Apprendre à apprendre dans un environnement incertain, et dynamique des réseaux corticaux pour la flexibilité comportementale

Maïlys Faraut

N° d'ordre : 309-2015 Année 2015

THÈSE DE L'UNIVERSITÉ DE LYON

Délivrée par

L'UNIVERSITÉ CLAUDE BERNARD LYON 1

ÉCOLE DOCTORALE Neurosciences et Cognition

DIPLÔME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le 15 décembre 2015

par

Mme FARAUT Maïlys

Apprendre à apprendre dans un environnement incertain, et dynamique des réseaux corticaux pour la flexibilité comportementale

Directeur de thèse : M. PROCYK Emmanuel

Co-directeur de thèse : M. WILSON Charlie

Jury :	M. APICELLA Paul	(CNRS Marseille)	rapporteur
	M. COUTUREAU Etienne	(CNRS Bordeaux)	rapporteur
	M. GERVAIS Rémi	(Université Lyon 1)	examinateur
	M. KOECHLIN Etienne	(INSERM, Paris)	examinateur
	M. LEVY Richard	(APHP et Université Paris 6, Paris)	examinateur

Thèse préparée à l'Institut Cellule Souche et Cerveau, 18 avenue du Doyen Lépine 69500 Bron

Résumé

Notre environnement est complexe et changeant, ce qui apporte de l'incertitude dans les décisions de tous les jours. La capacité de détecter et résoudre l'incertitude est cruciale pour un comportement flexible et adapté. Notre hypothèse est que l'efficacité et la flexibilité comportementale en situation d'incertitude dépendent de la façon dont l'individu a appris à apprendre. Dans une 1ère étude, trois singes ont acquis un learning set pour une tâche aux règles stochastiques et changeantes. Leur réactivité aux évènements inattendus a augmenté lors de l'apprentissage, suivant l'évolution du degré d'incertitude environnementale. Cela a permis un transfert sans coût à une tâche plus complexe partageant la même structure, suggérant que les singes ont appris à apprendre la structure statistique de l'environnement. Nous avons ensuite étudié les mécanismes cérébraux sous-jacents à ce comportement flexible. Deux animaux ont reçu un implant d'électrocorticographie, sur les aires frontales et pariétales. Nous montrons d'abord, avec les données d'un animal, que des potentiels évoqués au feedback sont sensibles à la valence et au degré de surprise du feedback, et prédisent la stratégie à venir. Ensuite, nous présentons des résultats préliminaires montrant que des oscillations dans les bandes beta et thêta sont présentes au moment du feedback et de la décision, et que leur puissance est modulée de manière différente par les facteurs de la tâche. Ces résultats contribuent à révéler la complexité du réseau frontal pour la flexibilité comportementale, et ouvrent la voie à de nouvelles expériences pour comprendre comment ces mécanismes sont façonnés au cours du processus d'apprendre à apprendre.

Mots clés : apprentissage, flexibilité, prise de décision, apprendre-à-apprendre, électrophysiologie, cortex frontal, singe

Abstract

Our environment is both complex and changing, which triggers uncertainty in every decision we make. The ability to detect and solve the resulting uncertainty is crucial for adapted and flexible behavior. Our hypothesis is that behavioral efficiency and flexibility in an uncertain environment depend on the way the agent has learnt to learn. In a first study, 3 macaque monkeys developed a learning-set for a task with stochastic and changing rules. Monkey's reactivity to unexpected feedback increased across learning and paralleled the evolution of the degree of environmental uncertainty. This enabled them to transfer, without cost, to a more complex task with the same structure, suggesting that they learned to learn the statistical structure of the environment. We then studied the cerebral mechanisms underpinning this flexible behavior. Two animals were implanted with an electrocorticography implant over the frontal and parietal areas. We first showed, using data from one animal, that feedback related potentials were sensitive to feedback valence and unexpectedness, and predictive of the upcoming behavioral strategy. Then, we present preliminary results showing that oscillations in the beta and theta bands can be recorded at the time of feedback and at the time of decision, and that their power is modulated differently depending on the various task factors. These results contribute to reveal the complexity of the frontal cortical network enabling behavioral flexibility and open new horizons for future research to understand how these mechanisms are *shaped* throughout the learning to learn process.

Key words : learning, flexibility, decision making, learning-to-learn, electrophysiology, frontal cortex, monkey

Remerciements

Au risque de défier la hiérarchie, je souhaite commencer ces remerciements par un petit message à l'attention de Pippa, Dali et Kate, mais aussi Elie et Mary : merci les p'tits loups, c'est vous qui avez fait tout le boulot !

Je remercie Manu, de m'avoir accueillie dans son équipe, malgré mon passé de biologiste ignorante de l'électrophysiologie (voire du cerveau !) et de la programmation. Je suis très fière d'avoir pu apprendre tout ça, grâce à ta confiance. Merci aussi d'être un exemple de scientifique droit et honnête. Je garde toujours en tête ce qui constitue presque ta devise : "les données sont les données"... même si elles ne nous plaisent pas comme elles sont !

Je remercie les membres de mon jury d'avoir accepté d'évaluer ce travail : Paul Apicella, Etienne Coutureau, qui ont dû lire toutes ces pages, et Richard Levy. Je mets une emphase particulière en direction d'Etienne Koechlin, avec qui nous avons collaboré et qui est le concepteur original de la tâche comportementale utilisée dans ce travail (avec Anne Collins), et aussi en direction de Rémi Gervais qui m'a fait découvrir la passion de lier le comportement et les neurosciences.

Merci à Charlie et Fred. Je ne sais pas combien de pintes je vous dois pour tout ce que vous avez fait pour moi durant ces années, autant sur le plan professionnel, qu'amical. Merci d'être et d'avoir été un soutien indéfectible et merci de m'avoir supportée, moi et mes crises de folie de fin de journée. Merci aussi, parce que vous m'avez tout appris :

— Charlie, merci de m'avoir appris l'art d'entrainer les singes, appris que "tout dépend de la question qu'on veut poser", merci de m'avoir poussée et soutenue, en me répétant qu'il faut toujours être positif, et ... de m'avoir appris à boire de la bière, car vendredi, c'est pub !

— Fred, merci de m'avoir appris à programmer et avoir débuggé patiemment tous mes scripts, appris à faire des belles figures, appris la critique constructive et déconstructive, de m'avoir soutenue, mais aussi de m'avoir montré qu'il est possible de changer de style tous les deux mois (avec plus ou moins de réussite, mais c'est ça, un vrai hipster).

Un remerciement spécial aux animaliers : Florianne, Marco, Muriel, Xavier, Sandra, Margaux, Brigitte. Merci d'avoir pris et de prendre toujours soin de nos petits, même les we et jours fériés !

Merci à Chandara, d'avoir si efficacement assuré avec autant de gentillesse toutes mes étourderies administratives.

Merci à l'équipe : Vincent, Kep Kee, Céline, Nicolas, Julien, Estelle... et Maxime.

Je n'oublie pas, bien sûr, mes partenaires préférés de bêtises : Marion et Pierre ; et toute la tripotée d'étudiants de l'institut, dont Loulou, Grégoire, Rita, Elodie et tous les autres, qui participent activement à l'entrain collectif !

Une grosse bise à la bande des Chilipodes : Marie, Mathilde et Judith, et Claire, mes amies de toujours ; à la ptite bande de l'ENS, en particulier, Marion, Clémentine, Agnès, Fabian, Emily, bientôt presque tous docteurs ! Merci à Jadou, et Adriou, colocs et amis.

Bien sûr, un grand merci aux membres de ma famille, pour leur foi en moi et leur soutien de toujours, et à ma belle-famille, qui commence à être présente à mes côtés depuis un bout de temps déjà !

Merci à Clément, mon double, parce que comme tu le dis, cette thèse est aussi un peu la tienne ! J'ai hâte de partir avec toi à l'aventure, pour le post-doc !

Table des matières

Abréviations

Ach : acetylcholine

BOLD : Blood Oxygenation Level Dependant

CCA : cortex cingulaire antérieur,

CCM : cortex cingulaire médian

CCP : cortex cingulaire postérieur

COF : cortex orbito-frontal

CPA : Couplage Phase-Amplitude

CPF : cortex préfrontal (dm : dorso-médian ; vm : ventro-médian ; dl dorso-latéral)

DDM : Drift Diffusion Model

EEG : ElectroEncephaloGramme

ERN : Error Related Negativity

FB : feedback

FRN : Feedback Related Negativity

FRP : Feedback Related Potential

IRMf : Imagerie par Résonance Magnétique fonctionnelle

IT : Identity Task

LIP : cortex intrapariétal latéral

PCL : potentiels de champs locaux

Pré-SMA : aire motrice pré-supplémentaire

PST : Problem Solving Task

RL : Reinforcement Learning

ST : Switch Task

STN : noyau sous-thalamique

WSLS : Win Stay – Lose Shift

Par souci de précision sur les concepts employés, certains termes, difficiles à traduire de l'anglais de manière satisfaisante, sont en français dans le texte, suivi du terme anglais entre parenthèses et en italique.

Problématique

Notre environnement est complexe et changeant, et pour survivre, que ce soit pour fuir un danger ou trouver de la nourriture, les organismes ont dû développer des moyens d'apprendre de leur environnement afin de produire un comportement adapté et flexible. Dans cette thèse, nous nous intéresserons aux processus comportementaux et mécanismes neurophysiologiques qui nous permettent d'apprendre de l'environnement. En particulier, nous nous concentrerons sur l'apprentissage qui s'effectue dans un *but* (que nous assimilerons le plus souvent à l'idée de *renforcement*). Dans ce cas, il consiste à rendre efficaces, ou de plus en plus efficaces, nos interactions avec l'environnement pour atteindre ce but.

La compréhension des processus comportementaux et mécanismes neurophysiologiques de l'apprentissage nécessite de détailler ce qui est à leur origine, à savoir les problèmes rencontrés lorsqu'on doit apprendre de l'environnement. Ces problèmes sont multiples et de nature diverse, c'est pourquoi nous allons nous concentrer sur une de leur source majeure qui est que l'environnement est *incertain* quant aux moyens d'atteindre le but.

L'incertitude est une caractéristique importante de nombreuses décisions de la vie de tous les jours. Elle survient dans les situations où l'agent dispose d'une quantité limitée ou incalculable d'information sur les résultats de son comportement (Huettel et al., 2005). La capacité de détecter, analyser et résoudre l'incertitude de manière adéquate est cruciale pour réaliser un comportement flexible et adapté.

