter les punitions et d'accumuler les récompenses. Ces activations au niveau de l'insula antérieure et du cortex cingulaire antérieur suite à des stimuli aversifs avaient déjà été mises en évidence en IRMf au cours d'un protocole de conditionnement (Büchel et al. 1998). Comme chez l'animal, l'insula antérieure humaine semble répondre à une large variété de stimuli aversifs : un mauvais goût (Nitschke et al. 2006), la douleur (Craig 2003), un effort (Meyniel et al. 2013), une perte d'argent (Pessiglione et al. 2006).

Grâce à une étude en neuroimagerie fonctionnelle, O'Doherty et ses collaborateurs ont montré que des régions distinctes du cortex orbitofrontal sont activées par les récompenses et les punitions monétaires (O'Doherty et al. 2001), et que ces activations reflètent la magnitude des renforcements reçus. Des travaux en neuroimagerie fonctionnelle ultérieurs se sont donc intéressés à l'existence de régions cérébrales dont l'activité, suite aux erreurs de prédiction des renforcements, dépendrait de la nature et de la valence de ces renforcements. C'est ainsi qu'il a été montré, au sein du cortex orbitofrontal (Kringelbach & Rolls 2004), qu'il existe un double gradient fonctionnel dans le traitement des erreurs de prédiction des renforcement. La valence des renforcements serait traitée selon un gradient médiolatéral avec les récompenses au niveau du vmPFC et les punitions dans le latOFC. La nature des renforcements serait traitée selon un gradient antéro-postérieur avec les renforcements plus complexes et plus abstraits (comme les gains et pertes d'argent) représentés plus antérieurement dans le cortex orbitofrontal que les renforcements plus simples comme le goût et la douleur.

Au niveau sous-cortical, l'implication des neurones dopaminergiques de l'habénula et du mésencéphale dans le traitement des évènements aversifs a été montré en imagerie fonctionnelle chez l'humain (Hennigan et al. 2015). Une activation robuste de l'habénula et de deux régions du mésencéphale (l'ATV et la SN) a été observée en réponse aux stimuli aversifs. Le traitement de ces stimuli aversifs induit une augmentation de la connectivité fonctionnelle entre l'ATV d'un côté et l'habénula, le putamen et le cortex préfrontal médian de l'autre. Les résultats de cette étude apportent des preuves de l'existence d'un réseau comprenant l'ATV, la SN, l'habénula et des structures mésocorticolimbiques au cours du traitement des stimuli aversifs chez l'humain.

Le thalamus semble avoir un rôle dans la détection d'évènements et de stimuli aversifs. Plusieurs études chez l'humain rapportent une activation du thalamus en réponse à la perte d'argent (Knutson et al. 2000), et plus précisément du noyau antérieur du thalamus en réponse à la peur (Conejo et al. 2007) et du noyau dorsomédian du thalamus en réponse à la douleur (Wang et al. 2015).

(2) Réponses aux erreurs de prédiction des punitions

RONGEUR ET PRIMATE NON-HUMAIN

Chez la souris, la délivrance de chocs électriques inattendus induit des réponses de l'amygdale plus fortes que celles évoquées par des chocs attendus (McHugh et al. 2014). Au cours du protocole utilisé, la magnitude de la réponse de l'amygdale à ces punitions prédisait les réponses comportementales observées le jour suivant, liant ainsi réponse de l'amygdale aux erreurs de prédictions des punitions et adaptation comportementale. L'omission d'un choc électrique attendu induisant une diminution de l'activité unitaire des neurones de l'amygdale, celle-ci encode également un signal d'erreur de prédiction négatif des évènements aversifs.

Chez le singe, un encodage des erreurs de prédiction des évènements aversifs a été retrouvé dans l'amygdale et dans le cortex cingulaire antérieur (Klavir et al. 2013) ainsi que dans l'habénula (Matsumoto & Hikosaka 2009b). Des enregistrements unitaires simultanés de neurones du cortex cingulaire antérieur et de l'amygdale au cours d'un protocole de conditionnement aversif par renversement suggèrent que la réponse de l'amygdale aux erreurs de prédiction non signées (peuvent être positives ou négatives) se propage au cortex cingulaire antérieur dorsal où se développe un encodage signé des erreurs de prédiction des punitions, qui est finalement redistribué à l'amygdale. Les erreurs signées (positives ou négatives) sont nécessaires à l'adaptation du comportement : un choix induisant une erreur positive est à reproduire quand celui induisant une erreur négative est à éviter. Dans le cas de l'habénula, des enregistrements unitaires chez le singe ont rapporté une réponse de ces neurones suite à la délivrance d'une punition mais aussi suite à la non-délivrance d'une réponse. Ces travaux ont rapporté à la fois un encodage des renforcements en eux-mêmes mais également un encodage des erreurs de prédiction des punitions dans l'habénula (Matsumoto & Hikosaka 2009a).

Toujours au niveau sous-cortical, le rôle de l'aire tegmentale ventrale dans l'encodage de la prédiction d'évènements aversifs a été mis en évidence chez le rat grâce à l'utilisation de l'optogénétique combinée à des enregistrements unitaires (Matsumoto et al. 2016). Les neurones dopaminergiques de l'ATV latérale répondaient aux stimuli aversifs, et ces

réponses étaient réduites lorsque les punitions étaient attendues, signe d'une réponse aux erreurs de prédiction des punitions.

Suite aux travaux préliminaires réalisés chez l'animal sur les noyaux thalamiques (NAT et NDMT) et leur rôle dans l'évitement des punitions, il n'y a à ce jour pas d'équivalent chez l'humain. Sweeney-Reed et ses collaborateurs ont cependant étudié le rôle du NAT, et dans une moindre mesure du NDMT, dans les processus de mémorisation, en raison de la connexion directe entre le NAT et l'hippocampe au sein du circuit de Papez et le rôle de ce circuit dans la mémoire (Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016; Sweeney-Reed et al. 2017). Ils ont ainsi mis en évidence un rôle fonctionnel proche de ces deux noyaux au cours des processus de mémorisation et une implication fonctionnelle forte des oscillations à basse fréquence thêta (4-7Hz) en particulier. De plus, ils ont rapporté des phénomènes de couplage d'amplitude de phase entre les oscillations de base fréquence thêta et les oscillations de haute fréquence gamma, laissant supposer un rôle plus complexe dans des processus cognitifs de haut niveau comme la prise de décision.

HUMAIN

Beaucoup de travaux se sont intéressés à la détection des erreurs au cours de la cognition. Ces travaux ont identifié des régions sous-corticales (habénula, noyau sous-thalamique) et corticales (insula et cortex cingulaire antérieur) impliquées dans la détection des erreurs. Ces travaux ont servi de point de départ à des études plus poussées sur les régions impliquées dans la détection des erreurs de prédiction des punitions. En effet, des punitions inattendues sont le résultat d'un échec de stratégie comportementale, et représentent en soit des erreurs.

C'est ainsi que dans une revue des études de neuroimagerie sur l'apprentissage par renforcement (Garrison et al. 2013), l'habénula et l'insula ont été identifiées comme signalant les erreurs de prédiction des évènements aversifs, c'est-à-dire des erreurs de prédiction des punitions.

Depuis plusieurs années, la dopamine et les projections mésolimbiques dopaminergiques ont été mises en évidence comme jouant un rôle central dans le traitement de l'erreur (Bromberg-Martin & Hikosaka 2012; Hester et al. 2010). Au niveau cortical, l'insula antérieure a très probablement un rôle essentiel dans la prise de décision et le traitement des erreurs. Des travaux de modélisation Bayésienne et de connectivité dynamique (DCM) (Ham et al. 2013) s'intéressant au réseau de saillance proposent un rôle de l'insula

antérieure droite dans la production d'un signal précoce de contrôle cognitif pour éviter les erreurs, influençant y compris le cortex cingulaire antérieur, pourtant reconnu comme étant un acteur clé du traitement des erreurs (Figure 28).

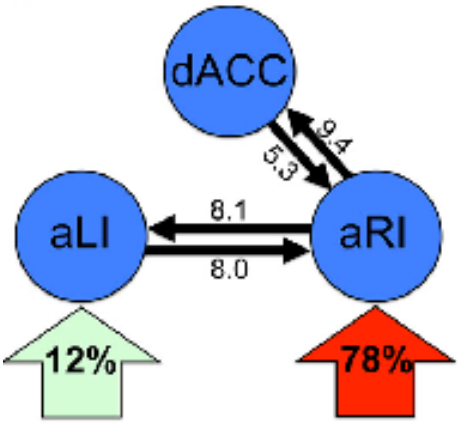


Figure 28 : Modélisation bayésienne moyenne du réseau de contrôle cognitif lors d'erreurs congruentes. Seules les connexions (flèches noires) entre les nœuds (cercles bleus) présentant une connectivité intrinsèque significative (force indiquée par les nombres sur les flèches, en milli Hertz) sont indiquées. Les larges flèches en-dessous représentent la probabilité que les nœuds soient la structure d'entrée du réseau. Figure issue de (Ham et al. 2013).

L'insula antérieure a été identifiée comme permettant la représentation des valeurs positives. Dans la littérature lésionnelle, il a été montré que des personnes présentant des lésions au niveau de l'insula antérieure notaient de manière atténuée les degrés d'excitation et de valeur que leur procuraient des stimuli visuels appétitifs, comparés à des sujets contrôles. De plus, les dommages à l'insula induisaient une réduction du degré d'appétitivité de substances auparavant addictives comme la cigarette et la cocaïne (Contreras et al. 2007; Naqvi et al. 2007). Concernant les paradigmes de prise de décision, il a été montré que des lésions diminuaient la sensitivité des individus au bénéfice global attendu dans le domaine des gains (Weller et al. 2009), de telle sorte que ces individus prenaient significativement moins de risques, y compris lorsqu'il était avantageux d'en prendre, comparé aux sujets contrôles. De plus, des lésions au niveau du striatum ventral et de l'insula antérieure induisaient toutes deux des difficultés à éviter les punitions, preuve de l'implication fonctionnelle de ces deux zones dans le traitement des valeurs des renforcements (Palminteri et al. 2012).

Il a été rapporté que les patients ayant une lésion au niveau du vmPFC avaient des difficultés à apprendre de conséquences négatives de leurs choix (Wheeler & Fellows 2008).

L'existence d'un lien fonctionnel entre l'insula antérieure et le striatum a également été proposée suite à des travaux de neuroimagerie fonctionnelle et lésionnels chez l'Homme, au cours de protocoles d'apprentissage par renforcement (Pessiglione et al. 2006; Palminteri et al. 2012). Dans le cas où les participants recevaient un renforcement appétitif (gain d'argent virtuel) suite au choix d'une option correcte et un renforcement aversif (une punition) sous la forme de la perte effective d'argent virtuel lorsqu'ils sélectionnaient des options incorrectes,

ceux-ci présentaient une activation à la fois au niveau du striatum lors de la présentation des options (ventral dans le domaine des gains, dorsal dans le domaine des pertes), et une activation de l'insula antérieure suite à la délivrance des punitions ((Palminteri et al. 2012) et Figure 29).

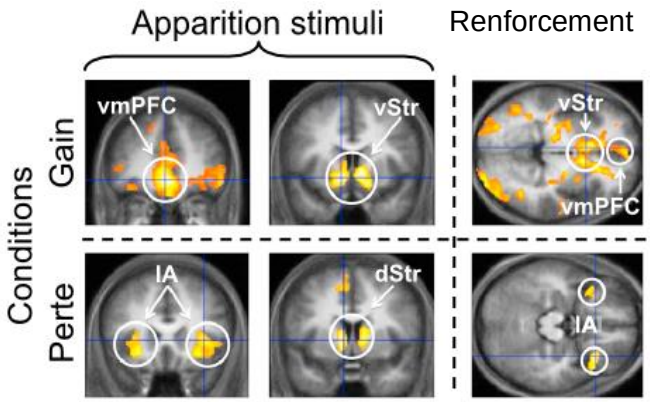


Figure 29 : Activation de deux systèmes distincts au cours de l'apprentissage par récompenses (haut : vmPFC et striatum ventral) et par évitement des punitions (bas : insula antérieure et striatum dorsal). Figure adaptée de (Palminteri et al. 2012).

L'ensemble de ces résultats met en évidence un possible rôle de l'insula antérieure à la fois dans l'encodage des valeurs des options mais aussi dans celui de la probabilité et de la magnitude des renforcements, qu'ils soient appétitifs ou aversifs. Cependant, un consensus n'a à ce jour pas encore été trouvé quant à la nature réelle des contributions de l'insula antérieure au cours de la prise de décision. C'est ce à quoi les travaux présentés dans ce manuscrit vont s'atteler.

Une étude en IRMf s'est intéressée à l'encodage des erreurs de prédiction des évènements saillants (appétitifs ou aversifs), pour faire écho au rôle identifié des neurones dopaminergiques dans cet encodage. Les résultats de cette étude de neuroimagerie fonctionnelle chez l'humain (Météreau & Dreher 2013) ont révélé que l'activité d'un réseau cérébral -comprenant le striatum, l'insula antérieure et le cortex cingulaire antérieur- covariait avec les erreurs de prédiction de la saillance suite à la délivrance de jus appétitifs et aversifs. De plus, l'activité de l'amygdale était corrélée avec ces deux renforcements (jus) et avec des images aversives, appuyant son rôle également dans l'encodage des erreurs de prédiction des renforcements aversifs. Ces résultats démontrent l'existence d'un réseau cérébral reflétant les erreurs de prédiction d'évènements saillants et dont l'activité dépend du type de renforcement.

c) Bases neurales du risque

RONGEUR ET PRIMATE NON-HUMAIN

Il a été rapporté que les neurones dopaminergiques du mésencéphale ventral permettent un encodage discret des probabilités de récompense et de l'incertitude, donc du risque (Fiorillo et al. 2003). L'incertitude est un paramètre critique pour évaluer précisément les prédictions de récompenses. La fréquence de décharge de ces neurones dopaminergiques covarie avec l'incertitude, supposant leur rôle dans la prise en compte du risque. Il a aussi été rapporté que ces neurones chez le singe créent des signaux combinant des informations concernant l'incertitude de récompense, la valeur et l'anticipation des punitions (Monosov et al. 2015). Des neurones du prosencéphale basal médian ont été identifiés chez le primate comme encodant à la fois l'incertitude de récompense, la valeur attendue de la récompense, la prédiction de punition et les récompenses et punitions inattendues.

Chez les rats, l'inactivation de l'insula antérieure réduit la préférence pour des options risquées quand les deux options risquée et sure ont la même valeur attendue (Ishii et al. 2012). Les travaux de doctorat de Eric Xu à l'Université de Dartmouth (Xu 2014) sur le rôle des neurones de l'insula antérieure au cours de la prise de décision en condition de gain en présence de risque chez le macaque rhésus rapportent (1) une absence d'effet de l'encodage intrinsèque de la valeur des récompenses sur la préférence au risque, (2) une préférence globale pour le risque plus que pour la nouveauté, bien que les singes soient plus attentifs aux renforcements de forte magnitude, et ce malgré un plus fort niveau de risque, et (3) l'absence de biais dû à la saillance des renforcements puisque les singes intégraient à la fois les probabilités et les valeurs des récompenses. Ce second résultat contredit l'idée selon laquelle la majorité des êtres humains ont une attitude neutre vis-à-vis du risque (Hayden & Platt 2009), c'est-à-dire que le risque n'influence pas significativement leurs choix.

HUMAIN

Parmi les aires cérébrales dont le rôle vis-à-vis du risque a été étudié se trouvent tout d'abord les aires associées au système de récompense. Ce circuit frontostriatal a été mis en évidence en neuroimagerie fonctionnelle pour son implication dans l'encodage distinct des valeurs des récompenses de l'encodage de l'incertitude associée à une attitude risquée au cours de la prise de décision (Tobler et al. 2006). Ainsi, le striatum encode les valeurs attendues élevées de manière indépendante de leur magnitude et de leur probabilité pendant que le cortex préfrontal voit son activité covarier avec le niveau d'incertitude et donc de risque : le latOFC est associé à une aversion au risque quand le mOFC est associé à une recherche de risque.

Ces résultats de neuroimagerie ont été complétés par une étude électrophysiologique invasive (stéréoEEG). Pour rappel, ces activations rapportées par le signal BOLD en neuroimagerie correspondent à des activités de champs locaux (LFP) de haute fréquence, associées à des processus cognitifs de haut niveau (Lachaux et al. 2012; Niessing et al. 2005), permettant de comparer directement ces résultats. Six patients épileptiques implantés en sEEG au niveau de l'OFC (Yansong Li et al. 2016) ont permis de mettre en évidence la dynamique cérébrale de l'encodage du risque et des valeurs associées aux récompenses dans l'OFC. Les travaux de Li se sont focalisés sur les erreurs de prédiction positives et incluent un faible nombre de patients, rendant complexe l'interprétation de ces résultats. Trois signaux successifs ont été identifiés : 1) environ 400ms après la présentation de stimuli prédisant les récompenses, une réponse associée à l'encodage de la probabilité de récompense, 2) un signal de risque a lieu pendant la fin de la phase d'anticipation de la récompense et dure jusqu'après la délivrance de la récompense, et 3) une réponse à la suite de la délivrance de la récompense encode la valeur de la récompense obtenue. De plus, une dissociation fonctionnelle a été observée au sein de l'OFC avec un encodage de la probabilité de récompense et du risque dans le latOFC et le mPFC alors que seul le latOFC encode la valeur obtenue. Les patients ayant une lésion focale au niveau du vmPFC présentent des difficultés à intégrer les probabilités, en particulier dans le cas de décisions incohérentes présentant du risque ou une ambiguïté (voir (Fellows 2011; Kennerley & Walton 2011) pour une revue complète).

Toujours dans le cortex préfrontal, le dlPFC est lui aussi impliqué dans la prise de décision risquée et morale, par exemple lorsqu'il s'agit de distribuer des ressources limitées (Greene et al. 2001). Le dlPFC est aussi activé quand les coûts et bénéfices de choix alternatifs sont considérés (Duncan & Owen 2000). Similairement, quand il y a plusieurs options alternatives parmi lesquelles il faut choisir, le dlPFC facilite une préférence envers l'option la plus équitable et supprime la tentation de maximiser un gain personnel (Knoch & Fehr 2007).

En neuroimagerie, l'activité de l'insula antérieure a également été mise en évidence comme précédant des comportements de choix sûr au cours de décisions d'investissement entre des stocks sûrs et des stocks plus risqués (Kuhnen & Knutson 2005; Knutson et al. 2007). De manière générale, l'insula semble avoir un rôle central dans le traitement du bénéfice subjectif et influence la prise de décision grâce à ses liens avec les systèmes affectifs. Cependant, certains résultats semblent indiquer un rôle dans le traitement des récompenses qui pourrait induire une augmentation de la prise de risque au cours de processus décisionnels, alors que d'autres indiquent un rôle de contrôle des centres de traitement des récompenses diminuant le niveau de risque au cours de la prise de décision.

Des études en IRMf chez l'humain ont rapporté une activation de l'insula précédant un choix risqué entre différentes options de valeur attendue équivalente (Xue et al. 2010). Inversement, il a également été montré que l'insula antérieure maintient et intensifie les représentations des valeurs négatives (aversives). Il a été rapporté (Berntson et al. 2011) que des personnes présentant des lésions à l'insula antérieure n'avaient pas uniquement une perception amoindrie de la valeur et de l'attrait (arousal) des stimuli visuels appétitifs, mais également pour les stimuli visuels aversifs. Il a été rapporté une diminution de la sensibilité à la probabilité des pertes lors de paris faits dans le domaine des pertes, suite à des lésions de l'insula antérieure (Clark et al. 2008; Weller et al. 2009). Ces individus avaient tendance à prendre plus de risques que les sujets contrôles, et n'ajustaient pas correctement leurs décisions selon la valeur attendue. L'insula antérieure a ainsi été associée à une aversion du risque puisqu'elle s'active lors de choix sûrs et d'erreurs causées par une aversion au risque, au contraire du noyau accumbens qui s'active lors de choix risqués et d'erreurs dues à une trop grande prise de risque (Kuhnen & Knutson 2005; Mohr et al. 2010). Cette activation plus forte de l'insula suite à des choix risqués que sûrs sert probablement à adapter plus efficacement le comportement (Paulus et al. 2003). Il a été montré que l'insula s'active pour refléter à la fois le risque que les erreurs de prédiction du risque ((Preuschoff et al. 2008) et Figure 30). La prédiction du risque module l'activité de l'insula antérieure puisqu'une augmentation de l'erreur de prédiction du risque est représentée sous la forme d'une activation plus forte de l'insula antérieure. Ces activations s'expliquent par le lien déjà connu de l'insula avec l'incertitude (Elliott et al. 2000; Critchley et al. 2001; Jones et al. 2010; Ernst et al. 2002; Hsu et al. 2005; Huettel et al. 2006) et les connexions bidirectionnelles entre l'insula et des structures impliquées dans la prise de décision et le traitement des récompenses (OFC, amygdale, ACC, noyau accumbens) (Reynolds & Zahm 2005). L'insula a également été liée au risque au niveau neurobiologique via l'implication d'un neurotransmetteur, la noradrénaline. Ce neurotransmetteur fabriqué au sein du locus coeruleus innervé fortement l'insula antérieure est connu pour son rôle dans la dilatation pupillaire. Une étude a mis en évidence une dilatation de la pupille suite à des erreurs de prédiction du risque, supportant indirectement le rôle de l'insula dans l'encodage des erreurs de prédiction du risque (Preuschoff et al. 2011).

En parallèle de l'activation de l'insula antérieure pour encoder les erreurs de prédiction et le risque (Preuschoff et al. 2008), d'autres régions cérébrales produisent des signaux explicites reflétant l'incertitude de la récompense : les neurones dopaminergiques encodent un signal de risque puisqu'il a été montré que leur activité peut covarier avec la déviation standard, c'est-à-dire la variance de la magnitude des renforcements (le risque). De plus, le striatum et

l'OFC émettent des signaux de risque associés à des récompenses monétaires qui covarient

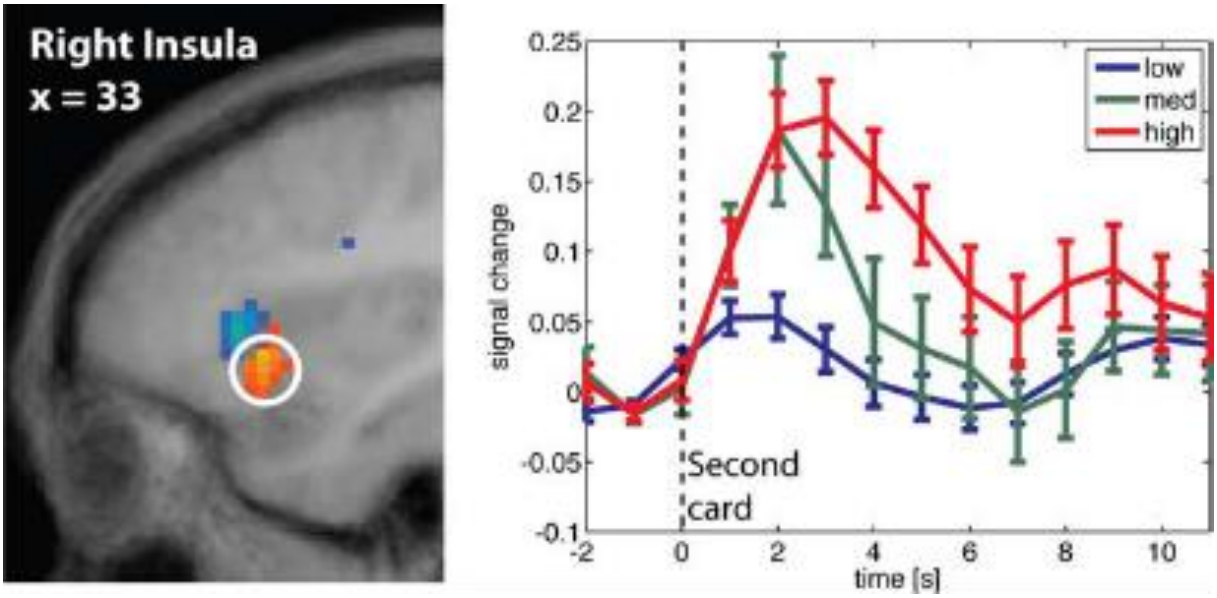


Figure 30 : L'insula antérieure encode la magnitude du risque et l'erreur de prédiction du risque. A gauche : localisation de l'activation au sein de l'insula antérieure sur cerveau MNI. A droite : réponses BOLD de l'insula antérieure suite à la délivrance de récompenses dans trois niveau de risque différents (bleu : faible, vert : moyen, rouge : élevé). Figure issue de (Preuschoff et al. 2008).

avec l'attitude envers le risque (Schultz et al. 2008).

Suite à ces résultats divergents dans la littérature actuelle, il semble important de s'intéresser au rôle fonctionnel de l'insula antérieure au cours de la prise de décision en situation de risque (McGuire et al. 2014; Behrens et al. 2007). Les travaux de Wright sur l'influence du risque sur la prise de décision mettent en évidence des incohérences entre de nombreuses études sur le risque et la prise de décision : certains rapportent une aversion au risque lors de la recherche de récompenses quand d'autres observent cette aversion lors de l'évitement de punitions. Wright et ses collaborateurs suggèrent que le contexte expérimental (différence de valeur entre les options, type de renforcements, probabilités des renforcements par exemple) peut être à l'origine de ces résultats variés sur l'influence du risque au cours de la prise de décision et de l'apprentissage par renforcement (Wright et al. 2012).

Cependant, il est important de noter que ces résultats s'intéressant aux régions ayant un rôle dans la prise de décision en situation de risque n'impliquent pas d'apprentissage, car le risque est toujours explicité. Afin d'étudier l'influence du risque sur l'apprentissage par renforcement, il nous est nécessaire d'utiliser un risque implicite, dont le niveau est à apprendre par essais et erreurs.

C. Questions soulevées et justifications méthodologiques

1. Dissociation fonctionnelle de l'apprentissage par récompense et par évitement des punitions

La revue de la littérature actuelle de la prise de décision et de l'apprentissage par renforcement en particulier a mis en évidence l'existence de débats concernant les bases neurales de l'apprentissage par récompense et de l'apprentissage par évitement des punitions. Malgré un consensus concernant le rôle du système dopaminergique dans ce que l'on appelle le « circuit de la récompense », plusieurs hypothèses coexistent concernant les corrélats neuronaux de l'apprentissage par récompenses ou par punitions : (1) l'existence de deux systèmes opposés impliquant des aires cérébrales corticales et sous-corticales distinctes, ou (2) l'existence d'une ségrégation fonctionnelle au sein même de ces régions cérébrales de la valence, ou encore (3) l'existence de populations neuronales distinctes mais co-localisées des informations relatives à la valence.

Les travaux présentés dans ce manuscrit se situent dans le prolongement de travaux de neuroimagerie fonctionnelle et permettront d'éclaircir en particulier les deux premières hypothèses. L'une des clés concernant les tâches utilisées est donc de dissocier expérimentalement l'apprentissage par récompense et par punition, en utilisant pour ce faire des GAINS et PERTES monétaires virtuels.

2. Dynamique de l'encodage des signaux de renforcement

Nous intéressant aux bases neurales et à la dynamique de l'apprentissage par renforcement en leur sein, chez l'Homme, il est nécessaire d'utiliser une technique ayant une bonne résolution temporelle telle que l'électroencéphalographie (de l'ordre de la milliseconde). De plus, un certain nombre de régions d'intérêt, y compris corticales, se situent en profondeur, loin de la surface du cerveau. C'est par exemple le cas du cortex préfrontal ventromédian, de l'insula antérieure, ou des noyaux du thalamus. Il est donc important de trouver une technique électrophysiologique permettant leur étude chez l'être humain, tout en combinant une bonne résolution spatiale et temporelle. Deux cas de figure sont présentés : le premier est l'utilisation de la technique d'électroencéphalographie stéréotaxique (sEEG) qui nous permettra d'accéder à ces aires cérébrales tout en conservant une bonne résolution temporelle (à l'inverse de l'IRMf, de l'ordre de la seconde). Cependant, il est évident qu'une technique telle que la sEEG est fortement invasive et ne peut être utilisée chez l'Homme qu'à des fins cliniques. Les participants ne sont donc pas des sujets sains mais des patients épileptiques présentant une épilepsie focale résistante aux médicaments et sont implantés

en sEEG dans leur cortex quelques semaines afin de localiser leur foyer épileptique dans l'optique d'une résection chirurgicale curative. Dans un second cas, la technique et les patients sont légèrement différents. Il s'agit de patients épileptiques pharmacorésistants ne pouvant subir une résection chirurgicale du foyer épileptique et présentant une résistance à un traitement par stimulation du nerf vagal pour traiter leurs crises. Ces patients sont implantés bilatéralement dans leur thalamus (NAT et NDMT) avec des macroélectrodes qui seront à terme utilisées pour appliquer une stimulation cérébrale profonde (SCP) de ces noyaux pour diminuer les crises (fréquence et intensité). Dans les deux cas, il est possible d'enregistrer les oscillations cérébrales grâce aux électrodes de sEEG (pendant toute l'hospitalisation) et de SCP (avant l'internalisation des câbles et le branchement du stimulateur) au cours d'un protocole d'apprentissage par renforcement. En raison de la pathologie des participants et de la technique utilisée, il n'est pas possible d'avoir un groupe de sujets sains comme contrôles. Les patients sont donc leurs propres contrôles, ce qui permet un passage des artefacts pathologiques dans l'activité basale enregistrée et leur disparition relative avec la multiplication des patients ayant tous des épilepsies différentes. Les protocoles cognitifs sont donc pensés de façon à ce que pour chaque essai, il y ait une période d'activité de base pouvant être déduite de celle au cours de l'essai afin que ne ressorte que l'activité physiologique.

En raison de notre intérêt pour l'encodage des signaux de renforcement au cours de l'apprentissage par renforcement, nous avons couplé les protocoles cognitifs avec un modèle computationnel de type Q-learning afin de suivre la magnitude des erreurs de prédiction des renforcements, des valeurs subjectives et tous les autres paramètres des algorithmes de Q-learning à prendre en compte pour expliquer le comportement et les activations observés.

3. Influence du risque sur l'apprentissage par renforcement

Pour étudier l'influence du risque sur la prise de décision, il est important de rappeler qu'il n'y a à ce jour aucune étude sur l'influence du risque sur l'apprentissage par récompenses ou par punitions. Pour commencer cette étude, nous avons donc fait appel à des sujets sains pour une étude comportementale. Cette étape est essentielle pour mettre au point des protocoles cognitifs et des algorithmes de modélisation computationnels capables d'isoler le risque, de mettre en évidence ses effets et d'expliquer la mécanistique impliquée. Cette partie constitue donc une troisième étude comportementale qui appelle à une discussion complète des effets rapportés avant de passer à l'étude électrophysiologique.

II. ETUDES EXPERIMENTALES

A. Etude 1 : Dynamique corticale de l'apprentissage par renforcement

Rewards and punishment learning differentially modulates intracerebral brain dynamics

Running title: Dynamics of reinforcement learning

Maëlle Camille Marie Gueguen[1,2], Jean-Philippe Lachaux[3], Philippe Kahane[2,4], Pablo Billeke[5], Sylvain Rheims[6], Mathias Pessiglione[7] and Julien Bastin[1,2]*

[1] Univ. Grenoble Alpes, F-38000 Grenoble, France.

[2] Inserm, U1216, F-38000 Grenoble, France.

[3] Lyon Neuroscience Research Center, Brain Dynamics and Cognition team, INSERM UMRS 1028, CNRS UMR 5292, Université Claude Bernard Lyon 1, Université de Lyon, F-69000, Lyon, France

[4] Neurology Department, CHU de Grenoble, Hôpital Michallon, F-38000 Grenoble, France.

[5] División de Neurociencia, Centro de Investigación en Complejidad Social (neuroCICS), Facultad de Gobierno, Universidad del Desarrollo, Santiago, Chile

[6] Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and Lyon 1 University, Lyon, France

[7] Motivation Brain & Behavior lab (MBB), Brain & Spine Institute (ICM), CNRS UMR 7225 - UPMC-P6 UMR S1127, Paris, France

*Correspondence to: julien.bastin@ujf-grenoble.fr (J.B.); Phone: (0033) 4 56 52 06 78; Fax: (0033) 4 56 52 05 98; Institut des Neurosciences de Grenoble, Bâtiment Edmond J. Safra des Neurosciences, Chemin Fortuné Ferrini, Université Joseph Fourier, Site Santé La Tronche, BP 170 38042 Grenoble Cedex 9, France.

ABSTRACT

Although the neural activities underlying reward-based and punishment-avoidance learning have been extensively studied, the neural dynamics underlying these forms of reinforcement learning remains unclear. We recorded intracerebral activity from patients with refractory epilepsy while they performed a reinforcement learning task. Predictions errors estimated from computational modeling were encoded in the broadband gamma activity (BGA) [50-150 Hz] of four brain regions: ventromedial prefrontal cortex (vmPFC), anterior insula (aIns), dorsolateral prefrontal cortex (dlPFC) and lateral orbitofrontal cortex (latOFC). Critically, we found a clear double dissociation: BGA in the vmPFC was modulated by reward prediction errors whereas BGA in aIns and dlPFC was modulated by punishment prediction errors. Granger causality analyses revealed a prominent drive from aIns toward latOFC, vmPFC and dlPFC at lower frequencies. These findings provide critical support for the hypothesis that opponent brain systems are recruited depending on the outcome valence — gain or loss — that has to be learned to reinforce behavior.

INTRODUCTION

Theoretically, reward-based learning and punishment-based learning can be modeled within a similar computational framework, in which the difference between the expected and actual outcome of a behavioral choice serves as a teaching signal to update the subjective value of that option (Sutton & Barto 1998). This framework provides a parsimonious account on how subjects learn a behavior that maximize rewards and minimize punishment and fits with well-documented neural observations such as the discovery of dopaminergic cells within the mesocorticolimbic reward circuit which encode reward prediction errors (Schultz 2016). A straightforward hypothesis would then be that reward-based and punishment-based learning have a common neural implementation such as an increase (decrease) of firing rate for positive (aversive) events (Matsumoto et al. 2016). An alternative proposal would be that these two form of learning signals involve functionally distinct subpopulation of neurons (Lammel et al. 2012; Matsumoto & Hikosaka 2009b), or even activate opponent brain circuits (Pessiglione et al. 2006).

Matsumoto et al. (2016) showed that the same dopaminergic cells are either inhibited by aversive events or excited by appetitive events, suggesting a common neural mechanism for reinforcement and punishment-learning. Yet, other studies using different experimental paradigms reported that dopaminergic cells either ignore, are excited or inhibited by aversive events (Brischoux et al. 2009; Cohen et al. 2012; Fiorillo 2013; Matsumoto et al. 2016). This lack of consensus might be due to differences in the reward context used by the authors (Matsumoto et al. 2016), in the type of inputs received by the recorded cells (Lammel et al. 2012) or in the precise anatomical location of the cells (Brischoux et al. 2009). Studies of the neocortex have also been globally inconclusive, with several conflicting observations in the orbitofrontal cortex (Morrison & Salzman 2009; Morrison et al. 2011; Roesch & Olson 2004).

Functional neuroimaging studies in humans support the view that two opponent learning circuits co-exist: reinforcement-based learning activates preferentially the ventromedial prefrontal cortex and the ventral striatum whereas punishment-based learning activates more the dorsal striatum and the anterior insula (Hare et al. 2008; Palminteri et al. 2015; Seymour et al. 2004; Pessiglione et al. 2006). Yet, the dissociation might not be clear-cut, since some of the existing results are inconsistent with this scheme. For instance, single-neuron recordings in monkeys and human fMRI studies suggest that the anterior insula may also respond to reward signals (Kirsch et al. 2003; Mizuhiki et al. 2012). In addition, negative outcomes also trigger responses in the lateral part of the orbitofrontal cortex (O'Doherty et al. 2001; Seymour et al. 2005; Jung et al. 2011), the dorsolateral prefrontal cortex (Ginther et al. 2016) or the amygdala (Yacubian 2006). More importantly, the opponent system hypothesis

is heavily challenged by results suggesting that a single network reacts to both gains and losses (Kim et al. 2006; Tom et al. 2007).

These partly diverging accounts in both the animal and human literature may be due to differences in the paradigms used, or in the techniques used to assess neural responses. Furthermore, asymmetries regarding data acquisition may also have contributed to obscure the debate, since for instance, no electrophysiological study has compared responses in the insular cortex and the orbitofrontal cortex during decision-making tasks. Furthermore, neither functional neuroimaging nor single-cell recordings are optimal to document the transient neural dynamics in response to feedback at a network level, including how information flows between brain regions during reinforcement learning (Hunt et al. 2015; Larsen & O'Doherty 2014).

Here, we sought to overcome some of these limitations with direct electrophysiological recordings (intracranial EEG) of several key human cortical structures of the reinforcement-learning network, using broadband gamma activity (BGA, 50–150 Hz) as a proxy of population-level spiking activity (Lachaux et al. 2007; Manning et al. 2009; Mukamel et al. 2005). Twenty patients performed an instrumental task designed to dissociate reward seeking from punishment avoidance learning (Pessiglione et al. 2006; Palminteri et al. 2012). In this task, patients were required to choose between two cues to maximize monetary gains (during reward-learning) or minimize monetary losses (during punishment-learning). We monitored choices to measure the performance, and calculated the reaction times for each session performed.

RESULTS

Intracerebral EEG data were collected from twenty patients with intractable epilepsy (see demographical details in Table 1 in supplementary materials and methods) while they performed an instrumental learning task during which reward and punishment conditions were matched in difficulty, as the same probabilistic contingencies were to be learned.

Behavior

Patients were able to learn the correct response over the 24 trials of the learning session. They tended to choose the most rewarding cue in the gain condition and avoided the most punishing cue in the loss condition (Figure 31; one-sample t-test versus chance (50%): $p<0.0001$ for both conditions, $t(19)=31.864$ for gain, $t(19)=76.412$ for loss). Their behavior matched observations in healthy subjects (Palminteri et al. 2015; Pessiglione et al. 2006) : patients were able to choose the correct option whether they learned from reward or punishments. We also found that reaction times were shorter in the gain condition than in the loss condition (839ms vs 1277ms, two-sample t-test, $p<0.0001$, $t(99)=10.291$).

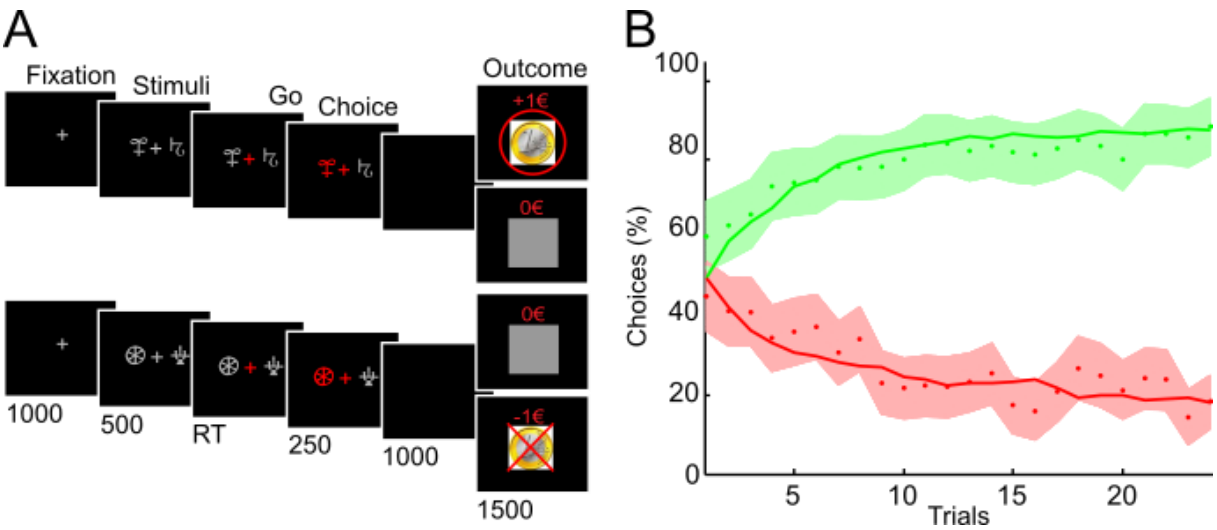


Figure 31: Behavioral task and results. (A) Successive screen of typical trials in the reward (top) and loss (bottom) conditions. Patients had to select one abstract visual stimulus among the two presented on each side of a central visual fixation cross, and subsequently observed the outcome. Duration are given in milliseconds. **(B)** Average learning curves (n=20 patients). Modeled behavioral choices (solid line) are superimposed over observed choices (average: dots; 99% confidence interval: shaded areas). Learning curves portray trial by trial, the average proportion of trials across patients corresponding to subjects choosing the 'correct' stimulus in the gain condition (green dots), and the 'incorrect' stimulus in the loss condition (red dots). Modeled learning curves (solid lines) represent the associated probabilities predicted by the computational model.

Intracerebral electro-encephalography

The aim of our analysis was to explore the computational principles and associated neural dynamics while learning to maximize rewards and avoid punishments. Our objectives were (1) to identify brain regions responding to prediction errors (independently from valence) and (2) to characterize in time the dissociation between brain processes underlying learning from rewards vs. learning from punishments.

We first focused the analyses on broadband gamma activity (BGA) [50-150 Hz], which is known to correlate with both spiking and fMRI activity (Niessing et al. 2005; Lachaux et al. 2007; Manning et al. 2009). We measured BGA from 1694 sEEG cortical sites across 17 patients (for three out of 20 patients, sampling rate was too low to estimate BGA). The location of sEEG contact-pairs is shown in Figure 32. Each sEEG contact was anatomically labeled as a function of individual gyri/sulci organization estimated by a cortical parcellation atlas (MarsAtlas, (Auzias et al. 2016). All 41 areas of MarsAtlas comprised at least one sEEG contact site. However, to increase the robustness and reproducibility of the reported results, we chose to restrict sEEG data analyses to 35 MarsAtlas' parcels that were sampled by at least ten sEEG contact-pairs recorded across at least three patients. This approach allowed us to conduct a pseudo-group level analysis in these 35 parcels (note that data from both hemispheres were collapsed to improve the power of the following analyses).

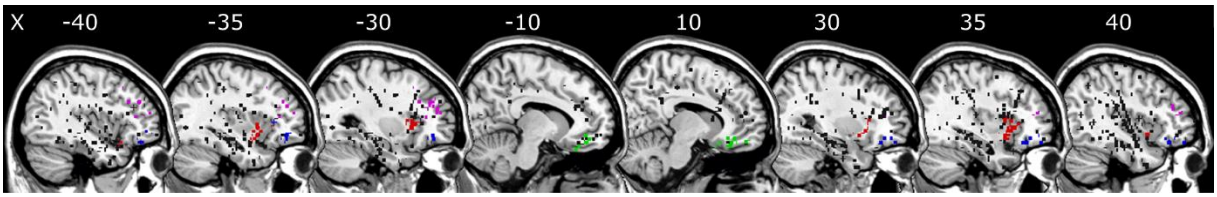


Figure 32: Anatomical locations of all intracerebral contact-pairs. Sagittal slices of a brain template over which each dot represents one sEEG contact-pair (n=1694). Red dots corresponds to sEEG sites located within the anterior insula; green dots: sEEG contact-pairs within the ventromedial prefrontal cortex; blue dots: sEEG contact-pairs within the lateral orbitofrontal cortex; purple dots: sEEG contact-pairs within the dorsolateral prefrontal cortex.

We next used the reinforcement-learning model to extract trial by trial prediction errors which were then used as parametric modulators of BGA time-locked to the onset of the outcome. For this analysis, BGA was computed for each sEEG contact-pair and each time-point and regressed again reward prediction errors and punishment prediction errors. For each brain region, we tested the significance of the regression for each sEEG contact-pair (see methods). We estimated a common statistical threshold for all electrodes by using multiple

permutation tests. Briefly, we first estimated a statistical threshold for each sEEG contact that took into account the multiple comparisons we did in the time dimension. Because the average and variance within and between brain region was negligible, we chose to use a common threshold corresponding to a corrected $p<0.01$ (i.e. the absolute value of the coefficient of correlation had to be superior to 0.24). This approach revealed a significant correlation between BGA and prediction errors in four brain regions: the anterior insula (aIns: n=41 out of 80 sEEG contact-pairs coding prediction errors), the dorsolateral prefrontal cortex (dlPFC: n=34 out of 81), the lateral orbitofrontal cortex (latOFC: n=18 out of 91) and the ventromedial prefrontal cortex (n=15 out of 67). We note that the proportion of sEEG contact-pairs that responded to prediction errors was significantly higher in aIns than in vmPFC or latOFC (Fisher contingency test, $p<0.05$), while the proportion of responding contact-pairs was also higher in the dlPFC relative to the latOFC ($p<0.05$, see Table 2 in supplementary materials).

In the following, we focus the analyses on these four regions of interest to examine whether reward vs. punishment-based learning differentially modulate these brain regions. We kept all sEEG contacts within each ROI in the following analyses, independently from their response to prediction errors (see Supplementary materials for similar analyses on only sEEG contacts responding to prediction errors).

Gamma activity encodes punishment prediction errors in the aIns and the dlPFC

To further investigate valence-related effects within the identified ROIs, we contrasted the strength of the reward vs. punishment prediction errors encoding. This analysis revealed that punishment prediction errors were negatively encoded by BGA in both the anterior insula and dorsolateral prefrontal cortex (Figure 33A and C). Critically, the strength of the correlation was significantly higher for punishment prediction errors than for reward prediction errors (aIns: two-tailed cluster-corrected paired t-test across all contacts: $p<0.01$ from 297ms to 594ms after outcome delivery, t(cluster)=-94.2454, Figure 33A; dlPFC: corrected $p<0.01$ from 141 ms to 594 ms after outcome delivery, t(cluster1)=-94.2454 and from 703ms to 2797ms, t(cluster2)=-503.2266; Figure 33C). Data from all sEEG-contacts within each ROI is provided in Figure 33. In particular, to examine putative differences between dlPFC and aIns, we compared the magnitude, onset and peak latencies of RPE and PPE encoding. We found that the onset of the peak of PPE encoding was significantly earlier in the anterior insula than in the dlPFC ($p=0.0190$).

Next, BGA time series were also regressed against the expected value of the chosen option modeled at the time of cue onset. We found that the value of the chosen cue was significantly encoded over long time intervals in both regions. Interestingly, this encoding of chosen Q value started around the timing of patients' choice and lasted until the offset of outcome display (and even later in the dlPFC, see Figure 39 in supplementary materials; cluster-corrected $p<0.01$ in both areas, aIns: from -1250ms to 1438ms after outcome delivery: t(cluster)=-657.5368; for dlPFC: from -1328ms to 2750ms: t(cluster)=-1425.2). Overall, model-based analyses of BGA suggest that punishment predictions errors and their predictive component (chosen values) are both negatively encoded in aIns and dlPFC.

Model-free analyses further support the hypothesis that BGA is modulated by the magnitude of punishment prediction errors in these areas. Hence, we also found that BGA was significantly higher after effective monetary losses compared to all other outcomes (Repeated measure ANOVA followed by Newman-Keuls post-hoc test: all p values <0.05, Figure 33B-D). This suggest that the outcome component of the encoded prediction error significantly modulate BGA in aIns and dlPFC. In addition, BGA increased more after the display of cues from loss pairs than from gain pairs (paired t-test; aIns: $p = 0.0021$, $t(79)=3.177$; dlPFC: $p<0.0001$, $t(80)=5.741$). This suggests that when patients expected an eventual loss, BGA activity was higher than in situations during which only neutral or gain outcomes were expected.

Overall, both model-based and model-free sEEG data analyses showed that aIns and dlPFC activities both contribute to punishment prediction error encoding, with significant modulations being relating to both outcome valence and to subject's expectations.

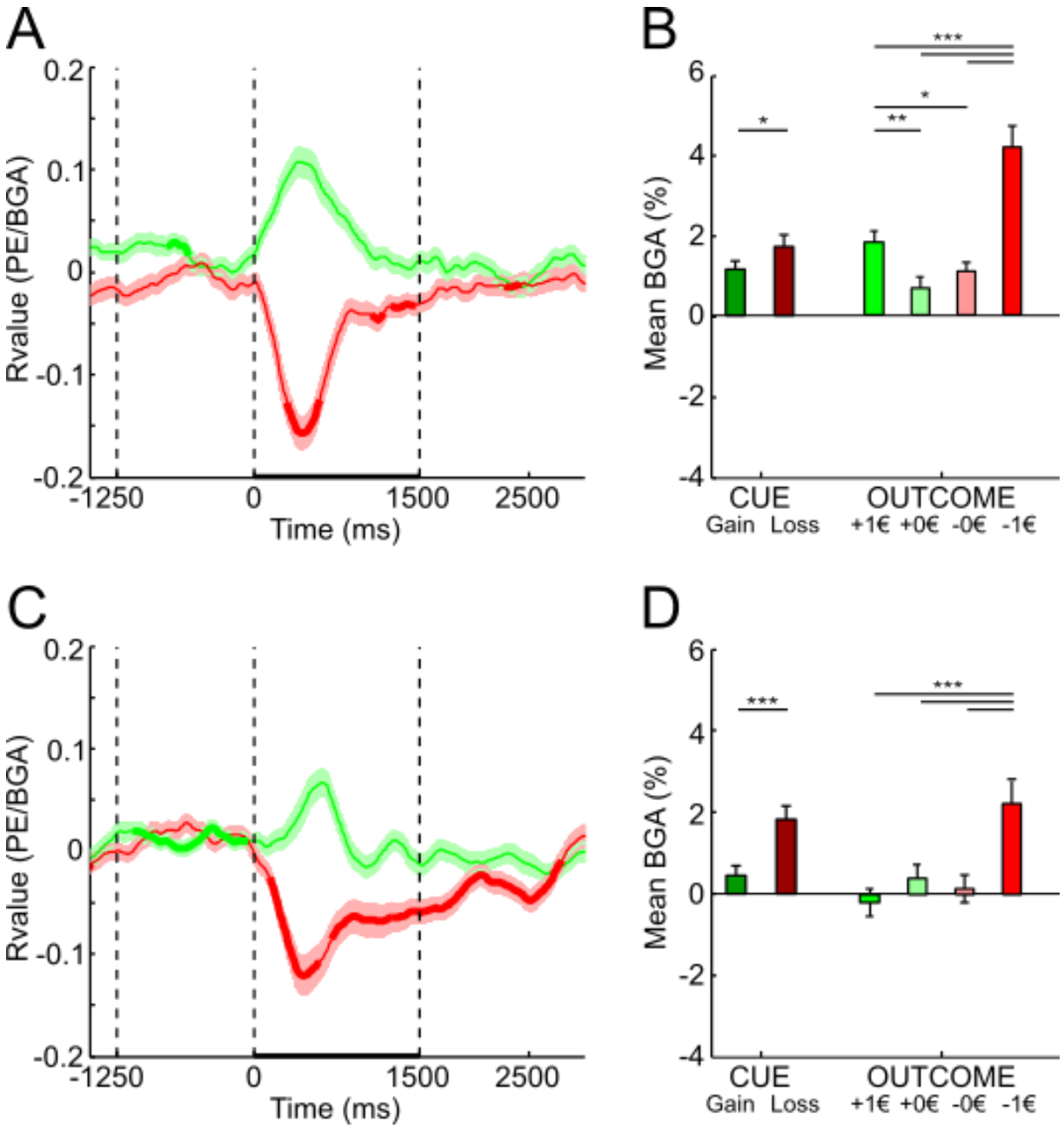


Figure 33: Gamma activity encodes punishment prediction errors in the anterior insula and in the dorsolateral prefrontal cortex. (A) Average aIns BGA correlation with prediction errors in gain (RPE, green) and loss (PPE, red) condition following outcome delivery (bold dark portion of x axis) (n=80 implanted contacts recorded from 11 patients). Mean PE encoding ± SEM across contacts (shaded areas). Bold lines represent time points at which the difference between RPE and PPE encoding reached significance ($p<0.01$ cluster corrected). **(B)** Average aIns BGA responses in the 800 ms following cue and outcome display in gain and loss condition (n=80 implanted contacts recorded from 11 patients). **(C)** Average dlPFC BGA correlation with prediction errors in gain (RPE, green) and loss (PPE, red) condition following outcome delivery (n=81 implanted contacts recorded from 8 patients). **(D)** Average dlPFC BGA responses in the 1000ms following cue and outcome display in gain and loss condition (n=8 implanted contacts recorded from 8 patients). Graphical conventions are identical to panel A-C and B-D.