Il est possible de distinguer 3 sources d'incertitude dans l'environnement (Payzan-LeNestour and Bossaerts, 2011). Tout d'abord, la première source d'incertitude émerge du fait qu'on ne connaisse pas toujours, dès le début, les meilleures op-

tions pour atteindre le but fixé. Dans cette thèse, nous en réfèrerons sous le terme d'*ambiguïté*. Ensuite, l'environnement peut être changeant, que ce soit de manière progressive ou soudaine ; et de manière prédictive ou non. Nous en parlerons comme du caractère *volatil* de l'environnement. Enfin, l'environnement n'est pas toujours fiable à 100%. Nous faisons ici référence à son caractère *stochastique*, et l'incertitude qui en découle, le *risque*. Ces trois types d'incertitude sont importants à différencier, car ils sont provoqués par différentes propriétés de l'environnement et nécessitent des comportements d'adaptation différents. L'ambiguïté peut être réduite à zéro après apprentissage des différentes options et de leurs résultantes, et nécessite donc de mettre en place des stratégies pour les découvrir. Le risque représente l'incertitude inhérente à un environnement non déterministe, incertitude qui ne pourra jamais disparaitre, même après le meilleur des apprentissages. Il faut donc trouver des moyens d'estimer son niveau global et de prendre en compte ce niveau dans ses choix, afin de ne pas les changer à chaque évènement surprenant. Au contraire, la volatilité est l'incertitude provoquée par des changements radicaux de l'environnement, et nécessite un changement des stratégies en cours. Dans le monde réel, ces trois types d'incertitude sont très souvent entremêlés, ce qui conduit à des difficultés supplémentaires à affronter. En effet, il est plus difficile de faire des choix adaptés car un évènement surprenant peut être dû au fait que l'environnement est stochastique, ou qu'il vient de changer, ou que nous n'avons pas encore appris toutes les conséquences des options à disposition. Des stratégies particulières pour estimer si un évènement surprenant est dû à l'une ou l'autre sont donc nécessaires. Nous verrons que cela représente un défi particulièrement élevé car notre perception du niveau de l'un peut être influencée par notre perception du niveau de l'autre (Payzan-LeNestour and Bossaerts, 2011). Cela suggère que l'étude des processus de prise de décision en contexte incertain doit combiner les trois types d'incertitude afin de comprendre au mieux comment ces processus ont évolué jusqu'à leur état actuel. Ainsi, dans une première partie, nous détaillerons d'abord des paradigmes expérimentaux qui ont permis d'étudier les processus comportementaux et les mécanismes physiologiques développés pour faire face à ces 3 types d'incertitude lorsqu'ils sont séparés, puis nous nous pencherons sur les études qui se sont intéressées à ces processus et mécanismes lorsque les

3 types d'incertitude sont entremêlés.

Prendre des décisions efficaces et flexibles dans ce genre d'environnement complexe et dynamique n'est pas tâche aisée. Comme nous le détaillerons, cela prend du temps et de nombreuses erreurs. Toute situation nouvelle nécessiterait de repartir de zéro et de tout réapprendre mais, chez les primates en particulier, ce n'est pas le cas. Ils possèdent en effet la capacité notable de rendre ces processus d'apprentissage de plus en plus efficaces. Il s'agit de la capacité d'*apprendre à apprendre*. Dans une deuxième partie, nous montrerons que le processus d'apprendre à apprendre est différent du processus d'apprendre, qu'il permet la mise en place d'un comportement efficace, flexible et généralisable. Ensuite, nous détaillerons les régions cérébrales qui semblent soutenir ce processus. Nous réfléchirons aux raisons pour lesquelles ce processus est particulièrement pertinent dans les environnements où s'entremêlent les trois types d'incertitude évoqués précédemment.

Chapitre 1

Apprendre de l'environnement : Processus comportementaux et mécanismes neurophysiologiques

Dans cette première partie d'introduction, nous détaillerons les processus comportementaux utilisés pour faire face à l'environnement incertain. Nous les décrirons sous leur forme la plus aboutie ; c'est-à-dire quand ces processus ont été acquis et sont performants (même s'ils peuvent ne pas être optimaux, ou rendus moins performants à cause de l'influence de certains facteurs, comme la fatigue par exemple), car c'est la manière la plus simple de les étudier et de comprendre comment ils se manifestent. Les processus décrits dans cette première partie se déroulent à l'échelle d'un problème, ou d'un nombre restreint de problèmes. Cela s'oppose à ce que nous décrirons dans la 2$^{\text{ème}}$ partie de cette introduction, à savoir le développement de ces processus à l'échelle d'un "grand nombre de problèmes", au cours desquels ils sont perfectionnés et deviennent généralisables à d'autres types de problèmes *via* le processus d'apprendre à apprendre (Partie 2).

1.1 Apprendre d'un environnement incertain : Processus comportementaux

1.1.1 L'ambiguïté : on ne connait pas toujours les meilleures options dès le début

Faire un choix adapté peut être un processus complexe. En particulier, une des premières sources d'incertitude émerge lorsqu'on ignore quelles actions peuvent conduire au but désiré, et avec quelles probabilités. C'est ce qu'on appelle l'*ambiguïté*. L'ambiguïté est intrinsèquement liée à l'apprentissage. Nous commencerons donc cette partie par nous intéresser aux sous-processus permettant l'apprentissage.

S'il n'est pas certain que ces processus existent en tant que tel dans le cerveau, ou soient effectivement dissociés, les définir a le mérite de permettre la conception de protocoles expérimentaux cherchant à les identifier, d'un point de vue comportemental, mais aussi de trouver leurs corrélats neurophysiologiques.

Voici un déroulé de ce que pourraient être ces processus. La première étape d'un comportement dirigé vers un but consiste à se représenter les différentes alternatives possibles qui pourraient conduire au but. Ensuite, en comparant son état interne actuel (par exemple, son niveau de faim) à l'état futur prédit qu'il serait possible d'atteindre en choisissant chacune des alternatives, on se crée une première idée sur les valeurs respectives de ces alternatives. Ensuite, il est nécessaire d'évaluer les variables externes qui pourraient potentiellement influencer ces valeurs des différentes alternatives (par exemple, le risque de ne pas arriver au but désiré, les délais pour l'atteindre...). Une étape suivante peut aussi consister à déterminer la valeur des actions associées à chaque alternative pour atteindre le but, en considérant, par exemple, les coûts, comme l'effort.La sélection finale d'une alternative pourrait ainsi se réaliser via la comparaison des différentes valeurs. Enfin, le choix fait, et l'action réalisée, une étape d'évaluation du résultat obtenu est nécessaire. Si cette valeur ne correspond pas à celle qui était attendue, les choix futurs sont changés, permettant un comportement adaptatif. Nous allons détailler ces différentes étapes et les processus qui ont permis de les caractériser.

L'attribution des valeurs : théories de l'apprentissage

Afin de résoudre un problème, une première étape pour l'agent consiste à se représenter le panel des différentes options possibles selon son propre état interne et l'état de l'environnement, puis de tester ces options afin de leur attribuer une valeur. Le concept de valeur a une pertinence comportementale puisque la valeur serait ce qui induit la motivation afin de provoquer l'action. La valeur d'une option consiste en l'ensemble des renforcements qu'il est possible d'obtenir à long terme en choisissant cette option. Mais la valeur de chaque option est rarement donnée telle quelle à l'agent qui doit donc utiliser des mécanismes pour l'apprendre ou l'estimer. Pour commencer, nous allons particulièrement parler du rôle des valeurs positives, à savoir des récompenses, dans l'apprentissage des valeurs des options (même s'il est vrai que les valeurs négatives provoquent aussi l'apprentissage comme nous l'évoquerons plus loin). Historiquement, les théories d'apprentissage associatif de Thorndike et Pavlov ont montré que les récompenses provoquent l'apprentissage, provoquant des changements du comportement (Thorndike - 1911, Pavlov - 1927). Le travail de Pavlov sur le conditionnement dit "classique" (ou stimulus-renforcement) a permis de montrer qu'un comportement change en réponse à un stimulus prédictif d'une récompense (en l'occurrence, un comportement de salivation en réponse à un son de cloche annonçant de la nourriture). Ce processus peut être opposé au conditionnement dit "opérant" (ou stimulus-réponse) proposé par Thorndike pour qui les changements de comportement sont le résultat de leurs conséquences après une *action*. Thorndike a montré en effet qu'un animal peut apprendre un nouveau comportement (en l'occurrence, trouver comment s'échapper d'une boite dite "puzzle box"), non pas grâce à une soi-disant perspicacité, mais par essai-erreur et lorsque le nouveau comportement est récompensé. La "loi de l'effet" de Thorndike propose que l'apprentissage consiste en la formation d'associations entre des stimuli et des réponses lorsque ces réponses amènent à des récompenses. Skinner (et d'autres comportementalistes) ont développé ces idées de conditionnement opérant, où chaque stimulus serait porteur d'un poids ou force associative caractérisant son degré de prédictibilité pour le renforcement. Une découverte par Herrnstein a été de montrer que ce sont bien les renforcements qui dirigent l'action : en effet, plus ces représentations seraient fortes,

plus elles conduiraient à la réalisation de l'action (Herrnstein, 1961).

De nombreux protocoles expérimentaux permettent d'étudier les représentations des valeurs des différentes actions. Par exemple, dans les expériences de dévaluation de la récompense utilisées chez le rat (Balleine and Dickinson, 1998), la valeur de l'association entre une action et sa récompense est détériorée en fournissant à l'animal cette récompense en surabondance ou en l'associant à un malaise digestif (Adams, 1982). Il est ensuite possible d'avoir une idée de la force de ces représentations pour l'animal, en étudiant à quel point il adapte ses actions suite à la dévaluation. Cela permet aussid'étudier les régions cérébrales qui sont critiques pour ces représentations, leur maintien ou leur formation, en couplant ces manipulations comportementales à des lésions, ou de la pharmacologie (par ex : (Coutureau and Killcross, 2003)). Il est ensuite intéressant de tester si les associations établies ont une nature causale ou non. Les paradigmes de dégradation de la contingence, par exemple, permettent d'étudier les représentations causales qu'un agent établit entre des actions et leurs récompenses,en donnant à l'animal des récompenses qui ne résultent plus de son action. Chez le rat, par exemple, cela se manifeste par une diminution progressive du taux de réponse à un levier utilisé pour obtenir des récompenses, lorsque les récompenses sont données de manière inversée ou aléatoire par rapport à l'enclenchement du levier par l'animal (par ex : (Coutureau et al., 2012).