Insular broadband gamma activity after monetary losses predicts inter-individual and within individual behavioral adaptations.

To further explore the functional role of BGA in aIns and dlPFC, we investigated the relationship between BGA and patients' performance. We found that the average magnitude of BGA following a punishment in the aIns (but not in the dlPFC) predicted patients' learning performance (n=41 aIns task-responsive sEEG contact-pairs, r=-0.429, p<0.0001, Figure 34A). To explore whether this (across patient's effect) could be traced within each patient's data, we next compared BGA after punishment delivery between "stay" trials (i.e. trials during which patients continued to choose the same cue in the subsequent choice despite a negative feedback in the previous one) and "switch trials" (i.e. trials during which patients switched to the other cue of the pair in the subsequent occurrence of this exact pair).

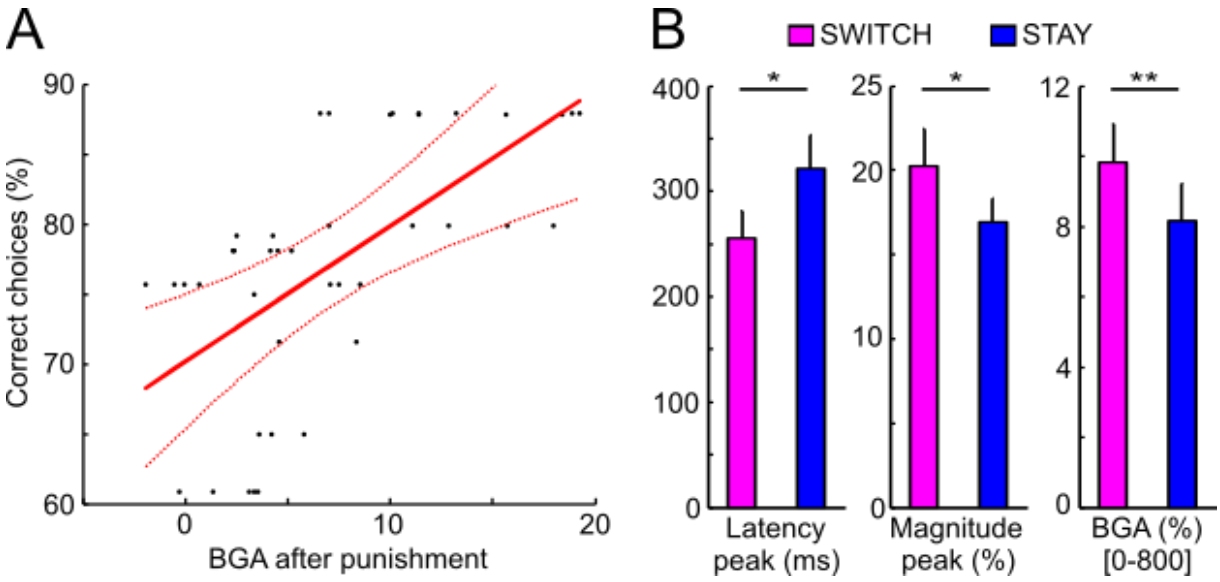


Figure 34: Gamma activity to punishment in the anterior insula is linked to performance. (A) Correlation between learning performance and post punishment BGA. The dots represent average across trials BGA of single sEEG contact pairs and the corresponding average learning performance, i.e. correct choices in the loss condition. The solid and dotted lines represent the linear regression ± STD. **(B)** Average latency and magnitude of the peak and average BGA response in the 800 ms following punishment delivery for all switch (pink) and stay (blue) trials. Data come from n=41 sEEG task-responsive contact-pairs within the aIns, recorded from 11 patients).

This analysis revealed that punished trials which led the patient to switch behavior exhibited a faster and larger increase of BGA activity than "stay" trials (peak latency: 256ms vs 322ms, p=0.0120, t(40)=2.633, paired t-test; peak magnitude: 20.3% vs 17.0%, p=0.0193, t(40)=2.437, paired t-test; average BGA in the [0-800ms] post outcome interval: 9.8% vs

8.2%, paired t-test: p=0.0040, t(40)=3.058). Thus, the magnitude of BGA in the aIns following monetary losses does not only predict learning performance across patients but also the behavior of individual patients on a trial to trial basis.

Gamma activity encodes reward prediction errors in the vmPFC

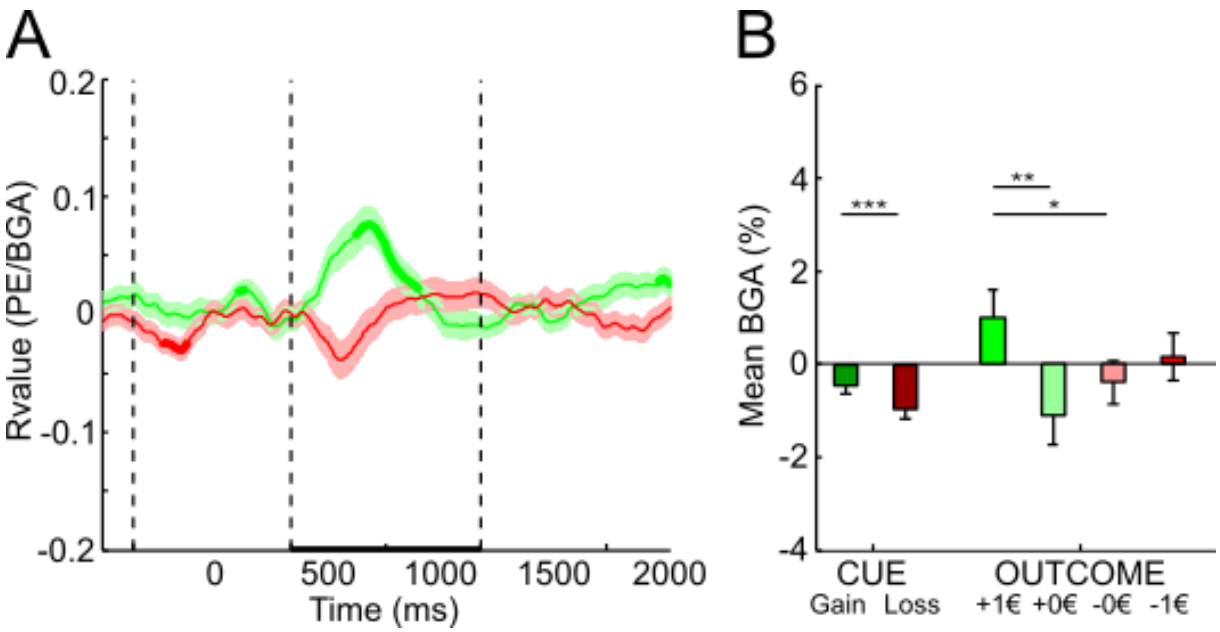


Figure 35: Gamma activity encodes reward prediction errors in the vmPFC. (A) Average vmPFC BGA correlation with prediction errors in gain (RPE, green) and loss (PPE, red) condition following outcome delivery (bold dark portion of x axis). Mean PE encoding ± SEM. Bold lines represent time points with significantly different RPE from PPE encoding ($p<0.01$ cluster corrected). **(B)** Average vmPFC BGA responses in the 800ms following cue display, response and outcome display in gain and loss condition (n=67 sEEG contact-pairs recorded from 11 patients).

The pattern of BGA observed in the vmPFC mirrored the effects observed in aIns and dlPFC: we found that vmPFC BGA positively encoded reward prediction errors (n=67 sEEG contact-pairs recorded across 11 patients). Critically, correlation coefficients computed between BGA and RPE were found to be significantly larger than correlation coefficients computed between BGA and PPE from 500 ms to 1016 ms after outcome delivery (Figure 35A; two-tailed cluster-corrected, $p<0.01$; paired t-test across all sEEG contact-pairs, $t(cluster)=95.4545$; see also Figure 40 in supplementary materials, that shows amplitude and latency data from all vmPFC contact-pairs).

In addition, vmPFC also positively encoded the value of the chosen cue (Q value) just after the onset of patient's choices and also during outcome display (see Figure 39 in supplementary materials, cluster-corrected $p<0.01$; one sample t-test across all vmPFC

electrodes, from 422ms to 1969ms after outcome delivery: t(cluster)=71.8962). Model-free analyses were consistent with these observations, since BGA in the vmPFC was relatively larger after the display of cues from gain pairs than from loss pairs (Figure 35B; p = 0.0001). Finally, BGA increase in the vmPFC were larger after effective monetary gains than after all other outcome types (neutral or effective losses, Figure 35B, all p values <0.05; RM ANOVA followed by a Newman-Keuls post-hoc test). We found no evidence regarding the relationship between BGA observed after an effective gain and patients' performance.

Thus, these results suggest that BGA in the vmPFC is modulated by the magnitude of reward prediction errors, corresponding to the encoding of both patients' reward expectation and the delivered reward.

Gamma activity encodes saliency in the latOFC.

Finally, in the latOFC (with n=91 contacts implanted across 12 patients), we found that BGA encoded a mixture of reward and punishment prediction error (Figure 37 in supplementary materials). Hence, during the early stages after the display of the outcome, the latOFC encoded preferentially punishment prediction errors (between [230 ms and 420] ms post outcome) whereas later on, reward prediction errors were preferentially encoded rather than punishment prediction errors (in the [890 – 953] and [1250 ms – 1391] ms time intervals) (two-tailed cluster-corrected paired t-test across all contacts: p<0.01 from 234ms to 422ms after outcome delivery, t(cluster1)=-75.4123, from 890ms to 953ms, t(cluster2)=8.0647 and from 1250ms to 1391ms, t(cluster3)=18.6047). A similar pattern was observed when chosen Q value were used as a regressor of BGA time-series: we found that Q values were initially negatively correlated with BGA whereas latter on, the encoding became positive (Figure 40D in supplementary materials, p<0.01 cluster corrected, from -266ms to 188ms after outcome display: t(cluster)=-72.0622 ; from 938ms to 1906ms: t(cluster)=196.2603). Model free analyses did not reveal any significant modulation of BGA at the time of cue display, while only effective gains triggered a significant response relative to neutral outcomes. That said, when the electrophysiological analyses were performed on sEEG contact-pairs that activity was modulated by prediction errors (task-responsive contact-pairs), a different pattern of result emerged, since it appeared clearly the gamma activity in the latOFC was used to significantly encode preferentially PPE from 328 ms to 469 ms after outcome delivery (Figure 37E in supplementary materials; n=18 task-responsive contact-pairs recorded across 9 patients, p<0.01 cluster corrected, t(cluster)=-21.0067).

Granger Causality

In order to examine how prediction errors were encoded from a network perspective, we next computed the Granger causality (GC) between the brain regions that were shown to be significantly modulated by prediction errors and that could be simultaneously recorded. We focused the analyses on the one second period following outcome display. We also tested whether GC influences during outcome processing were modulated by outcome valence (see Methods). GC analyses revealed a prominent drive from aIns toward latOFC, vmPFC and dlPFC (Figure 36). Interestingly, significant GC (wilcoxon and lme tests, p values < 0.05) occurred mostly at low frequency regimes. Furthermore, the causal influence of aIns on latOFC, and of latOFC on vmPFC was strongest after monetary punishments. A bidirectional interaction between aIns and dlPFC was also strengthened during punishments (wilcoxon test, $p<0.05$). These modulations of GC, thus likely reflect the transmission of punishment prediction error signals. Finally, GC influence between aIns and vmPFC was not modulated by outcome valence and might therefore reflect an unsigned prediction error signal.

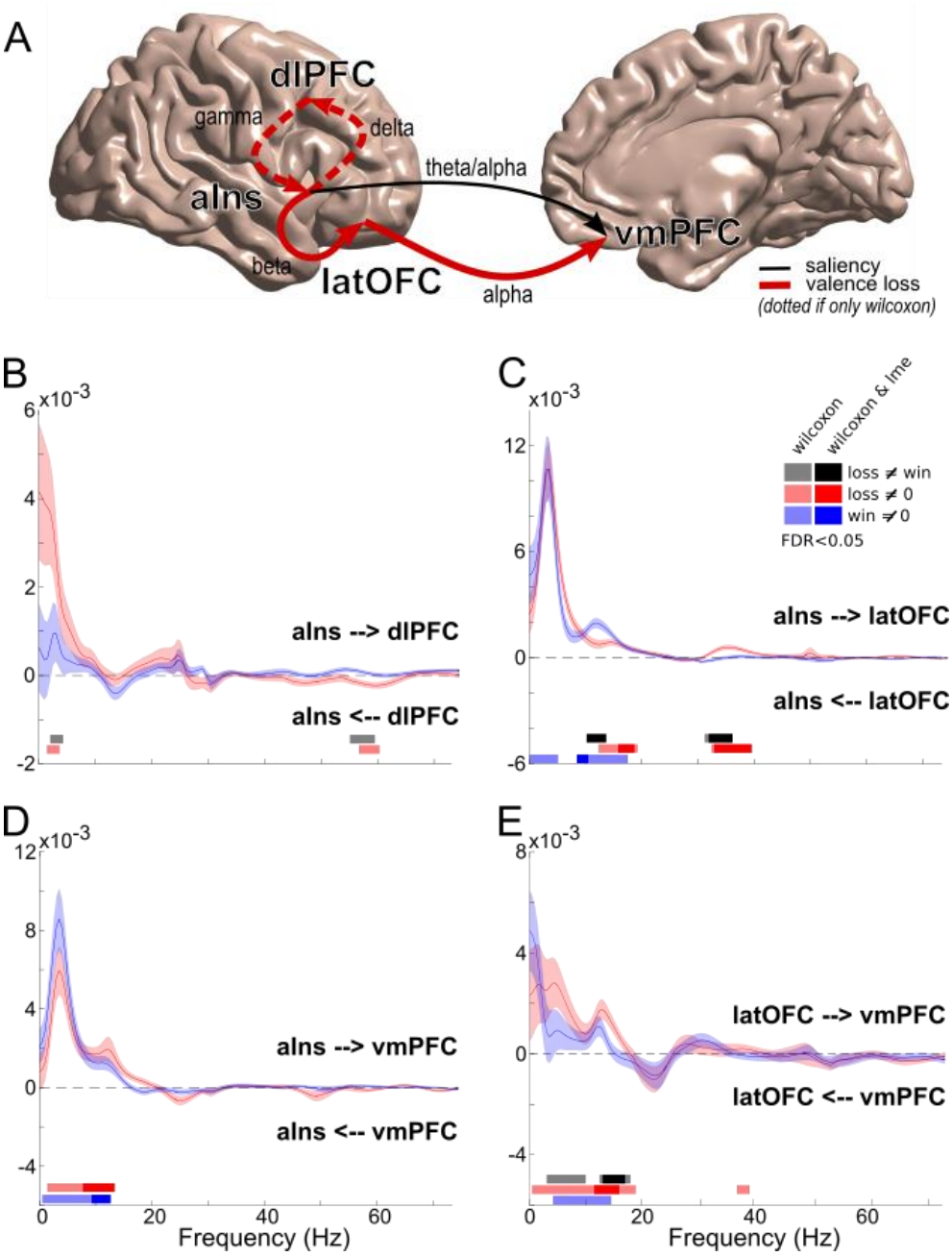


Figure 36 : Granger causality during reinforcement learning. (A) Summary of the causal influences found during RL between the aIns, dlPFC, vmPFC and latOFC. The influence were found to take place only during punishment avoidance learning (valence loss, bold red arrow) but also during both forms of RL (saliency, dark arrow). The frequency in which these links occur in written on the arrow. The solid and dotted lines represent the statistical models used to estimate these links. **(B)** The aIns influences the dlPFC in the delta range (~2Hz) in the loss condition. **(C)** The aIns influences the latOFC in the beta range (~35Hz) in the loss condition. **(D)** The aIns influences the vmPFC in the theta/alpha range (~6 to 15Hz). **(E)** The latOFC influences the vmPFC in the alpha range (~15Hz). Bars are displayed below each link studied to show significant effects of the Granger Causality over time, using one model

or the other. Lights bars represent the non-parametric Wilcoxon test, while the dark bars represent the mixed model.

DISCUSSION

We reached several new conclusions regarding how reinforcement signals are dynamically encoded in the human brain during reward and punishment-based learning, thanks to simultaneous SEEG recordings in the anterior insula and prefrontal cortex and computational modelling during instrumental learning. We found that prediction errors are reliably encoded in only four brain areas: the vmPFC, aIns, dlPFC and latOFC. Further, we demonstrate that reward prediction error signals are expressed in the vmPFC, while punishment prediction error signals are processed by the aIns and dlPFC. In addition, Granger causality analyses revealed a leading role of AI during punishment-avoidance learning.

One might object that our SEEG data are from patients with a severe neurological disorder, but we used a within subject design (each patient was its own control) to reveal task-related responses that were highly consistent with both previous human neuroimaging and animal electrophysiological studies. A second issue concerns that SEEG has a very sparse spatial sampling of the cortex. But this is precisely why we recorded from multiple patients to provide reasonable brain coverage across patients. Yet, we acknowledge that some brain regions may not have reached significance (false negatives) because they were insufficiently sampled in the current study. A third weakness is that we focused data analyses on BGA, because of the strong relationship between BGA and both BOLD and population-level spiking activity (Lachaux et al. 2007; Manning et al. 2009; Mukamel et al. 2005). Despite this theoretical argument, we ran exploratory analyses in the time-frequency place to reveal that theta activity (5-10 Hz) also played an interesting role during reinforcement learning in most of the brain regions identified, even if the functional interpretation of this signal across ROIs was not as straight-forward (we found that theta amplitude increased after monetary losses in all four ROIs). Fourth, one might question our choice to use secondary reinforcers, which might activate different regions than primary reinforcers (Prevost et al. 2010) or be not incentive enough to activate the reward and punishment brain systems. Yet, our results showing a double dissociation between vmPFC and aIns-dlPFC broadband activity at the outcome onset clearly show that this was not the case. Lastly, monetary outcomes also have several advantages: for example, rewards and punishments are more directly comparable than in animal studies that directly contrast brain activity following air-puff and fruit-juice delivery. It also allowed us to easily match task-difficulty by using identical probabilistic contingencies. Contrasting directly neural activities following effective gains and losses also exclude the possibility that BGA that differentiates these experimental conditions reflect more general processes such as saliency (Kahnt et al. 2014).

Prediction errors were found to modulate broadband gamma activity in only four brain regions. This result cannot be accommodated by the limited brain sampling inherent to sEEG recordings since we were able to record at least n=10 contact-pairs across three patients in 23 brain parcels defined by MarsAtlas (Auzias et al. 2016). This anatomical selectivity of prediction error signals does not support the view that these signals generate ubiquitous and heterogeneous reward and punishment activations throughout the human brain (Ramayya et al. 2015; Vickery et al. 2011). In contrast, the current electrophysiological data-set support the idea that only a limited set of brain regions encode reward and punishment prediction errors (Behrens et al. 2008; Seymour et al. 2004; Rutledge et al. 2010; Pessiglione et al. 2006).

To our knowledge, we provide the first direct electrophysiological data showing that in humans, the vmPFC and aIns/dlPFC play opponent functional roles during reward-seeking and punishment-avoidance learning (Liu et al. 2011; Palminteri et al. 2015; Pessiglione et al. 2006; Seymour et al. 2005; Yacubian 2006). This is an important step in the literature, given the controversy that still exists on this topic. While a consensus has been reached on the neural substrate underlying positive reward prediction errors (Bartra et al. 2013; Matsumoto et al. 2016; Seymour et al. 2004; Pessiglione et al. 2006; Roesch & Olson 2004; Schultz 2016), the debate is still open concerning punishment prediction-error, and even negative reward prediction errors (Bayer & Glimcher 2005; Hosokawa et al. 2007; Plassmann et al. 2010; Fiorillo 2013; Matsumoto & Hikosaka 2009b; Morrison & Salzman 2009). Here, punishment predictions errors reliably modulated broadband gamma activity in the aIns and dlPFC. We also extend previous fMRI findings that demonstrated that punishment prediction errors are implemented by a circuit mainly composed of anterior insula, lateral and dorsomedial prefrontal cortex and amygdala (Roy et al. 2014; Liu et al. 2011; O'Doherty et al. 2001; Seymour et al. 2005; Yacubian 2006) and with animal electrophysiological demonstration that BGA in the monkey's lateral PFC is higher after behavioral errors (Skoblenick et al. 2016).

Our results suggest that both of the algebraic components of the prediction error modulated BGA. More precisely, we found that BGA distinguished effective monetary losses and gains while at stimulus onset BGA was also significantly modulated by learning conditions. The modulation of BGA at the onset of the stimuli likely reflect subjects' anticipation of possible financial outcome and replicates the pattern of BOLD activity found during an identical task (Pessiglione et al. 2006). Critically, Q-value signals were also modulating BGA. This is consistent with previous studies that examined both components (Behrens et al. 2008; Roy et al. 2014); interestingly, we found that aIns, dlPFC and latOFC encoded negatively a

chosen value signal: the lower the expected value, the higher BGA, whereas vmPFC encoded positively subjects' expectations (Lefebvre et al. 2017; Skvortsova et al. 2014; Bartra et al. 2013). The latencies at which both of these effects were observed approximately corresponded to the timing of the participants' choices (i.e. 1250 ms before outcome onset). The timing of the chosen value signals revealed by SEEG may reflect a post-decision neural signal arising right after the participant's decision. In the context of the probabilistic learning task used in this study, this signal anticipates the outcome that may be subsequently delivered and is therefore crucial. Our results reveals latencies that appear much later than previously found with a model-based EEG analysis, as early as 250 ms post-stimulus onset in frontal electrodes (see Figure 38 in supplementary materials) (Fischer & Ullsperger 2013). This discrepancy may come from important differences between scalp and intracerebral recording and also from different neurophysiological bases of event-related potentials used as a neural proxy in Fischer & Ullsperger (that are thought to reflect the summed activity of post-synaptic potentials) vs. broadband gamma activity (used in this study) which is more related to the spiking activity of the neural population surrounding intracerebral electrode contacts. The predictive encoding of expected value in the anterior insula is consistent with anticipatory response to aversive stimuli that were shown to play a critical role in shaping subject's decisions (Büchel et al. 1998; Caria et al. 2010; Knutson et al. 2014; Wiech et al. 2010).

BGA in the anterior insula after effective monetary losses was found to predict inter-subject's learning performances. This means that BGA in the anterior insular cortex not only signals punishment prediction errors: it may also contribute to drive behavioral adaptations to improve learning performances. While the study design was suboptimal to establish whether post-punishment BGA in the aIns was specifically predictive of punishment learning across subjects, both fMRI (Samanez-Larkin et al. 2008) and lesion (Palminteri et al. 2012) evidence established a critical role of this structure during punishment-avoidance learning. BGA after monetary losses was significantly higher during trials followed by a behavioral switch (Sridharan et al. 2008). This suggest that aIns might strengthen aversiveness memorization, mandatory for a quick change in behavior. Such a high-level cognitive function would likely require the recruitment of additional brain resources as a new task-set would have to be updated to allow a switch in the action-outcome association. The dlPFC would be a natural candidate as an additional recruited region as it has previously been identified as important for behavioral changes when a change of task-set and/or computational model is required during feedback processing (Dehaene & Changeux 2000; Tanaka et al. 2004; Jung et al. 2010; Smith et al. 2015). The scheme is supported by Granger causality analyses in the

present study, as they revealed bi-directional information flows between aIns and dlPFC that were significantly stronger after monetary losses compared to gain outcomes.

We used granger causality to examine how prediction error encoding brain regions interacted during reinforcement learning and whether information flows were modulated by outcome valence. We found that aIns had a causal influence on all three other ROIs. Previous neuroimaging studies have not investigated how network dynamics was influenced by outcome valence and learning. One exception suggests that feedback-related information propagates from medial to lateral prefrontal cortex (Smith et al. 2015), but the anterior insular cortex could not be recorded in that study. This is important since anterior insular cortex was found to be the input structure within the saliency network (Ham et al. 2013) and we also reported that during error processing, aIns was driving both dACC and pre-SMA (Bastin et al. 2016), as predicted by an earlier study that established the involvement of these three areas during negative outcome monitoring (Jung et al. 2010). The specific modulation of granger influences by negative outcome in our study is also consistent with the view that information flows between brain regions are modulated by the decisional context (Hunt et al. 2012; Hunt et al. 2015).

In summary, we found intracerebral evidence that two brain systems are differentially involved during reward vs. punishment based learning. Broadband gamma activity in the vmPFC may signal selectively reward prediction errors, in accordance with its involvement in expected positive value. Conversely, aIns and dlPFC may convey punishment prediction error signals, in accordance with the involvement of these areas in the monitoring of negative outcomes. Within the network of brains regions shown to encode prediction errors, we found a striking drive from anterior insula to the vmPFC, dlPFC and latOFC that was revealed by Granger causality analyses, suggesting that aIns could be the input of this system. However, we could only characterize a limited part of the network given the spatial sampling limitations inherent to sEEG recordings. Further research is needed to simultaneously record and stimulate critical brain regions such as the striatum (Hare et al. 2008) or the periaqueductal gray (Roy et al. 2014) to further examine the exact functional relevance of information flows between distant brain regions during reinforcement learning.

EXPERIMENTAL PROCEDURES

Patients

The study was approved by the Ethics Committee for Biomedical Research of Grenoble Alpes University Hospital (ISD-sEEG 2009-A00239-48). All participants had normal or corrected to normal vision and provided written informed consent. Intracerebral recordings were obtained from 20 neurosurgical patients with intractable epilepsy (ten females aged 29.7 +- 4.25, and 10 males aged 37.8 ± 3.77 years) at the Epilepsy Department of the Grenoble Alpes University Hospital (patients' demographics and clinical details are summarized in Table 1 in supplementary materials). To localize epileptic foci that could not be identified through noninvasive methods, neural activity was monitored in lateral, intermediate, and medial wall structures in these patients using stereotactically implanted multilead electrodes (stereotactic electroencephalography, sEEG). Electrode implantation was performed according to routine clinical procedures, and all target structures for the presurgical evaluation were selected strictly according to clinical considerations with no reference to the current study (Isnard et al. 2000).

Behavioral task

Patients performed a probabilistic instrumental learning task adapted from previous studies (Pessiglione et al. 2006; Palminteri et al. 2012). Patients were provided with written instructions, which were reformulated orally if necessary so as to clarify that the aim of the task was to maximize their financial payoff and that to do so, they had to consider reward seeking and punishment avoidance as equally important (Figure 31).

Patients performed short training sessions to familiarize with task's timing and responses. Training procedure comprised a very short session with only two pairs of cues presented during 16 trials that was followed by 2 to 3 short sessions of five minutes so that at the end of the training procedure, all patients reached a threshold of 70 % correct choices during both the reward and punishment conditions. During sEEG recordings, patients performed three to six test sessions after the training. Each session was an independent task containing four new pairs of cues to be learned. Cues were abstract visual stimuli taken from the Agathodaimon alphabet. Each pair of cues was presented 24 times for a total of 96 trials. The four cue pairs corresponded to the two conditions (2 pairs of gain and 2 pairs of loss cues), which were respectively associated with different pairs of outcomes (winning 1€ versus nothing or losing 1€ versus nothing). Within each pair, the two cues were associated

to the two possible outcomes with reciprocal probabilities (0.75/0.25 and 0.25/0.75). On each trial, one pair was randomly presented and the two cues were displayed on a computer screen on the left and right of a central fixation cross, their relative position being counterbalanced across trials. The subject was required to choose the left stimulus or the right stimulus by using their left or right index to press the corresponding button on a joystick (Logitech Dual Action). Since the position on screen was counterbalanced, response (left versus right) and value (good versus bad cue) were orthogonal. The chosen cue was colored in red for 250 ms and then the outcome was displayed on the screen. In order to win money, patients had to learn by trial and error the cue–outcome associations, so as to choose the most rewarding cue in the gain condition and the less punishing cue in the loss condition.

sEEG data acquisition and preprocessing

Five to seventeen semirigid, multilead electrodes were stereotactically implanted in each patient. Electrodes had a diameter of 0.8 mm and, depending on the target structure, contained 8–18 contact leads 2-mm-wide and 1.5-mm-apart (DIXI Medical Instruments). All electrode contacts were identified on each patient's individual postimplantation MRI. Each subject's individual preimplantation MRI was coregistered with the postimplantation MRI (Carmichael et al. 2008) to determine the anatomical location of each contact. The MarsAtlas (Auzias et al. 2016) was used to label each electrode as a function of individual gyri/sulci organization. This labeling was further confirmed by visually inspecting patients' MRI and identify anatomical landmarks well established in the literature (Craig 2009b; Afif & Mertens 2010; Afif et al. 2010; Ongür & Price 2000; Bechara et al. 1998; Ursu & Carter 2005; Kringelbach & Rolls 2004). sEEG contact located in the white matter were also removed from the analysis. Only the 35 regions with at least 10 recording sites across 3 patients were retained for statistical analyses.

sEEG data were bandpass-filtered online from 0.1 to 200 Hz and sampled at 128 Hz (1 patient), 256 Hz (1 patient), 512 Hz (6 patients) or 1024 Hz (12 patients), using a reference electrode located in white matter. Each electrode trace was subsequently re-referenced with respect to its direct neighbor (bipolar derivations with a spatial resolution of 3.5 mm) to achieve high local specificity by cancelling out effects of distant sources that spread equally to both adjacent sites through volume conduction (Lachaux et al. 2003). 45 to 161 bipolar contact-pairs were recorded in each patient using a commercial video-sEEG monitoring system (System Plus, Micromed). Overall, 2083 recording sites were obtained over the twenty patients included in this study.

Behavioral and computational analyses

Percentage of correct choices and reaction times were monitored across time for all sessions. For each pair of cues, the consistency of choices was also monitored after trials associated with salient outcomes (effective rewards or punishments). If these trials were followed by an identical choice of cue, they were referred as “stay trials” whereas if the other cue was chosen on the following trial of the pair, they were referred as “switch trials”.

A standard Q-learning algorithm has been used to model action-value learning and fit to the observed behavior. For each pair of stimuli A and B, the model estimates the expected value of outcome if choosing A (Qa) and B (Qb), according to previous choices and outcomes. The expected values of both stimuli were set at 0 prior to learning, being the mean outcome value. After each trial t>0, the expected value of the chosen stimuli (say A) was updated according to the rule $Qa(t+1)=Qa(t)+\alpha*\delta(t)$. The outcome prediction error, $\delta(t)$, is the difference between the obtained and the expected outcomes, $\delta(t)=R(t)-Qa(t)$, with R(t) the reinforcement obtained among -1€, 0€ and -1€. Using the estimated values of reinforcement obtained after choosing each stimuli, the probability (or likelihood) of choosing each stimuli was estimated using the softmax rule: $Pa(t) = \exp[Qa(t)/\beta] / \{\exp[Qa(t)/\beta] + \exp[Qb(t)/\beta]\}$. The constant parameters alpha and beta respectively are the learning rate and the temperature (the likelihood of changing chosen stimulus after a surprise trial). They have been adjusted to fit to mean observed choices across the group so as to maximize the probability of the actual choices so that only two sets of parameters have been used across the group, corresponding to a specific α/β couple for each condition. Optimal parameters in the gain condition were $\alpha=0.2$ and $\beta= 0.35$ whereas in the loss conditions $\alpha=0.19$ and $\beta= 0.4$. Outcome prediction errors and Qvalues estimated by the model for each patient and each trial were then used as statistical regressors for sEEG data.

sEEG data analyses

Computation of single-trial broadband gamma envelopes. Broadband gamma activity (BGA) was extracted with the Hilbert transform of sEEG signals using custom Matlab scripts (Mathworks Inc., MA, USA) as follows. sEEG signals were first bandpass filtered in 10 successive 10-Hz-wide frequency bands (e.g. 10 bands, beginning with 50–60 Hz up to 140–150 Hz). For each bandpass filtered signal, we computed the envelope using standard Hilbert transform. The obtained envelope had a time resolution of 15.625 ms (64 Hz). Again,

for each band, this envelope signal (i.e. time-varying amplitude) was divided by its mean across the entire recording session and multiplied by 100 for normalization purposes. Finally, the envelope signals computed for each consecutive frequency bands (e.g. 10 bands of 10 Hz intervals between 50 and 150 Hz) were averaged together, to provide one single time-series (the BGA) across the entire session, expressed as percentage of the mean. This time-series were smoothed with a 250 ms sliding window to increase statistical power for inter-trial and inter-individual analyses of BGA dynamics.

Identification of sEEG sites responding to prediction errors. We used single-trial BGA responses and prediction error estimates from the computational model to identify sEEG contact-pairs that were significantly modulated by the magnitude of prediction errors. For each sEEG contact-pair, we extracted the time-course of BGA during a post outcome interval (-250 to 1000 ms, i.e. 81 time points) during each trial. We then correlated BGA at each time point to the estimated magnitude of single-trial prediction errors separately for reward prediction errors (RPE) and punishment prediction errors (PPE). This procedure provided a regression estimate per time point and per contact.

To estimate the statistical significance for each contact-pair ad each time-point, we produced a random distribution of surrogate regression coefficients for each sEEG contact by shuffling prediction errors across trials 1000 times for each time point and contact-pair (Maris & Oostenveld 2007). We then used this distribution to estimate the threshold above which correlation coefficients were significant. The significance threshold was set at p=0.01 (controlled for multiple comparisons). Once all responsive contacts were identified, a correction for clusters of response across time was performed across all responsive contacts. For this, a new random distribution of surrogate regression coefficients was obtained by shuffling prediction errors across trials 500 times for each contact at each time-point. The global surrogate distribution across all contacts of a given ROI went through a cluster correction process where 60000 random values were isolated to create surrogate clusters to be compared to actual clusters of response in our experimental data. Finally to test whether the parcel was primarily related to RPE or PPE encoding or both, a parametric 2-tailed paired t-test (p=0.01) corrected for multiple comparisons across time using the false discovery rate (Genovese et al. 2002) was applied over all contacts within a given parcel exhibiting both RPE and PPE encoding at each time-point.

Regions of interest encoding prediction errors were defined a posteriori as brain regions associated with PE encoding across at least 10 task-responsive sEEG contact-pairs across at least 3 patients. Note that to increase the number of exploitable regions of interest, we

pulled together small neighboring MarsAtlas parcels (e.g. PFrdls and PFrdli under the label ‘dlPFC’).

Once the responsive contacts identified for each ROI, gamma activity during the 1000ms following the cue (in gain and loss conditions) and the outcome delivery (+1€, +0€, -0€ and -1€) were compared (repeated measures ANOVA) to test whether these PE encoding where triggered by reinforcement magnitude or by expected values or both.

To test whether these responses to outcome delivery influenced the behavior, a linear regression has been performed between gamma responses to outcomes in the gain condition (+1€ and +0€) or in the loss condition (-0€ and -1€) respectively with the performance in the gain condition or in the loss condition based on the type of PE encoding exhibiting by the given ROI (RPE or PPE). In addition, the peak latency, peak magnitude and mean amplitude of the gamma response in the first 1000ms following the outcome delivery were compared (parametric 2-tailed paired t-test, p=0.05) between stay and switch trials. For RPE encoding ROIs, “switch” and “stay” trials were identified following rewarded trials (+1€) while for the PPE encoding ROIs, “switch” and “stay” trials were identified following punished trials (-1€).

Granger causality

In order to study the causality of one brain response onto another, a Granger Causality (GC) analysis has been performed on all recorded contacts within the four regions of interest in all patients implanted in at least 2 ROI (13 patients): 58 in the aIns (8 patients), 81 in the dlPFC (8 patients), 67 in the vmPFC (11 patients) and 91 in the latOFC (12 patients). For each pair of region, a theoretical link was studied (6 links in total). The first method was to compute a non-parametric Wilcoxon test, then correct the p-values using FDR $q<0.05$. The results of this analysis are presented in the Figure 36B under the label ‘Wilcoxon’ (light bars used to show significant effects). The second method was based on the use of a mixed model (label ‘wilcoxon & lme’ and dark bars in Figure 36B). For this model, the subject ID was used as a grouping factor as this allows to rule out differences driven by only one subject. Then, the p-values were corrected using FDR $q<0.05$. The consistent results between both methods are presented on the diagram in Figure 36A. For each link, various contrasts were studied to highlight valence related effects (gain condition vs loss condition) and simple signed effects (reward vs neutral in gain condition, punishment vs neutral in loss condition). Saliency related effects were revealed by significant signed effects in both gain and loss conditions but with no significant valence effect (see Table 3 in supplementary materials).

REFERENCES

Afif, A., and Mertens, P. 2010. Description of sulcal organization of the insular cortex. Surg. Radiol. Anat. SRA *32*, 491–498.

Afif, A., Minotti, L., Kahane, P., and Hoffmann, D. 2010. Anatomofunctional organization of the insular cortex: a study using intracerebral electrical stimulation in epileptic patients. Epilepsia *51*, 2305–2315.

Auzias, G., Coulon, O., and Brovelli, A. 2016. *MarsAtlas* : A cortical parcellation atlas for functional mapping: MarsAtlas. Hum. Brain Mapp. *37*, 1573–1592.

Bartra, O., McGuire, J.T., and Kable, J.W. 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage *76*, 412–427.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., Hoffman, D., Combrisson, E., Kujala, J., Perrone-Bertolotti, M., Kahanae, P., Lachaux, JP., Jerbi, K. 2016. Direct Recordings from Human Anterior Insula Reveal its Leading Role within the Error-Monitoring Network. Cereb. Cortex.

Bayer, H.M., and Glimcher, P.W. 2005. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron *47*, 129–141.

Bechara, A., Damasio, H., Tranel, D., and Anderson, S.W. 1998. Dissociation of working memory from decision making within the human prefrontal cortex. J. Neurosci.*18*, 428–437.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. 2008. Associative learning of social value. Nature *456*, 245–249.

Brischoux, F., Chakraborty, S., Brierley, D.I., and Ungless, M.A. 2009. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc. Natl. Acad. Sci. *106*, 4894–4899.

Büchel, C., Morris, J., Dolan, R.J., and Friston, K.J. 1998. Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron *20*, 947–957.

Caria, A., Sitaram, R., Veit, R., Begliomini, C., and Birbaumer, N. 2010. Volitional Control of Anterior Insula Activity Modulates the Response to Aversive Stimuli. A Real-Time Functional Magnetic Resonance Imaging Study. Biol. Psychiatry *68*, 425–432.

Carmichael, D.W., Thornton, J.S., Rodionov, R., Thornton, R., Mcevoy, A., Allen, P.J., and Lemieux, L. 2008. Safety of Localizing Epilepsy Monitoring Intracranial Electroencephalograph Electrodes Using MRI : Radiofrequency-Induced Heating. *1244*, 1233–1244.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88.

Craig, A.D.B. 2009. How do you feel--now? The anterior insula and human awareness. Nat. Rev. Neurosci. *10*, 59–70.

Dehaene, S., and Changeux, J.-P. 2000. Reward-dependent learning in neuronal networks for planning and decision making. pp. 217–229.

Fiorillo, C.D. 2013. Two Dimensions of Value: Dopamine Neurons Represent Reward But Not Aversiveness. Science *341*, 546–549.

Fischer, A.G., and Ullsperger, M. 2013. Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. Neuron *79*, 1243–1255.

Genovese, C.R., Lazar, N.A., and Nichols, T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage *15*, 870–878.

Ginther, M.R., Bonnie, R.J., Hoffman, M.B., Shen, F.X., Simons, K.W., Jones, O.D., and Marois, R. 2016. Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. J. Neurosci. Off. J. Soc. Neurosci. *36*, 9420–9434.

Ham, T., Leff, A., de Boissezon, X., Joffe, A., and Sharp, D.J. (2013). Cognitive Control and the Salience Network: An Investigation of Error Processing and Effective Connectivity. J. Neurosci. *33*, 7091–7098.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., and Rangel, A. 2008. Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. J. Neurosci. *28*, 5623–5630.

Hosokawa, T., Kato, K., Inoue, M., and Mikami, A. 2007. Neurons in the macaque orbitofrontal cortex code relative preference of both rewarding and aversive outcomes. Neurosci. Res. *57*, 434–445.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F.S., and Behrens, T.E.J. 2012. Mechanisms underlying cortical activity during value-guided choice. Nat. Neurosci.

15, 470–476.

Hunt, L.T., Behrens, T.E., Hosokawa, T., Wallis, J.D., and Kennerley, S.W. 2015. Capturing the temporal evolution of choice across prefrontal cortex. Elife *4*, e11945.

Isnard, J., Guénot, M., Ostrowsky, K., Sindou, M., and Mauguière, F. 2000. The role of the insular cortex in temporal lobe epilepsy. Ann. Neurol. *48*, 614–623.

Jung, J., Jerbi, K., Ossandón, T., Ryvlin, P., Isnard, J., Bertrand, O., Guénot, M., Mauguière, F., and Lachaux, J.-P. 2010. Brain responses to success and failure: Direct recordings from human cerebral cortex. Hum. Brain Mapp. *31*, 1217–1232.

Kahnt, T., Park, S.Q., Haynes, J.-D., and Tobler, P.N. 2014. Disentangling neural representations of value and salience in the human brain. Proc. Natl. Acad. Sci. *111*, 5000–5005.

Kim, H., Shimojo, S., and O'Doherty, J.P. 2006. Is Avoiding an Aversive Outcome Rewarding? Neural Substrates of Avoidance Learning in the Human Brain. PLoS Biol. *4*, e233.

Kirsch, P., Schienle, A., Stark, R., Sammer, G., Blecker, C., Walter, B., Ott, U., Burkart, J., and Vaitl, D. 2003. Anticipation of reward in a nonaversive differential conditioning paradigm and the brain reward system: NeuroImage *20*, 1086–1095.

Knutson, B., Katovich, K., and Suri, G. 2014. Inferring affect from fMRI data. Trends Cogn. Sci. *18*, 422–428.

Kringelbach, M.L., and Rolls, E.T. 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog. Neurobiol. *72*, 341–372.

Lachaux, J.-P., Chavez, M., and Lutz, A. 2003. A simple measure of correlation across time, frequency and space between continuous brain signals. J. Neurosci. Methods *123*, 175–188.

Lachaux, J.-P., Fonlupt, P., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., and Baciu, M. 2007. Relationship between task-related gamma oscillations and BOLD signal: New insights from combined fMRI and intracranial EEG. Hum. Brain Mapp. *28*, 1368–1375.

Lammel, S., Lim, B.K., Ran, C., Huang, K.W., Betley, M.J., Tye, K.M., Deisseroth, K., and Malenka, R.C. 2012. Input-specific control of reward and aversion in the ventral

tegmental area. Nature *491*, 212–217.

Larsen, T., and O'Doherty, J.P. 2014. Uncovering the spatio-temporal dynamics of value-based decision-making in the human brain: a combined fMRI-EEG study. Philos. Trans. R. Soc. B Biol. Sci. *369*, 20130473–20130473.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. 2017. Behavioural and neural characterization of optimistic reinforcement learning. Nat. Hum. Behav. *1*, 0067.

Liu, X., Hairston, J., Schrier, M., and Fan, J. 2011. Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev. *35*, 1219–1236.

Manning, J.R., Jacobs, J., Fried, I., and Kahana, M.J. 2009. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. J. Neurosci. *29*, 13613–13620.

Matsumoto, M., and Hikosaka, O. 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.

Matsumoto, H., Tian, J., Uchida, N., and Watabe-Uchida, M. 2016. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. eLife *5*, e17328.

Mizuhiki, T., Richmond, B.J., and Shidara, M. 2012. Encoding of reward expectation by monkey anterior insular neurons. J. Neurophysiol. *107*, 2996–3007.

Morrison, S.E., and Salzman, C.D. 2009. The Convergence of Information about Rewarding and Aversive Stimuli in Single Neurons. J. Neurosci. *29*, 11471–11483.

Morrison, S.E., Saez, A., Lau, B., and Salzman, C.D. 2011. Different Time Courses for Learning-Related Changes in Amygdala and Orbitofrontal Cortex. Neuron *71*, 1127–1140.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. 2005. Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science *309*, 951–954.

Niessing, J., Ebisch, B., Schmidt, K.E., Niessing, M., Singer, W., and Galuske, R.A.W. 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. Science *309*, 948–951.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. Nat. Neurosci. *4*, 95–102.

O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. 2004. Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Ongür, D., and Price, J.L. 2000. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. Cereb. Cortex *10*, 206–219.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., Pessiglione, M. 2012. Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. Neuron *76*, 998–1009.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. 2015. Contextual modulation of value signals in reward and punishment learning. Nat. Commun. *6*, 8096.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Plassmann, H., O'Doherty, J.P., and Rangel, A. 2010. Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. J. Neurosci. *30*, 10799–10808.

Prevost, C., Pessiglione, M., Metereau, E., Clery-Melin, M.-L., and Dreher, J.-C. 2010. Separate Valuation Subsystems for Delay and Effort Decision Costs. J. Neurosci. *30*, 14080–14090.

Ramayya, A.G., Pedisich, I., and Kahana, M.J. 2015. Expectation modulates neural representations of valence throughout the human brain. NeuroImage *115*, 214–223.

Roesch, M.R., and Olson, C.R. 2004. Neuronal Activity Related to Reward Value and Motivation in Primate Frontal Cortex. Science *304*, 307–310.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., and Wager, T.D. 2014. Representation of aversive prediction errors in the human periaqueductal gray. Nat. Neurosci. *17*, 1607–1612.

Rutledge, R.B., Dean, M., Caplin, A., and Glimcher, P.W. 2010. Testing the Reward

Prediction Error Hypothesis with an Axiomatic Model. J. Neurosci. *30*, 13525–13536.

Samanez-Larkin, G.R., Hollon, N.G., Carstensen, L.L., and Knutson, B. 2008. Individual differences in insular sensitivity during loss anticipation predict avoidance learning. Psychol. Sci. *19*, 320–323.

Schultz, W. 2016. Dopamine reward prediction-error signalling: a two-component response. Nat. Rev. Neurosci.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. 2004. Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. *8*, 1234–1240.

Skoblenick, K.J., Womelsdorf, T., and Everling, S. 2016. Ketamine Alters Outcome-Related Local Field Potentials in Monkey Prefrontal Cortex. Cereb. Cortex *26*, 2743–2752.

Skvortsova, V., Palminteri, S., and Pessiglione, M. 2014. Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. J. Neurosci. *34*, 15621–15630.

Smith, E.H., Banks, G.P., Mikell, C.B., Cash, S.S., Patel, S.R., Eskandar, E.N., and Sheth, S.A. 2015. Frequency-Dependent Representation of Reinforcement-Related Information in the Human Medial and Lateral Prefrontal Cortex. J. Neurosci. Off. J. Soc. Neurosci. *35*, 15827–15836.

Sridharan, D., Levitin, D.J., and Menon, V. 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Natl. Acad. Sci. *105*, 12569–12574.

Sutton, R.S., and Barto, A.G. 1998. Reinforcement Learning: an introduction.

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

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. 2007. The neural basis of loss aversion in decision-making under risk. Science *315*, 515–518.

Ursu, S., and Carter, C.S. 2005. Outcome representations, counterfactual comparisons and the human orbitofrontal cortex: Implications for neuroimaging studies of decision-making. Cogn. Brain Res. *23*, 51–60.

Vickery, T.J., Chun, M.M., and Lee, D. 2011. Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. Neuron *72*, 166–177.

Wiech, K., Lin, C. -s., Brodersen, K.H., Bingel, U., Ploner, M., and Tracey, I. 2010. Anterior Insula Integrates Information about Salience into Perceptual Decisions about Pain. J. Neurosci. *30*, 16324–16331.

Yacubian, J. 2006. Dissociable Systems for Gain- and Loss-Related Value Predictions and Errors of Prediction in the Human Brain. J. Neurosci. *26*, 9530–9537.

SUPPLEMENTARY MATERIALS

Patient	Sex	Age	Handedness	Nb bipoles	Freq Ech	EZ	Age onset	AED
1	F	29	Right	98	128	L frontopolar	19	LCM + LMT
2	F	42	Right	104	512	L temporal	13	CLB + LEV + ZON
3	F	20	Right	98	512	L temporal	11	CLB + LCM + LMT
4	M	48	Left	98	512	R temporal	12	OXC
5	M	37	Right	100	1024	L OFC	4	LCM + NZP
6	F	13	Right	80	1024	R PM	10	CBZ + VPA
7	F	34	Right	80	1024	L HPC + insular	5	LCM + LEV
8	M	32	Right	96	1024	Bilateral temporal	21	LTG + PER
9	F	15	Right	91	1024	R temporal + insular	7	CLB + LEV
10	M	28	Right	155	256	R fronto-temporal	11	CBZ + LCM
11	M	20	Right	161	1024	G frontal + temporal (2)	14	LCM + LMT + ZON

12	M	41	Right	135	512	R temporal	32	CLB + GBP
13	F	15	Right	99	1024	L temporo-occipital	12	CLB + LCM + LMT
14	M	40	Right	110	1024	R temporal	5	CBZ + LCM
15	F	47	Right	101	1024	R insular + L frontal	7	LCM + TPM
16	F	34	Right	90	1024	R OFC + temporal	28	CBZ + LCM + LMT
17	M	57	Right	107	1024	L temporal	37	CBZ + LMT + PER
18	M	41	Left	45	512	L temporal		CBZ + CLB
19	M	29	Right	99	1024	R OFC	25	LCM + LMT
20	F	48	Right	136	512	R temporal	19	LCM + LEV + LMT

Table 1 : Demographical and clinical details of the patients. All patients recorded in the study (n=20). Epileptic zone (EZ) abbreviations used: orbitofrontal OFC, premotor PM, hippocampus HPC left L and right R. Antiepileptic drugs (AED): alprazolam APZ, carbamazepine CBZ, clobazam CLB, gabapentin GBP, lacosamide LCM, lamotrigine LMT, levetiracetam LEV, nitrazepam NZP, oxcarbazepine OXC, perampanel PER, topiramate TPM, sodium valproate VPA, zonisamide ZON.

	aIns	dlPFC	vmPFC	latOFC
aIns (41/80)		0.4881	0.0166	0.0038
dlPFC (34/81)	ns		0.0941	0.0264
vmPFC (15/67)	*	ns		0.8471
latOFC (18/91)	**	*	ns	

Table 2: Relevance of the ROI based of their proportion of responding contacts. All pairs of ROI were compared using a Fisher contingency test on their relative proportions of responding contacts amongst all implanted contacts (e.g. 41 responding for 80 implanted in the anterior insula).

ROI #1	Nb bipoles ROI #1	ROI #2	Nb bipoles ROI #2	Nb of patients	Contrast	Frequency
aIns	42	vmPFC	41	6	+1€ vs +0€ -1€ vs -0€	Theta/alpha 6 to 15Hz
aIns	45	latOFC	52	7	-1€ vs -0€ Loss > gain	Beta 35Hz
latOFC	86	vmPFC	67	11	-1€ vs -0€ Loss > gain	Alpha 15Hz
aIns	41	dlPFC	39	5	-1€ vs -0€	Delta 2Hz
dlPFC	39	aIns	41	5	-1€ vs -0€	Gamma

						60Hz

Table 3: Significant links according to Granger Causality analysis. The regions forming each link and the number of bipoles corresponding are given across patients. The frequency in which the interactions occur are referenced according to their name and frequency range.

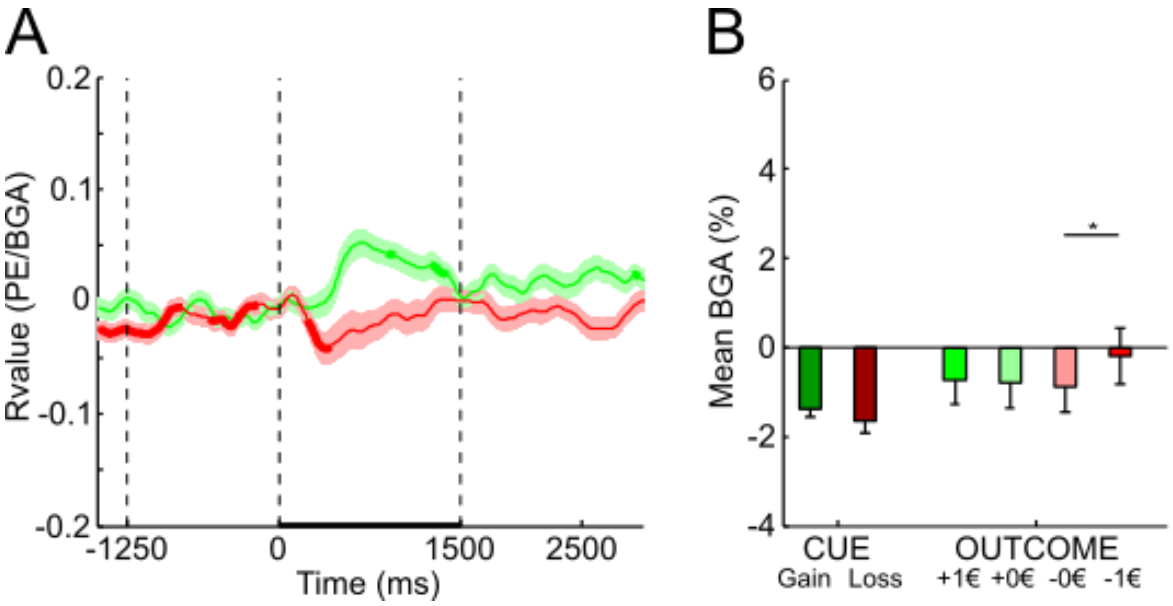


Figure 37: Gamma activity encodes saliency in the latOFC. (A) Average latOFC BGA correlation with prediction errors in gain (RPE, green) and loss (PPE, red) condition following outcome delivery (n=88 task-responsive contacts recorded from 12 patients). Mean PE encoding ± SEM. Bold lines represent time points with significatively different RPE from PPE encoding (p=0.01 cluster corrected). **(B)** Average latOFC BGA responses in the 1000ms following cue and outcome display in gain and loss condition (n=88 task-responsive contacts recorded from 12 patients).

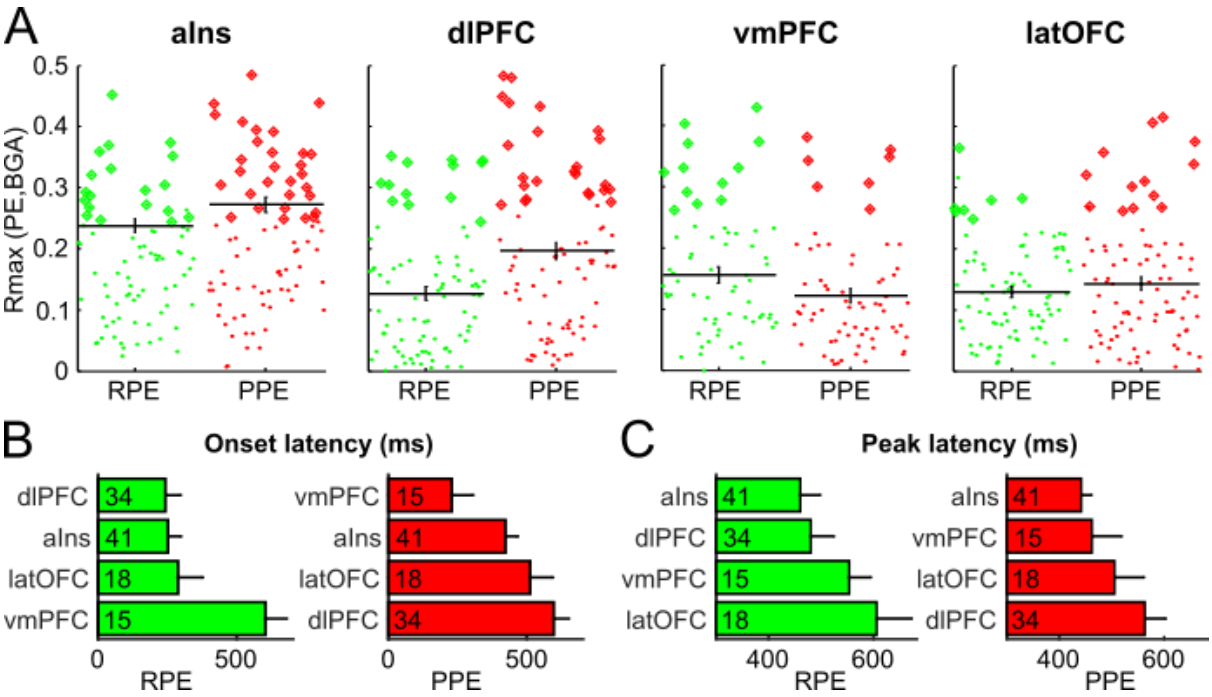


Figure 38: Distribution of RPE and PPE encoding in all 4 ROI. (A) Pattern of RPE (green) and PPE (red) encoding of all contacts within the 4 ROI. Mean PE encoding ± SEM. Diamonds represent contacts with significant PE encoding (p=0.01 cluster corrected). aIns = 80 implanted contacts recorded from 11 patients. dlPFC = 81 implanted contacts recorded from 8 patients. vmPFC = 67 implanted contacts recorded from 12 patients. latOFC = 91 implanted contacts recorded from 11 patients. **(B)** Onset latencies of RPE (green) and PPE (red) significant encoding of task-responsive contacts within the 4 ROI. **(C)** Peak latencies of RPE (green) and PPE (red) significant encoding of task-responsive contacts within the 4 ROI. **(B and C)** Mean latencies ± SEM. aIns = 41 task-responsive contacts recorded from 11 patients. dlPFC = 34 task-responsive contacts recorded from 8 patients. vmPFC = 15 task-responsive contacts recorded from 9 patients. latOFC = 18 task-responsive contacts recorded from 5 patients.

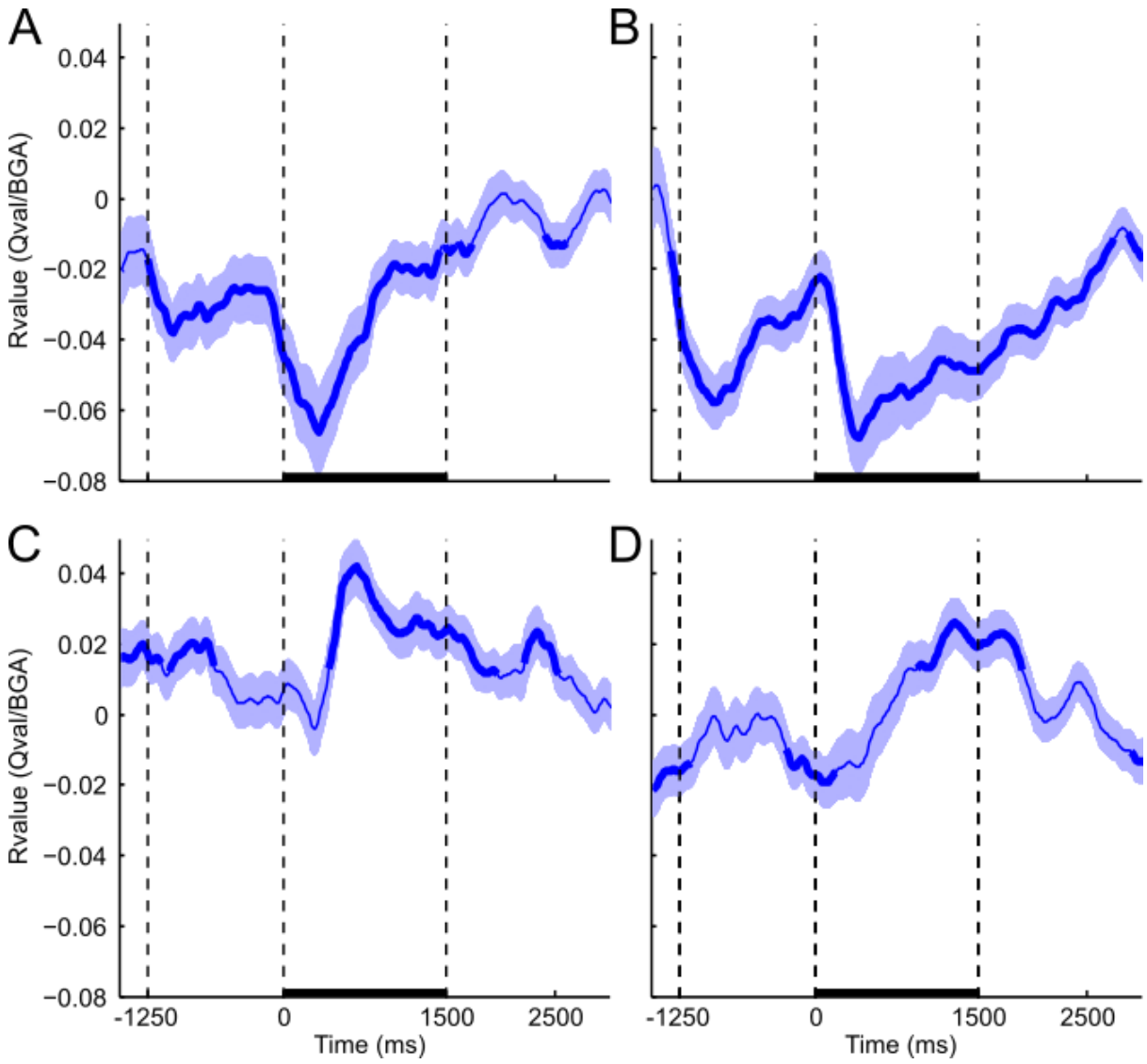


Figure 39: Gamma activity encoding of subjective values (Qval) during reinforcement learning. (A) Average aIns BGA correlation with Qvalues across all trials (blue) following outcome delivery (bold dark portion of x axis) (n=80 implanted contacts recorded from 11 patients). Mean Qval encoding ± SEM across contacts (shaded areas). Bold lines represent time points at which the Rvalue of Qvsal encoding in gamma reached significance ($p<0.01$ cluster corrected). **(B)** Average dlPFC BGA correlation with Qvalues across all trials (blue) following outcome delivery (n=81 implanted contacts recorded from 8 patients). **(C)** Average vmPFC BGA correlation with Qvalues across all trials (blue) following outcome delivery (n=67 implanted contacts recorded from 11 patients). **(D)** Average latOFC BGA correlation with Qvalues across all trials (blue) following outcome delivery (n=91 implanted contacts recorded from 11 patients). Graphical conventions are identical for all panels.

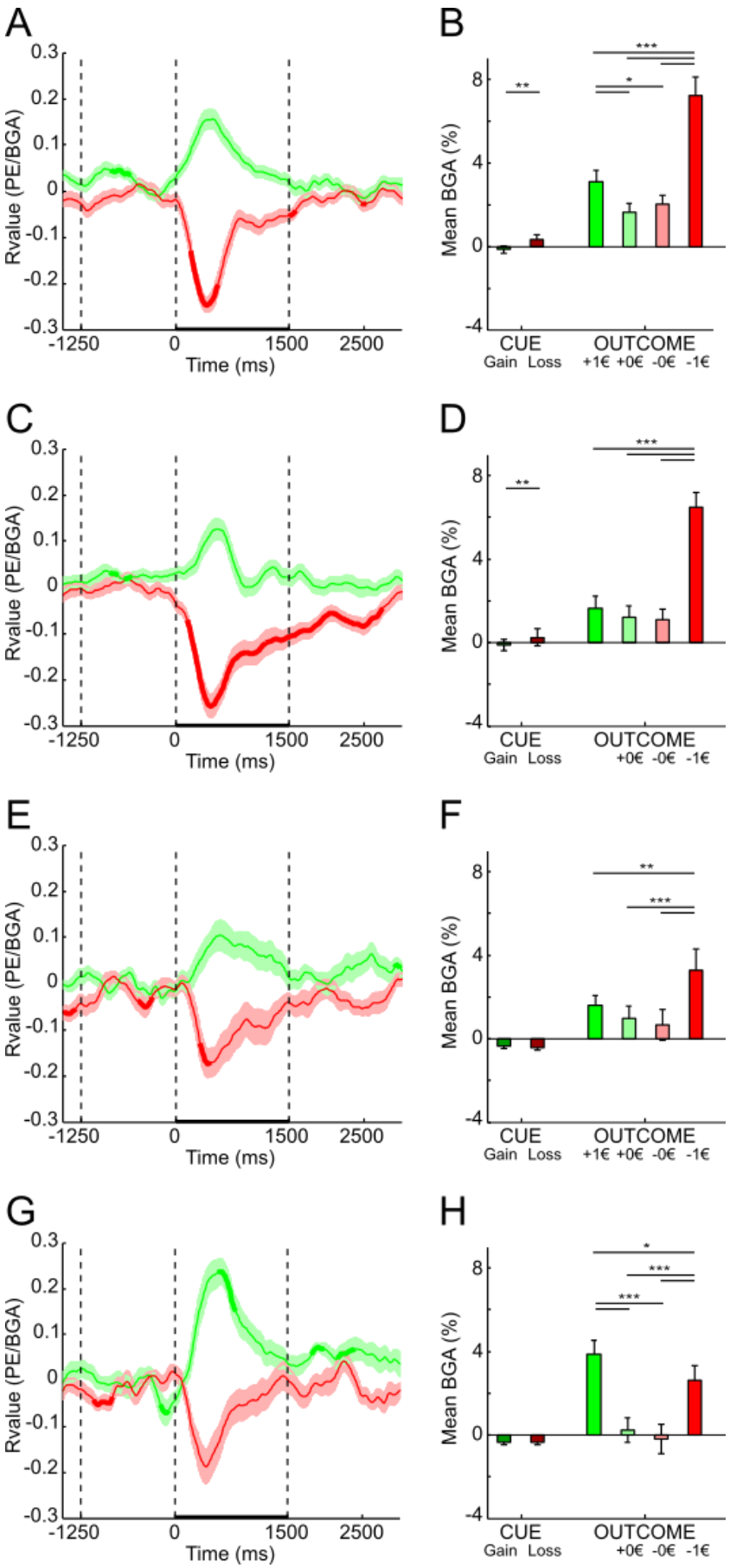

A
B
C
D
E
F
G
H
Rvalue (PE/BGA)
Time (ms)
-1250
0
1500
2500
Mean BGA (%)
CUE
Gain Loss
OUTCOME
+1€ +0€ -0€ -1€
+0€ -0€ -1€

Figure 40: Different encodings by gamma activity of the task-responsive contacts implanted in the aIns, dlPFC, latOFC and vmPFC. (A,C,E,G) Average BGA correlation with prediction errors in gain (RPE, green) and loss (PPE, red) condition following outcome delivery. Mean PE encoding ± SEM. Bold lines represent time points with a significant difference between RPE and PPE encoding (p=0.01 cluster corrected). **(B,D,F,H)** Average BGA responses in the 800ms following cue display, response and outcome display in gain and loss condition. **(A,B)** Average aIns BGA responses and correlation (n=41 task-responsive contacts recorded from 11 patients). **(C,D)** Average dlPFC BGA responses and correlation (n=34 task-responsive contacts recorded from 11 patients). **(E,F)** Average latOFC BGA responses and correlation (n=18 task-responsive contacts recorded from 9 patients). **(G,H)** Average vmPFC BGA responses and correlation (n=15 task-responsive contacts recorded from 5 patients).

B. Etude 2 : Dynamique sous-corticale de l'apprentissage par renforcement

Theta power is modulated in the human limbic thalamus during instrumental learning

Running title: Thalamic activity and reinforcement learning

Maëlle CM Gueguen[1,2], Jean-Philippe Lachaux[4], Lorella Minotti[3], Vincent Navarro[5,6], Philippe Kahane[2,4], Pablo Billeke[7], Mathias Pessiglione[5,8] and Julien Bastin[1,2*]

[1] Univ. Grenoble Alpes, F-38000 Grenoble, France.

[2] Inserm, U1216, F-38000 Grenoble, France.

[3] Neurology Department, CHU de Grenoble, Hôpital Michallon, F-38000 Grenoble, France.

[4] DYCOG Lab, Lyon Neuroscience Research Center, INSERM U1028, UMR 5292, University Lyon I, Lyon, France.

[5] ICM, INSERM UMRS 1127, CNRS UMR 7225, Université Pierre et Marie Curie UPMC-Paris 6, Paris, France

[6] Assistance Publique-Hôpitaux de Paris, Groupe Hospitalier Pitié-Salpêtrière, Paris, France

[7] División de Neurociencia, Centro de Investigación en Complejidad Social (NeuroCICS), Facultad de Gobierno, Universidad del Desarrollo, Santiago, Chile

[8] Motivation, Brain and Behavior team, Institut du Cerveau et de la Moelle épinière (ICM), Paris, France

*Correspondence to: julien.bastin@univ-grenoble-alpes.fr (J.B.); Phone: (0033) 4 56 52 06 78; Fax: (0033) 4 56 52 05 98; Institut des Neurosciences de Grenoble, Bâtiment Edmond J. Safra des Neurosciences, Chemin Fortuné Ferrini, Université Joseph Fourier, Site Santé La Tronche, BP 170 38042 Grenoble Cedex 9, France.

ABSTRACT

The dorsomedial (DMTN) and anterior thalamic nuclei (ATN) are thought to play a pivotal role during reinforcement learning, which is known to rely on a limbic ventral prefronto-striatal-thalamic circuit. Yet, the exact functional role of these thalamic nuclei during reward and punishment-based learning is unknown. To clarify this issue, we provide rare intracerebral stereotaxic electroencephalographic (sEEG) data acquired during reinforcement learning in five patients with intrathalamic depth electrodes implanted to treat severe refractory epilepsy. We demonstrate a modulation of theta activity related to outcome processing independently from its valence. Furthermore, a relative increase of beta oscillatory power was also found in both ATN and DMTN after monetary losses when directly compared to monetary gains, demonstrating a key role for the limbic thalamus in human instrumental learning processes.

INTRODUCTION

The dorsomedial (DMTN) and anterior thalamic nuclei (ATN) are believed to play a critical role in the ventral prefronto-striatal-thalamic limbic circuit central to instrumental learning, enabling one to dynamically update stimulus-action-response associations in order to produce flexible behaviors (Alcaraz et al. 2015; Balleine et al. 2015; Chakraborty et al. 2016; Gabriel et al. 1989). However, the exact functional role of these thalamic nuclei during reinforcement learning remains unclear in humans given the lack of circumscribed lesion data associated with the fact that noninvasive neuroimaging methods cannot distinguish signals that originate from small and deep areas within the thalamus. These limitations are important since animal studies suggest that ATN and DMTN could play dissociable roles during instrumental learning (Corbit et al. 2003). To clarify this issue, we had the exceptional opportunity to record intracerebral stereotaxic electroencephalographic (sEEG) data during reinforcement learning in four patients with intrathalamic depth electrodes implanted to treat severe refractory epilepsy (Fisher et al. 2010).

Interestingly, ATN and DMTN are both connected to the medial prefrontal cortex and to the cingular and insular cortices (Vertes et al. 2015). These cortical areas are thought to play critical roles during reward and punishment-based learning (Bartra et al. 2013; O'Doherty et al. 2001; Seymour et al. 2005). While human neuroimaging studies robustly found fMRI signals within the thalamus to encode the difference between predicted and actual outcomes (Chase et al. 2015; Garrison et al. 2013), the emerging view that limbic thalamic nuclei within this structure have dissociable roles during reinforcement learning is mainly supported by animal electrophysiological and lesion studies (Bradfield et al. 2013; Mitchell 2015). Neuronal activity in the ATN increased during fear conditioning (Conejo et al. 2007). Moreover, it was shown play a causal role during an aversive avoidance task (Gabriel et al. 1989) whereas instrumental behavior was unaffected by ATN lesion when outcomes were appetitive and when response-outcome associations had to be learn (Corbit et al. 2003). These data suggest that ATN could either be involved during pavlovian (rather than instrumental) processes, even if it remains to be firmly established, since an alternative account is that ATN could also be preferentially involved in instrumental learning contexts associated with aversive outcomes. This specific involvement of ATN during the processing of negative events is also supported by deep brain stimulation of ATN that slow-down patient's reaction time when they process threat-related information (Sun et al. 2015). The strong connections existing between DMTN and the medial prefrontal cortex may partly explain the accumulating evidence regarding its importance during instrumental behavior, (Parnaudeau et al. 2013;

Parnaudeau et al. 2015; Chakraborty et al. 2016; Corbit et al. 2003). The precise mechanism by which NAT and DMTN would modulate neural activity in the prefronto-striato-thalamic loops underlying reinforcement learning is unknown, but a possibility is that its oscillatory activity in the theta range could underlie modulations of information flows in this network (Ketz et al. 2015; Wright et al. 2015).

Together, these findings suggest that DMTN and ATN might be differentially involved during reinforcement learning processes in humans. We hypothesized that electrophysiological activity within these thalamic nuclei would reflect the psrocessing of prediction error signals and support efficient behavioral learning.

RESULTS

Intrathalamic EEG data were collected from five patients with intractable epilepsy (see demographical details in Table 4 in supplementary materials, and methods) while they performed an instrumental learning task during which reward and punishment conditions were matched in difficulty, as the same probabilistic contingencies were to be learned.

Behavior

Patients were able to learn the correct response over the 24 trials of a learning session. They tended to choose the most rewarding cue in the gain condition while they were also prone to avoid the most punishing cue in the loss condition (Figure 41); one-sample t-test versus chance (50%): global performance of 65.6% in gain vs 58.2% in loss, p=0.0002 and t(28)=4.2332 for gain ; p=0.0029 and t(28)=3.2610 for loss). Thus, as in previous studies in healthy subjects (Pessiglione et al. 2006; Palminteri et al. 2012), patients were able to choose the correct option when they had to learn from reward or from punishments. Interestingly, we also found that patients exhibited a symmetrical learning as they were as good at seeking rewards as at avoiding punishments (two-sample t-test: p=0.1025, t(56)=1.6601).

Patients also exhibited shorter reaction times in the gain condition than in the loss condition (1224ms vs 1746ms, two-sample t-test, p<0.0001, t(28)=5.268). However, analysis of individual performance profiles showed that only patient 1 and 2, and patient 5 to a lesser extent, managed to learn the task contingencies relatively well (Figure 41C, see Table 8 in supplementary materials). Nonetheless, when studying the 5 patients together, we found significant learning rates across the group compared to chance level (repeated measures ANOVA with a post-hoc Student-Newman-Keuls test: p=0.0406, F=2.393). For this reason, group behavioral analyses have been done on the 5 patients.

Based on these behavioral results that did not provide sufficient support in favor of consistent learning across subjects, we therefore restrained sEEG analysis to the characterization of feedback processing which interpretation did not rely on learning performance. Due to the very small number of patients (3 patients) having performed well enough during the cognitive task, we did not use a computational Qlearning model to predict the behavioral data we observed.

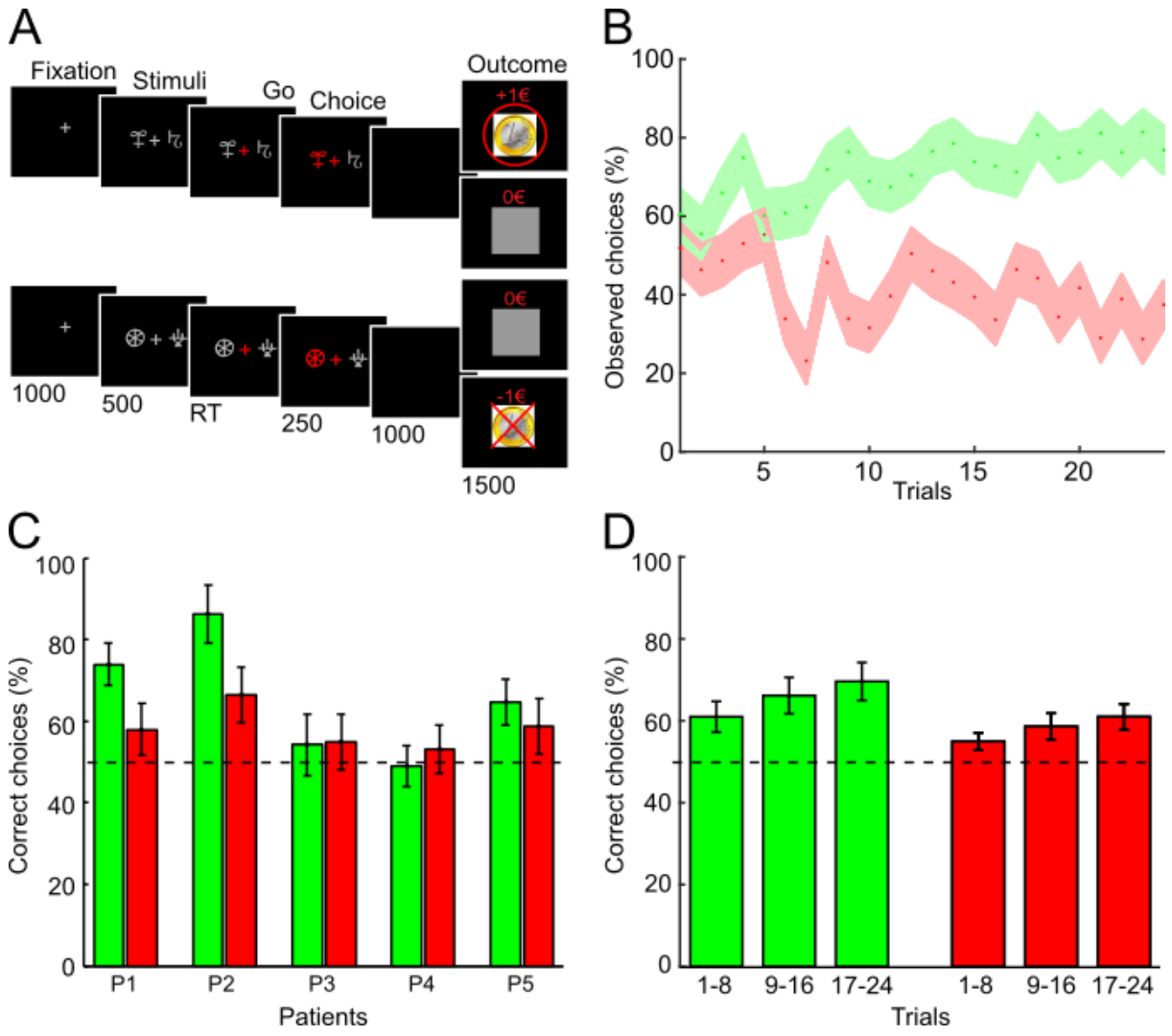


Figure 41: Behavioral task and results. (A) Successive screen of typical trials in the reward (top) and loss (bottom) conditions. Patients had to select one abstract visual stimulus among the two presented on each side of a central visual fixation cross, and subsequently observed the outcome. Duration are given in milliseconds. **(B)** Average learning curves (n=5 patients) (average: dots; 99% confidence interval: shaded areas). Learning curves portray trial by trial, the average proportion of trials across patients corresponding to subjects choosing the 'correct' stimulus in the gain condition (green dots), and the 'incorrect' stimulus in the loss condition (red dots). **(C)** Average performance (+/- SEM) of each patient across learning sessions in the gain (green) and loss (red) conditions. Average performance. Chance level is shown by the dotted black line. **(D)** Global performance of the group during early, middle and late learning in gain (green) and loss (red) conditions. Performance of trials 1 to 8, 9 to 16 and 17 to 24 was averaged, error bars indicate mean +/- SEM. Chance level is shown by the dotted black line. RT: reaction time.

Intracerebral electro-encephalography

The aim of data analysis was to explore both NAT and DMTN activities underlying learning to maximize rewards and/or learning to avoid punishments. Our objectives were (1) to identify regions of the thalamic nuclei that responded to outcome delivery during reinforcement learning, and (2) to highlight the possible behavioral effects of ANT-DBS on reinforcement learning.

We first performed a time-frequency analysis of the thalamic sEEG data. As the electrodes implanted sampled both DMTN and ATN parts of the limbic thalamus, we either averaged the neural activity across all contacts(Figure 43 upper line), as well as on the contact-pairs located in the ANT (Figure 43 middle line) or those located in the DMTN (Figure 43 lower line). The precise anatomical location of each contact was identified using the individual MRI images (see Figure 45 in supplementary materials) by the neurosurgeon in charge of the 3D reconstruction of the implantation for the whole STIC-France study. The electrophysiological results presented beslow are reported from only 4 patients over 5, as the data from the patient #4 have not been analyzed (yet).

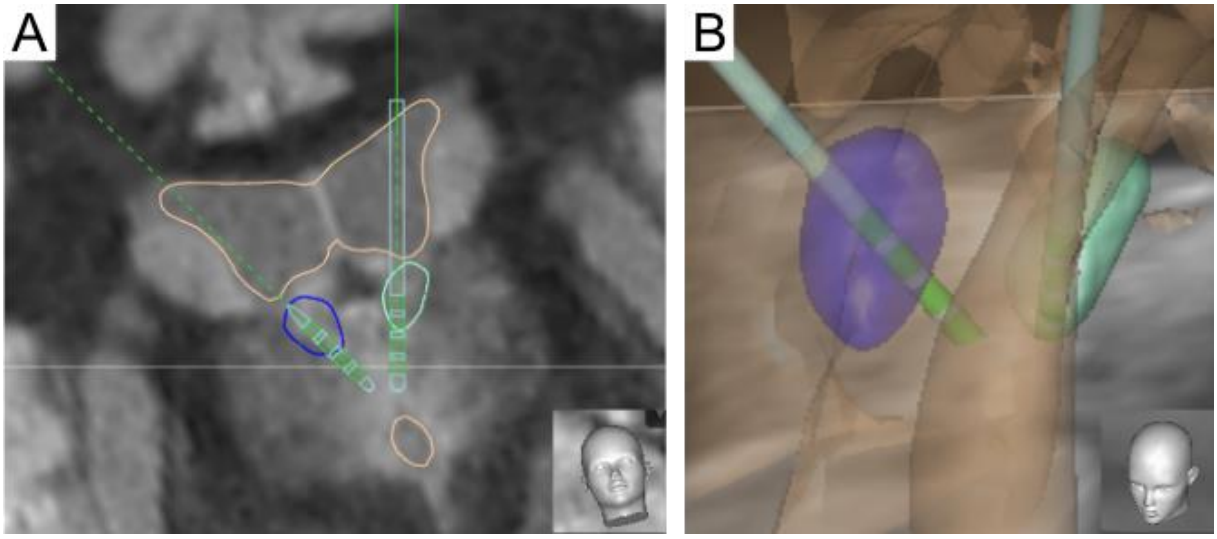


Figure 42 : Anatomical location of sEEG contacts within the limbic thalamic nucleus. Reconstruction of intrathalamic contact location using intraoperative X-ray image coordinates, superimposed on preoperative FGATIR MRI mask. Representative data from patient X: the upper contacts are clearly located within the ATN (segmented by the neurosurgeron for each patient) whereas the lower contacts were in the DMTN (see also Table 5 and Figure 45 in supplementary materials). The ventricular system/third ventricle is represented in orange, the anterior thalamic nuclei are in blue (right) and green (left), the

electrodes are shown in gray, their trajectories are in dotted green lines. **(A)** View on the MRI of the nucleus implantation, oriented alongside the electrode. **(B)** 3D reconstruction of the electrode within the nucleus. The frame of reference used to orient both reconstructions is indicated on the lower right corner of each panel.

Across all patients, the upper and most distal bipolar contacts have been located within the anterior thalamic nuclei (Figure 42, Table 5 and Figure 45 in supplementary materials) while the lower and most proximal bipolar contacts have been located within the dorsomedial nuclei of the thalamus. The results of the TF-analysis for these two selections of contacts are presented in the Figure 43 under the labels “NAT” and “DMTN” respectively.

We first averaged sEEG activity across all sEEG contact-pairs (“Whole NA”) to artificially increase the statistical power of these preliminary data obtained from four patients (patient #4 was not included in the sEEG analysis). We found a decrease of power in the theta/alpha band (4-13Hz) after the display of monetary outcomes (effective gain or loss). This decrease was maximal about 1s after outcome display (upper panels, statistical contrasts involving effective monetary gains or losses compared to the activity observed after the delivery of neutral outcomes: $p<0.01$, uncorrected for multiple comparisons). When we directly compared the neural activity during the gain versus the loss condition (top right panel), it appeared the power was significantly higher during the loss condition in the alpha (8-13Hz), beta (21-37Hz) and gamma (100-150Hz) bands ($p<0.01$). Interestingly, the latency at which these effects occurred appeared successively after the display of the outcome at around 500 ms (gamma), 700 ms (beta) and 900 ms (alpha).

When similar analyses were performed on ANT and DMTN, It turned out the NAT is the nucleus contributing the most to the late theta/alpha decrease (after 1s) in response to outcomes (middle left and central panels), while the DMTN showed an earlier theta decrease in response to aversive outcomes centered around 600ms post outcome onset (bottom central panel). The relative increase in gamma (500ms) and beta (700ms) following the outcome delivery in loss compared to gain trials seems to arise from the DMTN (bottom right panel).

Overall, the time-frequency analysis of the activity in the ANT and the DMNT during reinforcement learning suggest that a complex pattern of activity in theta/alpha (4-13Hz), beta (21-37Hz) and gamma (100-150Hz) was modulated by feedbacks. The responses to outcome delivery observed in the lowest frequencies were long-lasting (spanning approximately 1 second), while the responses to outcome delivery observed in higher bands

(beta and gamma) were more transient (i.e. lasted less than 500 ms). The responses to outcome arising from the ANT seem to be relying on low frequency activities, while those from the DMTN appear more complex with both low and high frequency influences.

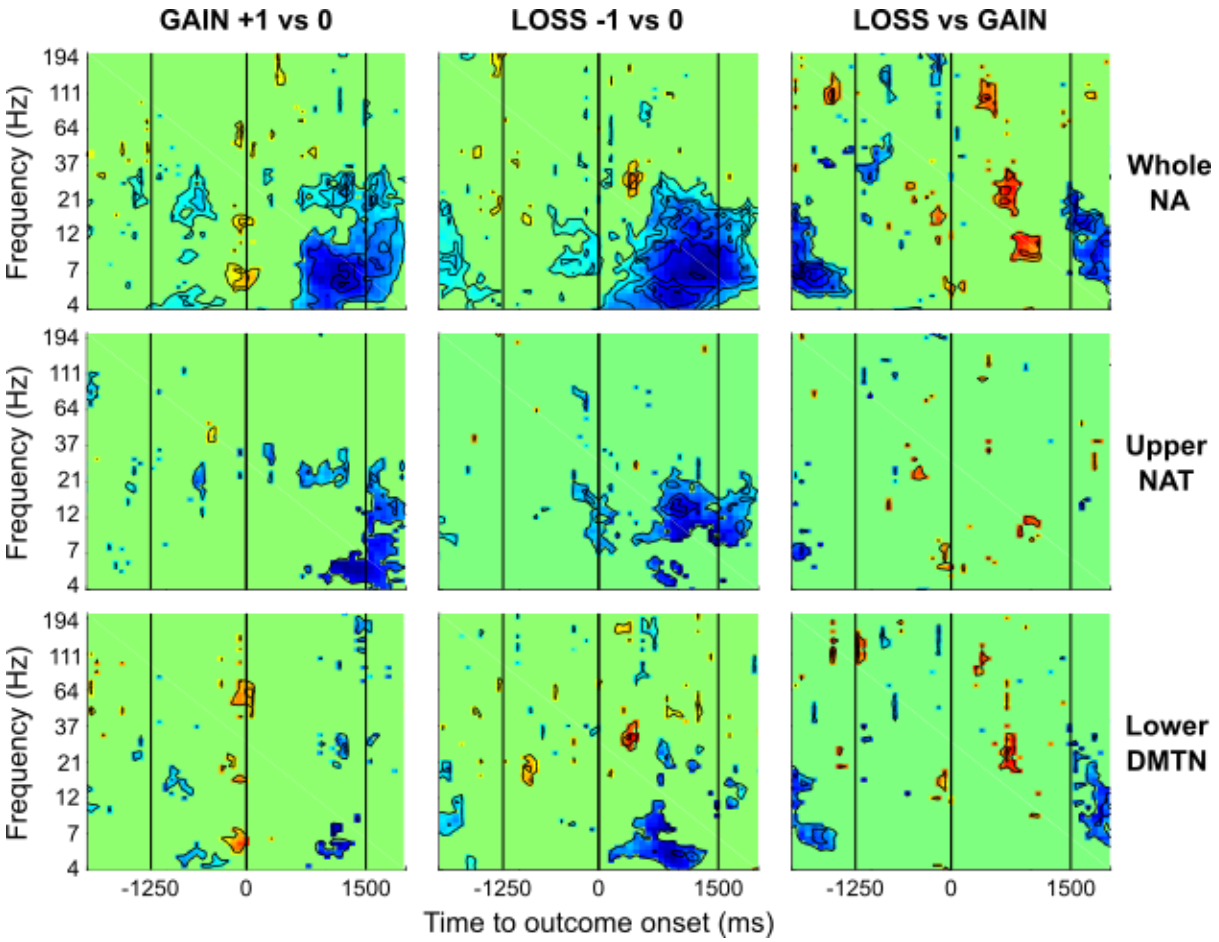


Figure 43 : Time-frequency analysis in the ANT and DMTN. (Upper line) All contacts (four on each electrode). **(Middle line)** Upper two contacts located within the ANT. **(Lower line)** Lower two contacts located within the DMTN. **(Right column)** Electrophysiological activity related to appetitive outcome display compared to neutral outcome display in the Gain condition. **(Central column)** Electrophysiological activity related to aversive outcome display compared to neutral outcome display in the Loss condition. **(Left column)** Electrophysiological activity related to outcome display during Loss condition compared to Gain condition. Power is color coded with hyper-activations in shades of red and hypo-activations in shades of blue, green being baseline level. Isocontours represent a p=0.01 non-corrected threshold. Choices are made 1250ms prior to the outcome onset, the outcome is displayed on the screen during 1500ms.

learning.

The first two patients included in this study have been performing the cognitive task twice more, three months apart, once their stimulation settings had been fixed. During this double-blind study, we focused on the behavioral changes induced by deep brain stimulation of the limbic thalamus on both forms of reinforcement learning, by contrasting the behavioral features (performance and reaction time) reported when the stimulation was turned ON and when it was turned OFF. These analyses have been done individually as the two patients P1 and P2 who underwent DBS stimulation during the protocol were stimulated in different nuclei of the thalamus and with different monopolar stimulation settings (see Table 5 in supplementary materials): P1 was stimulated in the NAT while P2 was stimulated in the DMTN.

Performance. We started by measuring the performance of the patients during the cognitive task in the gain condition and in the loss condition. As shown on Figure 44A and C, P1 had a good global performance close to 80% across all trials in both gain and loss conditions when the ANT-DBS was turned OFF. In contrast, P1 did not learn to select corrects symbols in neither gain nor loss condition when the stimulation was turned ON as the performance was close to but below 50% of correct choices. There was no significant bias on either gain or loss condition induced by the DBS as the performance was symmetrical (paired t-tests: $p=0.4391$ $t(5)=0.8402$ in DBS-ON and $p=0.1134$ $t(7)=1.809$ in DBS-OFF). DBS stimulation of the ANT seems to have prevented learning in both gain and loss conditions (unpaired t-tests: $p=0.0003$ $t(12)=5.076$ in gain and $p=0.0111$ $t(12)=2.997$ in loss). To sum this up, ANT-DBS significantly reduced performance in both learning conditions (one way ANOVA and post-hoc Student-Newman-Keuls test: $p<0.0001$, $F=12.177$). In comparison with P1, P2 received a bilateral monopolar stimulation of the DMTN, with a lower frequency than P1. In Figure 44B and C, we report no significant effect of the DMTN-DBS on P2 performance as it was close to 80% in all cases, hence revealing efficient learning (one way ANOVA and post-hoc Student-Newman-Keuls test: $p=0.9438$, $F=0.1265$).

Reaction time. Following the study of the individual performance of P1 and P2 DBS patients, we went on calculating the average reaction time across all sessions when the DBS was turned ON and OFF (see Table 5 and Table 7 in supplementary materials). P1 presented no significant difference in the reaction times reported during gain and loss conditions, no matter the DBS status (one way ANOVA and post-hoc Student-Newman-Keuls test: p0.7594, $F=0.3951$). On the contrary, P2 had fluctuating reaction times between all conditions (one way ANOVA and post-hoc Student-Newman-Keuls test: $p<0.0001$, $F=12.097$). P2 reaction times were significantly longer in loss than in gain conditions (paired t-test: $p=0.0074$

t(5)=4.338 in DBS-ON and p=0.0063 t(5)=4.518 in DBS-OFF). When the DMTN-DBS was turned ON, reaction times were significantly increased in both gain and loss condition than when the DMTN-DBS was turned OFF (paired t-test: p=0.0102 t(5)=4.014 in gain and p=0.0147 t(5)=3.650 in loss).

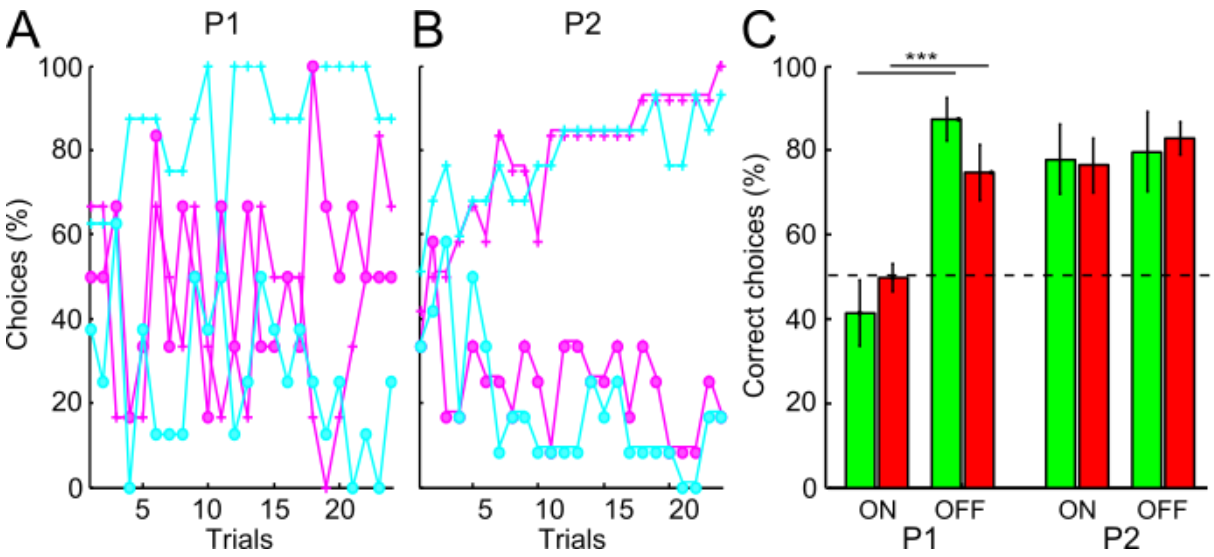


Figure 44 : Influence of deep brain stimulation of the anterior thalamic nucleus on reinforcement learning. Learning curves of **(A)** Patient 1 and **(B)** Patient 2. Observed choices when DBS stimulation was turned ON (magenta) and OFF (cyan). The learning curves portray, trial by trial, the average proportion of trials across patients corresponding to subjects choosing the 'correct' stimulus in the gain condition (crosses), and the 'incorrect' stimulus in the loss condition (circles). **(C)** Average performance of each patient over all sessions in DBS ON and DBS OFF (P1: 3 and 4 sessions; P2: 6 and 6 sessions). Average performance (correct choices) +/- SEM. Chance level is shown by the dotted line.

In conclusion, the deep-brain stimulation of thalamic nuclei has different effects across both our patients (P1 and P2) when we look at their performance and reaction time during reward-seeking learning and punishment-avoidance learning. These differential effects following DBS of the ANT for P1 and DBS of the DMTN for P2, with distinct stimulation parameters, will be discussed further below.

DISCUSSION

Thanks to intrathalamic electroencephalography signals recorded from epileptic patients while they performed a reinforcement learning task, we provide new insights on the functional implication of anterior and dorsomesial thalamic nuclei during reward and punishment avoidance learning. These preliminary data suggest that theta oscillations have a major contribution to outcome processing in the ANT especially, while higher-frequency oscillations are involved in outcome processing in the DMTN. With the use of an identical instrumental task during DBS, we may strengthen the empirical support by testing the causal role of the ANT and DMTN. Yet, we found that DBS of the ANT had unreliable effects across patients but results in a moderate worsening of their performance during reward-based learning. These preliminary results coming from a small cohort of patients are currently being consolidated by the inclusion of more patients. Furthermore, since previous attempts to relate thalamic sEEG activity to memory processes focused on timing, rather than amplitude analyses, such as phase-amplitude and cross-frequency couplings (Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016), these type of analyses would therefore be interesting in the context of reinforcement learning.

We reported a significant difference in the global performance observed in gain and in loss condition, with a better performance at seeking rewards than avoiding punishments. Importantly, even though some patients had close-to-chance-level performance, the group of 5 patients enrolled in the experiment succeeded in reaching a performance threshold of 60% of correct choices in both conditions by the end of their training (group performance of 69.6% in gain and 61.0% in loss in the last 8 trials). In addition, we found longer reaction times in loss than gain condition, suggesting that a higher cognitive may occur during the loss condition, as patients had to identify most aversive symbols so as to avoid them and select neutrally charged ones, while a neutral outcome in itself could be from a gain pair or a loss pair. A lengthening of reaction time can also be associated with a greater cognitive cost, especially during a Go/No-Go task (Benis et al. 2014; Chikazoe et al. 2009) and as a post-error slowing (Cavanagh & Frank 2014). This ambivalent character of the neutral outcome may well be the source of the longer RT and larger number of trials required to reach learning (70% of accurate responses) in the loss condition. In addition, this bias could also be related to the patients and perhaps a suboptimal functioning of the ANT as it is part of the Papez circuit, a key network for memory (Jankowski et al. 2013; Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016; Temel et al. 2012). Nevertheless, this interpretation seems unlikely since it would predict a deficit during both reward-based and punishment-avoidance learning.

While performing the time-frequency analysis on all contacts as well as on distinct ANT and DMTN bipolar contacts, we reported a reliable decrease of theta activity following outcome display. This theta response seemed to originate mostly from the ANT. Theta activity in the ANT has previously been reported to be involved in memory encoding higher cognitive processes (Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016), especially in the right ATN.

We reported differential effects of the ANT-DBS stimulation on punishment-avoidance learning as P1 was worse during the stimulation while P2 performance was improved when the stimulation was turned ON compared to the LFP recordings (see Table 7 in supplementary materials). They may be due to the difference of stimulation sites between the patients as P1 was stimulated in the ANT and P2 in the DMTN, and that their stimulation parameters were different (see Table 5 in supplementary materials). It has been reported in previous studies (Bradfield et al. 2013; Corbit et al. 2003) the ANT has no effect on performance in instrumental free-operant tasks. On the opposite, the DMTN has been found to play a specific and important role in the acquisition of goal-directed action (Smith et al. 2002; Corbit et al. 2003). As a result for P1, the drop of performance during punishment-avoidance learning could reflect a lesion-like effect of the DBS which would be in accord with the decrease of performance observed during avoidance learning of painful stimuli in previous studies following NAT lesions in rabbits (Gabriel et al., 1989). On the other hand with P2, the stimulation of the DMTN did not affect as much reinforcement learning performances.

In summary, our findings clearly need to be consolidated by including more patients so as to further establish whether NAT and DMTN play a significant role during reinforcement learning. This would broaden current understanding of the brain structures involved in reinforcement learning from a focus on the striatum and neocortex to recognition of a pivotal role for the ATN and DMTN.

EXPERIMENTAL PROCEDURES

Patients

All patients gave their written informed consent to participate in this study that was approved by our local ethics committee (Comité de Protection des Personnes Sud-Est I, protocol number: 2011-A00083-38). All participants had normal or corrected to normal vision and provided written informed consent. Intracerebral recordings were obtained from 5 neurosurgical patients with intractable epilepsy (2 females aged 29.7 +- 4.25, and 3 males aged 37.8 ± 3.77 years) at the Epilepsy Department of the Grenoble Alpes University Hospital or of the APHP - Pitie Salpetriere Hospital in Paris (see Table 4 in supplementary materials). These patients were implanted bilaterally in the limbic thalamic nuclei with two deep-brain stimulation electrodes (Medtronic DBS lead model 3389) as a surgical treatment to alleviate their seizures. Local field potentials were recorded from these macroelectrodes prior to the implantation of the pacemaker. Electrode implantation was performed according to clinical procedures as part of the clinical trial STIC-France (S.Chabardes), and all target structures for the presurgical evaluation were selected strictly according to clinical considerations with no reference to the current study. The patients involved in this study underwent the bilateral implantation of deep-brain stimulation electrodes in their anterior nuclei of the thalamus (ANT-DBS). Because of the small volume of the ANT, the electrodes were implanted through and through the nuclei, so as to ensure its maximal recording. The neurosurgeon targeted the nuclei so as to have the fourth (upper, most distal) contact located at the foremost dorsal edge of the ANT, allowing at least the two upper contacts to be inside the ANT and form a bipolar contact (see Figure 42 and Table 5). As a result, the more ventral proximal contacts pointed towards the nuclei located below the ANT along the implantation trajectory, in this case the dorsomedial nuclei of the thalamus (DMTN).