Ces expériences de dévaluation ont conduit à la distinction entre les comportements dits d'*habitude* et les comportements *dirigés vers un but* (Dickinson, 1985). Un conditionnement peut être soumis à extinction, c'est-à-dire, s'interrompre après que le stimulus ait été présenté de manière répétée sans le renforcement auquel il avait été associé. L'idée est que, après dévaluation, si le comportement est de type habitude, l'animal ne s'arrêtera pas d'effectuer l'action qui menait auparavant au renforcement. Dans ce cas, ce sont les associations entre les stimuli et les réponses qui dirigent les actions. Au contraire, lors d'un comportement de type dirigé vers un but, c'est la valeur du renforcement qui mène l'action, conduisant donc à l'extinction progressive des actions après dévaluation. Dans cette thèse, nous nous focalisons sur les comportements dirigés vers un but, mais il faut bien avoir à l'esprit que ce com-

portement peut alterner avec un comportement de type habitude, en particulier en fonction du type d'entrainement reçu par l'animal. Nous en parlerons plus en détail dans la 2ème partie de cette introduction.

Le rôle des feedback négatifs dans l'apprentissage

Les recherches des domaines des sciences cognitives et neurosciences se sont principalement construites autour de l'idée que seuls les *feedback* positifs ont un pouvoir motivant.. Cependant, des études ont montré que ce n'est pas toujours le cas. La nature du *feedback* semble avoir une influence différente sur la motivation à atteindre un but final en fonction du *degré d'engagement* de l'individu pour réaliser ce but. En effet, dans l'étude de Koo et Fishbach (Koo and Fishbach, 2008), les *feedback* positifs ("Vous avez accompli 50% du travail à faire") étaient efficaces pour motiver dans les situations où les individus étaient peu engagés pour atteindre le but, en permettant d'augmenter l'engagement. A contrario, les *feedback* négatifs ("Il vous reste 50% du travail à faire") étaient plus efficaces lorsque l'engagement de l'individu était grand, en permettant de signaler le décalage avec le but. Il existe également des formes d'apprentissage sans renforcement, qu'on appelle apprentissage latent (Tolman and Honzik, 1930). On peut présumer un rôle important des *feedback* négatifs dans ce genre d'apprentissage. Par exemple, dans l'expérience historique de Tolman et Honzik, des rats qui ont été entrainés à trouver leur chemin dans un labyrinthe sans renforcement à la clé ont de meilleures performances à partir du moment où les renforcements sont introduits que des rats contrôles qui ont appris avec des renforcements depuis le début. Ces résultats suggèrent que les rats du premier groupe ont appris la structure du labyrinthe malgré l'absence de renforcement. Une explication serait qu'ils ont appris des *feedback* négatifs, comme le fait de se retrouver face à un cul-de-sac. Nous reparlerons de cette étude dans la 2ème partie lorsque nous évoquerons le rôle de l'apprentissage latent pour l'apprentissage de la structure de l'environnement.

La sélection des actions

Les recherches sur le comportement de "*matching*" (correspondance) ont pour but d'expliquer les proportions des choix d'un sujet en fonction des stimuli et des

récompenses rencontrés. La loi de *matching* a été proposée par Herrnstein en 1961 (Herrnstein, 1961). En observant des pigeons, il a remarqué que leur fréquence de réponses aux différents stimuli était proportionnelle aux récompenses obtenues pour chaque stimulus. Cette loi a été généralisée pour expliquer le comportement d'un grand nombre d'espèces. Cependant, elle ne s'applique que dans certaines conditions, par exemple lorsque les réponses à réaliser pour obtenir les différements renforcements sont identiques (Prescott et al., 2007). Une autre loi a donc été proposée : la loi de "*matching des probabilités*", qui permet d'expliquer les comportements au cours desquels la distribution des réponses corrèle aux *probabilités* de récompense. L'équipe de Seth propose que ce comportement, parfois sous-optimal, est la conséquence de l'adaptation à un environnement compétitif (Prescott et al., 2007). L'idée proposée est qu'en raison de la compétition, un individu est contraint de distribuer ses réponses entre les différentes ressources, au lieu de se focaliser sur les ressources les plus riches (qui sont forcément plus investies par les compétiteurs). Les auteurs démontrent, en utilisant la modélisation, qu'en contexte sans compétition (agent seul), un comportement optimal pour exploiter les ressources est d'effectuer une sélection de type "tout ou rien" des ressources, alors qu'en présence de compétiteurs, il vaut mieux répartir ces réponses en fonction des probabilités. Différents modèles de sélection des actions ont été proposés, notamment par les neuroscientifiques (Cisek and Kalaska, 2010). Ils seront donc détaillés dans la partie sur les corrélats cérébraux.

Apprentissage par renforcement

De nombreux mécanismes computationnels ont été proposés pour formaliser les théories de conditionnement opérant. Une proposition serait que l'apprentissage de la valeur des options est basé sur une estimation de la différence entre ce qui avait été espéré, et ce qui a été obtenu.

L'apprentissage par renforcement (Sutton and Barto, 1960) propose une méthode par essai-erreur expliquant comment un agent apprend les valeurs des actions en interagissant avec l'environnement de façon à maximiser la quantité de récompense. Ce type de modèle permet bien de modéliser certains comportements d'apprentissage de valeurs d'actions chez les animaux et l'Homme (Khamassi et al.,

2013; Luksys et al., 2009; Payzan-LeNestour and Bossaerts, 2011; Samejima et al., 2005). Ces modèles d'apprentissage sont basés sur le concept de l'*erreur de prédiction*, qui mesure la divergence entre le résultat attendu et celui obtenu d'une action, et qui permet de mettre à jour les estimations de l'agent sur son environnement à chaque fois qu'il reçoit une nouvelle information (Rescorla and Wagner, 1972). Cette mise à jour des valeurs des actions est réalisée via la modulation de l'erreur de prédiction par un facteur, le *taux d'apprentissage*, indiquant à quelle vitesse une nouvelle information remplace une ancienne.

Selon la loi Rescorla, la valeur d'une action Vt menant au renforcement rt est mise à jour selon la formule $V_{t+1} = V_t + \alpha(r_t - V_t)$ dans laquelle α est le taux d'apprentissage, modulant la façon dont l'erreur de prédiction $r_t - V_t$ modifie la valeur de l'action V_t.

D'autres modèles ont ensuite émergé pour raffiner le modèle initial d'apprentissage par renforcement. Par exemple, les algorithmes d'apprentissage par différence temporelle ont permis d'intégrer le fait que la récompense est dévaluée au fur et à mesure du temps, grâce à un paramètre de *discount* temporel (Khamassi et al., 2013; Tanaka et al., 2004). Compris entre 0 et 1, lorsque ce paramètre est proche de 1, l'agent a un comportement orienté vers les récompenses à long terme, alors que si ce paramètre est proche de 0, l'agent se focalise sur les récompenses immédiates (Khamassi et al., 2013; Tanaka et al., 2004). Un rôle de la surprise, que nous détaillerons dans la partie sur la volatilité, a également été proposé comme moteur important de l'apprentissage (Courville et al., 2006; Pearce and Hall, 1980).

Une fois que les valeurs des actions sont mises à jour, la sélection des actions dépend d'une *balance* entre un comportement d'exploration et un comportement d'exploitation. La plupart du temps, il est pertinent de sélectionner l'action avec la plus forte valeur (exploitation), mais il peut être également nécessaire de sélectionner d'autres actions (exploration) afin de collecter possiblement de nouvelles informations. Savoir quand partir en exploration en particulier, n'est pas trivial : il faut décider d'abandonner un terrain connu pour un terrain inconnu dont on ignore s'il sera meilleur ou pas. Ces comportements sont regroupés sous le terme de *foraging*. Ce genre de situations constitue le défi quotidien de nombreuses espèces. En effet,

des comportements de *foraging* assez efficaces ont été trouvés même chez les invertébrés : du ver C.elegans (chez lequel on a trouvé les facteurs permettant l'alternance entre rester dans un patch nutritif et explorer l'environnement (Flavell et al., 2013)) à la drosophile (qui pond ses œufs en prenant en compte de manière très fine les coûts et bénéfices des milieux nutritifs proposés mettant en balance la concentration nutritive et les coûts de *foraging* pour les descendants (Yang et al., 2008)). Les algorithmes décrivant ces comportements sont très similaires entre les espèces, même s'ils ne sont probablement pas biologiquement encodés de la même façon (Pearson et al., 2014). Chez les mammifères, que ce soit chez les rongeurs ou les primates non-humains et humains, on retrouve les mêmes stratégies comportementales ainsi que les mêmes régions cérébrales impliquées dans la réalisation de ce comportement (Pearson et al., 2014).

Le processus de balance entre exploration et exploitation est particulièrement pertinent dans les environnements changeant, comme nous allons le voir dans la partie consacrée à la volatilité.

L'ambiguïté comme moteur de l'exploration ?