Behavioral task

Patients performed a probabilistic instrumental learning task adapted from previous imaging and patient studies (Palminteri et al., 2009, Palminteri et al., 2012; Pessiglione et al., 2006; Worbe et al., 2011). Patients were provided with written instructions, which were reformulated orally if necessary so as to clarify that the aim of the task was to maximize their financial payoff and that to do so, they had to consider reward seeking and punishment avoidance as equally important (Figure 1). Patients performed short training sessions to familiarize with task's timing and responses. Training procedure comprised a very short

session with only two pairs of cues presented during 16 trials that was followed by 2 to 3 short sessions of five minutes so that at the end of the training procedure, all patients reached a threshold of 70 % correct choices during both the reward and punishment conditions. During sEEG recordings, patients performed three to six test sessions after the training. Each session was an independent task containing four new pairs of cues to be learned. Cues were abstract visual stimuli taken from the Agathodaimon alphabet. Each pair of cues was presented 24 times for a total of 96 trials. The four cue pairs corresponded to the two conditions (2 pairs of gain and 2 pairs of loss cues), which were respectively associated with different pairs of outcomes (winning 1€ versus nothing or losing 1€ versus nothing). Within each pair, the two cues were associated to the two possible outcomes with reciprocal probabilities (0.75/0.25 and 0.25/0.75). On each trial, one pair was randomly presented and the two cues were displayed on a computer screen on the left and right of a central fixation cross, their relative position being counterbalanced across trials. The subject was required to choose the left stimulus or the right stimulus by using their left or right index to press the corresponding button on a joystick (Logitech Dual Action). Since the position on screen was counterbalanced, response (left versus right) and value (good versus bad cue) were orthogonal. The chosen cue was colored in red for 400 ms and then the outcome (either “nothing”, “gain”, or “loss”) was displayed on the screen. In order to win money, patients had to learn the cue–outcome associations by trial and error, so as to choose the most rewarding cue in the gain condition and the less punishing cue in the loss condition.

Behavioral and computational analyses

Percentage of correct choices and reaction times were monitored across time for all sessions (see Table 6 in supplementary materials). For each pair of cues, the consistency of choices was also monitored after trials associated with salient outcomes (effective rewards or punishments).

A standard Q-learning algorithm was to be use to model action-value learning and fit to the observed behavior. For each pair of stimuli A and B, the model estimates the expected value of outcome if choosing A (Qa) and B (Qb), according to previous choices and outcomes. The expected values of both stimuli were set at 0 prior to learning, being the mean outcome value. After each trial t>0, the expected value of the chosen stimuli (say A) was updated according to the rule $Qa(t+1)=Qa(t)+\alpha*\delta(t)$. The outcome prediction error, $\delta(t)$, is the difference between the obtained and the expected outcomes, $\delta(t)=R(t)-Qa(t)$, with R(t) the reinforcement obtained among -1€, 0€ and -1€. Using the estimated values of reinforcement

obtained after choosing each stimuli, the probability (or likelihood) of choosing each stimuli was estimated using the softmax rule: Pa(t) = exp[Qa(t)/β] / {exp[Qa(t)/β] + exp[Qb(t)/β]}. The constant parameters alpha and beta respectively are the learning rate and the temperature (the likelihood of changing chosen stimulus after a surprise trial). They have been adjusted to fit to mean observed choices across the group so as to maximize the probability of the actual choices so that only two sets of parameters has been used across the group, corresponding to a specific α/β couple for each condition (gain or loss). Outcome prediction errors and Qvalues estimated by the model for each patient and each trial were then used as statistical regressors for sEEG data.

sEEG data acquisition, preprocessing and analysis

Data acquisition. As in routine DBS procedure of Grenoble University Hospital, two semirigid, multilead electrodes (Medtronic DBS lead model 3389) were stereotactically implanted in each patient targeting their anterior thalamic nuclei. Placement of the thalamic electrodes was performed stereotactically. The angle of entry through the skull and the depth of each electrode was calculated based of MRI images of each patient's brain pre-operatively. The Model 3389 DBS lead features narrow (0.5 mm) spacing between each of the four electrodes (1.5 mm wide) at the distal end. The Model 3389 DBS lead provides electrodes spread over 7.5 mm. All electrode contacts were identified on each patient's individual postimplantation X-ray. Each subject's individual preimplantation MRI T1 and FGATIR (except for patient 1) sequences were coregistered with the postimplantation CT scan to determine the anatomical location of each contact, to compute all contact coordinates into the Montreal Neurological Institute (MNI) space using standard Statistical Parametric Mapping algorithms and to reconstruct the 3D structures of interest (ANT and ventricles). Coordinates of recording sites were computed as the MNI coordinates average of the two contacts used for each sEEG contact-pair.

Data preprocessing. A bipolar montage was used between adjacent electrode contacts. sEEG data were bandpass-filtered online from 0.1 to 200 Hz and sampled at 1024Hz (4 patients) or 2048 Hz (1 patients). Each electrode trace was subsequently re-referenced with respect to its direct neighbor (bipolar derivations with a spatial resolution of 2 mm) to achieve high local specificity by cancelling out effects of distant sources that spread equally to both adjacent sites through volume conduction (Lachaux et al. 2003). 6 bipolar contact-pairs (3 on each hemisphere) were recorded in each patient using a commercial video-sEEG monitoring system (System Plus, Micromed). Overall, 40 recording sites were obtained over the five

patients included in this study. Recording of intracranial signals was performed at the bedside in the two days following electrode implantation, before the electrodes were attached to a stimulator under the skin over the chest wall for epilepsy treatment, in a second operation approximately 3 days later. Stimulation via intracranial electrodes did not occur during the data recordings of local field potentials (LFP). No seizures took place during the testing sessions, and all patients were fully alert and cooperative throughout.

Computation of time-frequency maps. The electrophysiological encoding data were segmented into epochs 5 s pre-stimulus (cue and outcome display) to 5 s post-stimulus, because of the duration of one trial (see Figure 41) and as we noticed the patients tended to display long reaction times. Note that the time axes in all figures are set such that the stimulus was shown at time = 0 s. Each bipolar signal was processed using a Morlet wavelet transform to extract instantaneous power between 1 and 150 Hz. The number of cycles of the mother wavelet was set to 7, so that the support of the kernel was 350 ms at 20 Hz. Time-frequency power was log-transformed to improve Gaussianity of the data using the fixation period (average duration 1000 ms) as a baseline. The broad-spectrum of Morlet wavelet transform analyses was used for an extensive visual inspection of responses. Due to the implantation of the electrodes covering the ANT and the DMTN, the time-frequency analysis of the activity recorded in response to outcome display has been performed on three categories of contacts: 1) all the contacts recorded (2*3 bipoles per patient), 2) only the upper bipolar contacts located within the ANT and 3) only the lower bipolar contacts located within the DMTN.

REFERENCES

Alcaraz, F., Marchand, A.R., Vidal, E., Guillou, A., Faugère, A., Coutureau, E. Wolff, M. 2015. Flexible Use of Predictive Cues beyond the Orbitofrontal Cortex: Role of the Submedius Thalamic Nucleus. *Journal of Neuroscience*, 35(38), pp.13183–13193.

Balleine, B.W., Morris, R.W. & Leung, B.K., 2015. Thalamocortical integration of instrumental learning and performance and their disintegration in addiction. *Brain Research*, 1628, pp.104–116.

Bartra, O., McGuire, J.T. & Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, pp.412–427.

Benis, D., David, O., Lachaux, JP., Seigneret, E., Krack, P., Fraix, V., Chabardès, S., Bastin, J., 2014. Subthalamic nucleus activity dissociates proactive and reactive inhibition in patients with Parkinson's disease. *NeuroImage*, 91, pp.273–81.

Bradfield, L.A., Hart, G. & Balleine, B.W., 2013. The role of the anterior, mediodorsal, and parafascicular thalamus in instrumental conditioning. *Frontiers in Systems Neuroscience*, 7(October), p.51.

Cavanagh, J.F. & Frank, M.J., 2014. Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, 18(8), pp.414–421.

Chakraborty, S., Kolling, N., Walton, M.E., Mitchell, A.S. 2016. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *eLife*, 5(MAY2016), pp.1–23.

Chase, H.W., Kumar, P., Eickhoff, S.B., Doombrovski, A.Y. 2015. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2), pp.435–459.

Chikazoe, J., Jimura, K., Asari, T., Yamashita, K., Morimoto, H., Hirose, S., Miyashita, Y., Konishi, S. 2009. Functional dissociation in right inferior frontal cortex during performance of go/no-go task. *Cerebral Cortex*, 19(1), pp.146–152.

Conejo, N.M., González-Pardo, H., López, M., Cantora, R., Arias, J.L. 2007. Induction of c-Fos expression in the mammillary bodies, anterior thalamus and dorsal hippocampus after fear conditioning. *Brain Research Bulletin*, 74(1–3), pp.172–177.

Corbit, L.H., Muir, J.L. & Balleine, B.W., 2003. Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *European Journal of Neuroscience*, 18(5), pp.1286–1294.

Fisher, R., Salanova V, Witt T, Worth R, Henry T, Gross R, Oommen K, Osorio I, Nazzaro J, Labar D, Kaplitt M, Sperling M, Sandok E, Neal J, Handforth A, Stern J, DeSalles A, Chung S, Shetter A, Bergen D, Bakay R, Henderson J, French J, Baltuch G, Rosenfeld W, Youkilis A, Marks W, Garcia P, Barbaro N, Fountain N, Bazil C, Goodman R, McKhann G, Babu Krishnamurthy K, Papavassiliou S, Epstein C, Pollard J, Tonder L, Grebin J, Coffey R, Graves N; SANTE Study Group. 2010. Electrical stimulation of the anterior nucleus of thalamus for treatment of refractory epilepsy: Deep Brain Stimulation of Anterior Thalamus for Epilepsy. *Epilepsia*, 51(5), pp.899–908.

Gabriel, M., Sparenborg, S. & Kubota, Y., 1989. Anterior and medial thalamic lesions, discriminative avoidance learning, and cingulate cortical neuronal activity in rabbits. *Experimental Brain Research*, 76(2), pp.441–457.

Garrison, J., Erdeniz, B. & Done, J., 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience and biobehavioral reviews*, 37(7), pp.1297–310.

Gueguen, M.C.M., Lachaux, J.P., Kahane, P., Billeke, P., Pessiglione M., Bastin, J. Rewards and punishment learning differentially modulates intracerebral brain dynamics.

Jankowski, M.M., Ronnqvist KC, Tsanov M, Vann SD, Wright NF, Erichsen JT, Aggleton JP, O'Mara SM. 2013. The anterior thalamus provides a subcortical circuit supporting memory and spatial navigation. *Frontiers in systems neuroscience*, 7(August), p.45.

Ketz, N.A., Jensen, O. & O'Reilly, R.C., 2015. Thalamic pathways underlying prefrontal cortex–medial temporal lobe oscillatory interactions. *Trends in Neurosciences*, 38(1), pp.3–12.

Lachaux, J.-P., Chavez, M. & Lutz, A., 2003. A simple measure of correlation across time, frequency and space between continuous brain signals. *Journal of Neuroscience Methods*, 123(2), pp.175–188.

Mitchell, A.S., 2015. The mediodorsal thalamus as a higher order thalamic relay nucleus important for learning and decision-making. *Neuroscience & Biobehavioral Reviews*, 54, pp.76–88.

O'Doherty, J.P., Kringelbach ML, Rolls ET, Hornak J, Andrews C. 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience*, 4(1), pp.95–102.

Palminteri, S., Justo D, Jauffret C, Pavlicek B, Dauta A, Delmaire C, Czernecki V, Karachi C, Capelle L, Durr A, Pessiglione M. 2012. Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, 76(5), pp.998–1009.

Parnaudeau, S., O'Neill PK, Bolkan SS, Ward RD, Abbas AI, Roth BL, Balsam PD, Gordon JA, Kellendonk C. 2013. Inhibition of Mediodorsal Thalamus Disrupts Thalamofrontal Connectivity and Cognition. *Neuron*, 77(6), pp.1151–1162.

Parnaudeau, S., Taylor K, Bolkan SS, Ward RD, Balsam PD, Kellendonk C 2015. Mediodorsal Thalamus Hypofunction Impairs Flexible Goal-Directed Behavior. *Biological Psychiatry*, 77(5), pp.445–453.

Pessiglione, M., Seymour B, Flandin G, Dolan RJ, Frith CD 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106), pp.1042–5.

Seymour, B., O'Doherty JP, Koltzenburg M, Wiech K, Frackowiak R, Friston K, Dolan R 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, 8(9), pp.1234–1240.

Smith, D.M., Freeman JH Jr, Nicholson D, Gabriel M. 2002. Limbic thalamic lesions, appetitively motivated discrimination learning, and training-induced neuronal activity in rabbits. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(18), pp.8212–21.

Sun, L., Peräkylä J, Polvivaara M, Öhman J, Peltola J, Lehtimäki K, Huhtala H, Hartikainen KM 2015. Human anterior thalamic nuclei are involved in emotion-attention interaction. *Neuropsychologia*, 78, pp.88–94.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Esslinger C, Hinrichs H, Heinze HJ, Knight RT, Richardson-Klavehn A. 2014. Corticothalamic phase synchrony and cross-frequency coupling predict human memory formation. *Elife*, 3, p.e05352.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Richardson-Klavehn A, Hinrichs H, Heinze HJ, Knight RT, Rugg MD. 2016. Pre-stimulus thalamic

theta power predicts human memory formation. *NeuroImage*, 138, pp.100–108.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Hinrichs H, Heinze HJ, Rugg MD, Knight RT, Richardson-Klavehn A. 2015. Thalamic theta phase alignment predicts human memory formation and anterior thalamic cross-frequency coupling. *Elife*, 4(MAY), p.e07578.

Temel, Y., Hescham SA, Jahanshahi A, Janssen ML, Tan SK, van Overbeeke JJ, Ackermans L, Oosterloo M, Duits A, Leentjens AF, Lim L. 2012. *Neuromodulation in psychiatric disorders.*,

Vertes, R.P., Linley, S.B. & Hoover, W.B., 2015. Limbic circuitry of the midline thalamus. *Neuroscience & Biobehavioral Reviews*, 54, pp.89–107.

Wright, N.F., Vann SD, Aggleton JP, Nelson AJ. 2015. A Critical Role for the Anterior Thalamus in Directing Attention to Task-Relevant Stimuli. *Journal of Neuroscience*, 35(14), pp.5480–5488.

SUPPLEMENTARY MATERIALS

Patient	Sex of patients	Age at surgery	Sampling frequency
P1	F	35	2048
P2	F	34	1024
P3	M	28	1024
P4	M	42	1024
P5	M	54	1024

Table 4 : Demographical and clinical details of the patients. All patients recorded in the study (n=5).

Patient	Bipolar contacts used for LFP recordings		Location of the contacts within the thalamus		Total number of bipolar by patient	Monopolar contacts used for DBS		Stimulation parameters F(Hz) / P(ms) / A(V)	
	Left	Right	Left	Right		Left	Right	Left	Right
P1	2-3 1-2 0-1	2-3 1-2 0-1	NAT-NAT-NAT-DMTN	NAT-edge-DMTN - DMTN	6	2	2	130/60/4	130/60/4
P2	2-3 1-2 0-1	2-3 1-2 0-1	NAT-NAT-DMTN - DMTN	NAT-NAT-DMTN - DMTN	6	1	1	40/60/4	40/60/4
P3	2-3 1-2 0-1	2-3 1-2 0-1	NAT-NAT-DMTN - DMTN	NAT-NAT-DMTN - DMTN	6	-	-	-	-
P4	2-3 1-2 0-1	2-3 1-2 0-1	NAT-NAT-DMTN - DMTN	NAT-NAT-DMTN - DMTN	6	-	-	-	-
P5	2-3 1-2	2-3 1-2	NAT-NAT-DMTN	NAT-NAT-DMTN	6	-	-	-	-

	0-1	0-1	- DMTN	- DMTN					

Table 5 : Contacts information and stimulation parameters. The contact location was identified by the neurosurgeon using a 3D reconstruction software and based on X-ray and MRI images (T1 and FGATIR). The most proximal (tip of the electrode) contact is #0 while the most distal is #3.

Patients	LFP		ON-DBS		OFF-DBS	
	Number of sessions	Number of trials	Number of sessions	Number of trials	Number of sessions	Number of trials
P1	6	576	3	288	4	384
P2	6	576	6	576	6	576
P3	6	576	-	-	-	-
P4	5	480	-	-	-	-
P5	6	576	-	-	-	-

Table 6: Number of sessions and trials performed by all patients. Each session contains 4 pairs of symbols having 24 trials each, so there are 96 trials per session (48 gain trials and 48 loss trials).

Patients	LFP		ON-DBS		OFF-DBS	
	Gain	Loss	Gain	Loss	Gain	Loss
P1	74	58	41	49	86	74
P2	86	66	77	75	76	81

Table 7: Average performance of patients who underwent ANT-DBS. The performance corresponds to the rate of correct choices observed in each condition (in %).

Patient	Condition	P value	T value	Degrees of freedom
P1	Gain	0.0191	3.409	5
	Loss	0.0037	5.117	5
P2	Gain	0.0016	6.189	5

	Loss	0.0913	2.087	5
P3	Gain	0.5772	0.5958	5
	Loss	0.4650	0.7906	5
P4	Gain	0.7142	0.3932	4
	Loss	0.5275	0.6911	4
P5	Gain	0.1275	1.826	5
	Loss	0.2258	1.381	5

Table 8: Individual performance of patients compared to chance level. One sample t-tests compared to 50%.

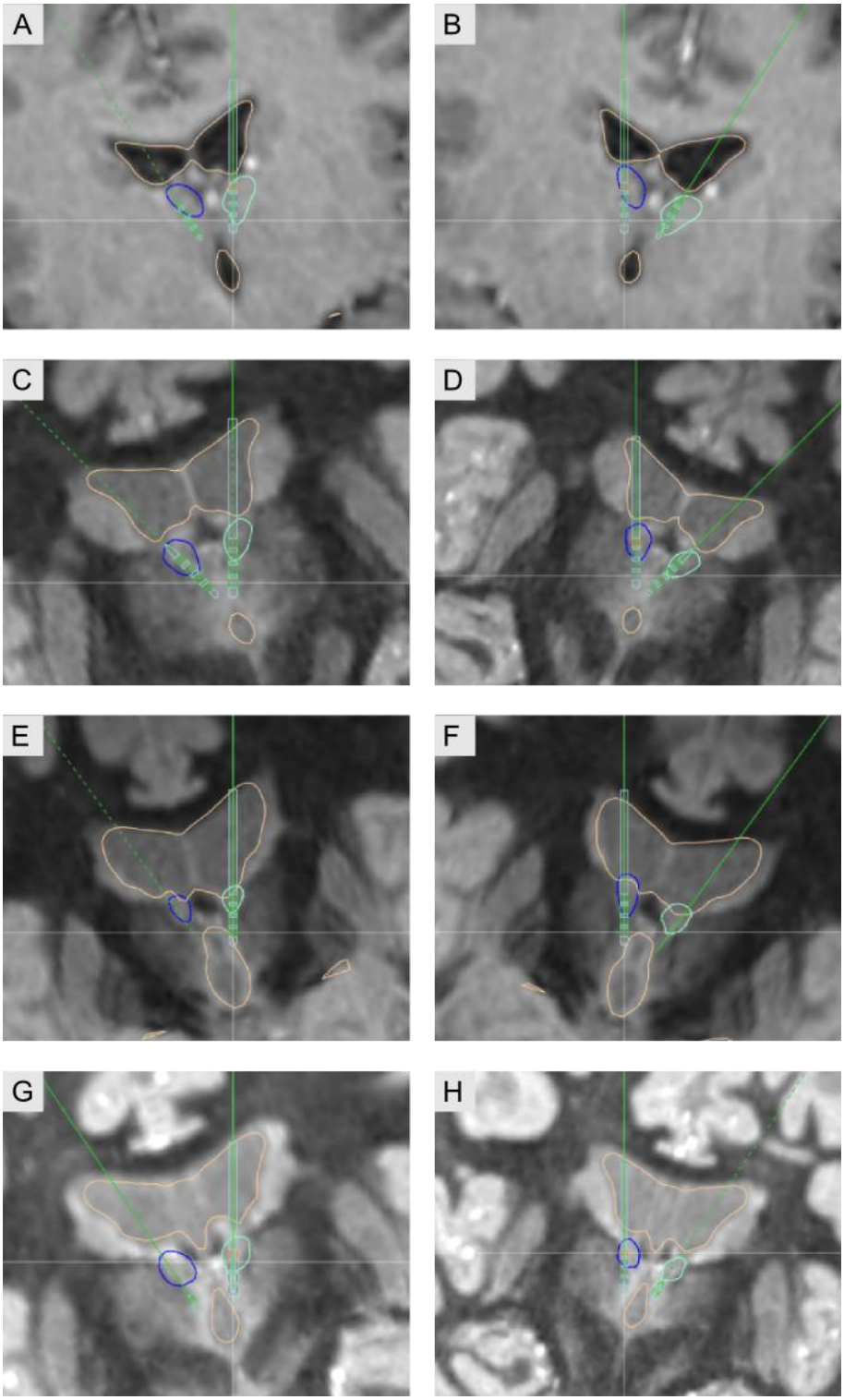


Figure 45: Individual anatomical location of the DBS electrodes within the thalamus of patients. Reconstruction of intrathalamic contact location using intraoperative X-ray image coordinates, superimposed on preoperative FGATIR MRI mask (T1 for P1 in A and B). View along the electrode. The ventricular system/third ventricle is represented in orange, the anterior thalamic nuclei are in blue (right) and green (left), the electrodes are shown in gray, their trajectories are in dotted green lines. **(A)** Left electrode of P1. **(B)** Right electrode of P1. **(C)** Left electrode of P2. **(D)** Right electrode of P2. **(E)** Left electrode of P3. **(F)** Right electrode of P3. **(G)** Left electrode of P5. **(H)** Right electrode of P5.

C. Etude 3 : Effet du risque sur l'apprentissage par renforcement

Reward and punishment learning systems are differentially affected by risk

Running title: Risk and reinforcement learning

Maëlle Camille Marie Gueguen[1,2], Stefano Palminteri[3] and Julien Bastin[1,2]*

[1] Univ. Grenoble Alpes, F-38000 Grenoble, France.

[2] Inserm, U1216, F-38000 Grenoble, France.

[3] Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France.

*Correspondence to: julien.bastin@ujf-grenoble.fr (J.B.); Phone: (0033) 4 56 52 06 78; Fax: (0033) 4 56 52 05 98; Institut des Neurosciences de Grenoble, Bâtiment Edmond J. Safra des Neurosciences, Chemin Fortuné Ferrini, Université Joseph Fourier, Site Santé La Tronche, BP 170 38042 Grenoble Cedex 9, France.

ABSTRACT

Economic choices are known to be strongly influenced by both outcome valence (i.e. whether an outcome entails a gain or a loss) and risk (defined here as the variance of the outcomes). Surprisingly, the behavioral and neural mechanisms underlying the possible influence of risk on reward and punishment based-learning have not yet been investigated. Here, to address this question, we used a novel instrumental learning task that was designed to manipulate independently risk and valence, while the difference of value between options was kept invariant. We found a striking interaction between the effects of risk and valence on learning: reward-based learning was not affected by the riskiness of the correct option (safe or risky) whereas subjects displayed a significant impairment at learning to choose the risky option during punishment avoidance learning. This finding is the exact opposite to the pattern of result predicted by the prospect theory (“reflection effect”), according to which subjects are risk seeker for losses and risk averse for gains. We discuss the possible computational and neural mechanisms that could underlie this behavioral pattern.

INTRODUCTION

Despite a considerable amount of research on how outcome valence and risk influence economic choices (Kolling, Wittmann et Rushworth 2014; Kuhnen et Knutson 2005; Preuschoff, Quartz et Bossaerts 2008; Pujara et al. 2015; Tom et al. 2007; Wright et al. 2012), how multiple outcome dimensions are integrated during decision-making remains unclear. Furthermore, the mechanisms through which we are able to learn implicitly these information through reinforcement learning processes did not yet receive much attention (Niv et al. 2012). Here, to address this issue, we compared learning performance of healthy subjects during a novel instrumental learning task during which outcome valence and risk were independently manipulated.

Indication that risk and valence critically influence choices comes mostly from forced-choice gambling tasks during which subjects are provided with explicit information about outcome variance (risk) and valence (gains/losses). A classic example central to the prospect theory is the 'reflection effect', according to which subjects tend to exhibit risky-seeking for gains and risk-aversion for losses (Kahneman & Tversky 1979). However, this view has been challenged since this pattern seems to be context-dependent, as opposite behavioral findings were identified when the choice context was varied during gambling (Wright et al. 2012). At the neural level, risk has been repeatedly associated with activations of the anterior insula and posterior cingulate cortex (McCoy & Platt 2005; Mohr et al. 2010; Wu et al. 2012; Preuschoff et al. 2008), gains were shown to reliably activate a ventromedial-prefronto-striatal network and losses were associated with anterior insula, dorsomesial and lateral prefrontal activations (Bartra et al. 2013). Yet, the neural implementation and behavioral mechanisms underlying our capacity to learn to integrate risk and valence dimensions implicitly by trial and error received less attention.

An interesting exception is the study by Niv et al. (Niv et al. 2012), during which subjects had to learn to choose between sure and risky options that had equal expected values. The authors showed a risk-sensitive reinforcement learning process implemented in the nucleus accumbens: NAC activity encoded reward-prediction error signals while stimulus value signals in NAC were predictive of behavioral risk aversion. This finding is intriguing since this pattern of neural activity does not fit with the absence of risk in the algorithm used to update stimulus value in classical reinforcement learning models. In addition, it is unclear how this reward-based risk-sensitive learning process translates when punishment-based reinforcement learning is considered. Hence, the interactions between valence and risk were never examined, neither at the behavioral, nor at the neural level during reward and

punishment-based learning that are thought to activate opponent brain systems (Palminteri et al. 2015; Pessiglione et al. 2006; Seymour et al. 2005; Bartra et al. 2013).

Here, we address this issue at the behavioral level by comparing the learning performance of healthy subjects during a novel instrumental learning task that independently manipulated outcome valence (gains and losses) and risk (safe and risky options). Thus, two features of the task served our purposes: first, the task contrasted reward seeking with punishment avoidance learning; second, we also manipulated the riskiness of the correct option (safe vs. risky). We found a striking interaction between outcome valence and risk during reinforcement learning by showing that risk modulated learning performance when subjects had to learn to choose the risky option during punishment-avoidance, thus exhibiting a risk aversion for losses. In a second experiment, we explored the role of the anterior insular cortex by directly recording its neural activity while epileptic patients implanted with deoth electrodes performed a learning task.

EXPERIMENTAL PROCEDURES

Participants

During experiment 1 (behavioral testing), we tested 22 healthy subjects (13 females and 9 males with an age mean of 36.8, age range: 22-58). All healthy subjects do not have any active neurological and psychiatric disease and they did not take any psychiatric treatment (healthy subjects' demographics details are summarized in Table 9).

During experiment 2, intracerebral recordings were subsequently obtained from 6 neurosurgical patients with intractable epilepsy (2 females aged 35 and 17 years ; 4 males aged 36.3 ± 7.10 years) at the Epilepsy Department of the Grenoble Alpes University Hospital (patients' demographics and clinical details are summarized in Table 10). To localize epileptic foci that could not be identified through noninvasive methods, neural activity was monitored in lateral, intermediate, and medial wall structures in these patients using stereotactically implanted multilead electrodes (stereotactic electroencephalography, sEEG). Electrode implantation was performed according to routine clinical procedures, and all target structures for the presurgical evaluation were selected strictly according to clinical considerations with no reference to the current study (Isnard et al. 2000).

All participants had normal or a corrected-to-normal view (if they have refraction trouble) and provided written informed consent. All procedures conformed to national and institutional guidelines for the use of human subjects, and to the Declaration of Helsinki. Experiments were approved by the Ethics Committee for Biomedical Research of Grenoble Alpes University Hospital (ISD-sEEG 2009-A00239-48).

Behavioral task

Learning task – experiment 1. Healthy subjects performed a probabilistic instrumental learning task adapted from previous studies (Palminteri et al. 2012; Pessiglione et al. 2006; Gueguen, Lachaux, et al. n.d.), as well as one epileptic patient (P1). They were provided with written instructions, which were reformulated orally if necessary so as to clarify that the aim of the task was to maximize their financial payoff and that to do so, they had to consider reward seeking and punishment avoidance as equally important (Figure 46). Participants performed short training sessions to familiarize with task's timing and responses. Training procedure consisted in two sessions (one TARGET and one CONTROL) with four pairs of cues presented during 20 trials which could be repeated once so that at the end of the

training procedure, all participants reached a threshold of 70 % correct choices during both the reward and punishment conditions for each session. For both training and testing sessions, symbols pairs were randomly formed to prevent any bias of symbol selection due to its shape or the repetition of the training session.

During testing itself, participants performed five test sessions after the training. Each session was an independent task containing four new pairs of cues to be learned. Cues were abstract visual stimuli taken from the Agathodaimon alphabet. Each pair of cues was presented 20 times for a total of 80 trials per session. The four cue pairs corresponded to the two conditions (2 pairs of gain cues and 2 pairs of loss cues), which were respectively associated with different pairs of outcomes (winning versus neutral in gain conditions or losing versus neutral in loss conditions). The behavioral task was designed to separate the effects of outcome valence (rewards or punishments) and the riskiness of the correct choice (risky or safe) on reinforcement learning. To this end, two types of sessions were used. In the TARGET sessions, pairs were formed of a safe symbol and a risky symbol, inducing a conflict between risk-seeking/risk-avoiding behavior and the strategy to select the best symbol in each pair. In the CONTROL sessions, this conflict was absent as safe symbols were paired together, as were risky symbols. The entire testing task was composed of three TARGET sessions alternated with two CONTROL sessions (TCTCT Figure 46C).

As shown in Figure 46B, in each TARGET condition, a safe symbol (100% of an intermediate reinforcement of ±0.5€) was paired with a risky symbol (a large reinforcement (±1€) balanced with a neutral outcome (0€) with reciprocal 80/20 or 20/80 likelihood depending on the riskiness of the correct symbol). As a consequence, four conditions were designed as follow (Figure 46B): a “gain risky” GR condition where the correct symbol provided a large reward (+1€) when selected but only in 80% of all trials, a “gain safe” GS condition where the correct symbol provided a small reward (+0.5€) each time it was selected, a “loss risky” LR condition where choosing the correct symbol helped avoiding a large punishment (-1€) in most trials (80%) and a “loss safe” LS condition where choosing the correct symbol lead to an intermediate punishment (-0.5€) each time.

Similarly, in each CONTROL condition, safe symbols (100% of an intermediate reinforcement of ±0.5€) were paired together, as well as risky symbols (one symbol associated with a large reinforcement (±1€) balanced with a neutral outcome (0€) with 70/30 or 30/70 likelihood, paired with a symbol associated with an intermediate reinforcement (±0.5€) balanced with a neutral outcome (0€) with 80/20 or 20/80 likelihood). As a consequence, four conditions were designed as follow (Figure 46B): a “gain risky” GR condition where both symbols were risky and provided rewards, a “gain safe” GS condition

where both symbols were safe and provided rewards, a “loss risky” LR condition where both symbols were risky and provided punishments and a “loss safe” LS condition where both symbols were safe and provided punishments. As the results presented in this study were solely obtained from TARGET sessions, the design of CONTROL sessions is summarized in (Figure 49) for comprehension purposes.

On each trial, one pair was randomly presented and the two cues were displayed on a computer screen on the left and right of a central fixation cross, their relative position being counterbalanced across trials (Figure 46B). After a buffer period of 500ms, the subject was required to choose the left or the right stimulus by using their left or right index to press the corresponding button on a joystick (Logitech Dual Action). Since the position on screen was counterbalanced, response (left versus right) and value (good versus bad cue) were orthogonal. The chosen cue was colored in red for 250ms and then the outcome was displayed on the screen for 1500ms. After a variable inter-trial interval of 800-1200ms, a new trial started with the presentation of the fixation cross. In order to win money, participants had to learn by trial and error the cue–outcome associations, so as to choose the most rewarding cue in the gain condition and the less punishing cue in the loss condition. At the end of each session, the global outcome in both gain and loss condition were presented to the

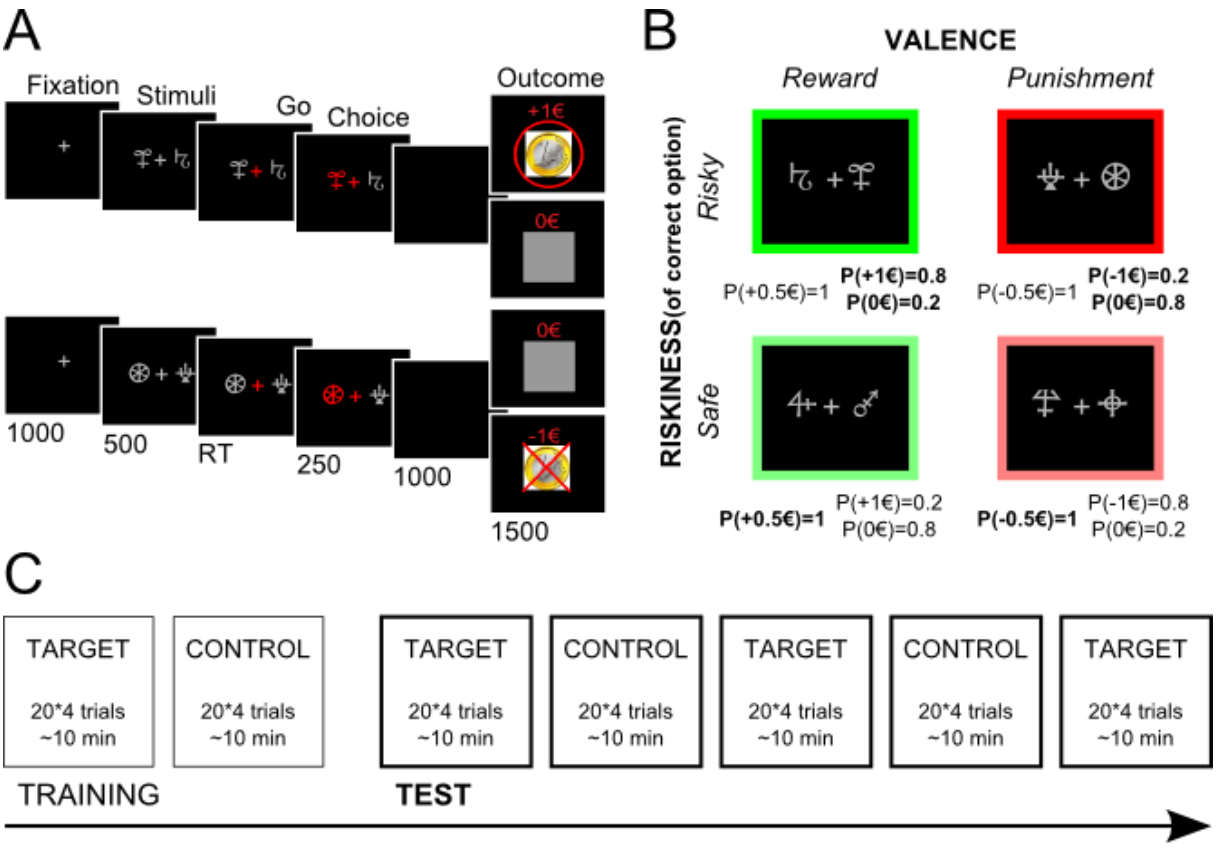


Figure 46: Behavioral task (A) Successive screen of typical trials in the reward (top) and loss (bottom) conditions. Patients had to select one abstract visual stimulus among the two presented on each side of a central visual fixation cross, and subsequently observed the outcome. Duration are given in milliseconds. **(B)** Four conditions were used to separate the effects of valence and risk on reinforcement learning. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red. Probabilities associated with the correct are written in bold. **(C)** Succession of TARGET and CONTROL sessions during the entire experiment.

participants.

In all experimental conditions, the most advantageous symbol in term of benefits depends on the difference of value (DV) between the two symbols. For each pair, the symbol with the higher value (magnitude of outcome * probability) was the most interesting. In this behavioral task, pairs of all conditions of both CONTROL and TARGET conditions had a DV of 0.3€ between the two symbols of the pair.

Modified cognitive task. In order to test the influence of the difference of value (DV) between the two symbols during the TARGET sessions, we developed a modified version of the cognitive task. Five epileptic patients (P2 to P6) performed a modified version of the learning task used in experiment 1. We replaced the CONTROL sessions by modified TARGET sessions (called T2 sessions). In the T2 sessions, reinforcement magnitudes and probabilities were modified so as to set DV=0.1€ during T2 sessions. The reinforcement magnitudes and probabilities of the T1 sessions were kept unchanged as the ones from the original TARGET session described above, setting DV=0.3€ in T1 sessions.

The precise design of the T2 sessions was therefore similar to T1, excepting a lower DV. Thus, a safe symbol (100% of an intermediate reinforcement of ±0.2€) was paired with a risky symbol (a large reinforcement of ±0.5€) balanced with a neutral outcome (0€) with 20/80 (GS and LR pairs) or 60/40 (GR and LS pairs) likelihood depending on the riskiness of the correct symbol). As a consequence, four conditions were designed as follows: a “gain risky” GR condition where the correct symbol provided a large reward (+0.5€) when selected but only in 60% of all trials, a “gain safe” GS condition where the correct symbol provided a small reward (+0.2€) each time it was selected, a “loss risky” LR condition where choosing the correct symbol helped avoiding a large punishment (-0.5€) in most trials (80%) and a “loss safe” LS condition where choosing the correct symbol lead to an intermediate punishment (-0.2) each trial. As the results presented in this study were solely obtained from T1 sessions, the design of T2 sessions is summarized in Figure 49B in supplementary materials, for comprehension purposes.

Behavioral analyses

Original cognitive task. CONTROL sessions were used to control individual bias by identifying participants according to their risk-seeking or risk-avoidance preference (design of CONTROL sessions in Figure 49A and behavioral results in Figure 50 in supplementary

materials). The following analyses were performed on TARGET sessions and the results reported in the article come from these TARGET sessions.

Modified cognitive task. T2 sessions were used to study the influence of DV on risky reinforcement learning (design of T2 sessions in Figure 49B (supplementary materials) but T2 data from 5 epileptic patients not shown in this study). The following analyses were performed on T1 sessions only, as they are identical to TARGET sessions from the original version of the cognitive task. The results reported in this study therefore come from only T1 sessions.

Observed choices were monitored in all conditions (GR, GS, LR and LS) and were used to obtain group learning curves once the data from all TARGET/T1 sessions had been averaged. In addition, the performance (rate of correct choices) was computed for all conditions, as well as the average reaction time (calculated from the onset of the GO screen display in Figure 46A). All statistical analyses were performed using group analyses. Performance and reaction times across the four conditions were compared using ANOVAs.

sEEG data acquisition and preprocessing

Five to seventeen semirigid, multilead electrodes were stereotactically implanted in each patient. Electrodes had a diameter of 0.8 mm and, depending on the target structure, contained 8–18 contact leads 2-mm-wide and 1.5-mm-apart (DIXI Medical Instruments). All electrode contacts were identified on each patient's individual postimplantation MRI. Each subject's individual preimplantation MRI was coregistered with the postimplantation MRI (Carmichael et al. 2008) to determine the anatomical location of each contact. The MarsAtlas (Auzias et al. 2016) was used to label each electrode as a function of individual gyri/sulci organization. This labeling was further confirmed by visually inspecting patients' MRI and identify anatomical landmarks well established in the literature (Craig 2009b; Afif & Mertens 2010; Afif et al. 2010; Ongür & Price 2000; Bechara et al. 1998; Ursu & Carter 2005; Kringelbach & Rolls 2004). sEEG contact located in the white matter were also removed from the analysis.

sEEG data were bandpass-filtered online from 0.1 to 200 Hz and sampled at 512 Hz (2 patients) or 1024 Hz (4 patients), using a reference electrode located in white matter. Each electrode trace was subsequently re-referenced with respect to its direct neighbor (bipolar derivations with a spatial resolution of 3.5 mm) to achieve high local specificity by cancelling out effects of distant sources that spread equally to both adjacent sites through volume

conduction (Lachaux et al. 2003). 84 to 157 bipolar contact-pairs were recorded in each patient using a commercial video-sEEG monitoring system (System Plus, Micromed). Overall, 647 recording sites were obtained over the twenty patients included in this study.

sEEG data analyses

Computation of single-trial broadband gamma envelopes. Broadband gamma activity (BGA) was extracted with the Hilbert transform of sEEG signals using custom Matlab scripts (Mathworks Inc., MA, USA) as follows. sEEG signals were first bandpass filtered in 10 successive 10-Hz-wide frequency bands (e.g. 10 bands, beginning with 50–60 Hz up to 140–150 Hz). For each bandpass filtered signal, we computed the envelope using standard Hilbert transform. The obtained envelope had a time resolution of 15.625 ms (64 Hz). Again, for each band, this envelope signal (i.e. time-varying amplitude) was divided by its mean across the entire recording session and multiplied by 100 for normalization purposes. Finally, the envelope signals computed for each consecutive frequency bands (e.g. 10 bands of 10 Hz intervals between 50 and 150 Hz) were averaged together, to provide one single time-series (the BGA) across the entire session, expressed as percentage of the mean. This time-series were smoothed with a 250 ms sliding window to increase statistical power for inter-trial and inter-individual analyses of BGA dynamics.

Identification of sEEG sites responding to reinforcement delivery. We used single-trial BGA responses to identify sEEG contact-pairs that were significantly modulated by the delivery of a reinforcement, whether it be appetitive (+1€), neutral (±0€) or aversive (-1€). For each sEEG contact-pair, we extracted the time-course of BGA during a post outcome interval (-250 to 1000 ms, i.e. 81 time points) during each trial. We then averaged the BGA modulation observed during the 800 ms following outcome delivery. At last, we performed an ANOVA with repeated measures with a Student-Newman-Keuls post-hoc test to identify significant differences in BGA response based on the type of outcome delivered. The anterior insula was our region of interest for its known role in risk encoding (Preuschoff et al. 2008). To identify reliable electrophysiological activities, at least 10 task-responsive sEEG contact-pairs across at least 3 patients were required.

RESULTS

Experiment 1: Behavior

Behavioral data were collected from 22 healthy subjects (see demographical details in Table 9 and methods) while they performed an instrumental learning task during which reward and punishment conditions were matched in difficulty, as the same probabilistic contingencies were to be learned (Figure 46).

Differential influence of risk on reward-based and punishment-avoidance learning

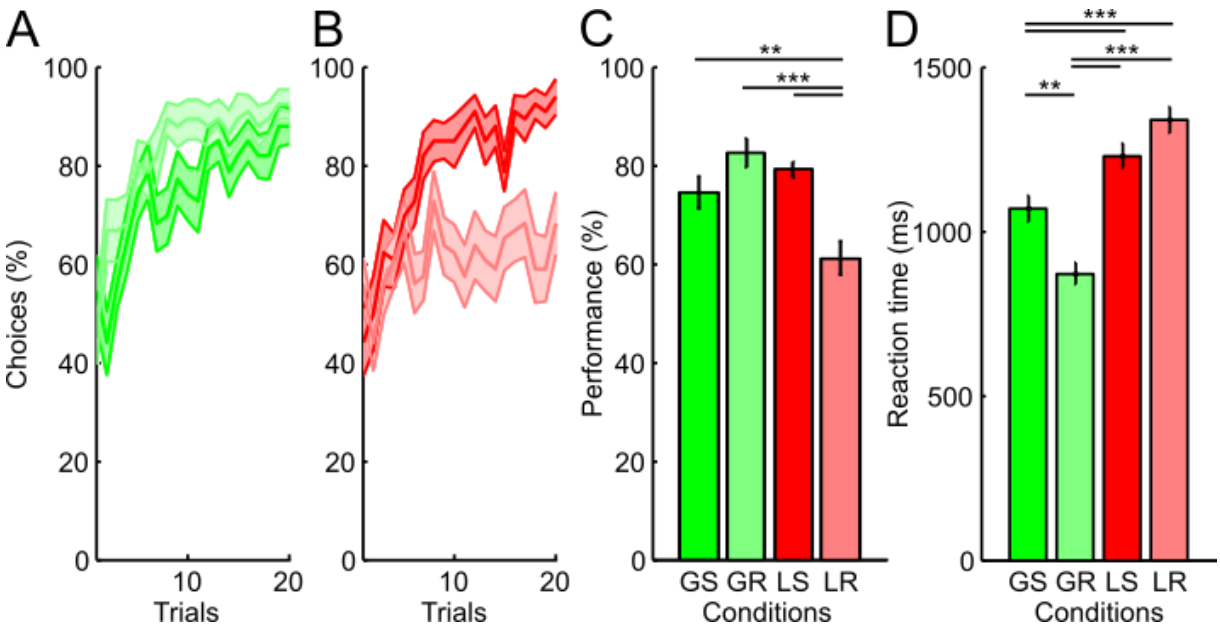


Figure 47: Behavioral results. Learning curves in the TARGET sessions. Group (n=22 subjects) learning curves portray, trial by trial, the average proportion of trials across patients corresponding to subjects choosing the 'correct' stimulus in the gain conditions **(A)**, and the 'correct' stimulus in the loss conditions **(B)**. Average choices (solid lines) are superimposed over 99% confidence interval (shaded areas). **(C)** Group performance in all conditions of the TARGET sessions. **(D)** Reaction time in all conditions of the TARGET sessions. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red.

Patients were able to learn the correct response over the 20 trials of a learning session. They tended to choose the most rewarding cue in the gain safe (GS) and gain risky (GR) conditions while they were also prone to avoid the most punishing cue in the loss safe (LS)

and loss risky (LR) conditions (Figure 47A and B; one-sample t-test versus chance (50%): $p<0.0001$ for GS, GR and LS, $t(21)=7.544$ for GS, $t(21)=11.260$ for GR, $t(21)=17.106$ for LS, $p=0.0045$ $t(21)=3.182$ for LR).

Thus, as in previous studies in healthy subjects (Palminteri et al. 2015; Pessiglione et al. 2006; Gueguen, Lachaux, et al. n.d.), patients were able to choose the correct option when they had to learn from reward or from punishments.

While the subjects managed to learn efficiently in all conditions, they presented a significant learning deficit in the LR condition compared to all others (rmANOVA, $p<0.0001$, $F=10.680$). No significant difference was reported between GS, GR and LS conditions during TARGET sessions, all reaching around 80% of correct choices. This is consistent with the performance observed during the CONTROL sessions, where we also observed a risk aversion during punishment-avoidance learning (paired-t-test $p=0.0159$, $t(21)=2.623$) (see Figure 50 in supplementary materials). As a consequence, risk seems to only affect punishment-avoidance learning (Figure 47C).

We also found that reaction times were shorter in the gain condition than in the loss condition (GS=1088ms and GR=935ms vs LS=1271ms and LR=1336ms, rmANOVA with post-hoc Student Newman Keuls test, $p<0.0001$, $F=29.944$, $p<0.001$ for GR vs LS/LR and GS vs LS/LR). In addition, risk induced significantly faster reaction time in the gain condition (GS=1088ms vs GR=935ms, post-hoc SNK test, $p<0.01$) but even though there is a tendency to increase reaction times in loss, risk did not significantly influenced the reaction times in the loss conditions (Figure 47D).

In conclusion, risk influenced differentially reward-seeking and punishment-avoidance learning. It induced a selective learning deficit in punishment-avoidance learning with a dramatically decreased performance. At the same time, risk reduced reaction times during reward-based learning.

Experiment 2: Intracerebral recordings from the anterior insula

The aim of our analysis was to explore the influence of risk on the activity of the anterior insular cortex while learning to maximize rewards and avoid punishments. To this end, we directly recorded electrophysiological brain activity from 6 patients with electrodes implanted in the anterior insula. Thus, in this small patients' sample, we did not find any reliable effect of risk on reinforcement learning during TARGET/T1 sessions (ANOVA with repeated

measures, p=0.9476 F=0.1189) (Figure 48A). Individual behavioral performance is reported in Figure 51 in supplementary materials.

Gamma activity in the anterior insula is modulated by risk during learning

We focused the analyses on broadband gamma activity (BGA) [50-150 Hz], which is known to correlate with both spiking and fMRI activity (Niessing et al. 2005; Lachaux et al. 2007; Manning et al. 2009). We measured BGA from 68 sEEG cortical sites across 6 patients. Each sEEG contact was anatomically labeled as a function of individual gyri/sulci organization estimated by a cortical parcellation atlas (MarsAtlas) (Auzias et al. 2016). Based on the pre-existing literature which identified the anterior insula as responding to risk during basic choices (Kuhnen & Knutson 2005; Preuschoff et al. 2008; Paulus 2009; Mohr et al. 2010; Wu et al. 2012), the region of interest of this electrophysiological study on the influence of risk on reinforcement learning is the anterior insula. Note that data from both hemispheres were collapsed to improve the power of the following analyses.

To examine whether risk influenced differentially the anterior insula during reward vs. punishment-based learning, we compared the average gamma band response in the anterior insula across all recorded contacts (n=68). Neural activity was time-locked to outcome delivery, and we chose to simply compare broadband gamma activity between the four conditions of the TARGET/T1 sessions (gain safe GS, gain risky GR, loss safe LS and loss risky LR).

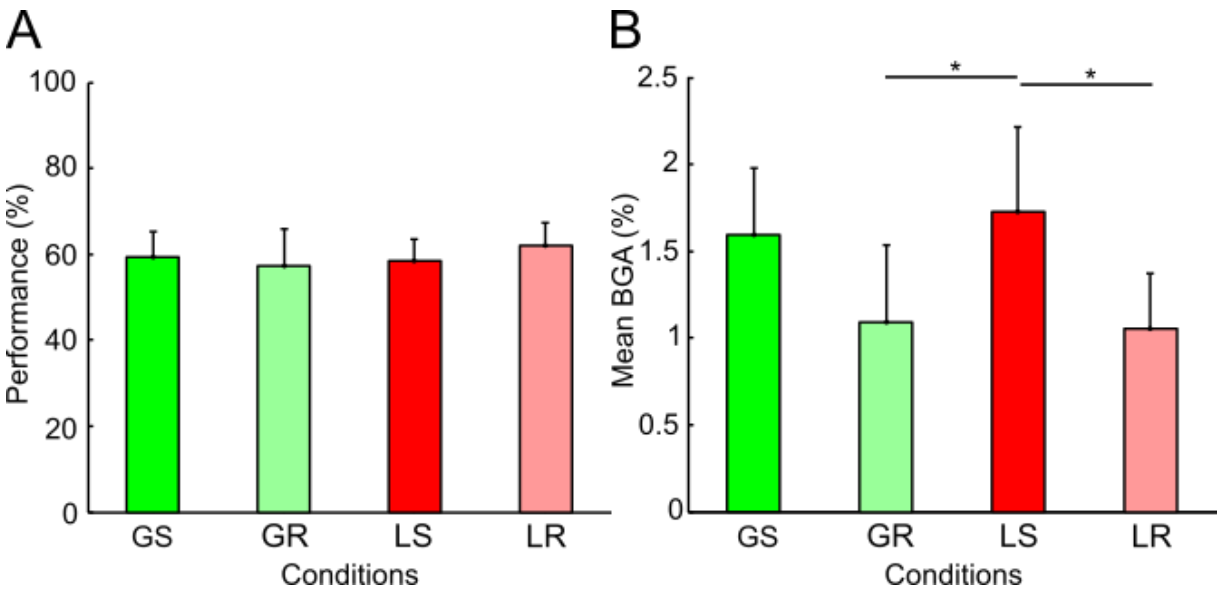


Figure 48: Performance and insular gamma activity during TARGET/T1 sessions in epileptic patients. (A) Group performance (n=6). Performance is described as the average proportion of trials where patients chose the 'correct' stimulus in the gain and loss conditions. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red. Patient 1 performed TARGET sessions as part of the original version of the task while patients 2 to 6 performed T1 sessions as part of the modified version of the task. **(B)**

Average BGA responses of the anterior insula in the 800 ms following outcome display in gain and loss condition (n=68 implanted contacts recorded from 6 patients).

This analysis revealed a significant effect of experismental conditions on the neural activity of the anterior insula ($F=3.63$, $p=0.013$) that was mainly driven by a lower neural response when the correct choice was rsisky ($GR<GS$ and $LR<LS$, $p<0.05$, post-hoc tests) (Figure 48B).

Overall, sEEG data analyses showed that risk influences the gamma activity of the anterior insula during reinforcement learning. Gamma activity of the anterior insula to outcome delivery was decreased when the correct choice was the risky one, in both reward-based learning and punishment-avoidance learning.

DISCUSSION

We investigated the effects of risk and valence on reinforcement learning by asking healthy subjects to perform a novel learning paradigm designed to independently manipulate these outcome dimensions. Note that because the difference of value between options was kept invariant, traditional Q-learning models would predict identical learning performances across the four experimental conditions obtained in the 2 by 2 factorial design of this study. In contrast, we found that subjects tended to choose the safe option during punishment learning, even when the riskier option was more advantageous –in terms of expected value- than the safe option. This risk-aversion was only observed during learning to avoid monetary losses, as risk did not affect reward-based learning performances. We discuss these results with reference to (i) context-dependent effects of risk and valence on choices, (ii) the possible computational processes that could underlie these behavioral findings and finally (iii) we also discuss the possible neural implementation of these effects.

The key behavioral finding of this study is that subjects were risk averse during punishment-based learning, but not during reward-based learning. To our knowledge, the possible interacting effects of risk and valence was not previously tested during learning. In contrast, when both outcome variance (risk) and valence (gains/losses) are explicitly indicated to the subjects, they were shown to be either risk averse for losses or for gains, depending on the context of the gambling tasks (Wright et al. 2012). More precisely, subjects displayed a striking risk aversion for losses when they had to choose between a sure loss and a gamble (four possible outcomes), a result which is consistent with the behavioral finding we report. However, they also found that subjects were preferentially risk averse for monetary gains when the task was to choose between two gambles, in line with the reflection effect from the prospect theory. This pattern of result strongly suggest that the combined effects of risk and valence are context-dependent given the reversal of the behavioral pattern observed between choice tasks (Wright et al. 2012). An interesting issue to expand this finding to a learning context would be to show a similar reversal when valence and risk are implicitly learned through experience, for example by showing that under some contextual circumstances, subjects are actually risk averse during reward-based learning, but not when they learn to avoid losses.

In a study that aimed at studying risk-sensitive reinforcement learning mechanisms (Niv et al. 2012), the authors compared subjects' choices when they had to choose between a sure gain of 20 cents or a 50 % gamble (0/40 cents). This procedure was used to match the options that had to be experientially learned through trial and error in terms of expected value because being sure to eventually get 20 cents is equivalent to having 40 cents on half of the

trials. This methodological trick has been repeatedly used to assess subjects' sensitivity to risk (independently from expected value), since a probability to choose the risky option below (beyond) 50 % would directly risk aversion (seeking). However, applied in the context of reinforcement learning, we argue that this methodological choice also strongly diminishes the ecological validity of examining how risk and variance are integrated during learning, as there is actually little to learn when both options are equated in terms of value. We therefore adopted a different approach by maintaining constant the difference of value between the two options that subjects had to learn between experimental condition, so that this "teaching signal" would be kept invariant (hence, the higher the difference, the easiest it is to learn for subjects). Another extension we made compared to the study of Niv et al. (Niv et al. 2012) is that we not only examined reward but also punishments. Behaviorally, the authors report that subjects were overall risk averse, as a majority of subjects preferred the sure option and that they also found a great variability in terms of risk sensitivity, both between subjects and even between sessions (within subjects). This absence of consistency between sessions is at odd with the replicability of the behavioral findings found for gambling tasks (Wright et al., 2012), while it is consistent with the variability found during tasks where there was actually no "good choice" (Hayden & Platt 2009; Huettel et al. 2006).

Another consistent finding was that subjects tended to be slower to choose options associated with losses than gains: this loss aversion indexed by longer reaction times is a classic finding reported both during reinforcement learning paradigms (Palminteri et al. 2015) or during choice tasks (Wright et al., 2012). This might reflect an approach/avoidance mechanism according to which subjects would tend to be slower to approach an aversive stimulus (such as a possible monetary loss following their choice) while they are faster to approach appetitive cues, such as possible monetary gains (Guitart-Masip et al. 2011). An interesting prediction from this theoretical account is that the extent to which subjects consider risk as appetitive (risk seekers) or aversive (risk averse) will give rise to opposite reaction time patterns, as subjects that are risk-averse (seeking) should be slower (faster) to approach risk.

From a computational point of view, we suggest that traditional reinforcement learning models that ignore the variance of the outcome and only estimate their mean should be updated to account for the present results and those presented previously (Niv et al. 2012). In their paper, the authors propose two possibilities to account for risk sensitivity during reinforcement learning. A first suggestion involves the implementation of nonlinear subjective utility functions to account for subject's sensitivity to risk. Thus, a concave (convex) nonlinear subjective utility function for outcomes would led to risk aversion (seeking). Another

possibility is that nonlinearity would be associated with prediction errors, for example by using different weights for positive and negative prediction errors (Niv et al. 2012). Accordingly, the degree of asymmetry between weights associated with positive and negative prediction errors should be closely related to risk sensitivity (Mihatsch & Neuneier 2002). Fitting these three different classes of models (classic Q-learning, utility and asymmetric prediction errors' models) on the present data-set would allow us to distinguish which type of assumption is more likely.

At the neural level a likely brain region that could implement the interaction between punishment and risk information is the anterior insula. This area was indeed showed to encode aversive information (Caria et al. 2010; Palminteri et al. 2012; Météreau & Dreher 2013; Hayes et al. 2014) and also risk (Preuschoff et al. 2008; Paulus 2009; Wu et al. 2012; Naqvi et al. 2014; Clark et al. 2014). To test this possibility, we used invasive recordings from the AI. Preliminary analyses suggest that a significant difference exist between risky and safe learning conditions, though independently from outcome valence (Figure 48). However, these preliminary data need to be cautiously interpreted, as we obtained so far inconsistent behavioral results in patients that displayed variable risk sensitivity (Figure 51 in supplementary materials), while we also failed to replicate the effect of valence which are massive in the anterior insula (as demonstrated in the first paper of this PhD thesis). To further investigate the functional involvement of anterior insula, we therefore plan to (1) simplify the task to decrease inter-individual variability in terms of risk-sensitivity and (2) to optimize data analysis to take advantage of this variability in parallel to examine the relationship between the anterior insula activity and each subject's sensitivity to risk.

In conclusion, this study provides a novel reinforcement-learning paradigm to distinguish how outcome valence and risk are integrated over time. However, given the novelty of the approach and the surprising behavioral results obtained here, we plan to conduct a replication study associated with computational analyses to confirm and strengthen the present results. Ultimately, this would provide novel experimental tools to further investigate the neural substrate of risk-sensitive learning mechanisms.

REFERENCES

Afif, A., Minotti, L., Kahane, P., and Hoffmann, D. 2010. Anatomofunctional organization of the insular cortex: a study using intracerebral electrical stimulation in epileptic patients. Epilepsia *51*, 2305–2315.

Afif, A. & Mertens, P., 2010. Description of sulcal organization of the insular cortex. *Surgical and radiologic anatomy : SRA*, 32(5), pp.491–8.

Auzias, G., Coulon, O. & Brovelli, A., 2016. MarsAtlas: A cortical parcellation atlas for functional mapping. *Human Brain Mapping*, 37(4), pp.1573–1592.

Bartra, O., McGuire, J.T. & Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, pp.412–427.

Bechara, A., Damasio, H., Tranel, D., and Anderson, S.W. 1998. Dissociation of working memory from decision making within the human prefrontal cortex. J. Neurosci. Off. J. Soc. Neurosci. *18*, 428–437.

Caria, A., Sitaram, R., Veit, R., Begliomini, C., and Birbaumer, N. 2010. Volitional Control of Anterior Insula Activity Modulates the Response to Aversive Stimuli. A Real-Time Functional Magnetic Resonance Imaging Study. Biol. Psychiatry *68*, 425–432.

Carmichael, D.W., Thornton, J.S., Rodionov, R., Thornton, R., Mcevoy, A., Allen, P.J., and Lemieux, L. 2008. Safety of Localizing Epilepsy Monitoring Intracranial Electroencephalograph Electrodes Using MRI : Radiofrequency-Induced Heating. *1244*, 1233–1244.

Clark, L., Studer B, Bruss J, Tranel D, Bechara A. 2014. Damage to insula abolishes cognitive distortions during simulated gambling. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16), pp.6098–103.

Craig, A.D.B., 2009. How do you feel--now? The anterior insula and human awareness. *Nature reviews. Neuroscience*, 10(1), pp.59–70.

Gueguen, M.C.M., Lachaux, J.P., Kahane, P., Billeke, P., Pessiglione M., Bastin, J. Rewards and punishment learning differentially modulates intracerebral brain dynamics.

Guitart-Masip, M., Fuentemilla L, Bach DR, Huys QJ, Dayan P, Dolan RJ, Duzel E. 2011. Action Dominates Valence in Anticipatory Representations in the Human Striatum and Dopaminergic Midbrain. *Journal of Neuroscience*, 31(21), pp.7867–7875.

Hayden, B.Y. & Platt, M.L., 2009. Gambling for Gatorade: risk-sensitive decision making for fluid rewards in humans. *Animal cognition*, 12(1), pp.201–7.

Hayes, D.J., Duncan NW, Xu J, Northoff G. 2014. A comparison of neural responses to appetitive and aversive stimuli in humans and other mammals. *Neuroscience & Biobehavioral Reviews*, 45, pp.350–68.

Huettel, S.A., Stowe CJ, Gordon EM, Warner BT, Platt ML 2006. Neural signatures of economic preferences for risk and ambiguity. *Neuron*, 49(5), pp.765–75.

Isnard, J., Guénot M, Ostrowsky K, Sindou M, Mauguière F. 2000. The role of the insular cortex in temporal lobe epilepsy. *Annals of Neurology*, 48(4), pp.614–623.

Kahneman, D. & Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica*, (47), pp.263–291.

Kolling, N., Wittmann, M. & Rushworth, M.F.S., 2014. Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron*, 81(5), pp.1190–1202.

Kringelbach, M.L. & Rolls, E.T., 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in neurobiology*, 72(5), pp.341–72.

Kuhnen, C.M. & Knutson, B., 2005. The neural basis of financial risk taking. *Neuron*, 47(5), pp.763–70.

Lachaux, J.-P., Fonlupt P, Kahane P, Minotti L, Hoffmann D, Bertrand O, Baciu M. 2007. Relationship between task-related gamma oscillations and BOLD signal: New insights from combined fMRI and intracranial EEG. *Human Brain Mapping*, 28(12), pp.1368–1375.

Lachaux, J.-P., Chavez, M. & Lutz, A., 2003. A simple measure of correlation across time, frequency and space between continuous brain signals. *Journal of Neuroscience Methods*, 123(2), pp.175–188.

Manning, J.R., Jacobs J, Fried I, Kahana MJ. 2009. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *Journal of Neuroscience*, 29(43), pp.13613–13620.

McCoy, A.N. & Platt, M.L., 2005. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature neuroscience*, 8(9), pp.1220–1227.

Metereau, E. & Dreher, J.-C., 2013. Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral cortex (New York, N.Y. : 1991)*, 23(2), pp.477–87.

Mihatsch, O. & Neuneier, R., 2002. Risk-Sensitive Reinforcement Learning. *Machine Learning*, 49(2/3), pp.267–290.

Mohr, P.N.C., Biele, G. & Heekeren, H.R., 2010. Neural Processing of Risk. *Journal of Neuroscience*, 30(19), pp.6613–6619.

Naqvi, N.H., Gaznick N, Tranel D, Bechara A. 2014. The insula: A critical neural substrate for craving and drug seeking under conflict and risk. *Annals of the New York Academy of Sciences*, 1316(1), pp.53–70.

Niessing, J., Ebisch B, Schmidt KE, Niessing M, Singer W, Galuske RA. 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science (New York, N.Y.)*, 309(5736), pp.948–951.

Niv, Y., Edlund JA, Dayan P, O'Doherty JP. 2012. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2), pp.551–562.

Ongür, D. & Price, J.L., 2000. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral cortex*, 10(3), pp.206–219.

Palminteri, S., Khamassi M, Joffily M, Coricelli G. 2015. Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6, p.8096.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., Pessiglione, M. 2012. Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. Neuron *76*, 998–1009.

Paulus, M.P., 2009. Gut-level choices:risk-taking and interoception...insula. *Power Point*, pp.1–45.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Preuschoff, K., Quartz, S.R. & Bossaerts, P., 2008. Human insula activation reflects risk prediction errors as well as risk. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(11), pp.2745–2752.

Pujara, M.S., Wolf RC, Baskaya MK, Koenigs M 2015. Ventromedial prefrontal cortex damage alters relative risk tolerance for prospective gains and losses. *Neuropsychologia*, 79, pp.70–75.

Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. *8*, 1234–1240.

Tom, S.M., Fox CR, Trepel C, Poldrack RA. 2007. The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, 315(5811), pp.515–8.

Ursu, S. & Carter, C.S., 2005. Outcome representations, counterfactual comparisons and the human orbitofrontal cortex: Implications for neuroimaging studies of decision-making. *Cognitive Brain Research*, 23, pp.51–60.

Wright, N.D., Symmonds M, Hodgson K, Fitzgerald TH, Crawford B, Dolan RJ 2012. Approach-Avoidance Processes Contribute to Dissociable Impacts of Risk and Loss on Choice. *Journal of Neuroscience*, 32(20), pp.7009–7020.

Wu, C.C., Sacchet, M.D. & Knutson, B., 2012. Toward an affective neuroscience account of financial risk taking. *Frontiers in neuroscience*, 6, p.159.

SUPPLEMENTARY MATERIALS

Healthy subject	Sex	Age	Handedness
1	F	25	Left
2	M	43	Right
3	F	24	Left
4	F	32	Right
5	F	31	Right
6	M	44	Right
7	M	42	Right
8	F	36	Right
9	F	56	Left
10	M	23	Right
11	F	27	Right
12	M	58	Right
13	F	43	Left
14	F	47	Right

15	M	24	Right
16	M	52	Right
17	M	30	Right
18	F	29	Right
19	M	23	Right
20	F	48	Right
21	F	50	Right
22	F	22	Right

Table 9 : Demographical details of the healthy subjects. All healthy subjects enrolled in the study (n=22).

Patient	Sex	Age	Handedness	Nb bipoles	Sampling frequency (Hz)	EZ
1	M	48	Right	91	1024	L temporal
2	F	35	Right	91	1024	R OFC + temporal
3	M	47	Right	139	512	Right SMA with

						dysplasic lesion
4	M	18	Right	157	512	Cryptogenic
5	M	32	Right	85	1024	R temporal
6	F	17	Right	84	1024	R temporal + insular

Table 10 : Demographical and clinical details of the patients. All patients recorded in the study (n=6). Epileptic zone (EZ)

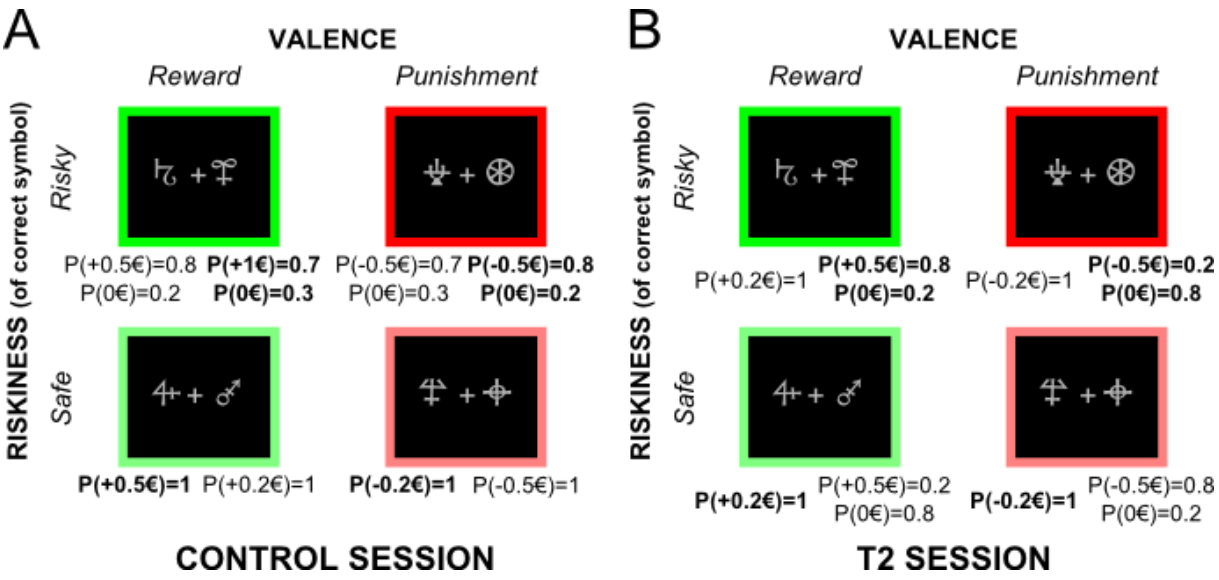


Figure 49**: Design of CONTROL and T2 sessions.** In both CONTROL sessions **(A)** and T2 sessions **(B)**, four conditions were used to separate the effects of valence and risk on reinforcement learning. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red. Probabilities associated with the correct are written in bold.

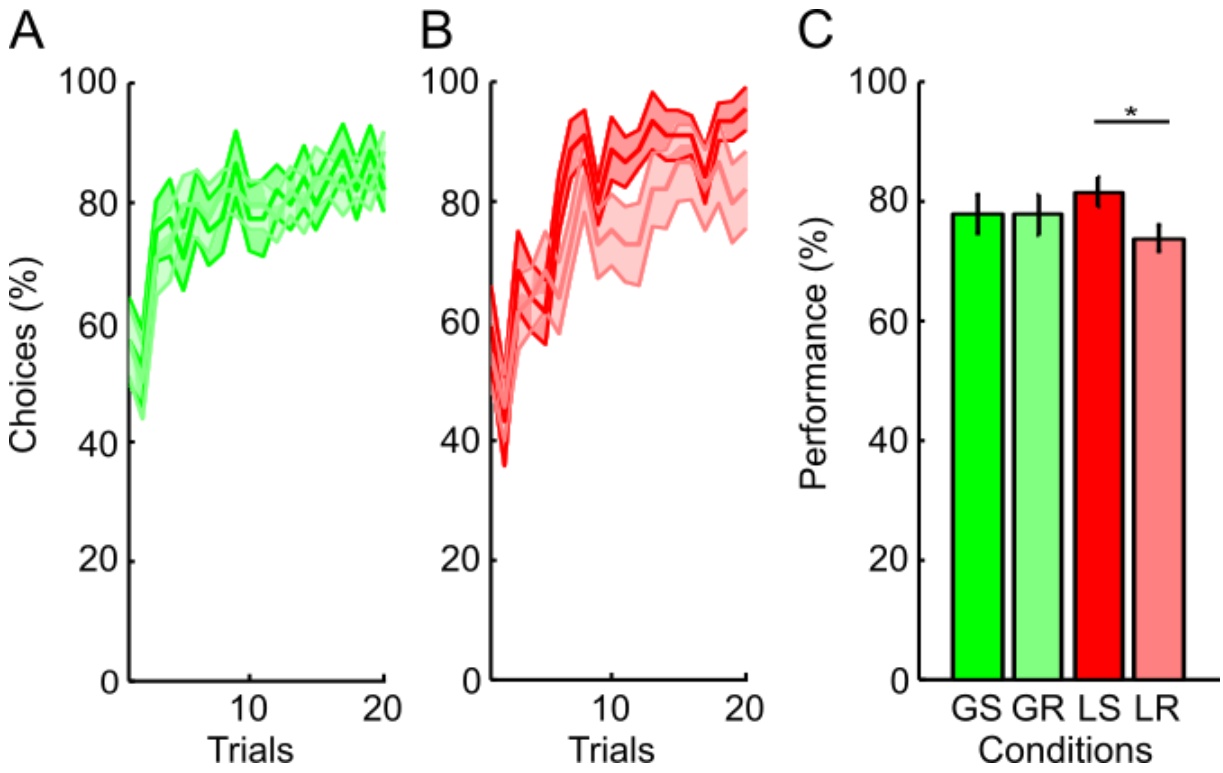


Figure 50 **: Risk aversion during punishment avoidance learning in CONTROL sessions.** Learning curves in the CONTROL sessions. Group (n=22 subjects) learning curves portray, trial by trial, the average proportion of trials across patients corresponding to subjects choosing the 'correct' stimulus in the gain conditions **(A)**, and the 'correct' stimulus in the loss conditions **(B)**. Average choices (solid lines) are superimposed over 99% confidence interval (shaded areas). **(C)** Group performance in all conditions of the CONTROL sessions. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red.

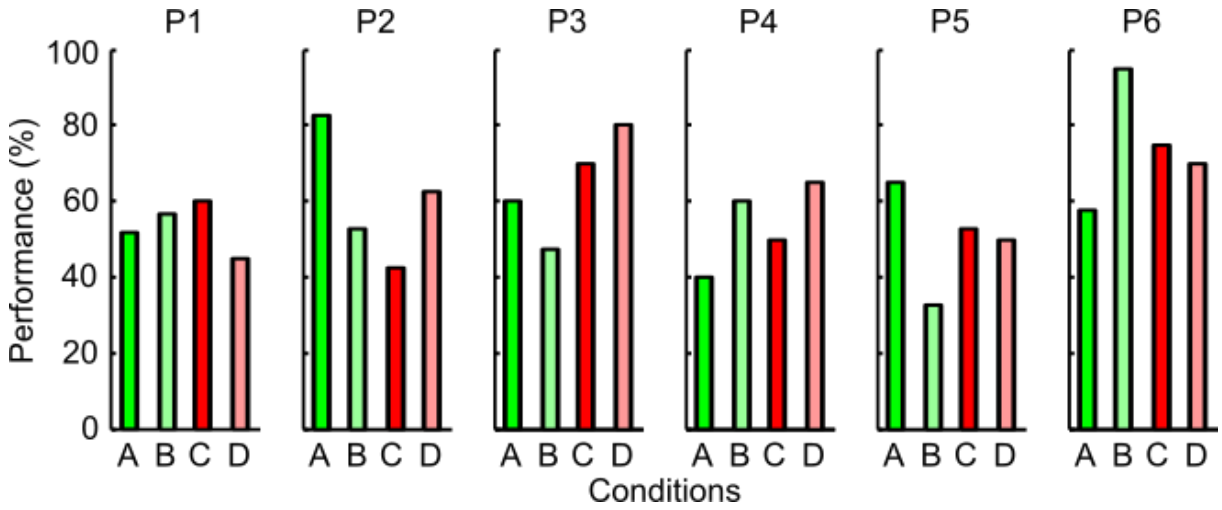


Figure 51**: Individual performance of epileptic patients during TARGET/T1 sessions.** Performance is described as the average proportion of trials where each insular patient (n=6) chose the 'correct' stimulus in the gain and loss conditions. Gain safe is in bright green, Gain risky in light green, Loss safe in bright red and Loss risky in light red. Patient 1 performed TARGET sessions as part of the original version of the task while patients 2 to 6 performed T1 sessions as part of the modified version of the task.

III. DISCUSSION

A. Vue d'ensemble de ces travaux de recherche

Le but premier des travaux de recherche présentés dans ce manuscrit était d'étudier la dynamique cérébrale de l'apprentissage par renforcement chez l'humain.

La première étude réalisée chez des patients épileptiques implantés au niveau cortical en sEEG s'est intéressée à la dynamique de l'encodage des signaux de renforcement, et en particulier à celui des erreurs de prédiction des récompenses et des punitions. L'enregistrement de potentiels de champs locaux a permis de mettre en évidence le rôle central de l'activité à haute-fréquence gamma (50-150Hz) dans ces encodages. Nous avons observé que la magnitude de ces réponses augmentait significativement suite à la présentation de récompenses au niveau du cortex préfrontal ventromédian, et suite à la délivrance de punitions dans l'insula antérieure, le cortex préfrontal dorsolatéral et le cortex orbitofrontal latéral. De plus, l'utilisation combinée des enregistrements électrophysiologiques et d'un modèle computationnel de Q-learning a souligné le rôle majeur de ces réponses gamma. En effet, pour chacune de ces quatre régions, nous avons relevé une corrélation significative entre les modulations gamma et la magnitude des erreurs de prédictions, signe d'un encodage possible de ces erreurs de prédictions des récompenses dans le vmPFC et des punitions dans l'insula antérieure, le dlPFC et le latOFC par l'activité oscillatoire gamma. De plus, nous avons également observé un encodage par l'activité gamma de la valeur subjective du symbole choisi au sein de ces régions. Pour finir, l'activité gamma de l'insula antérieure en réponse à une punition semble être un paramètre fonctionnel prédictif du comportement puisqu'une réponse gamma particulièrement rapide, forte et intense est associée avec une adaptation comportementale en vue d'éviter une future punition. En effet, nous avons vu que l'activité haute fréquence gamma de l'insula antérieure augmente plus lors des essais « switch » que des essais « stay » (voir Figure 34 issue de l'étude 1), ce qui suggère un rôle de la réponse gamma de l'insula antérieure aux punitions dans l'adaptation du comportement. Les résultats de cette première étude corroborent la théorie selon laquelle il existerait deux systèmes corticaux, l'un traitant de l'apprentissage par récompenses comprenant le cortex préfrontal ventromédian, l'autre de l'apprentissage par évitement de punitions grâce à l'action combinée de l'insula antérieure, du cortex préfrontal dorsolatéral et dans une moindre mesure du cortex orbitofrontal latéral. L'utilisation d'une analyse de causalité de Granger nous a permis de souligner les interactions entre ces régions corticales. L'insula antérieure semble ainsi avoir un rôle prépondérant au sein du circuit puisqu'elle influence causalement les réponses aux

renforcements aversifs au sein du dlPFC, du latOFC et du vmPFC par le biais d'activité basse-fréquence. Le dlPFC a la particularité d'influencer en retour l'activité de l'insula antérieure suite aux punitions, mais cette fois-ci grâce à une activité haute-fréquence gamma. Il semble donc y avoir un dialogue entre l'insula antérieure et le dlPFC. Au contraire, l'information allant de l'insula antérieure vers le cortex orbitofrontal apparaît comme monodirectionnelle. Suite à des punitions, l'activité de l'insula antérieure influence en cascade celle du latOFC puis celle du vmPFC. A noter que le lien causal allant de l'insula antérieure au vmPFC n'est pas modulé par la valence des renforcements, reflétant probablement un signal d'erreur de prédiction non signé. Il apparaît donc que bien qu'il existe deux systèmes au niveau cortical lors de l'apprentissage par renforcement, l'insula antérieure se présente comme une structure d'entrée au réseau avec un rôle fort de détection des punitions, et le vmPFC semble avoir un rôle de comparateur, voire de décisionnaire final en sortie du réseau. Ces interprétations seront discutées plus en détails par la suite.

Une fois nous être attelés au cortex dans la première étude, et en raison des circuits cortico-sous corticaux impliqués dans la détection des récompenses et punitions (voir I.B.2) il nous a paru important d'étudier également la dynamique de l'apprentissage par renforcement au niveau sous-cortical. Une région en particulier a attiré notre attention de par sa localisation anatomique à l'interface entre les ganglions de la base, dont le rôle dans la détection des récompenses a depuis longtemps été étudié (voir I.B.1), le cortex ayant été l'objet de la première étude. S'agissant d'une région cérébrale profonde, le thalamus reste très peu étudié chez l'humain, et ce malgré de premiers travaux lui supposant un rôle dans des processus cognitifs de haut-niveau comme la mémoire, la détection de récompenses et l'apprentissage instrumental (voir I.B.1.d) et I.B.2). Il est cependant possible d'enregistrer son activité électrophysiologique chez l'humain lorsque celui-ci est implanté à l'aide d'électrodes cérébrales profondes à visée clinique, ce qui a été notre cas grâce à l'étude France concernant l'utilisation de la SCP du NAT dans des formes d'épilepsie très sévères. Nous avons donc étudié la dynamique de l'activité cérébrale au sein du NAT et du NDMT au cours d'un protocole d'apprentissage par renforcement. L'enregistrement de potentiels de champs locaux a mis en évidence le rôle central des activités basse fréquence thêta dans la détection des renforcements, en particulier aversifs. Des activations plus haute fréquence (gamma et béta) ont également été observées avec réponses de plus faible durée. Ces résultats sont dans la lignée d'études antérieures suggérant un rôle du NAT et du NDMT dans l'apprentissage par évitement des punitions. Il faut cependant adresser le cas des fréquences variées des réponses oscillatoires enregistrées. En effet, la localisation sous-corticale de ces noyaux peut influer sur les fréquences des activités associées à des

processus cognitifs dits de haut-niveau. Des résultats préliminaires ont également permis d'étudier les effets causaux de la stimulation haute fréquence de ces noyaux thalamiques. Nous rapportons ici les premiers éléments qui permettront à l'avenir de tester le rôle causal des noyaux thalamiques antérieur et dorsomédian dans l'apprentissage par renforcement chez l'humain.

Enfin, une troisième étude présentée dans ce manuscrit s'est intéressée aux effets du risque sur l'apprentissage par renforcement. Il s'agit là des premiers résultats comportementaux puisqu'à ce jour, le risque et ses effets ont surtout été étudiés au cours de la prise de décision simple. Nous rapportons des effets différents de ceux observés au cours de la prise de décision simple, avec une aversion au risque présente lors de l'apprentissage par évitement des punitions sous la forme d'un déficit de performance. En parallèle, nous avons observé une diminution du temps de réaction uniquement lors de choix risqués permettant l'obtention de récompenses. Cela laisse supposer un comportement global tendant vers une aversion au risque lors de l'apprentissage par évitement des punitions et au contraire une attirance pour les choix risqués lors de l'apprentissage par récompenses. Ces résultats sont à l'inverse de ce qui est prédit par la Théorie des Perspectives de Kahneman et Tversky (1979), concernant l'influence du risque sur la prise de décision simple (voir I.A.3). Des études précédentes en imagerie fonctionnelle avaient proposé un rôle de l'insula antérieure dans l'aversion au risque au cours de la prise de décision simple. Nous apportons ici de premiers éléments corrélationnels concernant son implication au cours de l'apprentissage par renforcement grâce à des enregistrements électrophysiologiques intracérébraux.

Dans la suite de cette discussion générale, nous aborderons les aspects électrophysiologiques et fonctionnels ainsi que les aspects méthodologiques à prendre en considération concernant les travaux présentés dans ce manuscrit. Pour finir, nous discuterons les limitations et les perspectives soulevées par ces résultats.

B. Une vue plus complète du réseau cérébral sous-tendant l'apprentissage par renforcement et l'influence du risque

Les résultats des 3 études présentées dans ce manuscrit apportent des réponses concernant la dynamique intracérébrale de l'apprentissage par renforcement et de l'influence du risque sur celui-ci. Afin de mesurer les apports de ce travail par rapport à la littérature actuelle dans le domaine, un certain nombre de points ayant trait à la validité et aux contributions des résultats fonctionnels doivent être discutés. Pour commencer, je vais

discuter la validité du réseau fonctionnel proposé par les deux premières études. Les réponses apportées par ces résultats à la question de la ségrégation et/ou de l'intégration de la valence au cours de l'apprentissage par renforcement seront abordées dans un second temps. Troisièmement, le rôle du thalamus limbique au cours de ce processus décisionnel sera discuté. Finalement, les interactions entre le risque et la valence termineront cette discussion sur les aspects fonctionnels des travaux présentés dans ce manuscrit.

1. Apprentissage par renforcement : un réseau dynamique ancré anatomiquement

Au cours des deux premières études sur les corrélats neuronaux de l'apprentissage par renforcement au niveau cortical et thalamique, nous avons confirmé l'hypothèse selon laquelle il y aurait une dissociation corticale avec deux systèmes distincts traitant l'un de l'apprentissage par récompenses et l'autre de l'apprentissage par évitement des punitions. Nous avons également identifié deux nouveaux acteurs sous-corticaux à savoir les noyaux antérieurs et dorsomédians du thalamus. Ces études, basées sur des enregistrements intracérébraux de potentiels de champs locaux sous la forme d'activité extracellulaires haute fréquence, permettent d'avoir une vision plus globale du réseau anatomo-fonctionnel de l'apprentissage par renforcement. Cela permet aussi de lier entre eux les corrélats neuronaux déjà identifiés au sein des différentes régions impliquées que nous avions listées dans l'introduction.

En plus des résultats électrophysiologiques, donnant accès à une hypothèse corrélative de la structure du réseau, nous avons utilisé la méthode de causalité de Granger au cours de l'étude 1 sur le cortex, afin d'identifier des relations causales aux réponses observées. La causalité de Granger a déjà été utilisée avec succès pour étudier les relations intra-corticales sous-tendant la détection et le traitement des erreurs (Bastin et al. 2016; Billeke et al. n.d.). Dans le cadre de la première étude, la causalité de Granger nous a permis de proposer une structure au double réseau cortical sous-tendant la dissociation du traitement de l'apprentissage par récompenses et de l'apprentissage par évitement des punitions. Les acteurs principaux étant déjà identifiés (insula antérieure, vmPFC, latOFC et dlPFC), encore fallait-il les relier d'après leurs activités observées, de façon à rendre compte des influences réciproques au sein du réseau. L'analyse de Granger (voir article 1, Figure 36A) a donc identifié les connexions significatives suivantes. Lors de l'apprentissage par évitement des punitions, l'activité de l'insula antérieure influence celle du latOFC qui influence à son tour celle du vmPFC grâce à des activités extracellulaires d'assez basse fréquence (thêta, alpha

et bêta). Parallèlement, il existe un dialogue fonctionnel couvrant un spectre de plus haute fréquence (bêta et gamma) entre l'insula antérieure et le dlPFC, toujours au cours de l'évitement des punitions. Ce décours temporel de l'activité corticale au cours de l'apprentissage par évitement des punitions se retrouve au niveau des latences des réponses observées débutant tout d'abord au niveau de l'insula antérieure et du dlPFC (vers 300ms après le renforcement) et seulement ensuite au niveau du cortex orbitofrontal (latOFC vers 450ms). Lors de l'apprentissage par récompense, l'influence de l'insula antérieure sur le vmPFC se fait de manière plus directe, sans passer par le latOFC, et dans le spectre plutôt basse fréquence.

Il apparaît ainsi que l'insula antérieure a une place similaire à celle d'une structure d'entrée du réseau, en étant impliquée causalement dans les deux systèmes corticaux. Ceci est en lien avec le rôle connu de l'insula antérieure dans la détection des erreurs et des évènements saillants, quelle que soit leur valence (Ramautar et al. 2006; Hester et al. 2010; Bastin et al. 2016; Klein et al. 2013; Menon & Uddin 2010; Craig 2009a). Les connexions entre l'insula antérieure et dlPFC d'une part, et de l'insula antérieure vers le cortex orbitofrontal d'autre part, reposent sur des connexions anatomiques déjà identifiées grâce à l'utilisation de traceurs rétrogrades et antérogrades chez l'animal et à des études de tractographie chez l'humain (Bolstad et al. 2013; Hirose et al. 2016; Ghaziri et al. 2017; Kuramoto et al. 2017). De plus, chacune de ces régions a été identifiée comme recevant également des projections du NDMT et du NAT. L'analyse de la cytoarchitecture de ces différentes régions cérébrales a mis en évidence des projections formant deux réseaux ayant lieu majoritairement entre les régions agranulaires et le NAT d'un côté (insula antérieure, cortex orbitofrontal latéral et postérieur), et entre les régions granulaires et le NDMT de l'autre (insula postérieure, cortex orbitofrontal médian et antérieur) (Morecraft et al. 1992; Hof et al. 1995; Ray & Price 1993). A noter qu'il existe des connexions entre le NAT et le NDMT puisqu'il a été montré chez le rat qu'une lésion du NDMT induit une dénervation au niveau du NAT (Ouhaz et al. 2015). Au niveau cortical, similairement au gradient antéropostérieur granulaire-agranulaire retrouvé dans l'insula (Klein et al. 2013), il existe un gradient de granularité au niveau du cortex orbitofrontal avec des aires granulaires dans les régions antéro-médianes et des aires agranulaires dans les régions postéro-latérales (Morecraft et al. 1992; Hof et al. 1995). Tous ces éléments suggèrent la validité du modèle proposé par la causalité de Granger, puisque les liens identifiés comme permettant l'influence d'une région sur une autre sont tous rattachés à de véritables connexions anatomiques.

D'après le modèle proposé par la causalité de Granger, le vmPFC est présenté comme une structure de sortie du double réseau. Il reçoit ainsi des informations de la part de l'insula

antérieure lors de l'apprentissage par récompenses et des informations de la part du latOFC lors de l'apprentissage par évitement des punitions. Cette position de nœud de sortie est cohérente avec le rôle proposé du vmPFC comme un comparateur intégrant les informations des deux systèmes d'apprentissage afin de promouvoir la meilleure décision lors de l'émission d'une réponse cognitive (ici le choix de l'option). Des enregistrements de neurones unitaires dans l'OFC ont ainsi montré que « le traitement des stimuli appétitifs et aversifs converge au niveau du vmPFC, proposant une localisation anatomique aux processus exécutifs et émotionnels qui requièrent d'utiliser des informations provenant à la fois des systèmes appétitif et aversif » (Morrison & Salzman 2009). Cela implique que le vmPFC, tout comme ce qui était déjà connu pour le dlPFC, contribue à un apprentissage hybride basé sur les récompenses et les punitions (Abe et al. 2011).

Concernant l'utilisation de la causalité de Granger pour proposer un réseau cortical expliquant les résultats que nous rapportons dans l'étude 1, nous avons fait le choix de regarder les réponses moyennes au cours du temps, alors qu'il est également possible d'appliquer une causalité de Granger à la dynamique des réponses afin d'avoir une image de la dynamique réelle du réseau fonctionnel au cours de la cognition (Bastin et al. 2016). Cela pourrait ainsi permettre de mettre en évidence une probable dissociation temporelle des réseaux lors de l'apprentissage par récompenses et de l'apprentissage par évitement des punitions, et d'identifier s'il y a une intégration ou une ségrégation des réponses fonctionnelles d'un point de vue temporel.

Une autre technique que la causalité de Granger offre l'opportunité d'étudier la dynamique des réponses entre différentes régions cérébrales : la modélisation de la dynamique causale (DCM). Cette technique permet de tester des modèles plausibles de connectivité effective entre différentes régions formant un réseau potentiel. La technique de DCM a été utilisée pour tester un réseau dans une étude sur le rôle du noyau périaqueducal gris au cours d'une forme d'apprentissage par renforcement (Roy et al. 2014). Il s'agissait d'identifier la connectivité effective probable reliant des régions encodant les erreurs de prédiction à des régions encodant les valeurs afin de développer un modèle empirique du fonctionnement cérébral de l'apprentissage basé sur la douleur. En effet, malgré l'utilisation de modèles computationnels spécifiant des interactions dynamiques entre des régions encodant les renforcements, les valeurs subjectives et les erreurs de prédiction, les nombreuses études en IRMf réalisées n'avaient pas examiné la dynamique inter-régions, c'est-à-dire la connectivité effective. Bien que la technique de DCM puisse identifier des réseaux fonctionnels plausibles, elle reste complémentaire de la causalité de Granger pour se focaliser sur les fréquences ainsi que les latences des réponses observées. Cet aspect liant

la dynamique des réponses dans le réseau et les modèles computationnels utilisés est central pour identifier ce qui est réellement encodé par les régions au cours de l'apprentissage par renforcement : s'agit-il de l'attente du renforcement, d'une erreur de prédiction, d'une valeur subjective, de la détection d'un évènement saillant ou de tout autre signal impliqué dans un comportement dirigé vers un but ? (Bissonette & Roesch 2015).

2. Ségrégation et intégration de la valence au sein de ce réseau

L'un des objectifs des travaux présentés dans ce manuscrit était de confronter les hypothèses coexistantes concernant les corrélats neuronaux : 1) l'existence de deux systèmes opposés impliquant des aires cérébrales corticales et sous-corticales distinctes, ou au contraire 2) une ségrégation au sein même de ces régions cérébrales entre des neurones traitant d'une forme ou d'une autre de prise de décision, voire 3) un gradient fonctionnel de traitement de la valence au niveau neuronal. Grâce à la première étude en particulier, nous avons mis en évidence une **séparation** en deux systèmes de régions corticales, soutenant la première hypothèse. De précédents travaux sur les réponses corticales aux stimuli appétitifs et aversifs chez l'humain suggéraient déjà l'existence de ces deux systèmes, et les latences des réponses que nous avons enregistrées au niveau de l'insula antérieure, du vmPFC, du latOFC et du dlPFC sont cohérentes avec celles rapportées alors (Jung et al. 2010). Cependant, lorsque l'on s'intéresse aux différents bipoles enregistrés de manière individuelle au sein de chaque région d'intérêt, donc à de grands groupes de neurones pour chaque bipole, il y a en réalité une variété de réponses avec des bipoles encodant uniquement les erreurs de prédictions de récompenses, d'autres uniquement les erreurs de prédiction des punitions et d'autres les deux, ainsi que certains bipoles ayant une modulation significative de leur gamma suite à un renforcement (appétitif ou aversif) mais ne présentant pas d'encodage des erreurs de prédiction (voir Figure 52 pour les distributions au niveau du vmPFC et de l'insula antérieure). C'est d'ailleurs également le cas lorsque l'on regarde les encodages des valeurs subjectives.

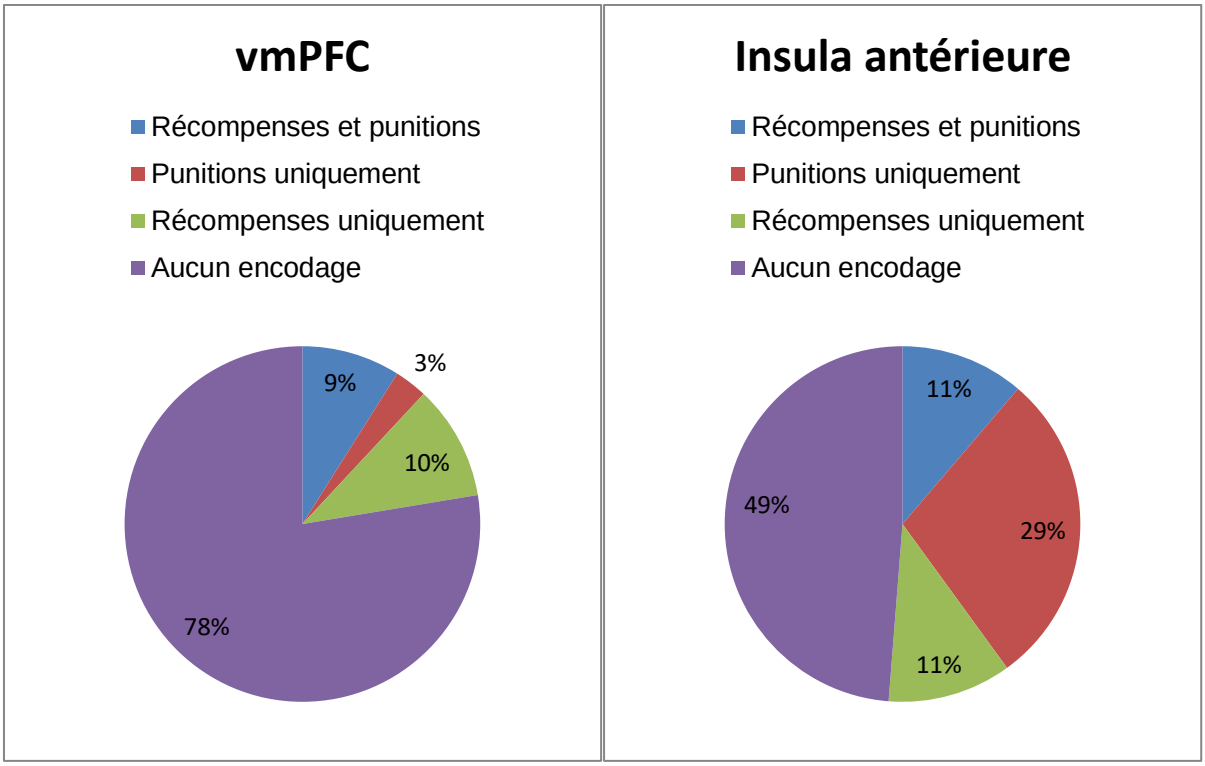


Figure 52 : Distribution de l'encodage des erreurs de prédiction parmi les bipoles au sein du vmPFC et de l'insula antérieure par une activité extracellulaire haute fréquence gamma. Pour chacun des bipoles enregistrés à travers les patients (67 bipoles dans le vmPFC et 80 dans l'insula antérieure, chez 11 patients dans les deux cas) nous avons identifié des bipoles dont l'activité gamma était significativement corrélée avec la magnitude des erreurs de prédiction des récompenses (vert) des punitions (rouge) des deux (bleu) ou sans encodage des erreurs de prédiction des renforcements (violet).

Ces résultats à l'échelle d'une région d'intérêt sont en lien avec l'hypothèse 2 d'une **ségrégation** au sein d'une même région des réponses aux renforcements appétitifs et aversifs, ce qui a également été retrouvé chez le singe au niveau de l'insula (Caruana et al. 2011). Toutefois, nous n'avons pas observé de gradient antéro-postérieur dans l'insula antérieure au niveau des réponses aux renforcements appétitifs et aversifs. Pour finir, la technique de sEEG consiste en l'enregistrement de potentiels de champs locaux, c'est-à-dire de l'activité de groupes de plusieurs dizaines de neurones, et ne permet donc pas d'avoir de réponse quant à l'existence de neurones étant capables de combiner ces différents encodages. Pour répondre à la question de la **combinaison** des réponses (hypothèse 3), il est nécessaire de réaliser des enregistrements unitaires de neurones. Cela a beaucoup été réalisé au niveau sous cortical chez l'animal et l'humain (Schultz 1997; Hayashi et al. 2015; Cohen et al. 2015), également au niveau cortical mais majoritairement chez l'animal (Markowitsch & Pritzel 1987; Asaad & Eskandar 2011; Silvetti et al. 2011) en raison de la difficulté de trouver des patients ayant des microélectrodes corticales (Sheth et al. 2012).

3. Le thalamus limbique, un nouvel acteur de l'apprentissage par renforcement ?

A ce jour, la communauté scientifique a une vision assez cortico-centrée de la prise de décision (Balleine & O'Doherty 2010; Gonzalez et al. 2015). Cela est dû à de multiples facteurs dont l'idée qu'un tel processus cognitif de haut niveau implique forcément (et majoritairement) le cortex préfrontal, en faisant une cible de choix, mais aussi la relative facilité d'étude du cortex en comparaison aux régions sous-corticales, en particulier chez l'humain. L'étude de ce processus est supposément d'autant plus subtile chez l'humain que chez les rongeurs, à la fois à cause de la taille relative plus imposante du cortex préfrontal mais aussi de la possibilité d'étudier des nuances complexes tout en pouvant échanger avec les sujets des études. Cependant, et comme pour nombre de processus cognitifs, le cortex n'est pas le seul acteur, appuyant ainsi l'intérêt de s'intéresser aux contributions d'origine sous-corticale.

Dans le cadre de l'étude 2, nous nous sommes intéressés au rôle du thalamus au cours de l'apprentissage par renforcement, et en particulier à celui des noyaux antérieurs et dorsomédian du thalamus. Pour cela, nous avons réalisé des enregistrements de l'activité extracellulaire haute fréquence (potentiels de champs locaux) grâce à des macro-électrodes ciblant le NAT afin de mettre en place sa stimulation cérébrale profonde pour traiter des formes résistantes d'épilepsie temporale (Fisher et al. 2010; Lehtimäki et al. 2016). De par la trajectoire de l'implantation bilatérale des électrodes, nous avons eu accès à l'activité extracellulaire des noyaux antérieurs ainsi que des noyaux dorsomédians situés dans le prolongement de cette trajectoire. Les patients inclus dans cette étude ont réalisé un protocole cognitif d'apprentissage par renforcement pendant les enregistrements en sEEG, réalisés avant l'internalisation des câbles et le raccordement au stimulateur.

Le thalamus a depuis longtemps été étudié par le biais d'études lésionnelles chez l'animal dans le cadre de processus cognitifs, avec un intérêt particulier pour les processus mnésiques et les phénomènes d'approche et d'évitement (Gabriel et al. 1977; Freeman et al. 1996; Sparenborg & Gabriel 1992). Depuis, de plus en plus de travaux ont été entrepris pour étudier différents processus cognitifs allant de l'établissement de la mémoire (Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016; Štillová et al. 2015) à la détection de renforcements (de Bourbon-Teles et al. 2014; Smith et al. 2002; Knutson et al. 2000; Thut et al. 1997; Bradfield et al. 2013; Ouhaz et al. 2015) en passant par la détection d'évènements saillants (Brázdil et al. 2007), le conditionnement à la peur et des processus visuo-moteurs (Bočková et al. 2015). Un point sur lequel s'accordent les études sur le rôle cognitif du thalamus est l'implication majeure des noyaux antérieurs et médians

lors de processus cognitifs impliquant des régions du cortex préfrontal (Bradfield et al. 2013; Ouhaz et al. 2015; Alcaraz et al. 2015; Dalrymple-Alford et al. 2015; Sun et al. 2015). Malgré l'absence d'études, à notre connaissance, portant sur l'activité du NAT et du NDMT au cours de l'apprentissage par renforcement, le nombre non négligeable d'études à notre disposition indique un rôle probable des activités extracellulaires gamma, bêta et thêta de ces noyaux pendant la cognition (Parnaudeau et al. 2013; Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016; Bočková et al. 2015; Rektor et al. 2016; Wang et al. 2015). Il est aussi très probable que ces noyaux soient impliqués dans l'apprentissage d'associations stimulus-stimulus (Bradfield et al. 2013; de Bourbon-Teles et al. 2014; Ouhaz et al. 2015), dans la détection d'évènements appétitifs (Thut et al. 1997; Parnaudeau et al. 2015) et aversifs (Knutson et al. 2000; Conejo et al. 2007; Kuramoto et al. 2017; Wang et al. 2015) et finalement dans l'apprentissage par renforcement (Smith et al. 2002). Notre but au cours de l'étude 2 était donc de développer une méthode permettant de confirmer ces hypothèses en identifiant le rôle des potentiels de champs locaux au sein du NAT et du NDMT pendant l'apprentissage par renforcement.