Plusieurs études se sont intéressées aux facteurs qui faisaient pencher la balance dans un sens ou dans l'autre. Par exemple, des études de modélisation ont proposé d'intégrer un paramètre appelé *taux d'exploration* aux modèles d'apprentissage par renforcement (Khamassi et al., 2013). Grâce à ce paramètre, même si les actions avec les plus grandes valeurs ont toujours une plus grande probabilité d'être sélectionnées, l'exploration est rendue possible en modulant les valeurs du taux d'exploration. D'autres auteurs ont par exemple développé un index (appelé *Index de Gittins*) permettant de calculer la stratégie optimale pour une tâche de bandit à plusieurs bras, impliquant un dilemme entre exploration et exploitation (Gittins, 1979). Cet index indique que le bras à explorer est celui offrant la plus grande quantité *future* de récompense qu'il est possible d'espérer. Cependant, cet index n'est pas très efficace dans les environnements très changeant. Une autre proposition a été de postuler que ce qui pousserait à l'exploration est l'incertitude liée à l'ambiguïté : une option dont les conséquences seraient incertaines "mériterait" d'être explorée. Pour modéliser cette idée, il a été proposé l'existence d'un "bonus d'exploration",

qui serait ajouté à la valeur d'une option incertaine, indiquant la nécessité d'apprendre (Daw et al., 2005; Kakade and Dayan, 2002). Cette hypothèse implique que les individus devraient partir en exploration plus souvent en contexte de forte ambiguïté. Pourtant, les études de neuroéconomie tendent à montrer le contraire : les individus sont très aversifs à l'ambiguïté. Par exemple, le paradoxe de Ellsberg montre qu'en situation de choix entre 2 options, les gens préfèrent celle dont la loi de probabilité est connue, par rapport à la situation ambiguë (Ellsberg, 1961). Payzan-Lenestour et collaborateurs ont utilisé une tâche de bandit à 6 bras où l'ambiguïté varie au cours du temps pour comprendre, entre autres, ce qui pousse les individus à l'exploration. Ils ont comparé deux modèles pour expliquer les choix des participants. Chaque modèle module la valeur de chaque option en fonction du degré d'ambiguïté mais dans des sens opposés : le premier modèle augmente cette valeur quand l'ambiguïté augmente (bonus d'exploration) alors que le deuxième modèle les diminue (pénalité correspondant à l'aversion à l'ambiguïté) (Payzan-LeNestour and Bossaerts, 2011). Le modèle intégrant le bonus d'exploration était le moins bon pour expliquer le comportement des sujets, argumentant contre l'idée de l'ambiguïté comme moteur de l'exploration. En fait, les auteurs ont montré par la suite que l'ambiguïté peut provoquer l'exploration mais dans les environnements *stables* (non volatils), où le sujet a vraiment intérêt à explorer puisque cela conduit à diminuer l'incertitude (par ex : (Cavanagh et al., 2011)) (Payzan-LeNestour and Bossaerts, 2012). Au contraire, dans les environnements où la volatilité est très élevée, comme dans l'étude avec le bandit à 6 bras, utiliser un fort taux d'apprentissage n'est pas très rentable puisque le sujet doit repartir de zéro régulièrement. Ainsi, ces résultats montrent que l'apprentissage n'est pas du tout le même selon que l'environnement est stable ou non, en particulier si, en supplément, l'environnement n'est pas fiable. La présence de volatilité et de stochasticité dans l'environnement implique en effet des mécanismes d'adaptation supplémentaires que nous allons maintenant détailler.

1.1.2 La stochasticité (et le risque) : l'environnement n'est pas toujours fiable

Dans les environnements non déterministes, un évènement donné ne mène pas forcément au même résultat dans 100% des cas. Cela réfère à la notion de *risque*, utilisée en économie, qui représente la variance dans la distribution des probabilités de la récompense (Knight, 1921). De manière triviale, le risque constitue l'incertitude quant au fait de gagner ou pas, dans une situation où l'agent connait les probabilités de gagner de chaque option (parce qu'elles lui ont été préalablement indiquées, ou que l'environnement est stable et qu'il a eu le temps de les découvrir) (Huettel et al., 2005). Ainsi, le risque dénote le degré d'incertitude inhérent à une distribution connue des probabilités (contrairement à l'ambiguïté référant à une distribution inconnue des probabilités).

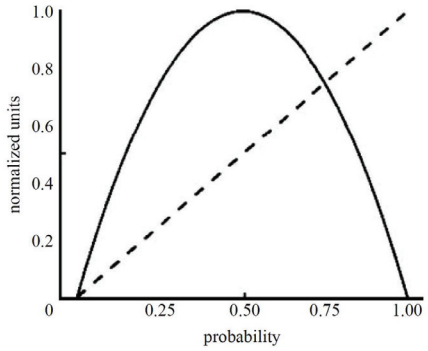

Figure 1.1 – Récompense attendue et risque en fonction des probabilités de récompense. La récompense attendue augmente de manière linéaire avec la probabilité de récompense p (ligne pointillée). La récompense attendue est minimale pour p=0 et maximale pour p=1. Le risque, mesuré comme la variance de la récompense, suit une fonction de probabilité en forme de U inversé, et est minimal pour p=0 et 1 et maximal pour p=0.5 (ligne pleine). D'après (Schultz et al., 2008).

Ainsi, le risque, défini comme la variance dans la distribution des probabilités de la récompense, suit une fonction des probabilités en U inversé (il est minimal à $p = 0$ et $p = 1$, et maximal à $p = 0.5$) (Figure 1.1). La valeur attendue de la récompense croit de manière linéaire avec les probabilités de récompense. Ainsi, des protocoles faisant varier les probabilités de récompense pour une magnitude fixe permettent de dissocier des corrélats neuronaux du risque de ceux liés à la valeur de la récompense.

Globalement, deux types d'approches s'intéressant à la prise de décision en condition de risque existent. La première approche est issue de l'économie et consiste à

étudier les décisions des individus lors d'un choix binaire, où toutes les informations sont disponibles (partie a). La deuxième approche est celle de domaines s'intéressant à ce comportement dans un contexte plus écologique comme le *foraging* où l'ambiguïté s'additionne à la stochasticité, puisque l'animal doit estimer lui-même le niveau de risque (partie b).

a) Prendre des décisions en contexte de risque "informé"

Evoluer en contexte de risque change la façon dont les personnes prennent des décisions binaires. Cela a beaucoup été étudié en économie, où, dans la majorité des cas, on demande au sujet de choisir entre deux options, dont les résultantes respectives sont décrites avec précision (par exemple, choisir entre "*gagner 4$ avec une probabilité de 80%, 0$ sinon" versus "gagner 3$ avec certitude*").

Une première idée a été que pour prendre une décision rationnelle dans ce genre de situation, il faut choisir l'option avec la plus grande valeur attendue (*expected value*, $EV = \sum p_i x_i$ où p_i et x_i sont respectivement la probabilité et la quantité de renforcement associés à chaque résultat possible (i) de l'option en question). Mais ce n'est pas ce que font la majorité des individus dans ce genre de situations. En effet, le paradoxe est que les sujets sont en fait prêts à donner relativement peu d'argent pour prendre un pari avec une très grande valeur attendue (revue (Weber and Johnson, 2009)). Ainsi, d'autres théories ont émergé par la suite. La théorie de l'utilité attendue (*expected utility*), originellement proposée par Bernoulli, puis reprise par Von Neuman et Morgenstern (1947), a démontré que les sujets choisissent les options avec la plus grande utilité plutôt que celles avec la plus grande valeur. Une augmentation d'argent de 0 à 1000 $ sera perçue comme plus grande qu'une augmentation de 2000 à 3000 $. Cependant, les modèles de l'utilité attendue ne sont toujours pas suffisants pour prédire parfaitement les comportements des sujets en situation d'incertitude, qui dévient systématiquement de 2 façons. Le premier facteur non pris en compte par ces modèles est le point de référence – expliquant que le même montant d'argent ne sera pas considéré de la même façon s'il constitue, par exemple, le meilleur ou le plus bas des prix à gagner. En effet, les individus semblent influencés par une variété de comparaisons relatives (Kahneman, 2003).

Il est maintenant bien défini que les individus comparent les résultats des options qu'ils ont choisies, avec ceux qu'ils auraient pu avoir en ayant sélectionné une autre option (Landman, 1997). Quand l'option non choisie se révèle être meilleure, l'agent expérimente le regret, qui a une très forte influence sur le comportement ultérieur. La théorie du regret proposée par (Loomes and Sugden, 1982) et (E and Bell, 1982) propose que les individus anticipent ce sentiment de regret et prennent leurs décisions de manière à maximiser l'utilité attendue et à minimiser le regret post-décisionnel. Enfin, la *Prospect Theory* de Kahneman et Tversky remplace la fonction d'utilité par une fonction de valeur qui est définie non en terme de résultats absolus, mais en terme de gains ou pertes relatifs (notamment en terme de changement par rapport à un point de référence) (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Une propriété de cette théorie est l'asymétrie entre les gains et les pertes (les individus ont une aversion au risque, plus ou moins forte, pour les situations concernant les gains, et une préférence pour le risque dans les situations de pertes) qui influence également beaucoup les choix. Toutes ces études suggèrent qu'il doit exister dans le cerveau des corrélats des mesures participant à cette estimation subjective des valeurs en situation de risque.

Ces biais dans la décision face au risque ne sont pas spécifiques aux humains.

Les macaques présentent la même influence du point de référence : ils n'ont pas de préférence absolue mais relative. Par exemple, Tinklepaugh a montré que des macaques n'acceptent pas une feuille de laitue s'ils s'attendent à une banane, alors qu'ils l'auraient acceptée si elle avait été présentée sans le point de référence constitué par la banane (Tinklepaugh, 1928). C'est un comportement qui s'explique très bien dans un cadre évolutif de *foraging* : un animal sera plus difficile sur la nourriture dans un environnement très riche. De même, des singes capucins vont préférer l'option pour laquelle on leur montre une récompense qu'ils reçoivent ensuite, par rapport à l'option où on leur montre deux récompenses mais n'en reçoivent qu'une (de la même taille que celle de la première option) (Chen et al., 2006). Les préférences des animaux face au risque varient. Dans une revue sur la sensibilité au risque chez 25 espèces, Bateson et Kacelnik ont trouvé que la plupart des espèces étaient soit aversives soit indifférentes au risque (Bateson and Kacelnik, 1998). Il semble que ce

comportement soit plus contrasté chez les primates : les études se contredisent chez le macaque (le présentant soit comme aversif soit comme recherchant le risque). Les tamarins, et bonobos éviteraient le risque, le marmoset y serait plutôt indifférent, et les chimpanzés le préféreraient (revue : (Stevens, 2010)).

Les chimpanzés pourraient être une des rares espèces à montrer une préférence pour le risque. Lorsqu'ils sont testés en laboratoire, ils seraient plutôt attirés par les choix à risque, au contraire des bonobos. Heilbronner et coll. proposent que cette différence a pour origine des différences de comportement dans leur habitat naturel (Heilbronner et al., 2008). En particulier, les bonobos se nourrissent principalement de plantes disponibles de manière abondante et sûre, contrairement aux chimpanzés qui préfèrent les fruits, une ressource plus variable en termes temporel et spatial. De plus, les chimpanzés sont plus exposés au risque car ils chassent des singes, contrairement aux bonobos, qui le font plus rarement. Les chimpanzés s'adonnent plus à la chasse, cette activité risquée, lors des périodes où les fruits sont les plus abondants (c'est-à-dire lorsqu'une partie au moins de l'alimentation est assurée). Ces données suggèrent que les différences de préférence face au risque entre les deux espèces ont peut-être été façonnées par la sélection naturelle en fonction de l'environnement.