Les résultats de l'analyse temps-fréquence de l'activité thalamique présentés dans l'étude 2 et obtenus chez 5 patients épileptiques rapportent des réponses au niveau du NAT et du NDMT lors de l'apprentissage par renforcement et en particulier suite à la délivrance de punitions. Ces réponses prennent la forme d'augmentation très rapide de l'activité extracellulaire haute-fréquence bêta et gamma ainsi qu'une augmentation plus soutenue dans le temps de l'activité thêta, bien que moins spécifique des punitions. Contrairement à l'équipe de Sweeney-Reed et à leurs travaux sur l'implication du NAT dans les processus mnésiques, nous avons procédé à une analyse des modulations de l'amplitude des réponses aux analyses de couplage phase/amplitude qu'ils avaient réalisées. Ces deux types d'analyses représentent deux visions différentes des activités enregistrées : soit avec un regard centré sur le noyau en lieu-même (analyse de l'amplitude des réponses), soit avec une vision du réseau en interaction (analyse des couplages de phase et d'amplitude entre régions). Une analyse réalisée par ce groupe en 2016 a mis en évidence une influence entre l'activité thêta du NDMT et l'activité gamma du NAT et du cortex préfrontal pendant la formation de mémoire. Cela confirme l'importance des activités basse (thêta) et haute (gamma) fréquence au niveau de ces noyaux du thalamus au cours de processus cognitifs impliquant le cortex préfrontal, comme la prise de décision. Leurs résultats et les nôtres concernant le NAT sont comparables dans la mesure où ils se complètent. En effet, la mémoire est un processus cognitif essentiel sous-tendant l'apprentissage par renforcement, puisqu'il est nécessaire de mémoriser les associations stimulus-renforcement afin d'identifier essai par essai les stimuli optimaux. Cela peut expliquer la présence de cette augmentation

prolongée de l'activité thêta suite à la délivrance d'un renforcement. Il est possible qu'elle soit le signe de la mise à jour de la valeur subjective du stimulus choisi ayant été mise en mémoire. Cependant, comme il semble y avoir une modulation de cette activité par la valence des renforcements, indiquée par une plus forte modulation thêta suite aux punitions, il est probable qu'elle ne reflète pas uniquement un processus mnésique pur, comme des travaux chez le singe le suggèrent (Chakraborty et al. 2016). La présence de modulations de l'activité haute fréquence gamma et bêta dans le NAT et le NDMT au cours de l'apprentissage par renforcement, bien que de très courte durée en comparaison des modulations thêta, est cohérente avec les travaux ayant mis en évidence un rôle de ces activités extracellulaires haute fréquence dans ces noyaux pendant des processus cognitifs impliquant le cortex préfrontal, comme c'est le cas pour l'apprentissage par renforcement avec le cortex orbitofrontal en particulier (Rektor et al. 2016; Bočková et al. 2015; Parnaudeau et al. 2013; Wang et al. 2015). En raison du faible nombre de patients inclus à ce jour dans cette étude, les statistiques présentées dans l'article 2 sont non corrigées. Afin d'augmenter la puissance statistique et de confirmer ces résultats, l'inclusion de patients supplémentaires est nécessaire et toujours en cours. Toutefois, pris ensemble, ces résultats confirment l'existence au niveau sous-cortical de deux nouveaux acteurs de la prise de décision : le noyau antérieur du thalamus et le noyau dorsomédian du thalamus.

4. Intégration fonctionnelle de la valence et de la variance du renforcement

Au cours de la troisième étude, nous nous sommes intéressés à l'influence qu'a le risque sur l'apprentissage par renforcement chez l'humain. Alors que les effets du risque sur la prise de décision simple ont été l'objet de plusieurs études chez l'animal et chez l'humain (O'Neill & Schultz 2013; Mohr et al. 2010; Wu et al. 2012), ils ne l'ont pas été au cours de l'apprentissage par renforcement. Seulement peu de travaux sur l'apprentissage par renforcement couplé au risque ont été réalisés (Niv et al. 2012), cette étude a été réalisée chez des sujets sains, donnant accès uniquement à des données comportementales. Nous utilisons la définition du risque comme étant la variance du renforcement, utilisée comme un standard dans le domaine de la finance (Markowitz 1952; Bossaerts 2010).

Kahneman et Tversky ont théorisé les effets du risque lors de choix simples, hors de toute condition d'apprentissage (voir dans l'introduction générale et (Kahneman & Tversky 1979) sous le nom de Théorie des Perspectives. En se basant sur des résultats expérimentaux, ils expliquent que le risque a une influence variable sur la prise de décision, puisqu'elle induit

une aversion au risque pour les gains et une attirance pour le risque pour les pertes, un phénomène aussi connu sous le nom d'effet de réflexion (Baucells & Villasís 2010).

Au cours de l'apprentissage par renforcement dans le cadre de l'étude 3, nous avons rapporté l'existence d'une aversion au risque lors de l'apprentissage par évitement des punitions, induisant un déficit d'apprentissage dans la condition perte risquée (où les choix corrects permettant d'éviter une perte d'argent fictif sont associés à un risque, sessions « Target 1 »). Contrairement aux effets observés pendant la prise de décision simple, nous n'avons pas observé d'influence du risque au cours de l'apprentissage par récompense. De plus, nous avons développé plusieurs versions du protocole d'apprentissage par renforcement face au risque. La version originale du protocole, décrite dans l'étude 3, consiste en l'alternance de sessions « Target 1 » (paires associant un symbole risqué et un symbole non risqué et dit « sûr ») et de sessions « Control » (appariement de deux symboles risqués ensemble ou de deux symboles sûrs ensemble). Dans la version modifiée, nous avons remplacé les sessions « Control » par des sessions « Target 1 » modifiées et renommées « Target 2 » où la différence de valeur entre les symboles a été augmentée en modifiant la magnitude des renforcements (±0.5€ au lieu de ±1€ et ±0.2€ au lieu de ±0.5€) ainsi que les probabilités associées (voir Figure 49 dans les matériels supplémentaires de l'étude 3). Dans les sessions « Target 1 » de la version modifiée du protocole, nous avons observé une aversion au risque pendant l'apprentissage par récompenses mais pas d'influence du risque pendant l'apprentissage par évitement des punitions, rappelant les effets du risque observés sur la prise de décision simple par Laury et ses collaborateurs avec une aversion au risque en gain et une relation neutre en perte (Laury & Holt 2005). En parallèle des effets du risque sur la performance pendant l'apprentissage par renforcement, nous avons également étudié les effets du risque sur les temps de réaction dans les différentes conditions. Nous avons reproduit les résultats obtenus en l'absence de risque avec une augmentation du temps de réaction lors de l'évitement des punitions par rapport à ceux mesurés lors de la recherche des récompenses, dans les deux premières études de ce manuscrit. Ceci est en lien avec des études précédentes indiquant que les individus sont plus lents lors de l'approche de stimuli aversifs et plus rapides lors de l'approche de stimuli appétitifs (Crockett et al. 2009; Guitart-Masip et al. 2011). Malgré la reproduction d'effets connus de la valence des renforcements sur les temps de réaction, nous avons eu des difficultés à mettre en évidence des effets cohérents du risque sur la performance à travers les différentes versions de notre protocole, bien que les sessions « Target 1 » soient exactement identiques entre les versions. Cette incohérence des effets du risque malgré des sessions identiques peut s'expliquer par un changement du contexte au sein duquel ces sessions ont eu lieu. Wright et ses collaborateurs ont étudié les effets variables du risque sur

la prise de décision simple rapportés dans différentes études, supposant que le contexte jouait un rôle essentiel dans l'influence du risque sur le comportement (Wright et al. 2012). Ils ont ainsi montré que le risque et la valence influencent indépendamment les choix. Cette dissociation n'était pas prédite par la Théorie des Perspectives. De plus, les travaux de Wright permettent d'expliquer à la fois les résultats classiques que ceux plus controversés à cause de leur incohérence avec les travaux classiques comme ceux de Laury. Pour cela, Wright et ses collaborateurs proposent l'existence d'un processus de prise en compte des choix passant de l'évaluation des options à la sélection de l'action (Corrado et al. 2009). En l'occurrence, l'évaluation des options implique des systèmes cérébraux séparés traitant la valence et le risque, quand la sélection de l'action suppose la contribution de mécanismes d'approche et d'évitement. En conclusion, il serait utile de développer un modèle computationnel adaptable pour modéliser les effets du risque sur l'apprentissage par renforcement au cours des différentes versions de notre protocole pour prendre en compte le contexte expérimental.

C. Deux techniques complémentaires pour étudier la dynamique intracérébrale

1. Apports des enregistrements de potentiels de champs locaux

De nombreuses techniques existant pour étudier les modulations de l'activité cérébrale au cours d'un processus cognitif, il est nécessaire de comparer ceux obtenus en sEEG dans nos travaux avec ceux de la littérature utilisant les techniques d'EEG de surface, d'IRMf et d'enregistrements unitaires.

a) Enregistrements électrophysiologiques de surface et intracrâniens chez l'humain

Classiquement, si l'on s'intéresse à la dynamique cérébrale humaine, la technique préférentiellement utilisée est l'EEG ou la MEG pour étudier l'activité nerveuse, mais elle suppose une faible résolution spatiale. Similairement, si l'on s'intéresse à des évènements nerveux ayant lieu au sein de régions cérébrales profondes, la technique de neuroimagerie de choix est l'IRMf, bien qu'elle soit associée à une perte de précision temporelle, ne permettant donc pas d'étudier correctement des processus dynamiques. Chez l'animal, les enregistrements électrophysiologiques permettent d'étudier la dynamique des processus

nerveux avec une grande précision spatiale, mais cette technique est hautement invasive puisqu'elle consiste en l'implantation d'électrodes directement à l'intérieur du cerveau. De plus, cette dernière technique ne peut que concerner l'enregistrement de l'activité d'un faible nombre de régions cérébrales à la fois ainsi qu'un échantillonnage restreint en termes de nombre de cellules enregistrées.

Comme nous avons pu le voir dans l'introduction concernant les corrélats neuronaux de l'apprentissage par renforcement, il y a une contradiction entre les résultats fournis par les études effectuées sur l'animal et chez l'humain concernant la dynamique et le type de réponses fonctionnelles associées aux différentes aires cérébrales impliquées dans l'apprentissage par renforcement, en particulier au niveau de l'insula antérieure et du cortex orbitofrontal. En effet, les études effectuées en EEG de surface ne révèlent pas toujours une implication de l'insula antérieure dans l'évitement des punitions, avec certains travaux suggérant son rôle dans la détection d'évènements saillants sans influence de leur valence (Menon & Uddin 2010; Ullsperger et al. 2010; Bastin et al. 2016).

Dans le cas des activités associées au cortex orbitofrontal (vmPFC et latOFC) et à l'insula antérieure par des études chez l'humain utilisant des enregistrements électrophysiologiques de surface (EEG), la localisation basée sur une reconstruction des sources induit une forte incertitude concernant des réponses produites par des sources profondes. Or, l'OFC et l'insula sont des aires cérébrales profondes. C'est ainsi que les réponses aux erreurs peuvent être rapportées comme provenant du cortex pariétal suite à la reconstruction des sources en EEG, sous forme d'une onde positive 300ms après l'erreur, nommée P300 (Fischer & Ullsperger 2013). En conséquence, les techniques d'EEG et de sEEG, bien que permettant toutes deux d'enregistrer l'activité électrophysiologique cérébrale et étant donc directement comparables, peuvent suggérer l'existence d'activités ayant des localisations contradictoires lorsque des aires profondes sont concernées.

b) Lien entre potentiels de champs locaux et signal BOLD

Afin d'étudier la dynamique des processus cérébraux de prise de décision, nous avons eu au cours de ma thèse l'opportunité d'utiliser des enregistrements électrophysiologiques (sEEG) chez des humains au cours d'un protocole cognitif pour lequel les réponses BOLD (IRMf) sont bien établies. En effet, l'apprentissage par récompense entraîne des réponses BOLD dans le vmPFC et l'apprentissage par évitement des punitions dans l'insula antérieure (Palminteri et al. 2012). Ainsi, les résultats obtenus au cours des études présentées dans les articles 1 (cortex) et 2 (thalamus) utilisant la sEEG pour enregistrer des potentiels de champs

locaux (LFP) peuvent être méthodologiquement comparés à ceux obtenus classiquement en IRMf.

Il a été montré que les signaux à haute fréquence enregistrés en potentiels de champs locaux sont corrélés avec les réponses BOLD évoquées (Logothetis et al. 2001; Gaglianese et al. 2016). Dans notre première étude, au cours de l'apprentissage par récompenses, nous avons observé une corrélation entre l'activité gamma et les erreurs de prédiction de récompenses dans le vmPFC. Au cours de l'apprentissage par évitement des punitions, nous avons observé une corrélation entre l'activité gamma et les erreurs de prédiction des punitions dans l'insula antérieure et le dlPFC. Ces résultats sont similaires à ceux retrouvés en IRMf dans le vmPFC et l'insula antérieure ((Pessiglione et al. 2006) et Figure 53) où une corrélation entre la magnitude des erreurs de prédiction et la réponse BOLD était observée (Jo & Jung 2016). Ces résultats similaires provenant de neuroimagerie et de potentiels de champs locaux laissent supposer qu'il est possible de comparer la dynamique des réponses fonctionnelles enregistrées sous forme de modulation du signal BOLD (IRMf) avec celles enregistrées sous forme d'activité haute-fréquence gamma (sEEG).

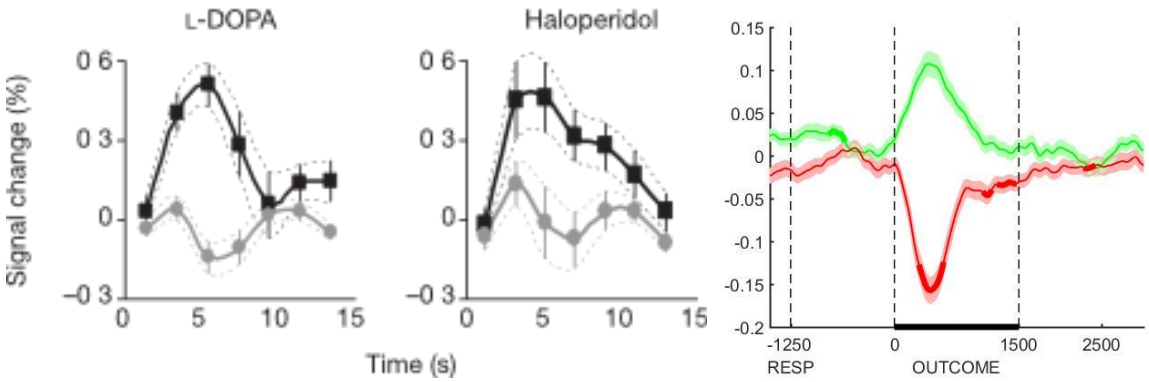


Figure 53: Encodage fonctionnel des erreurs de prédiction des renforcements au sein de l'insula antérieure rapporté en IRMf et en sEEG. (Gauche) La réponse BOLD observée au sein de l'insula antérieure en IRMf est corrélée avec la magnitude des erreurs de prédiction des punitions. Figure adaptée de (Pessiglione et al., 2006). Les corrélations de la réponse BOLD avec les erreurs de prédiction positives (en noir) et négatives (en gris) sont montrées pour les sujets sains hyper (L-DOPA) et hypo (Halopéridol) dopaminergiques. **(Droite)** Le même profil de réponse a été observé dans les réponses haute-fréquence gamma enregistrées en sEEG au niveau de l'insula antérieure (figure adaptée de la figure 3A de notre première étude en sEEG) En vert est représentée la corrélation entre l'activité gamma en condition gain avec la magnitude des erreurs de prédiction des récompenses, en rouge celle en condition perte avec la magnitude des erreurs de prédiction des punitions. Les sections épaissies représentent les encodages significatifs, c'est-à-dire les instants pour

lesquels il existe un effet significatif de la valence des erreurs de prédiction (erreurs de prédictions des récompenses ≠ erreurs de prédictions des punitions, $p<0.05$ corrigée au cluster).

Les études effectuées en IRMf ne révèlent pas toujours les mêmes régions cérébrales impliquées au cours de l'apprentissage par renforcement. Bien que le vmPFC soit régulièrement identifié comme impliqué dans l'apprentissage par récompense (Pessiglione et al. 2006; Vassena et al. 2014; Hampton et al. 2007), celle du latOFC dans l'apprentissage par évitement des punitions est bien plus rare, surtout en neuroimagerie (Ursu & Carter 2005). Bien que ces différences puissent être dues dans une certaine mesure à l'utilisation de différents modèles animaux, il a été montré qu'il existe une homologie anatomique à travers les mammifères (rongeurs, singes et humains) permettant de comparer les activités retrouvées au niveau de l'insula (Hayes et al. 2014) et du cortex orbitofrontal (Petrides 2005; Mackey & Petrides 2010; Mackey & Petrides 2014). Reste donc le rôle possible de la technique et du protocole utilisés dans la génération de ces contradictions.

Afin de pouvoir comparer objectivement les résultats des études de neuroimagerie fonctionnelle (IRMf) avec ceux des études en sEEG, il est nécessaire de savoir si la réponse BOLD enregistrée en IRMf et les activités haute-fréquence enregistrées en potentiels de champs locaux (sEEG) représentent les mêmes phénomènes physiologiques. Il a été montré que les activités haute-fréquence sont corrélées avec les réponses évoquées du signal BOLD (Logothetis et al. 2001). Dans notre première étude, nous avons observé une corrélation entre l'activité haute-fréquence gamma et les erreurs de prédiction des récompenses dans le vmPFC, et avec celles des punitions dans l'insula antérieure. Ces résultats sont en ligne directe avec ceux observés chez l'humain en IRMf au cours du même protocole cognitif (Pessiglione et al. 2006). Outre la similitude dans la localisation et la signification entre les réponses rapportées par Mathias Pessiglione et ses collaborateurs avec nos observations, il a été montré que les signaux hémodynamiques BOLD corrèlent fortement avec les activités haute fréquence gamma (Niessing et al. 2005). Cela nous permet donc de comparer directement nos résultats obtenus en sEEG avec l'intégralité de la littérature de neuroimagerie fonctionnelle. Bien que les activités haute-fréquence au niveau du cortex aient présenté un intérêt majeur en raison de leur lien démontré avec les processus cognitifs de haut-niveau (Lachaux et al. 2012), dont fait partie la prise de décision, les activités de plus basse fréquence (alpha, bêta et thêta) ont également un rôle à jouer. Ainsi, alors que les encodages fonctionnels sont réalisés au niveau local par des activités gamma (50-150Hz), les fonctions intégratives à longue distance exhibent des profils

spectraux plus variés, impliquant beaucoup la bande de fréquence bêta (13-35Hz) ((Donner & Siegel 2011) et Figure 54).

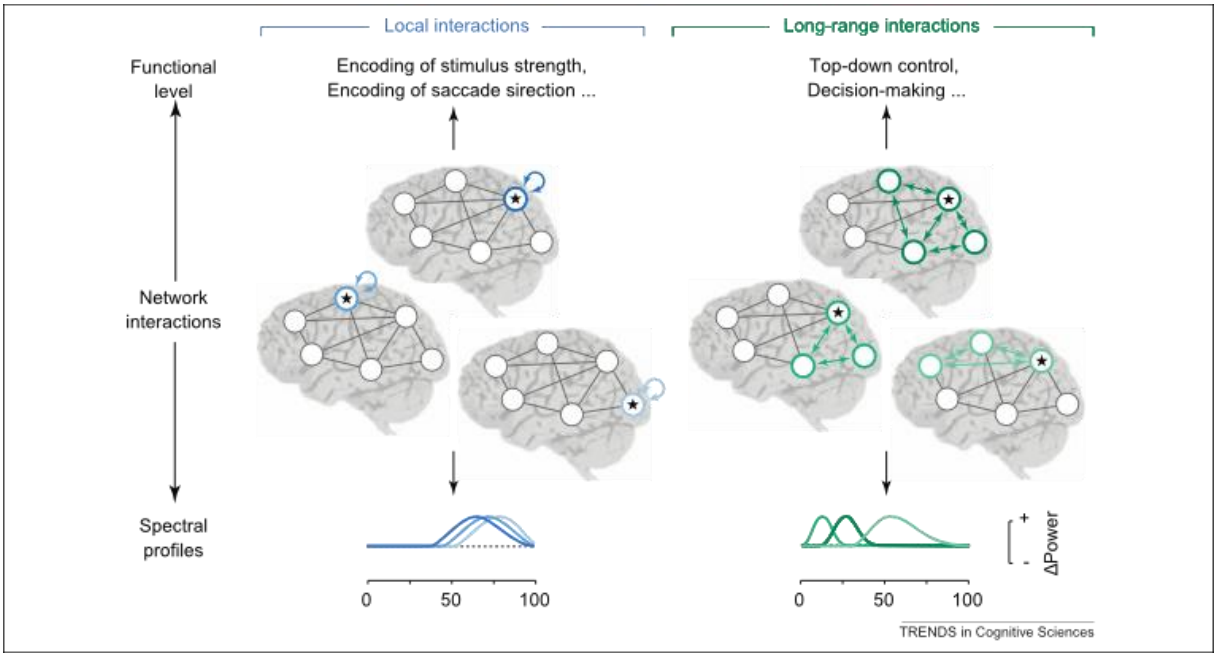


Figure 54 : Les interactions au sein des réseaux cérébraux influent sur le profil spectral des fonctions cognitives. Les populations neuronales locales sont impliquées dans deux types d'interactions (locales et à longue distance) au sein des réseaux cérébraux lors de l'encodage au niveau local et des fonctions intégratives, donc leur activité locale possède des profils spectraux différents. **(Milieu)** Six exemples hypothétiques de processus à longue distance et locaux. Chaque cercle correspond à un groupe local de neurones constituant un nœud d'un réseau à grande échelle. Les lignes et boucles représentent respectivement les connexions à distance et locales récurrentes. **(Haut)** Exemples de fonctions cognitives prises en charge par ce type d'interactions au sein du réseau. **(Bas)** Exemple des profils spectraux exprimés dans les réseaux locaux indiqués pas les astérisques. Alors que les fonctions d'encodage local reposent sur des activités de type gamma (<50Hz), les fonctions intégratives à longue distance impliquent des profils spectraux plus divers, dont beaucoup de type bêta. Figure issue de (Donner & Siegel 2011).

Ainsi, dans notre seconde étude portant sur le rôle de plusieurs noyaux du thalamus dans l'apprentissage par renforcement, nous avons effectué des analyses temps-fréquence afin d'identifier les bandes de fréquence particulièrement impliquées dans les activités observées. Cette analyse était d'autant plus importante à réaliser que les enregistrements directs de l'activité du thalamus au cours de processus cognitifs sont rares et que des activités basse fréquence (thêta, alpha et bêta) ont été rapportées au cours d'études des processus mnésiques (Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-

Reed et al. 2016). Nos résultats ont également mis en évidence une activité majeure de fréquence thêta (4-8Hz) ainsi que de plus faibles réponses de type alpha (8-13Hz) et bêta (13-35Hz) suite à la présentation de renforcements, en particulier aversifs. Peut-on alors mettre en lien ses activités basse fréquence avec les réponses BOLD rapportées dans le thalamus à des renforcements appétitifs ou aversifs (Kirsch et al. 2003; Knutson et al. 2000; Canessa et al. 2013) ? Une étude intéressante a investigué le lien des différentes fréquences d'activités enregistrées en potentiels de champs locaux (sEEG) avec l'amplitude et la dynamique du signal BOLD (Magri et al. 2012). Ils voulaient tester les prédictions d'une théorie suggérant qu'une augmentation de l'activité basse fréquence, alpha par exemple, pourrait induire une diminution du signal hémodynamique BOLD (Kilner et al. 2005). Après avoir testé et vérifié cette prédiction, Magri et ses collaborateurs sont allés plus loin en montrant que l'interaction entre les bandes alpha et gamma était reflétée dans l'amplitude du signal BOLD, alors que l'interaction entre les bandes bêta et gamma influait sur la latence de la réponse BOLD. Ainsi, au vu de ces résultats, nous pouvons supposer que les courtes réponses bêta et gamma observées dans le thalamus seraient à mettre en lien avec le rôle suggéré du thalamus dans la détection des renforcements par des études de neuroimagerie fonctionnelle, d'autant plus que leurs latences, de l'ordre de 500ms (voir figure 3 de l'étude 2 sur le thalamus), sont cohérentes avec la détection ponctuelle de tels évènements saillants (voir figures 3 et 5 de l'étude 1 en sEEG dans le cortex). Les activités très basse fréquence thêta que nous rapportons au cours de l'apprentissage par renforcement sont, quant à elles, cohérentes avec celles rapportées lors de processus mnésiques (Sweeney-Reed et al. 2014; Sweeney-Reed et al. 2015; Sweeney-Reed et al. 2016; Sweeney-Reed et al. 2017), validant leur implication probable dans ce processus cognitif. Par ailleurs, les activités équivalentes entre des régions corticales et des régions sous corticales impliquées dans des processus cognitifs identiques ont très probablement des fréquences différentes. Par exemple, les modulations de l'activité bêta ont été mises en évidence pour leur rôle dans le contrôle cognitif des réponses motrices au sein du noyau sous-thalamique (Benis et al. 2014; Bastin et al. 2014; Bastin et al. 2014). Ceci est en cohérence avec ce que suggéraient Lachaux et ses collaborateurs (Lachaux et al. 2012) : « les activités extracellulaires haute fréquence ne seraient pas uniquement le reflet de l'activité d'une population de neurones mais possèderaient également un rôle causal sur l'activité neuronale elle-même » (Fröhlich & McCormick 2010; Anastassiou et al. 2011).

c) Des enregistrements de neurones unitaires à ceux d'une population de neurones en sEEG

Pour finir, il a été montré qu'il est possible de comparer les résultats des études basées sur des enregistrements de potentiels de champs locaux et les résultats de celles centrées sur des enregistrements unitaires de neurones. Les enregistrements unitaires sont depuis longtemps utilisés chez l'animal, donnant accès à une large bibliographie concernant les corrélats neuronaux de la prise de décision et de la détection des renforcements (Schultz 1997; Thut et al. 1997; Yi Li et al. 2016; Mizuhiki et al. 2012; Matsumoto et al. 2016; Monosov & Hikosaka 2012). Ces enregistrements unitaires ont été particulièrement utilisés pour étudier l'activité électrophysiologique des régions et noyaux sous-corticaux au cours de processus cognitifs, permettant d'avoir une idée précise des activités individuelles des neurones de ces noyaux. Comme nous l'avons évoqué précédemment, il est suggéré que les activités extracellulaires haute fréquence, classiquement enregistrées en potentiels de champs locaux, soient liées à ces activités unitaires au niveau cortical (Lachaux et al. 2012; Fröhlich & McCormick 2010). Afin de confirmer le lien fonctionnel entre ces deux types d'enregistrements neuronaux, il est donc important de pouvoir les obtenir simultanément. Ces enregistrements simultanés de potentiels de champs locaux et d'activité unitaire ont mis en évidence une relation temporelle entre les activités haute fréquence et l'activité unitaire des neurones corticaux chez l'humain (Jacobs et al. 2007) indiquant que « les activités thêta (1-4Hz) et delta (<1Hz) facilitent l'encodage de phase quand les activités gamma aident à décoder les combinaisons de neurones actifs simultanément ». De plus, une corrélation a été observée entre les variations de fréquence des activités haute-fréquence et l'activité unitaire corticale chez l'humain (Manning et al. 2009). Ainsi, ces changements du spectre d'activité haute-fréquence prédisent de manière fiable l'activité unitaire, suggérant que l'intensité des potentiels de champs locaux fournit des informations importantes concernant l'activité neuronale.

2. Modèles computationnels et processus neuronaux

Dans l'étude présentée dans l'article 1, nous avons utilisé une approche basée sur un modèle pour analyser les signaux électrophysiologiques de potentiels de champs locaux enregistrés en sEEG et les associer avec les processus cognitifs sous-jacents impliqués dans l'apprentissage par renforcement au sein des différentes régions d'intérêt. Cette méthodologie nous a donné l'opportunité de révéler des paramètres cachés du comportement observé et d'ainsi développer des hypothèses quant aux mécanismes

cérébraux impliqués. L'utilisation d'un modèle computationnel est un outil pour examiner plus précisément les processus d'intérêt, bien que cela ne nous permette pas d'affirmer avec une complète certitude que le cerveau implémente réellement ce modèle exact. Il s'agit plus de proposer des mécanismes plausibles pour aller plus loin dans l'élaboration d'hypothèses concernant le comportement observé.

Ainsi, nous avons observé une corrélation entre les modulations de l'activité oscillatoire haute-fréquence gamma (50-150Hz) au sein de régions corticales avec nos variables d'intérêt, à savoir les erreurs de prédiction des renforcements et les valeurs subjectives. Nous avons ensuite interprété ces corrélations comme mettant en évidence des processus d'encodage fonctionnel de ces paramètres au sein des régions corticales en question. En raison de cette approche basée sur un modèle et utilisant des LFP, nos conclusions sont soumises à plusieurs limitations.

De manière générale, une approche basée sur un modèle consiste à construire un modèle computationnel explicite de la tâche cognitive que les participants ont à réaliser. Le but du modèle est d'apporter un lien entre les manipulations externes (la présentation de différents stimuli par exemple) et les réponses comportementales observées (Corrado & Doya 2007). Afin d'accomplir cette transformation, le modèle contient des variables cachées non observables. Pour générer de telles variables, le modèle utilise généralement des paramètres libres qui doivent être déterminés et ajustés afin de refléter au plus près les données observées (dans le cas de l'article 1, ces paramètres libres sont alpha et bêta). Ces variables sont ensuite entrées comme étant des modulateurs paramétriques pour prédire les données issues des enregistrements électrophysiologiques (EEG) ou de neuroimagerie fonctionnelle (IRMf). L'exemple le plus commun est d'utiliser le signal d'erreur de prédiction des récompenses estimé à partir d'un modèle d'apprentissage comme variable expliquant les données BOLD (O'Doherty et al. 2007). Dans notre étude présentée dans l'article 1, nous avons utilisé une approche similaire en corrélant les paramètres du modèle computationnel avec l'activité électrophysiologique de potentiels de champs locaux enregistrés en sEEG.

La force d'une approche basée sur un modèle en comparaison de méthodes plus simples de « soustraction » est que cela permet, non seulement d'identifier les régions « actives » au cours du protocole cognitif, mais aussi d'élaborer des prédictions à propos des fonctions portées par ces activations (O'Doherty et al. 2007). Malheureusement, lorsque l'on corrèle une activité électrophysiologique avec des variables cachées issues d'un modèle prédisant bien les données, il est inévitable de se retrouver face à une analyse restrictive ne considérant que des signaux très spécifiques. Pour rappel, la méthode de soustraction consiste à comparer les activations brutes observées suite à différents évènements d'intérêt

au cours du protocole cognitif, afin d'identifier les régions donc l'activité est significativement modulée par l'occurrence de ces évènements.

Dans notre cas, nous avons utilisé un algorithme de Q-learning permettant d'estimer les valeurs subjectives (Q values) des options ainsi que les erreurs de prédiction une fois les renforcements délivrés car ces erreurs de prédiction sont considérées comme le moteur de l'apprentissage par renforcement. Nous avons ensuite identifié les régions d'intérêt en cherchant les régions cérébrales dont l'activité gamma encode les erreurs de prédiction des renforcements. Nous avons également étudié l'encodage absolu de la valeur subjective du symbole choisi (Q choisi) au sein de nos régions d'intérêt. Il s'agissait de l'encodage le plus évident de la valeur de l'option choisie permettant de prendre la décision et celui le plus utilisé dans la littérature au cours de protocoles cognitifs similaires (Palminteri et al. 2012; Palminteri et al. 2009; Worbe et al. 2011; Fischer & Ullsperger 2013; Pessiglione et al. 2006). Il est cependant important de noter que d'autres études ont utilisé d'autres variables pour déterminer s'il y a un encodage des valeurs subjectives au sein du cerveau. C'est ainsi que certaines études se sont intéressées à la différence entre les valeurs subjectives des deux options, supposant que le choix de la meilleure option se base sur un encodage relatif et non absolu de la valeur subjective des options (Q choisi – Q non choisi), appelée « valeur d'état » ou « state value » (Palminteri et al. 2011; Palminteri et al. 2015; Setogawa et al. 2014; Niv, Joel, et al. 2006; Niv, Daw, et al. 2006; Guitart-Masip et al. 2014). L'utilisation d'un modèle proposant un encodage relatif des valeurs subjectives est supportée par des théories supposant que les valeurs du contexte du choix (l'état) sont apprises et représentées séparément dans le cerveau, permettant leur comparaison (Niv, Joel, et al. 2006; Guitart-Masip et al. 2014; Palminteri et al. 2015). Dans ce cas, le modèle relatif est centré autour de l'idée que la valeur d'état sert de point de référence qui sera à comparer avec le renforcement obtenu avant de mettre à jour la valeur de l'option choisie, rendant l'encodage de la valeur des options relatif et non plus absolu. Cela permet de capturer la valeur globale attendue pour une paire d'options donnée (ou de symboles), et ce de manière indépendante de la politique de choix des participants (Moutoussis et al. 2008; Maia 2010) pour une comparaison plus approfondie de ces modèles et la discussion de leurs différences).

Dans la première étude, pour étudier les encodages des erreurs de prédiction et des valeurs subjectives, nous avons isolé leurs valeurs pour étudier séparément leurs corrélations avec les activités extracellulaires à haute fréquence gamma. Cependant, les variances des signaux des erreurs de prédiction et des valeurs subjectives sont fortement corrélées de par la définition computationnelle même des erreurs de prédiction : la différence entre le renforcement obtenu et celui attendu (Equation 14 : $EP = R - Q$). Il est donc nécessaire

d'utiliser une corrélation multiple pour étudier les contributions séparées des erreurs de prédiction et des valeurs subjectives pour s'assurer d'étudier des encodages non biaisés. Ce facteur sera pris en considération dans les analyses futures des données issues de cette première étude.

Pour finir, lors d'une étude s'intéressant à identifier des réponses cérébrales reflétant un encodage fonctionnel d'une variable d'un modèle computationnel, des règles de conduite ont été énoncées par Roy et ses collaborateurs (Roy et al. 2014). Ces règles au nombre de 3 constituent l'approche axiomatique réelle, et permettent d'identifier des régions encodant un paramètre computationnel comme les erreurs de prédiction ou les valeurs subjectives. Bien que développée lors d'une étude en IRMf sur l'encodage des erreurs de prédiction de la douleur, cette approche axiomatique réelle est applicable à nos résultats en sEEG sur l'encodage des erreurs de prédiction des renforcements (voir IIIIII.C.1). Adaptée au cas de l'apprentissage par renforcement, cette approche implique les trois axiomes (vérités) suivants. (1) Le premier axiome (effet du renforcement) stipule que l'activité suite à la délivrance d'un renforcement d'intérêt devrait être supérieure à celle sans ce renforcement. Cet axiome a été testé par une simple ANOVA à mesures répétées entre les réponses gamma moyennes pour les quatre renforcements utilisés (voir figures 3 et 5, panneaux de droite). (2) Le second axiome (effet d'attente) stipule que l'activité devrait diminuer avec l'augmentation de la probabilité de renforcement. Cela est similaire à ce qu'avait observé Schultz au niveau des neurones dopaminergiques lors de la délivrance d'une récompense entièrement prédite (Schultz 1997). Pour tester l'axiome 2, il faudrait comparer les activités enregistrées suites à des punitions selon la valeur subjective du stimulus lorsqu'il est choisi. En effet, si une punition est entièrement prédite au cours de nos protocoles, cela veut dire que le patient sait qu'il perdra mais seulement dans une minorité d'essais. La valeur subjective du symbole correct dans la condition perte tend donc à être quasi-nulle. Dans le cas d'une punition attendue car entièrement prédite, la valeur subjective est donc proche de 0 et nous nous attendons à ce que la réponse de l'insula antérieure à cette punition soit très faible. Cet axiome, s'il était validé, prédirait l'observation d'un phénomène proche du conditionnement rapporté par Schultz en 1997 chez le rat et pour lequel les neurones dopaminergiques ne répondaient plus aux récompenses attendues après conditionnement. Enfin, l'axiome 3 (effet équivalent des surprises) stipule que les résultats totalement prévus devraient générer des réponses équivalentes.

Finalement, en écho aux considérations concernant l'échantillonnage du cerveau en sEEG (III.C.1), nous avons identifié, parmi les 35 parcelles MarsAtlas échantillonnées, seulement celles encodant les erreurs de prédictions par une activité gamma comme régions d'intérêt.

Ce choix, bien que valable d'un point de vue théorique en raison de rôle prépondérant des erreurs de prédiction dans l'apprentissage par renforcement, induit un biais de sélection. Il est en effet possible d'avoir une approche plus conservatrice en utilisant la méthode de soustraction pour identifier les régions d'intérêt. Nous avons ainsi également analysé nos données brutes en utilisant, comme filtre de sélection des régions d'intérêt, la présence d'une modulation significative de l'activité extracellulaire haute-fréquence gamma en réponse aux renforcements (récompenses et/ou punitions). Suite à cette analyse soustractive, les régions cérébrales identifiées comme répondant significativement aux renforcements sont : l'insula antérieure, le latOFC, le vmPFC, le dlPFC mais aussi le cortex cingulaire antérieur (CCA), l'amygdale et l'hippocampe. Ces nouvelles régions d'intérêt sont cohérentes avec la littérature actuelle sur l'encodage des erreurs de prédiction et la prise de décision ((Umemoto et al. 2014; Polli et al. 2008; Jung et al. 2010; Météreau & Dreher 2013) pour le CCA, (Esber & Holland 2014; Holland 2012; Rygula et al. 2014; Météreau & Dreher 2013) pour l'amygdale) mais aussi avec l'importance des processus mnésiques dans l'apprentissage par renforcement ((Dumont et al. 2010; Freeman et al. 1996) pour l'hippocampe). De plus, ces régions n'avaient pas été conservées comme régions d'intérêt d'après la sélection selon l'encodage des erreurs de prédiction car nous n'avions pas atteint le seul de reproductibilité et de significativité des effets (10 bipoles chez 3 patients) mais de très peu. L'inclusion de quelques patients supplémentaires pourrait permettre de dépasser ce seuil pour avoir un échantillonnage similaire de ces trois régions (CCA, amygdale et hippocampe) avec celui déjà observé pour l'insula antérieure, le latOFC, le dlPFC et le vmPFC.

D. Limites et perspectives des travaux présentés

1. Des réponses pathologiques ou physiologiques ?

Une question revient souvent lorsque l'on explique étudier des phénomènes physiologiques chez des patients. C'est le cas avec ces travaux pour lesquels il est essentiel de se demander si des enregistrements électrophysiologiques chez des patients épileptiques permettent d'enregistrer des activités physiologiques ou pathologiques au cours de protocoles cognitifs. En d'autres termes, un cerveau humain épileptique est-il un bon modèle de cerveau humain sain ? Cette question a obtenu une réponse fiable et parfaitement claire concernant la validité des travaux réalisés en sEEG chez ces patients (Lachaux et al. 2012) dont voici le contenu : « Parce que les enregistrements en sEEG sont toujours obtenus chez des patients souffrant de pathologies cérébrales majeures, il est légitime de se demander si

les conclusions de n'importe quelle étude intracérébrale peuvent être appliquées à des cerveaux sains. Les chercheurs utilisant la sEEG ont développé une série de règles afin de répondre à ces questionnements : (a) ne s'intéresser qu'aux sites d'enregistrement localisés loin du foyer épileptique et dépourvus d'artefacts d'activité épileptiforme, comme les pointes épileptiques ; (b) se concentrer sur les résultats de régions fonctionnellement intactes d'après les tests neuropsychologiques et les clichés de neuroimagerie ; (c) se focaliser sur les résultats pouvant être reproduits chez plusieurs patients, possiblement présentant des foyers épileptiques différents et suivant des traitements différents ; (d) favoriser les observations cohérentes avec les études de neuroimagerie antérieures faites chez des sujets sains. ».

Les travaux présentés dans ce manuscrit (étude 1 en particulier et étude 2) portant sur des résultats issus d'enregistrements en sEEG ont donc tous suivi ces 4 recommandations. Pour commencer, le protocole cognitif utilisé a été développé de façon à ce que chaque patient soit son propre contrôle, et ce à chaque essai, grâce à la soustraction de l'activité basale (entre deux essais) à celle en réponse aux stimuli d'intérêt (technique de z-score), au niveau de chaque bipole enregistré. Chaque patient a également effectué un très grand nombre d'essais dans chaque condition (entre 144 et 288 par patient en condition gain et en condition perte), dont la succession était soumise à un délai variable (fixation durant entre 800ms et 1200ms) afin d'éviter toute synchronisation entre le protocole et les activités non liées au protocole. Cela nous a permis de nous assurer que toute activité pathologique résiduelle non détectée lors des tests neuropsychologiques se retrouverait dans le bruit de fond. De plus, pour chaque région d'intérêt, nous rapportons des résultats similaires observés chez tous les patients implantés dans ces régions (4 dans le thalamus et jusqu'à 11 dans l'insula antérieure). Les patients inclus dans ces deux études présentaient des implantations différentes en sEEG (étude 1, voir Table 1 dans les matériels supplémentaires) et des traitements antiépileptiques différents (études 1 et 2, voir les données cliniques dans les matériels supplémentaires). Les résultats obtenus sont tous cohérents avec ceux rapportés par la littérature actuelle chez des sujets sains, que ce soit en termes de régions d'intérêt ou de réponses attendues. Nous avons donc très probablement affaire à des résultats décrivant des réponses physiologiques.

Un dernier aspect peut être soulevé concernant l'utilisation de la sEEG pour étudier la dynamique cérébrale de l'apprentissage par renforcement (étude 1) : le biais lié à la couverture du cerveau par les implantations réalisées. Nous avons donc dû nous assurer que nous nous trouvions face à une couverture la plus complète possible du cerveau humain, sous peine de devoir émettre des réserves quant aux régions non enregistrées et

pour lesquelles nous ne pourrions donc pas avancer de réponse sur leur rôle dans l'apprentissage par renforcement. L'étude 1 portant sur le cortex, nous avons utilisé l'atlas de parcellisation du cortex cérébral humain connu sous le nom de MarsAtlas (Auzias et al. 2016). Cet atlas comporte 41 parcelles couvrant l'intégralité du cortex. A travers nos 20 patients, nous avons enregistré 2083 bipoles couvrant les 41 parcelles de MarsAtlas. N'ayant pas d'hypothèse forte concernant une latéralisation des réponses, nous avons donc fixé un seuil de couverture minimale de 10 bipoles implantés chez 3 patients dans une même parcelle (Lachaux et al. 2012), après avoir collapsé les deux hémisphères ensemble. Cette opération a identifié 35 parcelles ayant été suffisamment implantées pour que l'on puisse en tirer des conclusions solides. Seules 6 parcelles ont donc été ignorées par cette étude (lobe pariétal : cortex pariétal supérieur SPC, cortex pariétal supérieur médian SPCm, cortex pariétal médian PCm, cortex somatosensoriel dorsomédian Sdm ; lobe frontal : cortex moteur dorsomédian Mdm et cortex prémoteur dorsomédian PMdm), dont la littérature n'a, à ce jour, pas démontré d'implication dans l'apprentissage par renforcement. En conclusion, les résultats présentés dans l'étude 1 reflètent donc les effets majeurs observables en sEEG au cours de l'apprentissage par renforcement.

2. Limites des études corrélationnelles

Comme nous l'avons déjà expliqué à plusieurs reprises, les études 1 et 2 reposent majoritairement sur des effets corrélationnels pour étudier la dynamique cérébrale de l'apprentissage par renforcement. Elles se basent en effet sur l'enregistrement de l'activité cérébrale pendant que des patients réalisent un protocole cognitif. Le but est donc de corréler les activités enregistrées avec les comportements observés au même moment. Cette approche corrélationnelle, bien que très intéressante, n'est pas une preuve irréfutable de l'implication de ces régions cérébrales dans le processus cognitif en question : l'apprentissage par renforcement.

Il est cependant possible d'affirmer qu'une région est impliquée ou non dans un processus cognitif en étudiant la causalité de l'activité d'une région sur le comportement. Il existe plusieurs types d'études permettant l'identification de ces effets causaux : les études lésionnelles (Palminteri et al. 2012; Bechara et al. 2000; Dalrymple-Alford et al. 2015; Smith et al. 2002; Corbit et al. 2003; Aggleton & Nelson 2015), l'utilisation de la stimulation magnétique transcrânienne (TMS) (Burke & Coats 2016), la stimulation cérébrale profonde (Krack et al. 2010; Fisher et al. 2010; David et al. 2010) et les études pharmacologiques (Pessiglione et al. 2006; Parnaudeau et al. 2015; Parnaudeau et al. 2013). Dans le cadre de

l'apprentissage par renforcement, nous avons vu que la grande majorité des régions cérébrales impliquées sont localisées profondément dans le cerveau et non en surface du crâne, hormis le dlPFC, ce qui rend la TMS inadaptée à son étude. Des études lésionnelles ont déjà été réalisées dans le cadre de l'apprentissage par renforcement et de la prise de décision chez l'animal comme chez l'humain (Bechara et al. 1998; Palminteri et al. 2012; Corbit et al. 2003), confirmant leur intérêt dans l'identification causale des régions impliquées puisque leur absence entraîne des déficits cognitifs très ciblés. Quant à la stimulation cérébrale, celle-ci peut être réalisée à court terme en peropératoire pour éliciter des réponses physiologiques (David et al. 2010) ou pathologiques selon la fréquence de stimulation, mais il est aussi possible de stimuler des régions cérébrales à long terme. Cette stimulation à long terme connue sous le nom de stimulation cérébrale profonde est particulièrement utilisée pour traiter des maladies neuropsychiatriques comme la maladie de Parkinson, les troubles obsessionnels compulsifs, l'obésité morbide ou encore des formes d'épilepsie sévère selon les cibles (Thompson et al. 2012; Krack et al. 2010; Chabardès et al. 2013; DeLong & Benabid 2014; Torres et al. 2012; Fisher et al. 2010). Il est également possible d'utiliser la stimulation cérébrale profonde pour étudier des processus cognitifs, comme ce que nous avons fait dans l'étude 2 sur le noyau antérieur du thalamus (Vignal et al. 2007; Sharp et al. 2010; Sun et al. 2015). En plus de l'étude de phénomènes causaux, l'utilisation de la stimulation cérébrale profonde peut permettre de s'intéresser au timing précis des encodages cérébraux permettant l'apprentissage par renforcement en stimulant précisément à certains moments du protocole cognitif (lors de la présentation des stimuli, pendant la phase de choix ou suite à la délivrance du renforcement par exemple) afin d'inhiber ou d'activer certaines régions cérébrales uniquement à une étape du processus décisionnel et ainsi identifier précisément son rôle de façon ponctuelle. Cette stimulation à demande selon le déroulement d'un protocole cognitif peut se faire pendant l'enregistrement de potentiels de champs locaux en sEEG, comme ce qui est déjà fait à Grenoble avec le projet EPI-STIM (stimulations de 1Hz et 50Hz) afin d'éliciter des réponses physiologiques ou épileptiformes et donc l'induction de crises. Comme nous avons mis en évidence des encodages des erreurs de prédiction des renforcements au moment de la délivrance du renforcement mais pas au moment de la présentation des stimuli (étude 1), et ce avec des latences très précises, il peut être intéressant de stimuler précisément des zones du cerveau à différents moments d'un essai du protocole, pour étudier les effets de l'excitation ou de l'inhibition de ces zones sur le comportement au cours de l'apprentissage par renforcement. Lors de l'utilisation de la stimulation cérébrale profonde à long terme comme celle du noyau antérieur du thalamus pour traiter l'épilepsie, il est important de prendre en considération que cette stimulation peut affecter plus que l'activité de la région stimulée. Il a par exemple été rapporté que la SCP du NAT à haute fréquence chez l'humain induit une activation de

régions corticales faisant partie du circuit de Papez comme l'hippocampe, mais active aussi d'autre régions corticales liées indirectement au NAT comme le cortex orbitofrontal, rendant les conséquences fonctionnelles de cette stimulation plus complexes qu'il n'y parait (Rektor et al. 2016). Enfin, des études pharmacologiques ont été réalisées chez l'humain ou l'animal pour mettre en évidence les effets causaux d'une hypofonction ou d'une hyperfonction de régions et systèmes cérébraux impliqués dans l'apprentissage par renforcement et les comportements dirigés vers un but. C'est ainsi que le rôle spécifique de la dopamine dans l'apprentissage par récompense a été mis en évidence (Pessiglione et al. 2006). Une autre étude pharmacologique visant à étudier l'effet de l'hypoactivation du noyau dorsomédian du thalamus a mis en évidence le rôle de la synchronisation de l'activité bêta du NDMT avec le cortex préfrontal lors de processus cognitifs impliquant le cortex préfrontal comme l'apprentissage à renversement (reversal learning) ou la sélection des actions basée sur les renforcements (Parnaudeau et al. 2013; Parnaudeau et al. 2015). Les études corrélationnelles et causales sont donc complémentaires pour acquérir une compréhension la plus complète possible des processus cognitifs comme l'apprentissage par renforcement et la prise de décision en général.

3. Apports cliniques et sociétaux

L'amélioration de la compréhension des mécanismes cérébraux permettant aux êtres humains de prendre des décisions adaptées afin d'atteindre leurs objectifs (récompenses) sans être pénalisés (punitions) va certainement influencer largement la société et l'économie. A ce sujet, un biais comportemental majeur réside en notre aversion aux punitions qui peut nous pousser à prendre des décisions sous-optimales par crainte d'être punis, ne serait-ce que très rarement. L'être humain a ainsi une appétence très forte pour les récompenses mais l'évolution lui a fait craindre les punitions de telle manière que la recherche de récompenses et l'évitement des punitions peuvent biaiser nos décisions au quotidien.

Dans le cadre de la santé publique, de plus en plus de pathologies neuropsychiatriques sont reconnues pour impliquer le système dopaminergique et donc le circuit de la récompense, comme c'est le cas de la maladie de Parkinson et des troubles obsessionnels compulsifs. Cela est flagrant chez les patients parkinsoniens traités en dopa-thérapie qui peuvent présenter des biais décisionnels majeurs induisant une hypersensibilité aux récompenses, pouvant se traduire par des addictions comportementales (jeux et achats compulsifs par exemple) (Voon et al. 2010; Clark 2010; Pessiglione et al. 2006). De même, des patients

atteints de la maladie de Huntington vont présenter des déficits progressifs atteignant l'apprentissage par évitement des punitions puis l'apprentissage par récompenses en raison d'une dégénérescence progressive du striatum dorsal puis ventral (Palminteri et al. 2012). Ces quelques exemples sont la preuve que l'étude de la dynamique cérébrale des mécanismes sous-tendant l'apprentissage par renforcement est importante pour permettre l'identification précise de déficits cognitifs subtils que peuvent présenter un nombre certain de patients atteints de maladies neurodégénératives ou de lésions cérébrales plus focales (tumeurs, gliomes et AVC par exemple). Cela constitue une étape importante dans la reconnaissance globale de ces pathologies et une amélioration de leur prise en charge par le personnel médical, les proches des malades et la société en général.

L'étude de la prise de décision économique et de l'apprentissage par renforcement est affiliée à la Neuroéconomie, un domaine actuellement en vogue. Cet intérêt au sein du monde économique et scientifique n'est cependant pas si récent puisque les travaux de Kahneman sur la prise de décision et le risque ont fait l'objet du prix Nobel d'économie en 2002, et que le prix Nobel d'économie 2017 vient d'être décerné à Richard Thaler pour ses travaux sur la finance comportementale. Il a entre-autres montré « combien l'aversion aux pertes peut expliquer pourquoi les individus accordent une plus grande valeur à une chose s'ils la possèdent que s'ils ne la possèdent pas », un phénomène appelé « l'aversion à la dépossession ». Une preuve de plus, si cela était nécessaire, du bienfondé de l'étude de l'apprentissage par renforcement pour mieux comprendre la société et les décisions prises, que ce soit à l'échelle de l'individu ou de manière globale.

BIBLIOGRAPHIE GENERALE

Abitbol, R., Lebreton M., Hollard G., Richmond B.J., Bouret S., Pessiglione M. 2015. Neural mechanisms underlying contextual dependency of subjective values: converging evidence from monkeys and humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(5), pp.2308–20.

Adams, R. & David, A.S., 2007. Patterns of anterior cingulate activation in schizophrenia: a selective review. *Neuropsychiatric disease and treatment*, 3(1), pp.87–101.

Afif, A., Minotti, L., Kahane, P., and Hoffmann, D. 2010. Anatomofunctional organization of the insular cortex: a study using intracerebral electrical stimulation in epileptic patients. Epilepsia 51, 2305–2315.

Afif, A., Becq, G. & Mertens, P., 2013. Definition of a Stereotactic 3-Dimensional Magnetic Resonance Imaging Template of the Human Insula. *Operative Neurosurgery*, 72(1 Suppl Operative), p.ons 35-ons 46.

Afif, A. & Mertens, P., 2010. Description of sulcal organization of the insular cortex. *Surgical and radiologic anatomy : SRA*, 32(5), pp.491–8.

Aggleton, J.P. & Nelson, A.J.D., 2015. Neuroscience and Biobehavioral Reviews Why do lesions in the rodent anterior thalamic nuclei cause such severe spatial deficits ? *Neuroscience and Biobehavioral Reviews*, 54, pp.131–144.

Alcaraz, F., Marchand, A.R., Vidal, E., Guillou, A., Faugère, A., Coutureau, E. Wolff, M. 2015. Flexible Use of Predictive Cues beyond the Orbitofrontal Cortex: Role of the Submedius Thalamic Nucleus. *Journal of Neuroscience*, 35(38), pp.13183–13193.

Alexander, G.E., Crutcher, M.D. & DeLong, M.R., 1990. Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, prefrontal and limbic functions. *Progress in brain research*, 85, pp.119–46.

Alpaydin, E., 2004. *Introduction to machine learning.*, MIT Press.

Anastassiou, C.A., Perin R, Markram H, Koch C. 2011. Ephaptic coupling of cortical neurons. *Nature Neuroscience*, 14(2), pp.217–223.

Anon, 2007. A quoi sert la dopamine ? Available at: https://www.lanutrition.fr/outils/a-quoi-sert-la-dopamine-.

Anon, 2013. Could time travel be a movement disorder? Available at: https://neuroendoimmune.wordpress.com/2013/06/04/great-scott-could-time-travel-be-a-movement-disorder-2/

Apicella, P., Ljungberg T, Scarnati E, Schultz W. 1991. Responses to reward in monkey dorsal and ventral striatum. *Experimental brain research*, 85(3), pp.491–500.

de Araujo, I.E., Kringelbach ML, Rolls ET, McGlone F. 2003. Human cortical responses to water in the mouth, and the effects of thirst. *Journal of neurophysiology*, 90(3), pp.1865–76.

Arbuthnott, G. & Garcia-Munoz, M., 2009. Dealing with the devil in the detail - some thoughts about the next model of the basal ganglia. *Parkinsonism & related disorders*, 15 Suppl 3, pp.S139-42.

Arias-Carrión, O., Stamelou M., Murillo-Rodríguez E., Menéndez-González M., Pöppel E. 2010. Dopaminergic reward system: a short integrative review. *International Archives of Medicine*, 3(1), p.24.

Arnsten, A.F., Paspalas CD, Gamo NJ, Yang Y, Wang M. 2010. Dynamic Network Connectivity: A new form of neuroplasticity. *Trends in cognitive sciences*, 14(8), pp.365–75.

Aron, A.R., Monsell S., Sahakian B.J., Robbins TW. 2004. A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain*, 127(7), pp.1561–1573.

Aron, A.R. & Poldrack, R.A., 2006. Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(9), pp.2424–33.

Asaad, W.F. & Eskandar, E.N., 2011. Encoding of Both Positive and Negative Reward Prediction Errors by Neurons of the Primate Lateral Prefrontal Cortex and Caudate Nucleus. *Journal of Neuroscience*, 31(49), pp.17772–17787.

Ashby, C.R. Jr, Rice O.V., Heidbreder C.A., Gardner E.L. 2015. The selective dopamine D 3 receptor antagonist SB-277011A significantly accelerates extinction to environmental cues associated with cocaine-induced place preference in male sprague-dawley rats. *Synapse*, 69(10), pp.512–514.

Astolfi, L., Cincotti F, Mattia D, De Vico Fallani F, Salinari S, Vecchiato G, Toppi J, Wilke C, Doud A, Yuan H, He B, Babiloni F. 2010. Imaging the social brain: multi-subjects EEG recordings during the “Chicken’s game.” In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, pp. 1734–1737.

Augustine, J.R., 1996. Circuitry and functional aspects of the insular lobe in primates including humans. *Brain research. Brain research reviews*, 22(3), pp.229–44.

Augustine, J.R., 1985. The insular lobe in primates including humans. *Neurological research*, 7(1), pp.2–10.

Auzias, G., Coulon, O. & Brovelli, A., 2016. MarsAtlas : A Cortical Parcellation Atlas for Functional Mapping. *Human brain mapping*, 37(4), pp.1573–92.

Badre, D., Kayser, A.S. & D’Esposito, M., 2010. Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, 66(2), pp.315–326.

Badre, D. & Wagner, A.D., 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia*, 45(13), pp.2883–901.

Bailey, P. & Von Bonin, G., 1957. Evolution of the cerebral cortex: organ of the mind. *What’s new*, (198), pp.13–9.

Balleine, B.W. & Dickinson, A., 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4–5), pp.407–19.

Balleine, B.W., Morris, R.W. & Leung, B.K., 2015. Thalamocortical integration of instrumental learning and performance and their disintegration in addiction. *Brain Research*, 1628, pp.104–116.

Balleine, B.W. & O’Doherty, J.P., 2010. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, 35(1), pp.48–69.

Barbas, H., Saha S., Rempel-Clower N., Ghashghaei T. 2003. Serial pathways from primate prefrontal cortex to autonomic areas may influence emotional expression. *BMC neuroscience*, 4(1), p.25.

Barbas, H. & Zikopoulos, B., 2007. The Prefrontal Cortex and Flexible Behavior. *The Neuroscientist*, 13(5), pp.532–545.

Bartra, O., McGuire, J.T. & Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, pp.412–427.

Bastin, J., Polosan, M., Piallat, B., Krack P., Bougerol T., Chabardès S., David O. 2014. Changes of oscillatory activity in the subthalamic nucleus during obsessive-compulsive disorder symptoms: Two case reports. *Cortex; a journal devoted to the study of the nervous system and behavior*, 60, pp.145–150.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., Hoffman, D., Combrisson, E., Kujala, J., Perrone-Bertolotti, M., Kahanae, P., Lachaux, JP., Jerbi, K. 2016. Direct Recordings from Human Anterior Insula Reveal its Leading Role within the Error-Monitoring Network. Cereb. Cortex.

Bastin, J., Polosan, M., Benis, D., Goetz L., Bhattacharjee M., Piallat B., Krainik A., Bougerol T., Chabardès S., David O. 2014. Inhibitory control and error monitoring by human subthalamic neurons. Translational psychiatry, 4(August 2013), p.e439.

Baucells, M. & Villasís, A., 2010. Stability of risk preferences and the reflection effect of prospect theory. *Theory and Decision*, 68(1–2), pp.193–211.

Bayer, H.M. & Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1), pp.129–41.

Baylis, L.L. & Gaffan, D., 1991. Amygdalectomy and ventromedial prefrontal ablation produce similar deficits in food choice and in simple object discrimination learning for an unseen reward. *Experimental brain research*, 86(3), pp.617–22.

Bechara, A., Damasio, H., Tranel, D., and Anderson, S.W. 1998. Dissociation of working memory from decision making within the human prefrontal cortex. J. Neurosci. Off. J. Soc. Neurosci. *18*, 428–437.

Bechara, A., Tranel, D. & Damasio, H., 2000. Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain : a journal of neurology*, 123 (Pt 1, pp.2189–2202.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. 2008. Associative learning of social value. Nature *456*, 245–249.

Behrens, T.E.J., Woolrich MW, Walton ME, Rushworth MF. 2007. Learning the value of

information in an uncertain world. *Nature neuroscience*, 10(9), pp.1214–21.

Bellmann, R., 1957. Markovian decision process.

Benis, D., David, O., Lachaux, JP., Seigneret, E., Krack, P., Fraix, V., Chabardès, S., Bastin, J., 2014. Subthalamic nucleus activity dissociates proactive and reactive inhibition in patients with Parkinson's disease. *NeuroImage*, 91, pp.273–81.

Bergson, C., Mrzljak L, Smiley JF, Pappy M, Levenson R, Goldman-Rakic PS. 1995. Regional, cellular, and subcellular variations in the distribution of D1 and D5 dopamine receptors in primate brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 15(12), pp.7821–36.

Berns, G.S., McClure SM, Pagnoni G, Montague PR. 2001. Predictability modulates human brain response to reward. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(8), pp.2793–8.

Berntson, G.G., Norman GJ, Bechara A, Bruss J, Tranel D, Cacioppo JT. 2011. The insula and evaluative processes. *Psychological science*, 22(1), pp.80–6.

Billeke, P., Ossandon, T., Perrone-Bertolotti, M., Kahane, P., Bastin, J., Lachaux, J.P., Fuentealba, P. *Beta oscillations in the human anterior insula encode performance feedback and relay unsigned prediction error to the medial refrontal cortex (submitted)*,

Bissonette, G.B. & Roesch, M.R., 2015. Neurophysiology of Reward-Guided Behavior: Correlates Related to Predictions, Value, Motivation, Errors, Attention, and Action. In *Current topics in behavioral neurosciences*. pp. 199–230.

Blackford, J.U., Buckholtz JW, Avery SN, Zald DH. 2010. A unique role for the human amygdala in novelty detection. *NeuroImage*, 50(3), pp.1188–93.

Blank, H., Biele G, Heekeren HR, Philiastides MG. 2013. Temporal characteristics of the influence of punishment on perceptual decision making in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(9), pp.3939–52.

Bočková, M., Jurák P, Chládek J, Chrastina J, Halámek J, Bočková M, Goldemundová S, Říha I, Rektor I. 2015. Complex Motor–Cognitive Factors Processed in the Anterior Nucleus of the Thalamus: An Intracerebral Recording Study. *Brain Topography*, 28(2), pp.269–278.

Bolstad, I., Andreassen OA, Reckless GE, Sigvartsen NP, Server A, Jensen J. 2013.

Aversive Event Anticipation Affects Connectivity between the Ventral Striatum and the Orbitofrontal Cortex in an fMRI Avoidance Task C. Soriano-Mas, ed. *PLoS ONE*, 8(6), p.e68494.

Bossaerts, P., 2010. Risk and risk prediction error signals in anterior insula. *Brain structure & function*, 214(5–6), pp.645–53.

de Bourbon-Teles, J., Bentley P., Koshino S., Shah K., Dutta A., Malhotra P., Egner T., Husain M., Soto D.. 2014. Thalamic Control of Human Attention Driven by Memory and Learning. *Current Biology*, 24(9), pp.993–999.

Bouret, S. & Richmond, B.J., 2010. Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(25), pp.8591–601.

Bradfield, L.A., Hart, G. & Balleine, B.W., 2013. The role of the anterior, mediodorsal, and parafascicular thalamus in instrumental conditioning. *Frontiers in Systems Neuroscience*, 7(October), p.51.

Brass, M. & Haggard, P., 2010. The hidden side of intentional action: the role of the anterior insular cortex. *Brain Structure and Function*, 214(5–6), pp.1–8.

Brázdil, M., Mikl M, Marecek R, Krupa P, Rektor I. 2007. Effective connectivity in target stimulus processing: A dynamic causal modeling study of visual oddball task. *NeuroImage*, 35(2), pp.827–835.

Breiter, H.C., Aharon I, Kahneman D, Dale A, Shizgal P. 2001. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30(2), pp.619–39.

Bridle, J.S., 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky, ed. *Advances in neural information processing systems*. Morgan Kaufmann.

Brischoux, F., Chakraborty, S., Brierley, D.I., and Ungless, M.A. 2009. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc. Natl. Acad. Sci. *106*, 4894–4899.

Bromberg-Martin, E.S. & Hikosaka, O., 2012. Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neurosciences*, 14(9), pp.1209–1216.

Bromberg-Martin, E.S., Matsumoto, M. & Hikosaka, O., 2010. Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, 68(5), pp.815–834.

Brown, R.G. & Pluck, G., 2000. Negative symptoms: the “pathology” of motivation and goal-directed behaviour. *Trends in neurosciences*, 23(9), pp.412–7.

Büchel, C., Morris, J., Dolan, R.J., and Friston, K.J. 1998. Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron *20*, 947–957.

Buckley, M.J., Mansouri FA, Hoda H, Mahboubi M, Browning PG, Kwok SC, Phillips A, Tanaka K. 2009. Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science (New York, N.Y.)*, 325(5936), pp.52–8.

Burke, M.R. & Coats, R.O., 2016. Dissociation of the rostral and dorsolateral prefrontal cortex during sequence learning in saccades: a TMS investigation. *Experimental Brain Research*, 234(2), pp.597–604.

Bussey, T.J, Muir JL, Everitt BJ, Robbins TW. 1996. Dissociable effects of anterior and posterior cingulate cortex lesions on the acquisition of a conditional visual discrimination: facilitation of early learning vs. impairment of late learning. *Behavioural brain research*, 82(1), pp.45–56.

Butter, C.M., McDonald, J.A. & Snyder, D.R., 1969. Orality, preference behavior, and reinforcement value of nonfood object in monkeys with orbital frontal lesions. *Science (New York, N.Y.)*, 164(3885), pp.1306–7.

Cachope, R. & Cheer, J.F., 2014. Local control of striatal dopamine release. *Frontiers in Behavioral Neuroscience*, 8, p.188.

Cai, X. & Padoa-Schioppa, C., 2014. Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. *Neuron*, 81(5), pp.1140–1151.

Calder, A.J., Keane J, Manes F, Antoun N, Young AW. 2000. Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, 3(11), pp.1077–1078.

Camille, N., Tsuchida, A. & Fellows, L.K., 2011. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(42), pp.15048–52.

Campbell-Meiklejohn, D.K., Kanai R, Bahrami B, Bach DR, Dolan RJ, Roepstorff A, Frith CD. 2012. Structure of orbitofrontal cortex predicts social influence. *Current Biology*, 22(4), pp.R123–R124.

Canessa, N., Crespi C, Motterlini M, Baud-Bovy G, Chierchia G, Pantaleo G, Tettamanti M, Cappa SF. 2013. The Functional and Structural Neural Basis of Individual Differences in Loss Aversion. *Journal of Neuroscience*, 33(36), pp.14307–14317.

Caria, A., Sitaram, R., Veit, R., Begliomini, C., and Birbaumer, N. 2010. Volitional Control of Anterior Insula Activity Modulates the Response to Aversive Stimuli. A Real-Time Functional Magnetic Resonance Imaging Study. Biol. Psychiatry *68*, 425–432.

Carlsson, A., 1959. Detection and assay of dopamine. *Pharmacological reviews*, 11(2, Part 2), pp.300–4.

Carmichael, D.W., Thornton, J.S., Rodionov, R., Thornton, R., Mcevoy, A., Allen, P.J., and Lemieux, L. 2008. Safety of Localizing Epilepsy Monitoring Intracranial Electroencephalograph Electrodes Using MRI : Radiofrequency-Induced Heating. *1244*, 1233–1244.

Carnicella, S., Drui G, Boulet S, Carcenac C, Favier M, Duran T, Savasta M. 2014. Implication of dopamine D3 receptor activation in the reversion of Parkinson's disease-related motivational deficits. *Translational psychiatry*, 4(6), p.e401.

Caruana, F., Jezzini A, Sbriscia-Fioretti B, Rizzolatti G, Gallese V. 2011. Emotional and social behaviors elicited by electrical stimulation of the insula in the macaque monkey. *Current biology : CB*, 21(3), pp.195–9.

Cauda, F., D'Agata F, Sacco K, Duca S, Geminiani G, Vercelli A. 2011. Functional connectivity of the insula in the resting brain. *NeuroImage*, 55(1), pp.8–23.

Cavada, C., Compañy T, Tejedor J, Cruz-Rizzolo RJ, Reinoso-Suárez F. 2000. The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cerebral cortex (New York, N.Y. : 1991)*, 10(3), pp.220–42.

Cavanagh, J.F. & Frank, M.J., 2014. Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, 18(8), pp.414–421.

Cerliani, L., Thomas RM, Jbabdi S, Siero JC, Nanetti L, Crippa A, Gazzola V, D'Arceuil H, Keysers C. 2012. Probabilistic tractography recovers a rostrocaudal trajectory of

connectivity variability in the human insular cortex. *Human brain mapping*, 33(9), pp.2005–34.

Chabardès, S., Polosan M, Krack P, Bastin J, Krainik A, David O, Bougerol T, Benabid AL. 2013. Deep Brain Stimulation for Obsessive-Compulsive Disorder: Subthalamic Nucleus Target. *World Neurosurgery*, 80(3–4), p.S31.e1-S31.e8.

Chakraborty, S., Kolling N, Walton ME, Mitchell AS. 2016. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *eLife*, 5(MAY2016), pp.1–23.

Chase, H.W., Swainson R, Durham L, Benham L, Cools R. 2011. Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of cognitive neuroscience*, 23(4), pp.936–46.

Chase, H.W., Kumar, P., Eickhoff, S.B., Doombrovski, A.Y. 2015. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2), pp.435–459.

Chikama, M., McFarland NR, Amaral DG, Haber SN. 1997. Insular cortical projections to functional regions of the striatum correlate with cortical cytoarchitectonic organization in the primate. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(24), pp.9686–705.

Chikazoe, J., Jimura, K., Asari, T., Yamashita, K., Morimoto, H., Hirose, S., Miyashita, Y., Konishi, S. 2009. Functional dissociation in right inferior frontal cortex during performance of go/no-go task. *Cerebral Cortex*, 19(1), pp.146–152.

Choi, E.Y., Yeo, B.T.T. & Buckner, R.L., 2012. The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 108(8), pp.2242–63.

Clark, L., Studer B, Bruss J, Tranel D, Bechara A. 2014. Damage to insula abolishes cognitive distortions during simulated gambling. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16), pp.6098–103.

Clark, L., 2010. Decision-making during gambling: an integration of cognitive and psychobiological approaches. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1538), pp.319–30.

Clark, L., Bechara A, Damasio H, Aitken MR, Sahakian BJ, Robbins TW. 2008. Differential effects of insular and ventromedial prefrontal cortex lesions on risky decision-making. *Brain*, 131(5), pp.1311–1322.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88.

Cohen, J.Y., Amoroso, M.W. & Uchida, N., 2015. Serotonergic neurons signal reward and punishment on multiple timescales. *eLife*, 4, pp.1–25.

Conejo, N.M., González-Pardo, H., López, M., Cantora, R., Arias, J.L. 2007. Induction of c-Fos expression in the mammillary bodies, anterior thalamus and dorsal hippocampus after fear conditioning. *Brain Research Bulletin*, 74(1–3), pp.172–177.

Contreras, M., Ceric, F. & Torrealba, F., 2007. Inactivation of the interoceptive insula disrupts drug craving and malaise induced by lithium. *Science (New York, N.Y.)*, 318(5850), pp.655–8.

Corbit, L.H., Muir, J.L. & Balleine, B.W., 2003. Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *European Journal of Neuroscience*, 18(5), pp.1286–1294.

Corrado, G.S., Sugrue LP, Seung HS, Newsome WT. 2005. Linear-Nonlinear-Poisson models of primate choice dynamics. *Journal of the experimental analysis of behavior*, 84(3), pp.581–617.

Corrado, G.S., Sugrue, L.P., Brown, J.W., Newsome, W.T. 2009. The trouble with choice: studying decision variables in the brain. In Glimcher P.W., Camerer, C.F., Fehr, E., Poldrack, R.A, eds. *Neuroeconomics: decision making in the brain*. Elsevier, pp. 463–480.

Corrado, G.S. & Doya, K., 2007. Understanding neural coding through the model-based analysis of decision making. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(31), pp.8178–80.

Craig, A.D.B., 2009a. Emotional moments across time: a possible neural basis for time perception in the anterior insula. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1525), pp.1933–42.

Craig, A.D.B., 2009b. How do you feel--now? The anterior insula and human awareness.

Nature reviews. Neuroscience, 10(1), pp.59–70.

Craig, A.D.B., 2003. Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, 13(4), pp.500–505.

Craig, A.D.B., 2010. Once an island, now the focus of attention. *Brain structure & function*, 214(5–6), pp.395–6.

Craig, A.D.B., Chen K, Bandy D, Reiman EM., 2000. Thermosensory activation of insular cortex. *Nature neuroscience*, 3(2), pp.184–90.

Critchley, H.D, Melmed RN, Featherstone E, Mathias CJ, Dolan RJ. 2001. Brain activity during biofeedback relaxation: a functional neuroimaging investigation. *Brain : a journal of neurology*, 124(Pt 5), pp.1003–12.

Crockett, M.J., Clark, L. & Robbins, T.W., 2009. Reconciling the Role of Serotonin in Behavioral Inhibition and Aversion: Acute Tryptophan Depletion Abolishes Punishment-Induced Inhibition in Humans. *Journal of Neuroscience*, 29(38), pp.11993–11999.

D'Ardenne, K., McClure SM, Nystrom LE, Cohen JD. 2008. BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. *Science*, 319(5867), pp.1264–1267.

Dalrymple-Alford, J.C., Harland B, Loukavenko EA, Perry B, Mercer S, Collings DA, Ulrich K, Abraham WC, McNaughton N, Wolff M. 2015. Anterior thalamic nuclei lesions and recovery of function: Relevance to cognitive thalamus. *Neuroscience and Biobehavioral Reviews*, 54, pp.145–160.

Damasio, A.R., 1994. *Descartes' Error: Emotion, Reason and the Human Brain*,

Damasio, H., Grabowski T, Frank R, Galaburda AM, Damasio AR. 1994. The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science (New York, N.Y.)*, 264(5162), pp.1102–5.

David, O., Bastin J, Chabardès S, Minotti L, Kahane P. 2010. Studying network mechanisms using intracranial stimulation in epileptic patients. *Frontiers in systems neuroscience*, 4(October), p.148.

Daw, N.D. & Abbott, L.F., 2005. *Theoretical neuroscience: computational and mathematical modeling of neural systems.*, MIT Press.

Daw, N.D. & Doya, K., 2006. The computational neurobiology of learning and reward. *Current opinion in neurobiology*, 16(2), pp.199–204.

Dayan, P. & Abbott, L.F., 2002. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. *Philosophical Psychology*, 15(4), pp.563–577.

Dayan, P. & Niv, Y., 2008. Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, 18(2), pp.185–196.

Deen, B., Pitskel, N.B. & Pelphrey, K.A., 2011. Three Systems of Insular Functional Connectivity Identified with Cluster Analysis. *Cerebral Cortex*, 21(7), pp.1498–1506.

Dehaene, S. & Changeux, J.-P., 2000. Reward-dependent learning in neuronal networks for planning and decision making. *Progress in brain research*, 126, pp.217–229.

DeLong, M.R. & Benabid, A.L., 2014. Discovery of High-Frequency Deep Brain Stimulation for Treatment of Parkinson Disease. *JAMA*, 312(11), p.1093.

DeYoung, C.G., Hirsh JB, Shane MS, Papademetris X, Rajeevan N, Gray JR. 2010. Testing Predictions From Personality Neuroscience. *Psychological Science*, 21(6), pp.820–828.

Dickinson, A., 1985. Actions and habits: the development of behavioural autonomy. *Philosophical transactions of the Royal Society of London*, 308, pp.67–78.

Dickinson, A., 1980. *Contemporary Animal Learning Theory*, Cambridge University Press.

Doll, B.B., Simon, D.A. & Daw, N.D., 2012. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6), pp.1075–81.

Donner, T.H. & Siegel, M., 2011. A framework for local cortical oscillation patterns. *Trends in cognitive sciences*, 15(5), pp.191–9.

Draganski, B., Kherif F, Klöppel S, Cook PA, Alexander DC, Parker GJ, Deichmann R, Ashburner J, Frackowiak RS. 2008. Evidence for Segregated and Integrative Connectivity Patterns in the Human Basal Ganglia. *Journal of Neuroscience*, 28(28), pp.7143–7152.

Drevets, W.C., Savitz, J. & Trimble, M., 2008. The subgenual anterior cingulate cortex in mood disorders. *CNS spectrums*, 13(8), pp.663–81.

Drui, G., Carnicella S, Carcenac C, Favier M, Bertrand A, Boulet S, Savasta M. 2014. Loss of dopaminergic nigrostriatal neurons accounts for the motivational and affective deficits in

Parkinson's disease. *Molecular Psychiatry*, 19(3), pp.358–367.

Dumont, J.R., Petrides, M. & Sziklas, V., 2010. Fornix and retrosplenial contribution to a hippocampo-thalamic circuit underlying conditional learning. *Behavioural brain research*, 209(1), pp.13–20.

Duncan, J. & Owen, A.M., 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in neurosciences*, 23(10), pp.475–83.

Elliott, R., Dolan, R.J. & Frith, C.D., 2000. Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cerebral Cortex*, 10, pp.308–317.

Ernst, M., Bolla K, Mouratidis M, Contoreggi C, Matochik JA, Kurian V, Cadet JL, Kimes AS, London ED. 2002. Decision-making in a Risk-taking Task A PET Study. *Neuropsychopharmacology*, 26(5), pp.682–691.

Esber, G.R. & Holland, P.C., 2014. The basolateral amygdala is necessary for negative prediction errors to enhance cue salience, but not to produce conditioned inhibition. *European Journal of Neuroscience*, 40(9), pp.3328–3337.

Falck, B. & Hillarp, N.A., 1959. On the cellular localization of catechol amines in the brain. *Acta anatomica*, 38, pp.277–9.

Falck, B., Hillarp, N.A. & Torp, A., 1959. A new type of chromaffin cells, probably storing dopamine. *Nature*, 183(4656), pp.267–8.

Favier, M., Carcenac, C., Drui, G., Vachez, Y., Boulet, S., Savasta, M., Carnicella, S. 2017. Implication of dorsostriatal D3 receptors in motivational processes: a potential target for neuropsychiatric symptoms in Parkinson's disease. *Scientific reports*, 7, p.41589.

Fellows, L.K., 2011. *The Neurology of Value*,

Fellows, L.K. & Farah, M.J., 2003. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain : a journal of neurology*, 126(Pt 8), pp.1830–7.

Ferré, S., Lluís C, Justinova Z, Quiroz C, Orru M, Navarro G, Canela EI, Franco R, Goldberg SR. 2010. Adenosine-cannabinoid receptor interactions. Implications for striatal function. *British Journal of Pharmacology*, 160(3), pp.443–453.

Fiorillo, C.D., 2013. Two Dimensions of Value: Dopamine Neurons Represent Reward But Not Aversiveness. *Science*, 341(6145), pp.546–549.

Fiorillo, C.D., Tobler, P.N. & Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, 299(5614), pp.1898–902.

Fischer, A.G. & Ullsperger, M., 2013. Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, 79(6), pp.1243–1255.

Fischer, A.G. & Ullsperger, M., 2013. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron*, 79(6), pp.1243–55.

Fisher, R., Salanova V, Witt T, Worth R, Henry T, Gross R, Oommen K, Osorio I, Nazzaro J, Labar D, Kaplitt M, Sperling M, Sandok E, Neal J, Handforth A, Stern J, DeSalles A, Chung S, Shetter A, Bergen D, Bakay R, Henderson J, French J, Baltuch G, Rosenfeld W, Youkilis A, Marks W, Garcia P, Barbaro N, Fountain N, Bazil C, Goodman R, McKhann G, Babu Krishnamurthy K, Papavassiliou S, Epstein C, Pollard J, Tonder L, Grebin J, Coffey R, Graves N; SANTE Study Group. 2010. Electrical stimulation of the anterior nucleus of thalamus for treatment of refractory epilepsy: Deep Brain Stimulation of Anterior Thalamus for Epilepsy. *Epilepsia*, 51(5), pp.899–908.

Fleming, S.M., Ryu J, Golfinos JG, Blackmon KE. 2014. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, 137(10), pp.2811–2822.

Freeman, J.H. Jr, Cuppernell C, Flannery K, Gabriel M. 1996. Limbic thalamic, cingulate cortical and hippocampal neuronal correlates of discriminative approach learning in rabbits. *Behavioural Brain Research*, 80(1–2), pp.123–136.

Fröhlich, F. & McCormick, D.A., 2010. Endogenous electric fields may guide neocortical network activity. *Neuron*, 67(1), pp.129–43.

Fu, J.-J., Tang JS, Yuan B, Jia H. 2002. Response of neurons in the thalamic nucleus submedius (Sm) to noxious stimulation and electrophysiological identification of on- and off-cells in rats. *Pain*, 99(1–2), pp.243–51.

Fudge, J.L., Breitbart MA, Danish M, Pannoni V 2005. Insular and gustatory inputs to the caudal ventral striatum in primates. *The Journal of comparative neurology*, 490(2), pp.101–18.

Funahashi, S., Bruce, C.J. & Goldman-Rakic, P.S., 1993. Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 13(4), pp.1479–97.

Fuster, J.M., 1997. Network memory. *Trends in neurosciences*, 20(10), pp.451–9.

Gabriel, M., Miller, J.D. & Saltwick, S.E., 1977. Unit activity in cingulate cortex and anteroventral thalamus of the rabbit during differential conditioning and reversal. *Journal of Comparative and Physiological Psychology*, 91(2), pp.423–433.

Gabriel, M., Sparenborg, S. & Kubota, Y., 1989. Anterior and medial thalamic lesions, discriminative avoidance learning, and cingulate cortical neuronal activity in rabbits. *Experimental Brain Research*, 76(2), pp.441–457.

Gaglianese, A., Vansteensel MJ, Harvey BM, Dumoulin SO, Petridou N, Ramsey NF. 2016. Correspondence between fMRI and electrophysiology during visual motion processing in human MT+. *NeuroImage*, 155(April), pp.480–489.

Gallay, D.S., Gallay MN, Jeanmonod D, Rouiller EM, Morel A. 2012. The insula of Reil revisited: multiarchitectonic organization in macaque monkeys. *Cerebral cortex (New York, N.Y. : 1991)*, 22(1), pp.175–90.

Garrison, J., Erdeniz, B. & Done, J., 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience and biobehavioral reviews*, 37(7), pp.1297–310.

Gaspar, P., Bloch, B. & Le Moine, C., 1995. D1 and D2 receptor gene expression in the rat frontal cortex: cellular localization in different classes of efferent neurons. *The European journal of neuroscience*, 7(5), pp.1050–63.

Genovese, C.R., Lazar, N.A. & Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4), pp.870–8.

Ghaziri, J., Tucholka A, Girard G, Houde JC, Boucher O, Gilbert G, Descoteaux M, Lippé S, Rainville P, Nguyen DK 2017. The Corticocortical Structural Connectivity of the Human Insula. *Cerebral Cortex*, 27(2), pp.1216–1228.

Gibbon, J., Baldock, M D, Locurto, C, Gold, L, Terrace, H S 1977. Trial and intertrial durations in autoshaping. *J Exp Psychol Anim Behav Proces*, pp.264–284.

Ginther, M.R., Bonnie RJ, Hoffman MB, Shen FX, Simons KW, Jones OD, Marois R. 2016.

Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 36(36), pp.9420–9434.

Gläscher, J., Hampton, A.N. & O'Doherty, J.P., 2009. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral cortex (New York, N.Y. : 1991)*, 19(2), pp.483–95.

Glickman, S.E. & Schiff, B.B., 1967. A biological theory of reinforcement. *Psychological review*, 74(2), pp.81–109.

Glimcher, P.W. & Fehr, E., 2014. *Neuroeconomics : decision making and the brain*, Elsevier Academic Press.

Goldberg, J.A. & Reynolds, J.N.J., 2011. Spontaneous firing and evoked pauses in the tonically active cholinergic interneurons of the striatum. *Neuroscience*, 198, pp.27–43.

Goldman-Rakic, P.S., 1988. Topography of cognition: parallel distributed networks in primate association cortex. *Annual review of neuroscience*, 11(1), pp.137–56.

Gonzalez, A., Hutchinson JB, Uncapher MR, Chen J, LaRocque KF, Foster BL, Rangarajan V, Parvizi J, Wagner AD. 2015. Electrocorticography reveals the temporal dynamics of posterior parietal cortical activity during recognition memory decisions. *Proceedings of the National Academy of Sciences*, 112(35), pp.11066–11071.

Gottfried, J.A., O'Doherty, J.P. & Dolan, R.J., 2003. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science (New York, N.Y.)*, 301(5636), pp.1104–7.

Graybiel, A.M., 2000. The basal ganglia. *Current biology : CB*, 10(14), pp.R509-11.

Greene, J.D., Sommerville RB, Nystrom LE, Darley JM, Cohen JD. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293(5537), pp.2105–2108.

Greengard, P., 2001. The neurobiology of dopamine signaling. *Bioscience reports*, 21(3), pp.247–69.

Gueguen, M.C.M., Lachaux, J.P., Kahane, P., Billeke, P., Pessiglione M., Bastin, J. Rewards and punishment learning differentially modulates intracerebral brain dynamics.

Gueguen, M.C.M., Palminteri, S. & Bastin, J., Rewards and punishment learning performance are differentially affected by risk.

Guitart-Masip, M., Fuentemilla L, Bach DR, Huys QJ, Dayan P, Dolan RJ, Duzel E. 2011. Action Dominates Valence in Anticipatory Representations in the Human Striatum and Dopaminergic Midbrain. *Journal of Neuroscience*, 31(21), pp.7867–7875.

Guitart-Masip, M., Duzel E, Dolan R, Dayan P. 2014. Action versus valence in decision making. *Trends in Cognitive Sciences*, 18(4), pp.194–202.

Haber, S.N., Kim KS, Mailly P, Calzavara R. 2006. Reward-Related Cortical Inputs Define a Large Striatal Region in Primates That Interface with Associative Cortical Connections, Providing a Substrate for Incentive-Based Learning. *Journal of Neuroscience*, 26(32), pp.8368–8376.

Haber, S.N., 2003. The primate basal ganglia: parallel and integrative networks. *Journal of chemical neuroanatomy*, 26(4), pp.317–30.

Hadland, K.A., Rushworth MF, Gaffan D, Passingham RE. 2003a. The anterior cingulate and reward-guided selection of actions. *Journal of neurophysiology*, 89(2), pp.1161–4.

Hadland, K.A., Rushworth MF, Gaffan D, Passingham RE. 2003b. The effect of cingulate lesions on social behaviour and emotion. *Neuropsychologia*, 41(8), pp.919–31.

Häggström, M., 2017. Basal ganglia. *Wikipedia*.

Ham, T., Leff, A., de Boissezon, X., Joffe, A., and Sharp, D.J. 2013. Cognitive Control and the Salience Network: An Investigation of Error Processing and Effective Connectivity. J. Neurosci. *33*, 7091–7098.

Hampton, A.N., Adolphs R, Tyszka MJ, O'Doherty JP. 2007. Contributions of the Amygdala to Reward Expectancy and Choice Signals in Human Prefrontal Cortex. *Neuron*, 55(4), pp.545–555.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., and Rangel, A. 2008. Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. J. Neurosci. *28*, 5623–5630.

Hare, T.A., Camerer, C.F. & Rangel, A., 2009. Self-control in decision-making involves modulation of the vmPFC valuation system. *Science (New York, N.Y.)*, 324(5927), pp.646–8.

Harlé, K.M., Chang LJ, van 't Wout M, Sanfey AG. 2012. The neural mechanisms of affect infusion in social economic decision-making : A mediating role of the anterior insula. *NeuroImage*, 61(1), pp.32–40.

Hayashi, K., Nakao, K. & Nakamura, K., 2015. Appetitive and Aversive Information Coding in the Primate Dorsal Raphe Nucleus. *Journal of Neuroscience*, 35(15), pp.6195–6208.

Hayden, B.Y., Heilbronner SR, Nair AC, Platt ML. 2008. Cognitive influences on risk-seeking by rhesus macaques. *Judgment and decision making*, 3(5), pp.389–395.

Hayden, B.Y., Pearson, J.M. & Platt, M.L., 2009. Fictive reward signals in the anterior cingulate cortex. *Science (New York, N.Y.)*, 324(5929), pp.948–950.

Hayden, B.Y. & Platt, M.L., 2009. Gambling for Gatorade: risk-sensitive decision making for fluid rewards in humans. *Animal cognition*, 12(1), pp.201–7.

Hayden, B.Y. & Platt, M.L., 2010. Neurons in Anterior Cingulate Cortex Multiplex Information about Reward and Action. *Journal of Neuroscience*, 30(9), pp.3339–3346.

Hayes, D.J., Duncan NW, Xu J, Northoff G. 2014. A comparison of neural responses to appetitive and aversive stimuli in humans and other mammals. *Neuroscience & Biobehavioral Reviews*, 45, pp.350–68.

Hellwig, B., 1993. How the myelin picture of the human cerebral cortex can be computed from cytoarchitectural data. A bridge between von Economo and Vogt. *Journal fur Hirnforschung*, 34(3), pp.387–402.

Hennigan, K., D'Ardenne, K. & McClure, S.M., 2015. Distinct Midbrain and Habenula Pathways Are Involved in Processing Aversive Events in Humans. *Journal of Neuroscience*, 35(1), pp.198–208.

Hester, R., Murphy K, Brown FL, Skilleter AJ 2010. Punishing an error improves learning: the influence of punishment magnitude on error-related neural activity and subsequent learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(46), pp.15600–15607.

Hikosaka, K. & Watanabe, M., 2000. Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral cortex (New York, N.Y. : 1991)*, 10(3), pp.263–71.

Hirose, S., Osada T, Ogawa A, Tanaka M, Wada H, Yoshizawa Y, Imai Y, Machida T,

Akahane M, Shirouzu I, Konishi S. 2016. Lateral–Medial Dissociation in Orbitofrontal Cortex–Hypothalamus Connectivity. *Frontiers in Human Neuroscience*, 10, p.244.

Hof, P.R., Mufson, E.J. & Morrison, J.H., 1995. Human orbitofrontal cortex: Cytoarchitecture and quantitative immunohistochemical parcellation. *The Journal of Comparative Neurology*, 359(1), pp.48–68.

Holland, P.C., 2012. Role of amygdala central nucleus in feature negative discriminations. *Behavioral neuroscience*, 126(5), pp.670–80.

Hollerman, J.R. & Schultz, W., 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience*, 1(4), pp.304–9.

Hollerman, J.R., Tremblay, L. & Schultz, W., 2000. Involvement of basal ganglia and orbitofrontal cortex in goal-directed behavior. *Progress in brain research*, 126, pp.193–215.

Holmes, N.M., Marchand, A.R. & Coutureau, E., 2010. Pavlovian to instrumental transfer: A neurobehavioural perspective. *Neuroscience and Biobehavioral Reviews*, 34(8), pp.1277–1295.

Hosokawa, T., Kato K, Inoue M, Mikami A. 2007. Neurons in the macaque orbitofrontal cortex code relative preference of both rewarding and aversive outcomes. *Neuroscience Research*, 57(3), pp.434–445.

Houde, J.F., Nagarajan SS, Sekihara K, Merzenich MM 2002. Modulation of the auditory cortex during speech: an MEG study. *Journal of cognitive neuroscience*, 14(8), pp.1125–38.

Howard, R.A., 1960. *Dynamic programming and Markov decision processes*,

Hsu, M., Bhatt M, Adolphs R, Tranel D, Camerer CF. 2005. Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, 310(5754), pp.1680–1683.

Huettel, S.A., Stowe CJ, Gordon EM, Warner BT, Platt ML 2006. Neural signatures of economic preferences for risk and ambiguity. *Neuron*, 49(5), pp.765–75.

Hull, C.L., 1943. *Principles of Behaviour*, Appleton-Century-Crofts.

Hunt, L.T., Behrens, T.E., Hosokawa, T., Wallis, J.D., and Kennerley, S.W. 2015. Capturing the temporal evolution of choice across prefrontal cortex. Elife *4*, e11945.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F.S., and Behrens, T.E.J. 2012. Mechanisms underlying cortical activity during value-guided choice. Nat. Neurosci. *15*, 470–476.

Hutchison, W.D., Davis KD, Lozano AM, Tasker RR, Dostrovsky JO. 1999. Pain-related neurons in the human cingulate cortex. *Nature Neuroscience*, 2(5), pp.403–405.

Ibáñez-Sandoval, O., Xenias HS, Tepper JM, Koós T. 2015. Dopaminergic and cholinergic modulation of striatal tyrosine hydroxylase interneurons. *Neuropharmacology*, 95, pp.468–76.

Ishii, H., Ohara S, Tobler PN, Tsutsui K, Iijima T. 2012. Inactivating anterior insular cortex reduces risk taking. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(45), pp.16031–9.

Isnard, J., Guénot M, Ostrowsky K, Sindou M, Mauguière F. 2000. The role of the insular cortex in temporal lobe epilepsy. *Annals of Neurology*, 48(4), pp.614–623.

Isoda, M. & Hikosaka, O., 2008. Role for Subthalamic Nucleus Neurons in Switching from Automatic to Controlled Eye Movement. *Journal of Neuroscience*, 28(28), pp.7209–7218.

Ito, M. & Doya, K., 2015a. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(8), pp.3499–514.

Ito, M. & Doya, K., 2015b. Parallel Representation of Value-Based and Finite State-Based Strategies in the Ventral and Dorsal Striatum. O. Sporns, ed. *PLoS computational biology*, 11(11), p.e1004540.

Izquierdo, A., Suda, R.K. & Murray, E.A., 2004. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(34), pp.7540–8.

Jacobs, J., Kahana MJ, Ekstrom AD, Fried I. 2007. Brain oscillations control timing of single-neuron activity in humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(14), pp.3839–3844.

Jacobsen, C.F., 1935. Functions of Frontal Association Area in Primates. *Archives of*

Neurology And Psychiatry, 33(3), p.558.

Jakab, A., Molnár PP, Bogner P, Béres M, Berényi EL. 2012. Connectivity-based parcellation reveals interhemispheric differences in the insula. *Brain topography*, 25(3), pp.264–71.

Jankowski, M.M., Ronnqvist KC, Tsanov M, Vann SD, Wright NF, Erichsen JT, Aggleton JP, O'Mara SM. 2013. The anterior thalamus provides a subcortical circuit supporting memory and spatial navigation. *Frontiers in systems neuroscience*, 7(August), p.45.

Jo, S. & Jung, M.W., 2016. Differential coding of uncertain reward in rat insular and orbitofrontal cortex. *Nature Publishing Group*, (September 2015), pp.1–13.

Jocham, G., Klein, T.A. & Ullsperger, M., 2014. Differential Modulation of Reinforcement Learning by D2 Dopamine and NMDA Glutamate Receptor Antagonism. *Journal of Neuroscience*, 34(39), pp.13151–13162.

Jones, C.L., Ward, J. & Critchley, H.D., 2010. The neuropsychological impact of insular cortex lesions. *Journal of neurology, neurosurgery, and psychiatry*, 81(6), pp.611–8.

Jung, J., Jerbi, K., Ossandón, T., Ryvlin, P., Isnard, J., Bertrand, O., Guénot, M., Mauguière, F., and Lachaux, J.-P. 2010. Brain responses to success and failure: Direct recordings from human cerebral cortex. Hum. Brain Mapp. *31*, 1217–1232.

Jung, J., Bayle D, Jerbi K, Vidal JR, Hénaff MA, Ossandon T, Bertrand O, Mauguière F, Lachaux JP. 2011. Intracerebral γ modulations reveal interaction between emotional processing and action outcome evaluation in the human orbitofrontal cortex. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 79(1), pp.64–72.

Kable, J.W. & Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience*, 10(12), pp.1625–33.

Kacelnik, A., 1997. Normative and descriptive models of decision making: time discounting and risk sensitivity. In *Ciba Foudation Symposium*. p. 208: 51-70.

Kahneman, D. & Frederick, S., 2005. *A model of heuristic judgment*,

Kahneman, D., Lovallo, D. & Sibony, O., 2011. Before you make that big decision... *Harvard business review*, 89(6), pp.50–60, 137.

Kahneman, D., Slovic, P. & Tversky, A., 1982. *Judgment Under Uncertainty: Heuristics and*

Biases, Cambridge University Press.

Kahneman, D. & Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica*, (47), pp.263–291.

Kahnt, T., Park, S.Q., Haynes, J.-D., and Tobler, P.N. 2014. Disentangling neural representations of value and salience in the human brain. Proc. Natl. Acad. Sci. *111*, 5000–5005.

Kamin, L.J., 1969a. Predictability, surprise, attention, and conditioning. In B. A. Camp, ed. *Punishment and Aversive Behavior*. pp. 279–296.

Kamin, L.J., 1969b. Selective association and conditioning. In N.J. Mackintosh and W.K. Honig, ed. *Fundamental issues in associative learning*. Dalhousie University, Halifax: Halifax: Dalhousie University Press, pp. 42–64.

Kaplan, J.T., Gimbel, S.I. & Harris, S., 2016. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, 6(1), p.39589.

Kennerley, S.W., Behrens, T.E.J. & Wallis, J.D., 2011. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature neuroscience*, 14(12), pp.1581–9.

Kennerley, S.W. & Walton, M.E., 2011. Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies. *Behavioral neuroscience*, 125(3), pp.297–317.

Ketz, N.A., Jensen, O. & O'Reilly, R.C., 2015. Thalamic pathways underlying prefrontal cortex–medial temporal lobe oscillatory interactions. *Trends in Neurosciences*, 38(1), pp.3–12.

Kienast, T. & Heinz, A., 2006. Dopamine and the diseased brain. In *CNS Neurological Disorders Drug targets*. p. 5: 109-131.

Kilner, J.M., Mattout J, Henson R, Friston KJ. 2005. Hemodynamic correlates of EEG: A heuristic.

Kim, H., Shimojo, S. & O'Doherty, J.P., 2006. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS biology*, 4(8), p.e233.

Kirsch, P., Schienle, A., Stark, R., Sammer, G., Blecker, C., Walter, B., Ott, U., Burkart, J.,

and Vaitl, D. 2003. Anticipation of reward in a nonaversive differential conditioning paradigm and the brain reward system: NeuroImage *20*, 1086–1095.

Klavir, O., Genud-Gabai, R. & Paz, R., 2013. Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning. *Neuron*, 80(5), pp.1290–300.

Klein, T.A., Ullsperger, M. & Danielmeier, C., 2013. Error awareness and the insula: links to neurological and psychiatric diseases. *Frontiers in human neuroscience*, 7(February), p.14.

Knoch, D. & Fehr, E., 2007. Resisting the Power of Temptations: The Right Prefrontal Cortex and Self-Control. *Annals of the New York Academy of Sciences*, 1104(1), pp.123–134.

Knutson, B., Adams CM, Fong GW, Hommer D. 2001. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(16), p.RC159.

Knutson, B., Westdorp A, Kaiser E, Hommer D. 2000. FMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage*, 12(1), pp.20–27.

Knutson, B., Rick S, Wimmer GE, Prelec D, Loewenstein G. 2007. Neural predictors of purchases. *Neuron*, 53(1), pp.147–56.

Knutson, B. & Greer, S.M., 2008. Anticipatory affect: neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511), pp.3771–3786.

Knutson, B., Katovich, K. & Suri, G., 2014. Inferring affect from fMRI data. *Trends in Cognitive Sciences*, 18(8), pp.422–428.

Koechlin, E. & Summerfield, C., 2007. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6), pp.229–235.

Kolling, N., Wittmann, M.K. & Rushworth, M.F.S., 2014. Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron*, 81(5), pp.1190–1202.

Kozlovskiy, S.A., Vartanov, A.V., Pyasik, M.M., Velichkovsky, B.M. 2012. The Cingulate Cortex and Human Memory Process. *Psychology in Russia: State of Art*, 5(1), p.231.

Krack, P., Hariz MI, Baunez C, Guridi J, Obeso JA. 2010. Deep brain stimulation: from

neurology to psychiatry? *Trends in Neurosciences*, 33(10), pp.474–484.

Kralik, J.D., Xu ER, Knight EJ, Khan SA, Levine WJ 2012. When less is more: evolutionary origins of the affect heuristic. E. Addessi, ed. *PloS one*, 7(10), p.e46240.

Krigolson, O.E., Hassall, C.D. & Handy, T.C., 2014. How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. *Journal of cognitive neuroscience*, 26(3), pp.635–44.

Kringelbach, M.L., O'Doherty J, Rolls ET, Andrews C. 2003. Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral cortex (New York, N.Y. : 1991)*, 13(10), pp.1064–71.

Kringelbach, M.L., 2005. The human orbitofrontal cortex: linking reward to hedonic experience. *Nature reviews. Neuroscience*, 6(9), pp.691–702.

Kringelbach, M.L. & Rolls, E.T., 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in neurobiology*, 72(5), pp.341–72.

Kuhnen, C.M. & Knutson, B., 2005. The neural basis of financial risk taking. *Neuron*, 47(5), pp.763–70.

Kuramoto, E., Iwai H, Yamanaka A, Ohno S, Seki H, Tanaka YR, Furuta T, Hioki H, Goto T. 2017. Dorsal and ventral parts of thalamic nucleus submedius project to different areas of rat orbitofrontal cortex: a single neuron-tracing study using virus vectors. *Journal of Comparative Neurology*.

Kurth, F., Zilles K, Fox PT, Laird AR, Eickhoff SB. 2010. A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure and Function*, 214(5–6), pp.1–16.

Lachaux, J.-P., Axmacher N, Mormann F, Halgren E, Crone NE. 2012. High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Progress in neurobiology*, 98(3), pp.279–301.