Une des théories établies dans la littérature animale du *foraging* en situation de risque, la règle du budget énergétique (*Energy budget rule*, (Caraco, 1981)), similaire à la *Prospect theory* chez l'Homme, prédit une aversion au risque lorsque les animaux ne sont pas en danger de sous-alimentation (domaine des gains) et une préférence pour le risque dans le cas contraire (domaine des pertes). Selon cette règle, un animal devrait préférer les options de *foraging* avec une variance plus grande dans les situations où le gain énergétique espéré est plus haut que celui nécessaire à la survie ; et préférer les options avec une variance plus grande si le gain espéré est inférieur. La raison serait qu'une plus grande variance dans les résultats obtenus augmente les chances d'obtenir le gain calorique nécessaire à la survie. Cependant, les niveaux de sensibilité au risque, à la fois chez l'Homme et l'animal, dévient des prédictions de ces modèles (Weber and Johnson, 2009).

b) Prendre des décisions quand on doit estimer soi-même le risque

Les études historiques d'économie décrites précédemment ne se sont pas penchées sur la façon dont les gens prennent des décisions dans des contextes où ils doivent estimer eux-mêmes les résultantes des différentes options, via l'expérience. Dans ces situations, 2 types d'incertitude (ambiguïté et le risque) s'entremêlent.

Décisions à partir d'instructions vs décisions à partir de l'expérience

L'étude de Hertwig et collaborateurs (Hertwig et al., 2004) a montré que deux situations expérimentales, dans lesquelles les décisions sont prises à partir d'instructions ou à partir de l'expérience, ne conduisaient pas les sujets à faire les mêmes choix. La première situation implique des décisions élaborées à partir des informations obtenues par des instructions (externes) alors que la deuxième implique des décisions construites sur la base d'informations recueillies par l'expérience de l'environnement, par exemple par essai-erreur. Contrairement aux individus dans les expériences classiques d'économie (avec des instructions), les sujets prenant des décisions à partir de l'expérience ne surestiment pas les évènements avec une petite probabilité (les évènements rares), mais au contraire, semblent les sous-estimer. Dans l'exemple de pari que nous avons cité auparavant (*"gagner 4$ avec une probabilité de 80%, 0$ sinon"* versus *"gagner 3$ avec certitude"*), 20% des sujets choisissent la première option dans l'étude de Kahneman et Tversky (1979) qui proposait le pari de cette façon ; contre 63% des sujets dans une situation de décision par expérience (Barron and Erev, 2003). Deux facteurs ont été montrés comme pouvant jouer un rôle dans la prise de décision : le *mode d'information* (description symbolique versus expérience) et le *nombre de décisions* (unique ou répété). Hertwig et coll. proposent que la différence entre les deux études réside dans la façon dont les sujets apprennent la vraisemblance avec laquelle les évènements rares (et les autres) peuvent être attendus. Dans leur étude, les participants du groupe "décisions à partir de l'expérience" étaient autorisés à explorer les options jusqu'à ce qu'ils décident par eux même d'arrêter la recherche pour faire un choix. Un élément important pour l'estimation des évènements rares est donc le *temps passé en recherche* : en effet, une recherche rapide diminue la probabilité d'observer les évènements rares et

conduit donc à sous-estimer leur probabilité d'occurrence. Un autre effet entre aussi en jeu : l'effet de récence (*recency effect*), étant donné que les évènements rares ont une probabilité plus grande d'avoir été observés il y a plus longtemps, ils ont moins d'influence sur la décision.

Les décisions à partir d'instructions ou à partir de l'expérience ont également pu être comparées chez l'animal, en particulier le macaque (Heilbronner and Hayden, 2015). Dans ce cas, les instructions représentaient le niveau de risque de manière symbolique à l'aide de jauges, plus ou moins remplies. Les auteurs ont reporté le même biais que chez l'homme : les singes étaient moins aversifs au risque dans les cas de paris appris par expérience par rapport à ceux représentés visuellement à l'aide de jauges. Ces résultats renforcent le fait que le macaque est un bon modèle, puisqu'il possède les même biais que l'homme ; mais surtout, que les décisions à partir d'instruction ne sont pas forcément un bon modèle pour comprendre les processus permettant d'apprendre dans un environnement naturel stochastique.

Tâches probabilistes et foraging dans un monde probabiliste

Un meilleur modèle serait donc les tâches probabilistes, qui se rapprochent plus des situations naturelles de *foraging* où l'animal doit tester plusieurs fois une option afin d'en estimer le niveau de risque. De nombreuses études se sont penchées sur ces situations où le sujet doit prendre des décisions dans un contexte expérimental où il doit estimer lui-même le risque et/ou les autres sources d'incertitude, car ces informations ne lui sont pas données d'entrée de jeu. Les paradigmes utilisés pour créer un contexte expérimental de risque consistent à diminuer la prédictibilité des associations stimulus-réponses-résultat apprises en faisant varier leurs probabilités : de 100% pour un environnent complètement certain à 50% pour un environnement complètement incertain. Les animaux (du rat au singe) sont capables d'apprendre les probabilités de récompense associés à différents stimuli (Amiez et al., 2006; Dalton et al., 2014). Par exemple, dans une étude chez le macaque où des stimuli étaient prédictifs de différentes probabilités de récompense après un délai, Fiorillo et collaborateurs ont montré que le comportement de léchage (en anticipation de la récompense liquide) augmentait avec la probabilité de récompense (Fiorillo et al., 2003).

Les situations de *foraging* peuvent représenter des situations où il est nécessaire d'apprendre les probabilités des différentes options. En effet, la qualité des environnements peut varier, rendant le choix entre explorer et exploiter plus difficile. Les animaux prennent en comptent le niveau d'incertitude de l'environnement dans leur comportement de recherche de nourriture, ainsi que le degré de fiabilité des indices qui permettent de prédire l'environnement. Par exemple, McLinn et Stephens ont conçu une expérience avec des geais, où l'incertitude quant à la qualité des différents environnements proposés était manipulée (dans la condition de haute incertitude, la qualité était très variable) mais de manière prédictible grâce à un stimulus coloré (McLinn and Stephens, 2006). Dans les conditions de faible incertitude, les geais ignoraient les indices prédictifs de la qualité du patch (en allant explorer dans les patchs, quelle que soit leur couleur), alors que dans un environnement très incertain (où la qualité des patchs diffère beaucoup), ils prennaient beaucoup plus en compte ces indices.

Ces études montrent que les animaux et les êtres humains ont développé des stratégies comportementales pour faire face à l'incertitude liée au risque, et qu'ils sont soumis à des biais similaires. De manière intéressante, les études éthologiques montrent que l'environnement dans lequel l'individu se développe joue un rôle crucial dans les attitudes adoptées face au risque. Cela pointe de nouveau l'influence cruciale du degré d'incertitude de l'environnement d'apprentissage dans la mise en place des processus de décision pour réagir de manière appropriée à l'incertitude. L'implication de ces résultats dans ce travail de thèse a été que nous avons délibérément choisi d'entrainer nos singes dans un environnement de *foraging* probabiliste afin que les processus de décision face au risque, développés au cours de l'apprentissage, soient réellement liés à la structure de l'environnement.

Une autre vision du risque

Le concept de risque, tel que défini par les économistes comme la variance dans la distribution des probabilités, n'est pas le même que celui utilisé par les psychologues cliniciens. Les cliniciens qualifient un comportement de "risqué", lorsqu'il qui peut nuire à l'individu lui-même ou aux autres (l'utilisation de drogues, les rapports sexuels non protégés, l'alpinisme...), c'est-à-dire le risque comme la possibilité de

conséquences négatives. La critique des cliniciens concernant le concept économique de risque est qu'il n'est pas très performant pour prédire les différences individuelles dans les situations naturelles de prise de risque (Tom Schonberg et al., 2011). En particulier, ils considèrent la composante émotionnelle de la prise de risque très importante à prendre en compte, qu'elle soit négative (peur et anxiété) mais aussi positive (exaltation).

Deux protocoles ont permis de prédire avec succès les comportements de prise de risque naturalistes. Le développement de ce genre de protocoles est crucial car, si les tâches utilisées ne reflètent pas un comportement naturel, il y a de fortes chances que les processus et mécanismes mis en évidence ne reflètent pas complètement, ou de manière artefactuelle, les processus et mécanismes réels tels qu'ils existent dans le cerveau. Le premier protocole est le test de l'*Iowa Gambling Task* (Bechara et al., 1994). Dans ce test, deux "mauvaises" options offrent de grandes récompenses dans la plupart des essais mais lorsqu'on perd, les pertes sont aussi plus hautes. La valeur globale de ces options est donc plus faible par rapport à celle des deux autres options, dites "bonnes" options, qui offrent des récompenses plus petites mais sont associées à des coûts moins grands, résultant au final en un meilleur gain global. Les participants apprennent la nature des différentes options par essai-erreur. Plusieurs études ont montré que des patients avec diverses lésions conduisant à des comportements "risqués" dans la vie de tous les jours ont des déficits comportementaux dans cette tâche. Le problème de ce genre de tâches est qu'elles comportent plusieurs facteurs confondant qui rendent difficiles l'isolation de l'effet du risque. Par exemple, si les "mauvaises" options sont bien plus "risquées" dans le sens économique, l'augmentation de la variance est ici confondue avec une valeur attendue plus faible. De plus, la tâche requiert d'apprendre les valeurs attendues à long terme des différentes options, rendant impossible de déterminer si les différences comportementales individuelles dans cette tâche reflètent des différences dans l'apprentissage, l'attitude envers le risque ou la sensibilité à la magnitude gains/pertes. Un second paradigme est la tâche de *Balloon Analogue Risk taking*, au cours de laquelle le participant "pompe" pour faire gonfler un ballon simulé sans savoir à quel moment il va exploser (Lejuez et al., 2002). Chaque coup de pompe augmente la récompense

potentielle mais également la probabilité d'explosion. Dans la plupart des études, le participant ignore la distribution des probabilités d'explosion et doit l'apprendre par essai-erreur. Cependant, la tâche peut être également réalisée en fournissant au participant la probabilité d'explosion, ce qui permet d'éliminer le facteur confondant de l'ambiguïté.

Ainsi, les protocoles où les sujets doivent estimer eux même le risque possèdent cette faiblesse de l'existence d'un facteur confondant entre la décision en réaction au risque et l'apprentissage du risque, ce qui peut poser problème si l'on veut distinguer les influences spécifiques ou les corrélats neurophysiologiques de l'un ou l'autre de ces facteurs. Cependant, dans les environnements naturels, les deux sont mixés et au cours de l'évolution, les animaux ont dû développer des mécanismes pour gérer leur présence simultanée et leur interaction. Ce genre de protocole est donc pertinent parce qu'il modélise une situation dans lesquels les organismes ont évolué.

Dans cette section, nous avons vu que, même si l'environnement n'est pas fiable, les mécanismes d'apprentissage permettent d'estimer le niveau de stochasticité de l'environnement. C'est pourquoi l'incertitude provoquée par la stochasticité est qualifiée d'incertitude "attendue" : on peut apprendre à l'estimer, et à prendre des décisions adaptées en prenant en compte son niveau. Cependant, ceci n'est vrai que si l'environnement est stable. Or, les environnements naturels peuvent potentiellement être soumis à des changements, brusques ou progressifs. La présence de la volatilité crée ainsi une deuxième source d'incertitude qui peut nécessiter de relancer l'apprentissage à tout moment. Cette nouvelle source d'incertitude, dite "inattendue", car potentiellement non prévisible, implique des mécanismes permettant un haut niveau de flexibilité comportementale, que nous allons maintenant détailler.

1.1.3 La volatilité : l'environnement peut changer, progressivement ou brusquement

L'environnement peut changer, progressivement ou soudainement, et ce, de manière indiquée ou prédictible ou de manière imprédictible ; et de manière rare ou fréquente. Même dans les cas les plus simples, où les changements sont signalés par exemple, le fait que l'environnement puisse changer impose une certaine flexibilité

comportementale.

Concept de "Task-set"

Le concept de "*task-set*" a été proposé pour rendre compte de la flexibilité comportementale. Un *task-set* est le groupe de règles ou de stratégies (par exemple, tous les processus perceptuels, attentionnels, mnémoniques et moteurs) nécessaires pour réaliser une tâche (Sakai, 2008). Un *task-set* n'est pas spécifique du stimulus car il peut s'appliquer à tout stimulus, tant qu'il appartient au set des stimuli pertinents pour la tâche. Le concept de '*task-set*' est particulièrement pertinent pour expliquer les comportements dans les environnements volatils, car il propose que les sujets basculent d'une version du *task-set* à l'autre, lors des changements (Collins and Koechlin, 2012).

Pour qu'un *task-set* soit utile, il doit être maintenu en tant que *task-set* "acteur", tant qu'il reste pertinent pour le contexte et l'environnement. Ainsi, la compréhension d'un comportement flexible peut être pensée comme la compréhension de ce que cela signifie de "maintenir" un *task-set*, et de comment celui-ci est changé. Une fois qu'un *task-set* a été sélectionné, il doit être testé dans le contexte en cours, évalué en direct, et, s'il est insatisfaisant, remplacé par un autre plus pertinent. Ce *task-set* alternatif peut être un ancien *task-set*, extrait de la mémoire car on sait qu'il a fonctionné ultérieurement, ou, si aucun ancien *task-set* n'est satisfaisant, un nouveau *task-set* créé, et ajouté aux précédents (Collins and Koechlin, 2012).

a) Réagir au changement indiqué : paradigmes de *task-switching*

Les paradigmes de *task-switching* (alternance entre des tâches) ont permis l'étude des mécanismes déployés en réaction à un changement de tâche ou des règles en cours (Monsell, 2003). Par exemple, le sujet doit trouver le bon stimulus parmi deux qui lui sont présentés, et la règle est indiquée de manière alternative par la forme ou la couleur du stimulus (Buschman et al., 2012). Dans la plupart de ces études, l'environnement est déterministe et les changements sont signalés, ou s'ils sont non signalés, ils sont prédictibles (par exemple, un changement tous les x essais). Dans ce genre d'environnement, l'effet du *task-switching* peut être étudié en comparant les temps de réaction et les performances après un changement, en comparaison

à un essai de répétition. Cela peut-être informatif des mécanismes déployés dans les paradigmes de *task-switching*, tels que la gestion en direct des *task-sets* et leur reconfiguration après un changement.

La sélection d'un task-set

Lors d'un changement de tâche ou de règles, on peut considérer qu'il est nécessaire de sélectionner un nouveau *task-set*. Cette étape est consommatrice de temps, comme le montrent classiquement les augmentations de temps de réaction après un changement par rapport à un essai de répétition (Rogers and Monsell, 1995). Cette étape est également associée à un taux d'erreurs plus élevé (Rogers and Monsell, 1995). Ces altérations des performances après un changement sont appelées "coût de changement" (*switch cost*). Ces coûts peuvent être réduits de manière importante en laissant un temps de préparation plus long au sujet, ou en lui fournissant plus d'information sur la tâche à venir par exemple, ce qu'on appelle l'effet de préparation. Une interprétation que l'on pourrait donner concernant les coûts de changement serait qu'ils reflètent le temps occupé par la reconfiguration du *task-set*. Ceci pourrait regrouper divers processus, comme la redirection de l'attention vers les nouveaux attributs pertinents des stimuli, ou la redéfinition du but à atteindre, des règles d'action, des processus moteurs (Monsell, 2003)... Une question est de savoir si ce processus de reconfiguration de *task-set* est démarré de manière endogène ou exogène (Rogers and Monsell, 1995). La réduction des coûts de changement permettant d'adopter un *task-set* en avance, sans connaissance préalable du stimulus, suggère que le déclenchement de ce processus serait, en partie du moins, endogène. Cependant, de nombreux expérimentateurs ont tenté d'éliminer complètement les coûts de changement, en vain, diminuant jusqu'à une asymptote d'environ 600ms de préparation (même en laissant jusqu'à 5 min de préparation, des coûts résiduels subsistent) (Kimberg et al., 2000; Sohn et al., 2000). Cela suggère qu'une partie de la reconfiguration des *task-sets* ne peut se dérouler qu'une fois le stimulus affiché, indiquant donc un déclenchement de manière exogène. Et en effet, il existe des situations (comme dans les tâches de *Stroop* (MacLeod, 1991) ; ou des cas pathologiques de "comportement d'utilisation" (Lhermitte, 1983)), où le stimulus lui-même est capable d'activer ou d'évoquer le *task-set* habituellement associé à ce stimulus, sans

intention première de le faire.

La gestion des différents task-sets

Gérer les *task-sets* pendant l'exécution d'une tâche est également complexe, comme en témoignent les augmentations de temps de réaction quand des sujets doivent alterner entre des blocs d'essais d'une autre tâche par rapport à des blocs d'essais d'une même tâche (Rogers and Monsell, 1995). C'est ce que appelle les "coûts de mixing" et peuvent être interprétés comme le coût de maintenir plusieurs *task-sets* en mémoire (Monsell, 2003). De plus, cette difficulté est accrue par le phénomène d'interférence (Cohen et al., 1990). Les coûts d'interférence sont le résultat de la distraction produite par les propriétés non-pertinentes des stimuli (comme dans les paradigmes avec des essais congruents et incongruents). Ainsi, les coûts d'interférence seraient superposés aux coûts de changement, ce qui rend difficile leur différenciation. Il existerait d'ailleurs des différences inter-spécifiques concernant la manière de gérer ces deux phénomènes. Stoet et Snyder ont montré que, dans un paradigme de *task-switching*, des singes avaient relativement peu de coûts de changement en comparaison aux coûts d'interférence ; mais c'était l'inverse chez des sujets humains (Stoet and Snyder, 2004). Cette différence ne pouvait pas être entièrement due aux différences de temps de pratique puisqu'elle subsistait même après un entrainement intensif des sujets humains (Rogers and Monsell, 1995). Une interprétation possible serait que les singes, à défaut des humains, seraient biaisés pour la vitesse dans la balance vitesse-précision. Ces coûts de changement ont été montrés dans des paradigmes déterministes, où les sujets sont sûrs de la réponse à fournir. Dans les paradigmes non déterministes, dans lequel le sujet n'est pas sûr si un changement a réellement eu lieu, les coûts de changement devraient augmenter en longueur et complexité. Par exemple, l'absence d'indice indiquant le changement induirait un niveau d'inertie du *task-set* en cours, mesurable par des réponses persévératives ; et les réponses suivantes seraient caractérisées par une variété de réponses exploratoires et persévératives (Collins and Koechlin, 2012). La proportion de chaque donnerait des indications sur le degré avec lequel le sujet cherche et teste des nouveaux *task-sets*. Dans ce genre de paradigmes, il est possible que la différenciation endogène/exogène concernant la reconfiguration de *task-set* soit moins pertinente. En effet, pour obte-

nir des performances optimales après un changement de contingences, l'information exogène (telles que les erreurs, les indices...) ET les processus endogènes acquis au travers de l'expérience dans le paradigme doivent être combinés. Par exemple, des signaux d'erreur observés de manière aléatoire peuvent être considérés comme du bruit, alors qu'une série de signaux d'erreur peut être interprétée comme le signe d'un changement. Ainsi, des processus endogènes sont requis pour agir sur les signaux d'erreurs exogènes afin de différencier les deux situations. Ce point crucial sera développé dans la partie suivante, dans laquelle nous détaillerons les processus utilisés pour répondre à ce genre de défi.

b) Détecter un changement non signalé

Dans un environnement changeant mais non déterministe, un des défis consiste à détecter le changement. En effet, il peut ne pas être indiqué, ou non-évident si l'environnement est probabiliste. Courville et collaborateurs proposent un rôle de la surprise dans la détection du changement, permettant l'apprentissage (Courville et al., 2006).

Le modèle de Pearce-Hall (Pearce and Hall, 1980) reprend les idées de l'apprentissage associatif mais explique en supplément comment les évènements surprenants permettent un apprentissage plus rapide. Selon ce modèle, les forces associatives sont mises à jours plus ou moins rapidement selon un attribut appelé *associabilité*. L'associabilité est modulée par la surprise (la différence entre le renforcement précédent et ce qui avait été prédit pour le stimulus considéré) car elle dépend de la précision avec laquelle la force associative du stimulus a permis de prédire le renforcement.

Courville et collaborateurs proposent une explication de la raison pour laquelle les évènements surprenants provoquent un apprentissage plus rapide en utilisant un cadre Bayésien (Courville et al., 2006). Les approches Bayésiennes interprètent les réponses de l'animal comme émergeant d'un raisonnement statistique sur la probabilité d'obtenir un renforcement étant donné l'historique des renforcements passés. Selon les modèles (discriminant vs génératif) les agents apprennent des probabilités de l'environnement en se basant respectivement 1) sur les stimuli conditionnés seulement ou 2) sur l'ensemble des stimuli conditionnés (stimuli neutres) et incondi-

tionnés (renforcements). Plusieurs expériences chez l'animal cherchent à mettre en évidence quels types d'évènements surprenants (renforcements ou stimulus neutres) provoquent un apprentissage, afin de faire pencher la balance en faveur de l'un ou l'autre de ces modèles.