Lachaux, J.-P., Fonlupt P, Kahane P, Minotti L, Hoffmann D, Bertrand O, Baciu M. 2007. Relationship between task-related gamma oscillations and BOLD signal: New insights from combined fMRI and intracranial EEG. *Human Brain Mapping*, 28(12), pp.1368–1375.

Lachaux, J.-P., Chavez, M. & Lutz, A., 2003. A simple measure of correlation across time, frequency and space between continuous brain signals. *Journal of Neuroscience Methods*, 123(2), pp.175–188.

Lammel, S., Lim, B.K., Ran, C., Huang, K.W., Betley, M.J., Tye, K.M., Deisseroth, K., and Malenka, R.C. 2012. Input-specific control of reward and aversion in the ventral tegmental area. *Nature*, 491(7423), pp.212–217.

Landauer, T.K., 1969. Reinforcement as consolidation. *Psychological review*, 76(1), pp.82–96.

Landmann, C. & Langlais, V., 2014. Anatomie et Organisation fonctionnelle du Thalamus. Available at: http://slideplayer.fr/slide/1310617/.

Larsen, T. & O'Doherty, J.P., 2014. Uncovering the spatio-temporal dynamics of value-based decision-making in the human brain: a combined fMRI-EEG study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655), pp.20130473–20130473.

Laury, S. & Holt, C., 2005. Further reflections on prospect theory. *Andrew Young School of Policy Studies Research Paper Series*, pp.06–11.

Lee, S.W., Shimojo, S. & O'Doherty, J.P., 2014. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), pp.687–99.

Lee, T.G., Blumenfeld, R.S. & D'Esposito, M., 2013. Disruption of Dorsolateral But Not Ventrolateral Prefrontal Cortex Improves Unconscious Perceptual Memories. *Journal of Neuroscience*, 33(32), pp.13233–13237.

Leech, R. & Sharp, D.J., 2014. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1), pp.12–32.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., Palminteri, S. 2017. Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 67(March), pp.1–9.

Lehéricy, S., Ducros M, Van de Moortele PF, Francois C, Thivard L, Poupon C, Swindale N, Ugurbil K, Kim DS. 2004. Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. *Annals of Neurology*, 55(4), pp.522–529.

Lehtimäki, K., Möttönen T, Järventausta K, Katisko J, Tähtinen T, Haapasalo J, Niskakangas

T, Kiekara T, Öhman J, Peltola J. 2016. Outcome based definition of the anterior thalamic deep brain stimulation target in refractory epilepsy. *Brain Stimulation*, 9(2), pp.268–275.

Leong, Y.C., Radulescu A, Daniel R, DeWoskin V, Niv Y. 2017. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2), pp.451–463.

Levy, B.J. & Wagner, A.D., 2011. Cognitive control and right ventrolateral prefrontal cortex: reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences*, 1224(1), pp.40–62.

Levy, R. & Dubois, B., 2006. Apathy and the Functional Anatomy of the Prefrontal Cortex–Basal Ganglia Circuits. *Cerebral Cortex*, 16(7), pp.916–928.

Li, Y., Zhong W, Wang D, Feng Q, Liu Z, Zhou J, Jia C, Hu F, Zeng J, Guo Q, Fu L, Luo M. 2016. Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nature communications*, 7, p.10503.

Li, Y., Vanni-Mercier G, Isnard J, Mauguière F, Dreher JC. 2016. The neural dynamics of reward value and risk coding in the human orbitofrontal cortex. *Brain*, 139(4), pp.1–15.

Liu, X., Hairston, J., Schrier, M., and Fan, J. 2011. Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 35(5), pp.1219–1236.

Ljungberg, T., Apicella, P. & Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of neurophysiology*, 67(1), pp.145–63.

Logothetis, N.K., Pauls J, Augath M, Trinath T, Oeltermann A. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843), pp.150–7.

Losecaat, A.B., Boksem, M.A.S. & Sanfey, A.G., 2014. Neural mechanisms underlying context-dependent shifts in risk preferences. *NeuroImage*, 103, pp.355–363.

Luna, B. & Sweeney, J.A., 2004. The Emergence of Collaborative Brain Function: fMRI Studies of the Development of Response Inhibition. *Annals of the New York Academy of Sciences*, 1021(1), pp.296–309.

Mackey, S. & Petrides, M., 2014. Architecture and morphology of the human ventromedial prefrontal cortex. *The European journal of neuroscience*, 40(5), pp.2777–96.

Mackey, S. & Petrides, M., 2010. Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains. *The European journal of neuroscience*, 32(11), pp.1940–50.

Mackintosh, N.J., 1975. A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological review*, pp.276–298.

Macmillan, M., 2000. An Odd Kind of Fame: Stories of Phineas Gage. In MIT Press, pp. 116–119, 307–333.

Macmillan, M., 2008. Phineas Gage - Unravelling the myth. *The Psychologist - British Psychological Society*, 21(9), pp.828–831.

MacPherson, S.E., Phillips, L.H. & Della Sala, S., 2002. Age, executive function, and social decision making: a dorsolateral prefrontal theory of cognitive aging. *Psychology and aging*, 17(4), pp.598–609.

Magri, C., Schridde U, Murayama Y, Panzeri S, Logothetis NK. 2012. The amplitude and timing of the BOLD signal reflects the relationship between local field potential power at different frequencies. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(4), pp.1395–407.

Maia, T. V, 2010. Two-factor theory, the actor--critic model, and conditioned avoidance. *Learning & Behavior*, 38(1), pp.50–67.

Mander, B.A., Rao V, Lu B, Saletin JM, Lindquist JR, Ancoli-Israel S, Jagust W, Walker MP. 2013. Prefrontal atrophy, disrupted NREM slow waves and impaired hippocampal-dependent memory in aging. *Nature neuroscience*, 16(3), pp.357–64.

Manning, J.R., Jacobs, J., Fried, I., and Kahana, M.J. 2009. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *Journal of Neuroscience*, 29(43), pp.13613–13620.

Maris, E. & Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), pp.177–190.

Markowitsch, H.J. & Pritzel, M., 1987. Single unit activity in cat prefrontal and parietal cortex during performance of a symmetrically reinforced go-no go task. *The International journal of neuroscience*, 32(3–4), pp.719–46.

Markowitz, H., 1952. Portofolio selection. *Journal of finance*, (7), pp.77–91.

Matsumoto, H., Tian, J., Uchida, N., and Watabe-Uchida, M. 2016. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. *eLife*, 5, p.e17328.

Matsumoto, M. & Hikosaka, O., 2009a. Representation of negative motivational value in the primate lateral habenula. *Nature Neuroscience*, 12(1), pp.77–84.

Matsumoto, M. & Hikosaka, O., 2009b. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248), pp.837–841.

McBride, W.J., Murphy, J.M. & Ikemoto, S., 1999. Localization of brain reinforcement mechanisms: intracranial self-administration and intracranial place-conditioning studies. *Behavioural brain research*, 101(2), pp.129–52.

McClure, S.M. & D'Ardenne, K., 2009. Comutational neuroimaging monitoring reward learning with blood flow. In J.-C. Dreher & L. Tremblay, eds. *Handbook of reward and decision making*. Academic Press - Elsevier.

McCoy, A.N. & Platt, M.L., 2005. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature neuroscience*, 8(9), pp.1220–1227.

McDonald, M.L., MacMullen C, Liu DJ, Leal SM, Davis RL. 2012. Genetic association of cyclic AMP signaling genes with bipolar disorder. *Translational psychiatry*, 2(10), p.e169.

McGuire, J.T., Nassar MR, Gold JI, Kable JW. 2014. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, 84(4), pp.870–881.

McHugh, S.B., Barkus C, Huber A, Capitão L, Lima J, Lowry JP, Bannerman DM. 2014. Aversive Prediction Error Signals in the Amygdala. *Journal of Neuroscience*, 34(27), pp.9024–9033.

Menon, V. & Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. *Brain structure & function*, 214(5–6), pp.655–667.

Mesulam, M.M. & Mufson, E.J., 1982. Insula of the old world monkey. I. Architectonics in the insulo-orbito-temporal component of the paralimbic brain. *The Journal of comparative neurology*, 212(1), pp.1–22.

Météreau, E. & Dreher, J.-C., 2013. Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral cortex (New York, N.Y. : 1991)*, 23(2), pp.477–87.

Meyers, C.A., Berman SA, Scheibel RS, Hayman A. 1992. Case report: acquired antisocial personality disorder associated with unilateral left orbital frontal lobe damage. *Journal of psychiatry & neuroscience : JPN*, 17(3), pp.121–5.

Meyniel, F., Sergent C, Rigoux L, Daunizeau J, Pessiglione M. 2013. Neurocomputational account of how the human brain decides when to have a break. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7), pp.2641–6.

Mihatsch, O. & Neuneier, R., 2002. Risk-Sensitive Reinforcement Learning. *Machine Learning*, 49(2/3), pp.267–290.

Miller, B.L. & Cummings, J.L., 2007. *The human frontal lobes : functions and disorders*, Guilford Press.

Miller, E.K., Freedman, D.J. & Wallis, J.D., 2002. The prefrontal cortex: categories, concepts and cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 357(1424), pp.1123–36.

Milner, P.M., 1988. The neural basis of reward and reinforcement. In *Neuroscience and biobehavioral reviews*. pp. 59–186.

Mirenowicz, J. & Schultz, W., 1994. Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of neurophysiology*, 72(2), pp.1024–7.

Mirenowicz, J. & Schultz, W., 1996. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379(6564), pp.449–51.

Mitchell, A.S., 2015. The mediodorsal thalamus as a higher order thalamic relay nucleus important for learning and decision-making. *Neuroscience & Biobehavioral Reviews*, 54, pp.76–88.

Mizuhiki, T., Richmond, B.J. & Shidara, M., 2012. Encoding of reward expectation by monkey anterior insular neurons. *Journal of Neurophysiology*, 107(11), pp.2996–3007.

Mohr, P.N.C., Biele, G. & Heekeren, H.R., 2010. Neural Processing of Risk. *Journal of Neuroscience*, 30(19), pp.6613–6619.

Monosov, I.E. & Hikosaka, O., 2012. Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(30), pp.10318–30.

Monosov, I.E., Leopold, D.A. & Hikosaka, O., 2015. Neurons in the Primate Medial Basal Forebrain Signal Combined Information about Reward Uncertainty, Value, and Punishment Anticipation. *Journal of Neuroscience*, 35(19), pp.7443–7459.

Montague, P.R., 1996. A Framework for Mesencephalic Predictive Hebbian Learning. In p. 76: 1936-1947.

Mora, F., Avrith, D.B. & Rolls, E.T., 1980. An electrophysiological and behavioural study of self-stimulation in the orbitofrontal cortex of the rhesus monkey. *Brain research bulletin*, 5(2), pp.111–5.

Morecraft, R.J., Geula, C. & Mesulam, M.M., 1992. Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. *The Journal of Comparative Neurology*, 323(3), pp.341–358.

Morrens, J., 2014. Dopamine neurons coding prediction errors in reward space, but not in aversive space: a matter of location? *Journal of neurophysiology*, 112(5), pp.1021–4.

Morris, G., Arkadir D, Nevet A, Vaadia E, Bergman H. 2004. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, 43(1), pp.133–143.

Morris, R.W., Dezfouli A, Griffiths KR, Balleine BW. 2014. Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. *Nature Communications*, 5, p.4390.

Morrison, S.E., Saez, A., Lau, B., and Salzman, C.D. 2011. Different Time Courses for Learning-Related Changes in Amygdala and Orbitofrontal Cortex. *Neuron*, 71(6), pp.1127–1140.

Morrison, S.E. & Salzman, C.D., 2009. The Convergence of Information about Rewarding and Aversive Stimuli in Single Neurons. *Journal of Neuroscience*, 29(37), pp.11471–11483.

Moss & Simmon, 2013. Dorsolateral prefrontal cortex. *Psychlopedia*.

Moutoussis, M., Bentall, R.P., Williams, J., Dayan, P. 2008. A temporal difference account of avoidance learning. *Network (Bristol, England)*, 19(2), pp.137–60.

Mowrer, R.R. & Klein, S.B., 2000. *Handbook of Contemporary Learning Theories*, Lawrence Erlbaum Associates Inc.

Mufson, E.J., Mesulam, M.M. & Pandya, D.N., 1981. Insular interconnections with the amygdala in the rhesus monkey. *Neuroscience*, 6(7), pp.1231–48.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. 2005. Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science (New York, N.Y.)*, 309(5736), pp.951–954.

Murray, E.A., Wise, S.P. & Graham, K.S., 2016. *The Evolution of Memory Systems*, Oxford: Oxford University Press.

Nambu, A., 2004. A new dynamic model of the cortico-basal ganglia loop. In *Progress in brain research*. pp. 461–466.

Nambu, A., Tokuno, H. & Takada, M., 2002. Functional significance of the cortico-subthalamo-pallidal “hyperdirect” pathway. *Neuroscience research*, 43(2), pp.111–7.

Namburi, P., Beyeler A, Yorozu S, Calhoon GG, Halbert SA, Wichmann R, Holden SS, Mertens KL, Anahtar M, Felix-Ortiz AC, Wickersham IR, Gray JM, Tye KM. 2015. A circuit mechanism for differentiating positive and negative associations. *Nature*, 520(7549), pp.675–678.

Naqvi, N.H., Rudrauf D, Damasio H, Bechara A. 2007. Damage to the insula disrupts addiction to cigarette smoking. *Science (New York, N.Y.)*, 315(5811), pp.531–4.

Naqvi, N.H., Gaznick N, Tranel D, Bechara A 2014. The insula: A critical neural substrate for craving and drug seeking under conflict and risk. *Annals of the New York Academy of Sciences*, 1316(1), pp.53–70.

Naqvi, N.H. & Bechara, A., 2009. The hidden island of addiction: the insula. *Trends in neurosciences*, 32(1), pp.56–67.

Naqvi, N.H. & Bechara, A., 2010. The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure and Function*, pp.1–16.

Nelson, C.A. & Collins, M.L., 2008. *Handbook of developmental cognitive neuroscience*, MIT Press.

Nestler, E.J., 2013. Cellular basis of memory for addiction. *Dialogues in clinical neuroscience*, 15(4), pp.431–43.

Newsome, W.T., Britten, K.H. & Movshon, J.A., 1989. Neuronal correlates of a perceptual

decision. *Nature*, 341(6237), pp.52–54.

Niessing, J., Ebisch B, Schmidt KE, Niessing M, Singer W, Galuske RA. 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science (New York, N.Y.)*, 309(5736), pp.948–951.

Nieuwenhuys, R., 2012. The insular cortex. In *Progress in brain research*. pp. 123–163.

Nishi, A., Kuroiwa, M. & Shuto, T., 2011. Mechanisms for the Modulation of Dopamine D1 Receptor Signaling in Striatal Neurons. *Frontiers in Neuroanatomy*, 5, p.43.

Nitschke, J.B., Dixon GE, Sarinopoulos I, Short SJ, Cohen JD, Smith EE, Kosslyn SM, Rose RM, Davidson RJ. 2006. Altering expectancy dampens neural response to aversive taste in primary taste cortex. *Nature neuroscience*, 9(3), pp.435–42.

Niv, Y., Edlund JA, Dayan P, O'Doherty JP. 2012. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2), pp.551–562.

Niv, Y., Daw, N.D. & Dayan, P., 2006. Choice values. *Nature neuroscience*, 9(8), pp.987–8.

Niv, Y., Joel, D. & Dayan, P., 2006. A normative perspective on motivation. *Trends in cognitive sciences*, 10(8), pp.375–81.

Niv, Y. & Schoenbaum, G., 2008. Dialogues on prediction errors. *Trends in cognitive sciences*, 12(7), pp.265–72.

O'Doherty, J.P., Kringelbach ML, Rolls ET, Hornak J, Andrews C. 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience*, 4(1), pp.95–102.

O'Doherty, J.P., Rolls ET, Francis S, Bowtell R, McGlone F, Kobal G, Renner B, Ahne G. 2000. Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *Neuroreport*, 11(2), pp.399–403.

O'Doherty, J.P., Hampton, A.N. & Kim, H., 2007. Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences*, 1104(1), pp.35–53.

O'Neill, M. & Schultz, W., 2010. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, 68(4), pp.789–800.

O'Neill, M. & Schultz, W., 2013. Risk Prediction Error Coding in Orbitofrontal Neurons. *Journal of Neuroscience*, 33(40), pp.15810–15814.

Olds, J., 1956. A preliminary mapping of electrical reinforcing effects in the rat brain. *Journal of comparative and physiological psychology*, 49(3), pp.281–5.

Olds, J. & Milner, P.M., 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, 47(6), pp.419–27.

Olsen, C.M., 2011. Natural rewards, neuroplasticity, and non-drug addictions. *Neuropharmacology*, 61(7), pp.1109–1122.

Ongür, D. & Price, J.L., 2000. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral cortex*, 10(3), pp.206–219.

Ostlund, S.B. & Balleine, B.W., 2007. The contribution of orbitofrontal cortex to action selection. *Annals of the New York Academy of Sciences*, 1121, pp.174–92.

Ouhaz, Z., Ba-M'hamed S, Mitchell AS, Elidrissi A, Bennis M. 2015. Behavioral and cognitive changes after early postnatal lesions of the rat mediodorsal thalamus. *Behavioural Brain Research*, 292, pp.219–232.

Padoa-Schioppa, C., 2007. Orbitofrontal cortex and the computation of economic value. *Annals of the New York Academy of Sciences*, 1121, pp.232–53.

Padoa-Schioppa, C., 2009. Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(44), pp.14004–14014.

Padoa-Schioppa, C. & Assad, J.A., 2006. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090), pp.223–6.

Padoa-Schioppa, C. & Assad, J.A., 2008. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature neuroscience*, 11(1), pp.95–102.

Padoa-Schioppa, C. & Cai, X., 2011. The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives. *Annals of the New York Academy of Sciences*, 1239(1), pp.130–7.

Palminteri, S., Boraud T, Lafargue G, Dubois B, Pessiglione M. 2009. Brain hemispheres selectively track the expected value of contralateral options. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(43), pp.13465–72.

Palminteri, S., Khamassi M, Joffily M, Coricelli G. 2015. Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6, p.8096.

Palminteri, S., Justo D, Jauffret C, Pavlicek B, Dauta A, Delmaire C, Czernecki V, Karachi C, Capelle L, Durr A, Pessiglione M. 2012. Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, 76(5), pp.998–1009.

Palminteri, S., Lebreton M, Worbe Y, Hartmann A, Lehéricy S, Vidailhet M, Grabli D, Pessiglione M. 2011. Dopamine-dependent reinforcement of motor skill learning: Evidence from Gilles de la Tourette syndrome. *Brain*, 134(8), pp.2287–2301.

Pandya, D.N., Van Hoesen, G.W. & Mesulam, M.M., 1981. Efferent connections of the cingulate gyrus in the rhesus monkey. *Experimental brain research*, 42(3–4), pp.319–30.

Papez, J.W., 1995. A proposed mechanism of emotion. 1937 [classical article]. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 7(1), pp.103–112.

Parnaudeau, S., O'Neill PK, Bolkan SS, Ward RD, Abbas AI, Roth BL, Balsam PD, Gordon JA, Kellendonk C. 2013. Inhibition of Mediodorsal Thalamus Disrupts Thalamofrontal Connectivity and Cognition. *Neuron*, 77(6), pp.1151–1162.

Parnaudeau, S., Taylor K, Bolkan SS, Ward RD, Balsam PD, Kellendonk C. 2015. Mediodorsal Thalamus Hypofunction Impairs Flexible Goal-Directed Behavior. *Biological Psychiatry*, 77(5), pp.445–453.

Paulus, M.P., 2009. Gut-level choices:risk-taking and interoception...insula. *Power Point*, pp.1–45.

Paulus, M.P., Rogalsky C, Simmons A, Feinstein JS, Stein MB. 2003. Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage*, 19(4), pp.1439–48.

Pavlov, I.P., 1927. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.*, Oxford: Oxford University Press.

Pavlov, I.P., 1928. *Lectures on Conditioned Reflexes*, International Publishers.

Pearl, J., 1984. *Heuristics: intelligent search strategies for computer problem solving*, Addison-Wesley.

Pena-Garijo, J., Barros-Loscertales A, Ventura-Campos N, Ruipérez-Rodríguez MÁ, Edo-Villamon S, Ávila C. 2011. [Involvement of the thalamic-cortical-striatal circuit in patients with obsessive-compulsive disorder during an inhibitory control task with reward and punishment contingencies]. *Revista de neurologia*, 53(2), pp.77–86.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106), pp.1042–5.

Peterson, R.L., 2005. The neuroscience of investing: fMRI of the reward system. *Brain research bulletin*, 67(5), pp.391–7.

Petrides, M., 2005. Lateral prefrontal cortex: architectonic and functional organization. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(April), pp.781–795.

Philiastides, M.G., Biele G, Vavatzanidis N, Kazzer P, Heekeren HR. 2010. Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage*, 53(1), pp.221–32.

Plassmann, H., O'Doherty, J.P. & Rangel, A., 2010. Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. *Journal of Neuroscience*, 30(32), pp.10799–10808.

Platt, M.L. & Huettel, S.A., 2008. Risky business: the neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, 11(4), pp.398–403.

Plotkin, H.C. & Oakley, D.A., 1975. Backward conditioning in the rabbit (Oryctolagus cuniculus). *Journal of comparative and physiological psychology*, 88(2), pp.586–90.

Polli, F.E., Barton JJ, Thakkar KN, Greve DN, Goff DC, Rauch SL, Manoach DS. 2008. Reduced error-related activation in two anterior cingulate circuits is related to impaired performance in schizophrenia. *Brain : a journal of neurology*, 131(Pt 4), pp.971–86.

Preuschoff, K., 't Hart, B.M. & Einhäuser, W., 2011. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, 5(September), pp.1–12.

Preuschoff, K., Quartz, S.R. & Bossaerts, P., 2008. Human insula activation reflects risk prediction errors as well as risk. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(11), pp.2745–2752.

Prevost, C., Pessiglione, M., Metereau, E., Clery-Melin, M.-L., and Dreher, J.-C. Separate Valuation Subsystems for Delay and Effort Decision Costs. *Journal of Neuroscience*, 30(42), pp.14080–14090.

Price, J.L., 1999. Prefrontal cortical networks related to visceral function and mood. *Annals of the New York Academy of Sciences*, 877, pp.383–96.

Proverbio, A.M., Riva, F. & Zani, A., 2010. When neurons do not mirror the agent's intentions: sex differences in neural coding of goal-directed actions. *Neuropsychologia*, 48(5), pp.1454–63.

Pujara, M.S., Wolf RC, Baskaya MK, Koenigs M. 2015. Ventromedial prefrontal cortex damage alters relative risk tolerance for prospective gains and losses. *Neuropsychologia*, 79, pp.70–75.

Ramautar, J.R., Slagter HA, Kok A, Ridderinkhof KR. 2006. Probability effects in the stop-signal paradigm: the insula and the significance of failed inhibition. *Brain research*, 1105(1), pp.143–54.

Ramayya, A.G., Pedisich, I. & Kahana, M.J., 2015. Expectation modulates neural representations of valence throughout the human brain. *NeuroImage*.

Rangel, A., Camerer, C.F. & Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. *Nature reviews. Neuroscience*, 9(7), pp.545–56.

Ray, J.P. & Price, J.L., 1993. The organization of projections from the mediodorsal nucleus of the thalamus to orbital and medial prefrontal cortex in macaque monkeys. *The Journal of comparative neurology*, 337(1), pp.1–31.

Redgrave, P., Rodriguez M, Smith Y, Rodriguez-Oroz MC, Lehericy S, Bergman H, Agid Y, DeLong MR, Obeso JA. 2010. Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience*, 11(11), pp.760–772.

Rektor, I., Doležalová I, Chrastina J, Jurák P, Halámek J, Baláž M, Brázdil M 2016. High-

Frequency Oscillations in the Human Anterior Nucleus of the Thalamus. *Brain Stimulation*, 9(4), pp.629–631.

Rescorla, R.A., 1969. Conditioned inhibition of fear resulting from negative CS-US contingencies. *Journal of comparative and physiological psychology*, 67(4), pp.504–9.

Rescorla, R.A., 1967. Pavlovian conditioning and its proper control procedures. *Psychological review*, 74(1), pp.71–80.

Rescorla, R.A. & Wagner, A.R., 1972. A Theory of Pavlovian Conditioning : Variations in the Effectiveness of Reinforcement and Nonreinforcement.

Reynolds, S.M. & Zahm, D.S., 2005. Specificity in the projections of prefrontal and insular cortex to ventral striatopallidum and the extended amygdala. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(50), pp.11757–67.

Richfield, E.K., Penney, J.B. & Young, A.B., 1989. Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. *Neuroscience*, 30(3), pp.767–77.

Richfield, E.K., Young, A.B. & Penney, J.B., 1989. Comparative distributions of dopamine D-1 and D-2 receptors in the cerebral cortex of rats, cats, and monkeys. *The Journal of comparative neurology*, 286(4), pp.409–26.

Ridderinkhof, K.R., 2014. Neurocognitive mechanisms of perception–action coordination: A review and theoretical integration. *Neuroscience & Biobehavioral Reviews*, 46, pp.3–29.

Robbins, T.W. & Arnsten, A.F.T., 2009. The Neuropsychopharmacology of Fronto-Executive Function: Monoaminergic Modulation. *Annual Review of Neuroscience*, 32(1), pp.267–287.

Roesch, M.R. & Olson, C.R., 2004. Neuronal Activity Related to Reward Value and Motivation in Primate Frontal Cortex. *Science*, 304(5668), pp.307–310.

Roesch, M.R., Taylor, A.R. & Schoenbaum, G., 2006. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron*, 51(4), pp.509–20.

Rolls, E.T., Critchley HD, Mason R, Wakeman EA. 1996. Orbitofrontal cortex neurons: role in olfactory and visual association learning. *Journal of neurophysiology*, 75(5), pp.1970–81.

Rolls, E.T., Burton, M.J. & Mora, F., 1980. Neurophysiological analysis of brain-stimulation reward in the monkey. *Brain research*, 194(2), pp.339–57.

Romo, R. & Schultz, W., 1990. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of neurophysiology*, 63(3), pp.592–606.

Rosenkilde, C.E., Bauer, R.H. & Fuster, J.M., 1981. Single cell activity in ventral prefrontal cortex of behaving monkeys. *Brain research*, 209(2), pp.375–94.

Rothschild, M. & Stiglitz, J.E., 1970. Increasing risk: I. A definition. *Journal of Economic Theory*, *2*(3), pp.225–243.

Rouault, M., 2015. *Integration of beliefs and affective values in human decision-making*.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., and Wager, T.D, 2014. Representation of aversive prediction errors in the human periaqueductal gray. Nat. Neurosci. *17*, 1607–1612.

Rushworth, M.F.S., Hadland KA, Gaffan D, Passingham RE. 2003. The effect of cingulate cortex lesions on task switching and working memory. *Journal of cognitive neuroscience*, 15(3), pp.338–53.

Rushworth, M.F.S. & Behrens, T.E.J., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4), pp.389–97.

Rutledge, R.B., Dean, M., Caplin, A., and Glimcher, P.W. 2010. Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. *Journal of Neuroscience*, 30(40), pp.13525–13536.

Rygula, R., Clarke HF, Cardinal RN, Cockcroft GJ, Xia J, Dalley JW, Robbins TW, Roberts AC. 2014. Role of Central Serotonin in Anticipation of Rewarding and Punishing Outcomes: Effects of Selective Amygdala or Orbitofrontal 5-HT Depletion. *Cerebral cortex (New York, N.Y. : 1991)*, p.bhu102.

Samanez-Larkin, G.R., Hollon, N.G., Carstensen, L.L., and Knutson, B. 2008. Individual differences in insular sensitivity during loss anticipation predict avoidance learning. *Psychological science*, 19(4), pp.320–3.

Schmajuk, N.A., 2008. Classical conditioning. *Scholarpedia*, 3(3), p.2316.

Schmajuk, N.A., Lamoureux, J.A. & Holland, P.C., 1998. Occasion setting: a neural network approach. *Psychological review*, 105(1), pp.3–32.

Schmidt, L., Lebreton M, Cléry-Melin ML, Daunizeau J, Pessiglione M. 2012. Neural mechanisms underlying motivation of mental versus physical effort. *PLoS biology*, 10(2), p.e1001266.

Schoenbaum, G., Takahashi Y, Liu TL, McDannald MA. 2011. Does the orbitofrontal cortex signal value? *Annals of the New York Academy of Sciences*, 1239(1), pp.87–99.

Schoenbaum, G., Roesch, M.R. & Stalnaker, T.A., 2006. Orbitofrontal cortex, decision-making and drug addiction. *Trends in neurosciences*, 29(2), pp.116–24.

Schultz, W., 1997. A Neural Substrate of Prediction and Reward. *Science*, 275(5306), pp.1593–1599.

Schultz, W., 2016. Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*.

Schultz, W., Preuschoff K, Camerer C, Hsu M, Fiorillo CD, Tobler PN, Bossaerts P. 2008. Explicit neural signals reflecting reward uncertainty. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1511), pp.3801–11.

Schultz, W., O'Neill M, Tobler PN, Kobayashi S. 2011. Neuronal signals for reward risk in frontal cortex. *Annals of the New York Academy of Sciences*, 1239(1), pp.109–117.

Schultz, W., 2007. Reward. *Scholarpedia*, 2(3), p.1652.

Schultz, W., Apicella, P., Ljungberg T, Romo R, Scarnati E., 1993. Reward-related activity in the monkey striatum and substantia nigra. *Progress in brain research*, 99, pp.227–35.

Schultz, W., Apicella, P. & Ljungberg, T., 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 13(3), pp.900–13.

Schultz, W. & Dickinson, A., 2000. Neuronal coding of prediction errors. *Annual review of neuroscience*, 23(1), pp.473–500.

Semendeferi, K., Lu A, Schenker N, Damasio H. 2002. Humans and great apes share a large frontal cortex. *Nature Neuroscience*, 5(3), pp.272–276.

Setogawa, T., Mizuhiki T, Matsumoto N, Akizawa F, Shidara M. 2014. Self-choice enhances value in reward-seeking in primates. *Neuroscience research*, 80, pp.45–54.

Seymour, B., O'Doherty JP, Koltzenburg M, Wiech K, Frackowiak R, Friston K, Dolan R 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, 8(9), pp.1234–1240.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. 2004. Temporal difference models describe higher-order learning in humans. *Nature*, 429(June), pp.664–667.

Seymour, B., Singer, T. & Dolan, R.J., 2007. The neurobiology of punishment. *Nature Reviews Neuroscience*, 8(4), pp.300–311.

Sharp, D.J., Bonnelle V, De Boissezon X, Beckmann CF, James SG, Patel MC, Mehta MA 2010. Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(4), pp.50–52.

Sheth, S.A., Mian MK, Patel SR, Asaad WF, Williams ZM, Dougherty DD, Bush G, Eskandar EN. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature*, 488(7410), pp.218–21.

Shimamura, A.P., 2000. The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, 28(2), pp.207–218.

Shizgal, P., 1997. Neural basis of utility estimation. *Current opinion in neurobiology*, 7(2), pp.198–208.

Silberberg, A., Roma PG, Huntsberry ME, Warren-Boulton FR, Sakagami T, Ruggiero AM, Suomi SJ. 2008. On loss aversion in capuchin monkeys. *Journal of the experimental analysis of behavior*, 89(2), pp.145–55.

Silvetti, M., Seurinck, R. & Verguts, T., 2011. Value and Prediction Error in Medial Frontal Cortex: Integrating the Single-Unit and Systems Levels of Analysis. *Frontiers in Human Neuroscience*, 5, p.75.

Simon, D., Craig KD, Miltner WH, Rainville P. 2006. Brain responses to dynamic facial expressions of pain. *Pain*, 126(1–3), pp.309–18.

Skinner, F.B., 1938. *The Behavior of Organisms: An Experimental Analysis.*, Appleton-

Century-Crofts.

Skoblenick, K.J., Womelsdorf, T. & Everling, S., 2016. Ketamine Alters Outcome-Related Local Field Potentials in Monkey Prefrontal Cortex. *Cerebral Cortex*, 26(6), pp.2743–2752.

Skvortsova, V., Palminteri, S. & Pessiglione, M., 2014. Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. *Journal of Neuroscience*, 34(47), pp.15621–15630.

Small, D.M., Small DM, Zatorre RJ, Dagher A, Evans AC, Jones-Gotman M. 2001. Changes in brain activity related to eating chocolate: from pleasure to aversion. *Brain : a journal of neurology*, 124(Pt 9), pp.1720–1733.

Smith, A.D. & Bolam, J.P., 1990. The neural network of the basal ganglia as revealed by the study of synaptic connections of identified neurones. *Trends in neurosciences*, 13(7), pp.259–65.

Smith, D.M., Freeman JH Jr, Nicholson D, Gabriel M. 2002. Limbic thalamic lesions, appetitively motivated discrimination learning, and training-induced neuronal activity in rabbits. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(18), pp.8212–21.

Smith, E.H., Banks, G.P., Mikell, C.B., Cash, S.S., Patel, S.R., Eskandar, E.N., and Sheth, S.A. 2015. Frequency-Dependent Representation of Reinforcement-Related Information in the Human Medial and Lateral Prefrontal Cortex. *Journal of Neuroscience*, 35(48), pp.15827–15836.

Sparenborg, S. & Gabriel, M., 1992. Local norepinephrine depletion and learning-related neuronal activity in cingulate cortex and anterior thalamus of rabbits. *Experimental Brain Research*, 92(2), pp.267–285.

Sridharan, D., Levitin, D.J. & Menon, V., 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, 105(34), pp.12569–12574.

Staddon, J.E.R., 1983. *Adaptive Behaviour and Learning*, Cambridge University Press.

Staddon, J.E.R. & Niv, Y., 2008. Operant conditioning. *Scholarpedia*, 3(9), p.2318.

Štillová, K., Jurák P, Chládek J, Chrastina J, Halámek J, Bočková M, Goldemundová S, Říha

I, Rektor I. 2015. The Role of Anterior Nuclei of the Thalamus: A Subcortical Gate in Memory Processing: An Intracerebral Recording Study. *PloS One*, 10(11), p.e0140778.

Sun, L., Sun, L., Peräkylä J, Polvivaara M, Öhman J, Peltola J, Lehtimäki K, Huhtala H, Hartikainen KM 2015. Human anterior thalamic nuclei are involved in emotion-attention interaction. *Neuropsychologia*, 78, pp.88–94.

Sutton, R.S., 1988. Learning to predict by the method of temporal differences. In *Machine Learning*. pp. 9–44.

Sutton, R.S. & Barto, A.G., 1998. Reinforcement Learning: an introduction.

Sutton, R.S. & Barto, A.G., 1981. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, 88(2), pp.135–70.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Borchardt V, Walter M, Hinrichs H, Heinze HJ, Rugg MD, Knight RT. 2017. Anterior Thalamic High Frequency Band Activity Is Coupled with Theta Oscillations at Rest. *Frontiers in Human Neuroscience*, 11, p.358.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Esslinger C, Hinrichs H, Heinze HJ, Knight RT, Richardson-Klavehn A. 2014. Corticothalamic phase synchrony and cross-frequency coupling predict human memory formation. *Elife*, 3, p.e05352.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Richardson-Klavehn A, Hinrichs H, Heinze HJ, Knight RT, Rugg MD. 2016. Pre-stimulus thalamic theta power predicts human memory formation. *NeuroImage*, 138, pp.100–108.

Sweeney-Reed, C.M., Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, Hinrichs H, Heinze HJ, Rugg MD, Knight RT, Richardson-Klavehn A. 2015. Thalamic theta phase alignment predicts human memory formation and anterior thalamic cross-frequency coupling. *Elife*, 4(MAY), p.e07578.

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. 2004. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature neuroscience*, 7(8), pp.887–93.

Tang, J.-S., Qu, C.-L. & Huo, F.-Q., 2009. The thalamic nucleus submedius and ventrolateral orbital cortex are involved in nociceptive modulation: A novel pain modulation pathway.

Progress in Neurobiology, 89(4), pp.383–389.

Taylor, S.B., Lewis, C.R. & Olive, M.F., 2013. The neurocircuitry of illicit psychostimulant addiction: acute and chronic effects in humans. *Substance abuse and rehabilitation*, 4, pp.29–43.

Temel, Y., Hescham SA, Jahanshahi A, Janssen ML, Tan SK, van Overbeeke JJ, Ackermans L, Oosterloo M, Duits A, Leentjens AF, Lim L. 2012. *Neuromodulation in psychiatric disorders.*

Thompson, A., Morishita, T. & Okun, M.S., 2012. *DBS and electrical neuro-network modulation to treat neurological disorders.* 1st ed., Elsevier Inc.

Thorndike, E.L., 1933. A Theory of the action of the after-effects of a connection upon it. *Psychological review*, pp.434–439.

Thorndike, E.L., 1911. *Animal intelligence: experimental studies*, Macmillan Publishers Limited, part of Springer Nature.

Thorpe, S.J., Rolls, E.T. & Maddison, S., 1983. The orbitofrontal cortex: neuronal activity in the behaving monkey. *Experimental brain research*, 49(1), pp.93–115.

Thut, G., Schultz W, Roelcke U, Nienhusmeier M, Missimer J, Maguire RP, Leenders KL. 1997. Activation of the human brain by monetary reward. *Neuroreport*, 8(5), pp.1225–8.

Tobler, P.N., O'Doherty JP, Dolan RJ, Schultz W. 2006. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of neurophysiology*, 97(2), pp.1621–1632.

Tobler, P.N., Fiorillo, C.D. & Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. *Science (New York, N.Y.)*, 307(5715), pp.1642–5.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. 2007. The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, 315(5811), pp.515–8.

Torres, N., Chabardes S, Piallat B, Devergnas A, Benabid AL. 2012. Body fat and body weight reduction following hypothalamic deep brain stimulation in monkeys: an intraventricular approach. *International Journal of Obesity*, 36(12), pp.1537–1544.

Tremblay, L. & Schultz, W., 2000. Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of neurophysiology*, 83(4),

pp.1877–85.

Tremblay, L. & Schultz, W., 1999. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729), pp.704–8.

Triarhou, L.C., 2013. The cytoarchitectonic map of Constantin von Economo and Georg N. Koshkinas. *Microstructural Parcellation of the Human Cerebral Cortex*, pp.33–54.

Tsujimoto, S., Genovesio, A. & Wise, S.P., 2012. Neuronal Activity during a Cued Strategy Task: Comparison of Dorsolateral, Orbital, and Polar Prefrontal Cortex. *Journal of Neuroscience*, 32(32), pp.11017–11031.

Türe, U., Yaşargil DC, Al-Mefty O, Yaşargil MG 1999. Topographic anatomy of the insular region. *Journal of neurosurgery*, 90(4), pp.720–33.

Ullsperger, M., Harsay HA, Wessel JR, Ridderinkhof KR. 2010. Conscious perception of errors and its relation to the anterior insula. *Brain structure & function*, 214(5–6), pp.629–43.

Umemoto, A., Lukie CN, Kerns KA, Müller U, Holroyd CB. 2014. Impaired reward processing by anterior cingulate cortex in children with attention deficit hyperactivity disorder. *Cognitive, affective & behavioral neuroscience*.

Ursin, H., Rosvold, H.E. & Vest, B., 1969. Food preference in brain lesioned monkeys. *Physiology & Behavior*, 4(4), pp.609–612.

Ursu, S. & Carter, C.S., 2005. Outcome representations, counterfactual comparisons and the human orbitofrontal cortex: Implications for neuroimaging studies of decision-making. *Cognitive Brain Research*, 23, pp.51–60.

Vallone, D., Picetti, R. & Borrelli, E., 2000. Structure and function of dopamine receptors. *Neuroscience and biobehavioral reviews*, 24(1), pp.125–32.

Vassena, E., Krebs RM., Silvetti M., Fias W., Verguts T. 2014. Dissociating contributions of ACC and vmPFC in reward prediction, outcome, and choice. *Neuropsychologia*, 59, pp.112–123.

Vermeer, A.B.L. & Sanfey, A.G., 2015. The Effect of Positive and Negative Feedback on Risk-Taking across Different Contexts. *PloS one*, pp.1–13.

Vertes, R.P., Linley, S.B. & Hoover, W.B., 2015. Limbic circuitry of the midline thalamus.

Neuroscience & Biobehavioral Reviews, 54, pp.89–107.

Vickery, T.J., Chun, M.M. & Lee, D., 2011. Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. *Neuron*, 72(1), pp.166–177.

Vignal, J.-P., Maillard L, McGonigal A, Chauvel P 2007. The dreamy state: hallucinations of autobiographic memory evoked by temporal lobe stimulations and seizures. *Brain : a journal of neurology*, 130(Pt 1), pp.88–99.

Volman, S.F., Lammel S, Margolis EB, Kim Y, Richard JM, Roitman MF, Lobo MK. 2013. New Insights into the Specificity and Plasticity of Reward and Aversion Encoding in the Mesolimbic System. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(45), pp.17569–17576.

Volz, K.G., Rübsamen, R. & von Cramon, D.Y., 2008. Cortical regions activated by the subjective sense of perceptual coherence of environmental sounds: a proposal for a neuroscience of intuition. *Cognitive, affective & behavioral neuroscience*, 8(3), pp.318–28.

Voon, V., Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, Hallett M. 2010. Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron*, 65(1), pp.1–14.

Walker, F.O., 2007. Huntington's disease. *The Lancet*, 369(9557), pp.218–228.

Wallis, J.D., 2011. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nature Neuroscience*, 15(1), pp.13–19.

Wallis, J.D. & Kennerley, S.W., 2011. Contrasting reward signals in the orbitofrontal cortex and anterior cingulate cortex. *Annals of the New York Academy of Sciences*, 1239, pp.33–42.

Walton, M.E., Behrens TE, Buckley MJ, Rudebeck PH, Rushworth MF. 2010. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, 65(6), pp.927–39.

Wang, J., Huo FQ, Li YQ, Chen T, Han F, Tang JS. 2005. Thalamic nucleus submedius receives GABAergic projection from thalamic reticular nucleus in the rat. *Neuroscience*, 134(2), pp.515–23.

Wang, J., Cao B., Yu TR, Jelfs B, Yan J, Chan RH, Li Y. 2015. Theta-frequency phase-

locking of single anterior cingulate cortex neurons and synchronization with the medial thalamus are modulated by visceral noxious stimulation in rats. *Neuroscience*, 298, pp.200–210.

Watkins, C.J.C.H. & Dayan, P., 1992. Technical Note Q , -Learning. , 292, pp.279–292.

Weber, E.U., Shafir, S. & Blais, A.-R., 2004. Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*, 111(2), pp.430–445.

Weller, J., Levin IP, Shiv B, Bechara A. 2009. The effects of insula damage on decision-making for risky gains and losses. *Social Neuroscience*, 4(4), pp.347–358.

Wheeler, E.Z. & Fellows, L.K., 2008. The human ventromedial frontal lobe is critical for learning from negative feedback. *Brain : a journal of neurology*, 131(Pt 5), pp.1323–31.

Wiech, K., Lin, C.S., Brodersen, K.H., Bingel, U., Ploner, M., and Tracey, I. 2010. Anterior Insula Integrates Information about Salience into Perceptual Decisions about Pain. *Journal of Neuroscience*, 30(48), pp.16324–16331.

Wise, R.A., 2002. Brain reward circuitry: insights from unsensed incentives. *Neuron*, 36(2), pp.229–40.

de Wit, S. & Dickinson, A., 2009. Associative theories of goal-directed behaviour: a case for animal-human translational models. *Psychological research*, 73(4), pp.463–76.

Worbe, Y., Palminteri S, Hartmann A, Vidailhet M, Lehéricy S, Pessiglione M. 2011. Reinforcement Learning and Gilles de la Tourette Syndrome. *Archives of General Psychiatry*, 68(12), p.1257.

Wright, N.F., Vann SD, Aggleton JP, Nelson AJ. 2015. A Critical Role for the Anterior Thalamus in Directing Attention to Task-Relevant Stimuli. *Journal of Neuroscience*, 35(14), pp.5480–5488.

Wright, N.D., Symmonds M, Hodgson K, Fitzgerald TH, Crawford B, Dolan RJ. 2012. Approach-Avoidance Processes Contribute to Dissociable Impacts of Risk and Loss on Choice. *Journal of Neuroscience*, 32(20), pp.7009–7020.

Wu, C.C., S Samanez-Larkin GR, Katovich K, Knutson B. 2014. Affective traits link to reliable neural markers of incentive anticipation. *NeuroImage*, 84, pp.279–89.

Wu, C.C., Sacchet, M.D. & Knutson, B. 2012. Toward an affective neuroscience account of financial risk taking. *Frontiers in neuroscience*, 6, p.159.

Xu, E.R., 2014. *Affective decision-making in rhesus monkeys: reward order, risk and the roles of the anterior insula and premotor cortex*. Dartmouth College.

Xue, G., Lu Z, Levin IP, Bechara A. 2010. The impact of prior risk experiences on subsequent risky decision-making: the role of the insula. *NeuroImage*, 50(2), pp.709–16.

Yacubian, J., 2006. Dissociable Systems for Gain- and Loss-Related Value Predictions and Errors of Prediction in the Human Brain. *Journal of Neuroscience*, 26(37), pp.9530–9537.

Yager, L.M., Garcia A.F., Wunsch A.M., Ferguson S.M. 2015. The ins and outs of the striatum: role in drug addiction. *Neuroscience*, 301, pp.529–41.

Yang, Y. & Raine, A., 2009. Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: A meta-analysis. *Psychiatry Research: Neuroimaging*, 174(2), pp.81–88.

Yaxley, S., Rolls, E.T. & Sienkiewicz, Z.J., 1988. The responsiveness of neurons in the insular gustatory cortex of the macaque monkey is independent of hunger. *Physiology & behavior*, 42(3), pp.223–9.

Yelnik, J., Bardinet E, Dormont D, Malandain G, Ourselin S, Tandé D, Karachi C, Ayache N, Cornu P, Agid Y. 2007. A three-dimensional, histological and deformable atlas of the human basal ganglia. I. Atlas construction based on immunohistochemical and MRI data. *NeuroImage*, 34(2), pp.618–638.

Yelnik, J., 2002. Functional anatomy of the basal ganglia. *Movement disorders : official journal of the Movement Disorder Society*, 17 Suppl 3, pp.S15-21.

Yin, H.H. & Knowlton, B.J., 2006. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6), pp.464–476.

Résumé de la thèse :

Chaque jour, nous prenons des décisions impliquant de choisir les options qui nous semblent les plus avantageuses, en nous basant sur nos expériences passées. Toutefois, les mécanismes et les bases neurales de l'apprentissage par renforcement restent débattus. D'une part, certains travaux suggèrent l'existence de deux systèmes opposés impliquant des aires cérébrales corticales et sous-corticales distinctes lorsque l'on apprend par la carotte ou par le bâton. D'autres part, des études ont montré une ségrégation au sein même de ces régions cérébrales ou entre des neurones traitant l'apprentissage par récompenses et celui par évitement des punitions. Le but de cette thèse était d'étudier la dynamique cérébrale de l'apprentissage par renforcement chez l'homme. Pour ce faire, nous avons utilisé des enregistrements intracérébraux réalisés chez des patients épileptiques pharmacorésistants pendant qu'ils réalisaient une tâche d'apprentissage probabiliste. Dans les deux premières études, nous avons investigué la dynamique de l'encodage des signaux de renforcement, et en particulier à celui des erreurs de prédiction des récompenses et des punitions. L'enregistrement de potentiels de champs locaux dans le cortex a mis en évidence le rôle central de l'activité à haute-fréquence gamma (50-150Hz). Les résultats suggèrent que le cortex préfrontal ventro-médian est impliqué dans l'encodage des erreurs de prédiction des récompenses alors que pour l'insula antérieure, le cortex préfrontal dorsolatéral sont impliqués dans l'encodage des erreurs de prédiction des punitions. De plus, l'activité neurale de l'insula antérieure permet de prédire la performance des patients lors de l'apprentissage. Ces résultats sont cohérents avec l'existence d'une dissociation au niveau cortical pour le traitement des renforcements appétitifs et aversifs lors de la prise de décision. La seconde étude a permis d'étudier l'implication de deux noyaux limbiques du thalamus au cours du même protocole cognitif. L'enregistrement de potentiels de champs locaux a mis en évidence le rôle des activités basse fréquence thêta dans la détection des renforcements, en particulier dans leur dimension aversive. Dans une troisième étude, nous avons testé l'influence du risque sur l'apprentissage par renforcement. Nous rapportons une aversion spécifique au risque lors de l'apprentissage par évitement des punitions ainsi qu'une diminution du temps de réaction lors de choix risqués permettant l'obtention de récompenses. Cela laisse supposer un comportement global tendant vers une aversion au risque lors de l'apprentissage par évitement des punitions et au contraire une attirance pour le risque lors de l'apprentissage par récompenses, suggérant que les mécanismes d'encodage du risque et de la valence pourraient être indépendants. L'amélioration de la compréhension des mécanismes cérébraux sous-tendant la prise de décision est importante, à la fois pour mieux comprendre les déficits motivationnels caractérisant plusieurs pathologies neuropsychiatriques, mais aussi pour mieux comprendre les biais décisionnels que nous pouvons exhiber.

Summary of this PhD thesis:

We make decisions every waking day of our life. Facing our options, we tend to pick the most likely to get our expected outcome. Taking into account our past experiences and their outcome is mandatory to identify the best option. This cognitive process is called reinforcement learning. To date, the underlying neural mechanisms are debated. Despite a consensus on the role of dopaminergic neurons in reward processing, several hypotheses on the neural bases of reinforcement learning coexist: either two distinct opposite systems covering cortical and subcortical areas, or a segregation of neurons within brain regions to process reward-based and punishment-avoidance learning. This PhD work aimed to identify the brain dynamics of human reinforcement learning. To unravel the neural mechanisms involved, we used intracerebral recordings in refractory epileptic patients during a probabilistic learning task. In the first study, we used a computational model to tackle the brain dynamics of reinforcement signal encoding, especially the encoding of reward and punishment prediction errors. Local field potentials exhibited the central role of high frequency gamma activity (50-150Hz) in these encodings. We report a role of the ventromedial prefrontal cortex in reward prediction error encoding while the anterior insula and the dorsolateral prefrontal cortex encoded punishment prediction errors. In addition, the magnitude of the neural response in the insula predicted behavioral learning and trial-to-trial behavioral adaptations. These results are consistent with the existence of two distinct opposite cortical systems processing reward and punishments during reinforcement learning. In a second study, we recorded the neural activity of the anterior and dorsomedial nuclei of the thalamus during the same cognitive task. Local field potentials recordings highlighted the role of low frequency theta activity in punishment processing, supporting an implication of these nuclei during punishment-avoidance learning. In a third behavioral study, we investigated the influence of risk on reinforcement learning. We observed a risk-aversion during punishment-avoidance, affecting the performance, as well as a risk-seeking behavior during reward-seeking, revealed by an increased reaction time towards appetitive risky choices. Taken together, these results suggest we are risk-seeking when we have something to gain and risk-averse when we have something to lose, in contrast to the prediction of the prospect theory. Improving our common knowledge of the brain dynamics of human reinforcement learning could improve the understanding of cognitive deficits of neurological patients, but also the decision bias all human beings can exhibit.