Tout d'abord, des expériences sur l'inhibition latente permettent de montrer que les renforcements surprenants accélèrent l'apprentissage. L'inhibition latente est le fait que l'acquisition de l'association entre un stimulus et un renforcement est ralentie si ce stimulus a été préalablement présenté de manière répétitive sans renforcement. Dans une expérience réalisée chez le rat (Hall and Pearce, 1979), l'acquisition stimulus-renforcement est aussi retardée si le stimulus est préalablement pré-exposé à un renforcement faible (comme un faible choc) avant d'être associé à un renforcement plus grand (choc plus fort). Mais de manière intéressante, si la préexposition consiste en 2 phases : d'abord une présentation du stimulus avec un faible renforcement (faible choc), puis une présentation de ce stimulus en absence de renforcement (inhibition latente), l'acquisition de l'association du stimulus avec le renforcement plus fort (choc fort) est accélérée. Une interprétation est que la surprise causée par le changement entre la préexposition avec le faible renforcement à la préexposition sans renforcement serait à l'origine du l'apprentissage accéléré.

Une façon de trancher entre les modèles discriminant vs génératif est de regarder si des stimuli neutres surprenants accélèrent aussi l'apprentissage. C'est ce que Blaisdell et collaborateurs ont montré grâce à une expérience d' "*overshadowing*", faisant pencher la balance en faveur du modèle génératif (Blaisdell et al., 1998). L'*overshadowing* consiste en l'introduction d'un nouveau stimulus B pendant la phase de renforcement du stimulus A. Cette inclusion d'un stimulus neutre retarde l'apprentissage. Mais lorsque cette phase d'*overshadowing* succède à une phase d'inhibition latente, le déficit d'apprentissage est réduit. La surprise créée par l'inclusion d'un stimulus neutre B signalerait une augmentation du taux de changement, facilitant l'acquisition et contrant l'inhibition latente.

Le rôle de la surprise dans l'apprentissage, et notamment dans la modulation du taux d'apprentissage a été formalisé dans différents modèles, comme, par exemple, le modèle *delta rule* Bayésien proposé par Nassar et collaborateurs (Nassar et al.,

2010). Dans cette étude, les auteurs montrent que l'influence d'un *feedback* dépend à la fois de l'erreur faite lors de la prédiction de ce *feedback* mais également du nombre de *feedback* similaires rencontrés auparavant. Le modèle *delta rule* décrit l'adaptation mise en place par les sujets dans les environnements dynamiques où le passé ne prédit pas toujours le futur et dans lequel les croyances doivent parfois s'adapter rapidement, en particulier après un *feedback* inattendu. Dans ce papier, les auteurs montrent que les sujets reconnaissent les points de changement, et que leur taux d'apprentissage varie en essai par essai en fonction des *feedback* surprenants, et que sa magnitude dépend du degré de confiance.

Les *renforcements stochastiques* constituent également un autre type de renforcement surprenant. Mais la surprise surviendrait seulement en cas d'évidence de changement des probabilités. Des variations aléatoires liées à des probabilités constantes ne seraient pas surprenantes et ne provoqueraient donc pas d'apprentissage accéléré. Cela a été montré notamment via des expériences d'extinction par renforcement partiel. Gallistel et Gibbon ont testé la vitesse d'extinction à des stimuli préalablement partiellement renforcés vs complétement renforcés (Gallistel and Gibbon, 2000). Lors de l'extinction, l'absence de renforcement est moins surprenante dans les cas où les stimuli avaient été partiellement renforcés. Cela montre les effets des différents types d'incertitude sur l'apprentissage : la volatilité provoque une accélération de l'apprentissage contrairement à la stochasticité, si elle est stable. Cela suggère que les différentes formes d'incertitude ont différentes conséquences sur l'apprentissage. Il est donc intéressant de regarder la façon dont les organismes réagissent lorsque les différentes sources d'incertitude sont présentes dans l'environnement.

Ces études montrent aussi l'importance du contexte d'apprentissage, et notamment de sa structure statistique, sur les stratégies d'apprentissage utilisées par la suite. Ainsi, le fait que le modèle interne de l'animal prenne en compte que le monde puisse changer ou non est crucial pour l'apprentissage. Cet aspect va nous intéresser particulièrement dans la deuxième partie (Partie II) de cette introduction, et pour les résultats présentés dans cette thèse.

1.1.4 L'environnement peut être à la fois stochastique, volatil et ambigu !

Dans le monde réel, l'environnement combine souvent les 3 types d'incertitude. Ainsi, un changement dû à la volatilité peut consister en un changement du niveau de stochasticité. Dans ces situations, il faut savoir estimer le niveau de volatilité afin de s'adapter de manière adéquate aux changements de stochasticité qui en résultent (partie a) et être capable de gérer tous les types d'incertitude à la fois (partie b). Enfin, face à ces changements, un comportement flexible nécessite d'avoir la capacité de proposer d'autres stratégies (partie c).

a) Savoir estimer la probabilité d'un changement : mécanismes pour estimer le niveau de volatilité de l'environnement et s'adapter aux changements de volatilité.

Savoir détecter un changement et réagir en conséquent est crucial pour la flexibilité comportementale. Cependant, dans les environnements hautement dynamiques, la surprise déclenchée par un *feedback* inattendu peut être provoquée par la stochasticité ou par la volatilité. Estimer le niveau relatif de chaque source d'incertitude est alors crucial pour réagir de manière appropriée.

Les théories de statistiques Bayésiennes formalisent l'idée qu'une inférence optimale et l'apprentissage dépendent de manière critique de la façon de représenter et intégrer les diverses sources d'incertitude associées à un contexte comportemental (Yu and Dayan, 2005). Dans le cas de la volatilité, il a été suggéré que les êtres humains s'adaptent à un environnement décisionnel volatil en suivant des règles Bayésiennes (Behrens et al., 2007; Nassar et al., 2010), en modulant les différents éléments des modèles d'apprentissage que nous avons évoqués précédemment. En effet, dans les modèles d'apprentissage par renforcement, comme nous l'avons vu, les valeurs des options sont mises à jour de manière proportionnelle à l'erreur de prédiction qui est la différence entre le résultat attendu et obtenu. L'erreur de prédiction doit être multipliée par un facteur, le taux d'apprentissage, pour déterminer à quel degré la valeur de l'action est mise à jour. Le taux d'apprentissage reflète

donc le taux auquel toute nouvelle information remplace une ancienne. Les modèles Bayésiens d'apprentissage proposent des stratégies formelles pour mettre à jour de manière optimale les croyances sur l'environnement chaque fois qu'une nouvelle information est observée. Dans ces modèles, il est suggéré que le taux d'apprentissage doit dépendre du niveau d'incertitude dans les estimations de la valeur de l'action (Behrens et al., 2007). L'incertitude, quant à elle, est déterminée selon les statistiques des renforcements. Ainsi, quand l'expérience récente est plus prédictive du future que l'expérience plus lointaine, le taux d'apprentissage doit être grand. C'est le cas notamment dans les environnements volatils. A contrario, dans les situations de stabilité, où l'historique est la dimension pertinente, l'agent doit considérer ses expériences sur une longue période de temps, ce qui correspond à une petite valeur du taux d'apprentissage.

L'agent devrait en fait paramétrer son taux d'apprentissage de telle sorte à maximiser sa capacité à prédire les futurs résultats, ce qui est le but du processus d'apprentissage. Les données expérimentales allant dans ce sens proviennent notamment d'études chez le singe et le rat (Gallistel et al., 2001; Kennerley et al., 2006; Sugrue et al., 2004). Dans ces études, la capacité de détecter les changements dans les taux de la récompense dépend du niveau de changement de son historique d'apprentissage. Une étude de Behrens et collaborateurs a eu pour but de manipuler directement le taux d'apprentissage des sujets en manipulant la volatilité de l'environnement (Behrens et al., 2007). Dans cette étude, les sujets exécutent une tâche de bandit à un bras, dans laquelle ils doivent choisir entre des stimuli bleus et verts. L'expérience consiste d'abord en une période où les probabilités liées au stimulus bleu sont à 75% (c'est-à-dire, un environnement certain car stable), puis une période où les probabilités de récompenses alternent entre 80% pour le bleu, et 80% pour le vert, tous les 30-40 essais (créant ainsi un environnement incertain car volatil). Les auteurs ont estimé le taux d'apprentissage de leurs sujets en utilisant un modèle d'apprentissage *delta-rule* (cette règle permet d'estimer les probabilités de l'essai suivant en utilisant les probabilités prédites et l'erreur de prédiction de l'essai en cours, qui sont modulés par le taux d'apprentissage). Les sujets étaient plus réactifs à toute nouvelle information dans la phase de l'expérience où la volatilité était forte

par rapport à la phase stable, montrant bien une adaptation de leur taux d'apprentissage en fonction du niveau de volatilité de l'environnement. En particulier, cette étude démontre que les participants utilisent un modèle Bayésien pour estimer la volatilité de manière optimale et ajuster leurs décisions en fonction, pour obtenir les résultats les plus avantageux.

A la différence du modèle de Behrens, le modèle de Gallistel n'utilise pas le taux d'apprentissage mais "apprend" une nouvelle estimation seulement quand il perçoit une divergence entre l'estimation en cours et l'expérience récente (Gallistel et al., 2014). Cette nouvelle estimation n'est basée que sur l'historique d'une période qu'il considère stationnaire (la période depuis le dernier changement détecté). Ainsi, très souvent, la nouvelle estimation est basée sur un set complètement différent de la précédente. Ce modèle utilise la détection des erreurs de prédiction mais ne l'utilise pas via une règle *delta-rule* : il initie à la place un processus de sélection Bayésien qui décide de ce qui a causé l'erreur. L'estimation qui s'en suit dépend entièrement de cette décision.

Ainsi, plusieurs modèles ont été proposés pour expliquer comment les individus et les animaux étaient capables d'estimer la probabilité d'un changement. Une façon de trancher entre ces modèles pourrait peut-être de regarder comment ces processus de détection des changements se mettent en place, en étudiant, depuis le début, le développement des réponses à l'incertitude au cours de l'apprentissage d'une tâche, plutôt que lorsque la tâche est déjà apprise et maitrisée.

b) Evaluer tous les types à la fois

Quand l'environnement combine tous les types d'incertitude, des mécanismes pour les dissocier deviennent nécessaires, afin de déterminer si un résultat inattendu est juste le fait de la stochasticité de l'environnement ou la conséquence d'un changement fondamental de l'environnement.

Nassar et collaborateurs proposent un mécanisme pour résoudre ce problème en utilisant, une fois encore, un modèle Bayésien (Nassar et al., 2010). Ce modèle permet d'ajuster l'influence des nouveaux résultats observés en fonction de l'estimation en cours de l'incertitude et des probabilités d'un changement fondamental dans le

processus via lequel ces résultats sont générés. Ainsi, les résultats qui sont inattendus à cause d'un changement fondamental de l'environnement ont plus d'influence que ceux qui sont inattendus à cause de la stochasticité environnementale persistante.

Payzan-Lenestour et Bossaerts proposent également un modèle Bayésien permettant le suivi simultané des 3 types d'incertitude dans une tâche de bandit à 6 bras : ambiguïté (ou incertitude d'estimation), risque (incertitude attendue, liée à l'environnement probabiliste) et incertitude inattendue (volatilité avec des changements rares) (Payzan-LeNestour and Bossaerts, 2011). Les auteurs calculent les estimations des différents types d'incertitude de la façon suivante. Le risque peut être mesuré via l'entropie des probabilités des résultats (sous-entendu qu'un grand niveau de risque correspond à des probabilités des résultats très variables). L'ambiguïté reflète la dispersion de la distribution postérieure des probabilités des résultats. Ceci peut être estimé comme la variance de la distribution postérieure, ou son entropie. L'incertitude inattendue est la vraisemblance que les probabilités des résultats changent d'un coup. Le niveau de volatilité évolue au cours du temps, en fonction des fluctuations dans l'évidence des sauts. Grâce à ces estimations, les auteurs ont étudié comment les différents niveaux d'incertitude affectent le taux d'apprentissage, et comment la perception de chacune influence les autres. Un résultat important pour ce travail de thèse est que l'estimation du niveau de risque par les sujets est moins bonne pour les bras avec une grande probabilité de changement (grande volatilité). Ainsi, *lorsque la volatilité est trop forte, les sujets surestiment la stochasticité*. Cela montre l'interaction antagoniste entre les perceptions de ces 2 types d'incertitude.

En fait, les auteurs démontrent que les perceptions des niveaux respectifs des trois types d'incertitude s'influencent fortement les unes les autres, en jouant sur le taux d'apprentissage ; et ils proposent un rôle pivot de la volatilité. Leurs conclusions peuvent être résumées ainsi. Tout d'abord, la volatilité affecte l'ambigüité et par conséquent, le taux d'apprentissage. En effet, quand la volatilité est faible, il n'est plus nécessaire d'apprendre (diminution de l'ambiguïté), ainsi, le taux d'apprentissage diminue. A contrario, quand la volatilité est forte, l'ambiguïté est forte, ce qui induit une augmentation du taux d'apprentissage. Deuxièmement, la perception de la volatilité et la perception du risque s'influencent aussi l'une et l'autre.

En effet, si un résultat particulier est attendu avec une faible probabilité et qu'il se réalise, alors il est plus probable qu'il se soit réalisé à cause d'un saut dans les contingences. A contrario, si un résultat a une haute probabilité *a priori*, alors le fait qu'il se réalise sera vraisemblablement moins attribué à un saut dans les contingences. De cette façon, les probabilités des résultats, parce qu'elles permettent d'évaluer la probabilité de changement, ont également un effet sur le taux d'apprentissage. Ces résultats montrent que les 3 niveaux d'incertitude influencent le taux d'apprentissage via le rôle central de la volatilité. Cela montre l'importance de mêler tous les types d'incertitude dans un protocole expérimental pour étudier la façon dont on prend des décisions en conditions réelles.

Un autre résultat intéressant est que la façon dont les sujets réalisent cette tâche dépend en fait des informations fournies *a priori* aux sujets, notamment par les consignes. En effet, dans une expérience supplémentaire, les auteurs n'ont pas précisé aux participants que l'environnement était sujet à changement (pas d'information donnée sur la possibilité que l'environnement soit volatil). Dans cette situation, les participants ont été incapables de détecter lorsque les probabilités effectuaient un saut et leur comportement était alors mieux modélisé par un modèle d'apprentissage par renforcement que par un modèle Bayésien. Les résultats des questionnaires suggèrent que les participants attribuaient une séquence de mauvais résultats à une mauvaise période de risque, plutôt qu'à un changement des probabilités. Leur représentation de l'environnement excluait de fait la non-stationnarité. Ce résultat est important dans le cadre de cette thèse. En effet, il montre que l'information *a priori* sur l'environnement détermine la façon dont les décisions vont être prises. Cela nous oriente vers la deuxième partie (Partie II) où nous discuterons de l'impact de ce qu'un individu a appris sur son environnement sur la façon dont il prend des décisions dans cet environnement.

c) Avoir d'autres stratégies à proposer quand il y a un changement

Réagir de manière flexible à un changement de règles nécessite d'être capable de détecter si la stratégie utilisée n'est plus valide mais aussi de proposer d'autres stratégies le cas échéant. Plusieurs modèles ont pour vocation de proposer des mé-

canismes pour réaliser ces différents processus.

Les modèles d'apprentissage par renforcement utilisant le concept de *task-set* postulent que le *task-set* permettant le comportement en cours (appelé le *task-set acteur*) est ajusté en fonction des résultats de l'action, en particulier quand ceux-ci sont récompensés (Sutton and Barto, 1960)). Les modèles de suivi de l'incertitude (*Uncertainty Monitoring (UM) model*) complètent le modèle précédent en rajoutant l'idée qu'il est nécessaire de tester la fiabilité du *task-set* en cours, afin de le changer s'il cesse d'être optimal (Yu and Dayan, 2005). Les modèles combinant l'apprentissage par renforcement et le suivi de l'incertitude postulent le suivi d'un nombre fixe de *task-sets* en compétition, et que la sélection du plus pertinent s'effectue en comparant leur fiabilité respective (Samejima and Doya, 2007).

Collins et Koechlin ont créé le modèle PROBE qui prend en compte l'idée, très plausible, qu'aucun des *task-sets* du groupe n'est satisfaisant, nécessitant la création d'un nouveau *task-set* (Collins and Koechlin, 2012). Ainsi, le modèle PROBE contrôle la sélection et l'ajustement du *task-set* en cours et permute ce *task-set* avec un autre plus pertinent en le piochant dans la mémoire à long terme, permettant de créer et de tester de nouveaux *task-sets* quand il est nécessaire d'ajuster les actions. D'après ce modèle, un *task-set* consiste en un *mapping sélectif* encodant l'association entre le stimulus et la réponse, un *mapping prédictif* encodant les résultats attendus des actions étant donné les stimuli, et un *mapping contextuel* encodant les indices externes prédisant la fiabilité du *task-set*.

Collins et Koechlin présupposent qu'un nombre limité de *task-sets* est suivi en parallèle et que chacun est comparé au *task-set* acteur. Dans leur modèle, de nouveaux *task-sets* sont créés si aucun des *task-sets* du réservoir n'est assez bon. En utilisant l'inférence Bayésienne, la fiabilité de N *task-sets* est suivie et estimée à deux moments : avant et après l'action. La fiabilité est testée en direct avant l'action en utilisant l'information amenée par le contexte et le degré de volatilité de l'environnement perçu par les sujets (grâce au *mapping contextuel*), permettant le choix d'actions (grâce au *mapping sélectif*). La fiabilité est ensuite testée après l'action en évaluant les résultats obtenus. Cela permet de mettre à jour la valeur du *task-set* en cours via 3 processus : l'ajustement du *mapping sélectif* acteur via l'apprentis-

sage par renforcement standard ; la mise à jour de la prédiction des résultats par le *mapping prédictif* ; et la mise à jour de la fiabilité du *task-set* via l'ajustement du *mapping contextuel*. La sélection des *task-sets* est basée sur l'estimation que le *task-set* en cours est plus fiable que non-fiable. Quand deux *task-sets* sont suivis et évalués, si un des deux remplit cette condition, il est automatiquement sélectionné, alors que si aucun des deux ne remplit cette condition, un nouveau *task-set* est créé et testé pour voir s'il remplit les conditions d'un bon acteur. Un nouveau *task-set* inclut à la fois le *mapping sélectif* et le *mapping prédictif*. Ils sont créés à partir d'une mixture des tous les *mappings* internes stockés dans la mémoire à long terme, de manière biaisée par les informations contextuelles de manière à ce que ce qui est retiré corresponde à l'information en cours dans la tâche. Ce modèle permet de reproduire le comportement humain de manière très précise, lorsqu'il est supposé que 3 ou 4 *task-sets* sont suivis en même temps. De manière intéressante, cette étude a aussi montré des différences inter-individuelles dans la façon d'utiliser les *task-sets*. Pour deux tiers des sujets, dits "exploiteurs", la répétition de *task-sets* déjà rencontrés conduisaient à de meilleures performances, suggérant qu'ils réutilisaient les *task-sets* déjà appris. Au contraire, pour le tiers des sujets restant, dits "explorateurs", la répétition des *task-sets* n'était pas reliée à une amélioration des performances. Les performances des explorateurs étaient en fait meilleures que les exploiteurs dans les conditions où aucun *task-set* n'était répété. Cela montre qu'il n'existe pas de "meilleure" stratégie, car l'une et l'autre sont adaptées à des contextes différents. Un point intéressant serait de comprendre l'origine de ces différences inter-individuelles. Une hypothèse serait qu'elles seraient liées aux a-priori des sujets sur la variabilité de l'environnement. Dans ce travail de thèse, nous proposerons que cet a-priori est influencé par le degré de variabilité de l'environnement dans lequel on apprend à apprendre.

Dans une étude suivante, ce modèle a été utilisé pour lier l'activité en IRMf de certaines régions cérébrales avec l'évolution en direct de la fiabilité du *task-set* acteur, mais également de la fiabilité des autres stratégies en concurrence avec l'acteur ((Donoso et al., 2014), résultats détaillés dans la partie consacrée aux corrélats neurophysiologiques).

1.2 Corrélats neurophysiologiques

Dans cette partie, nous n'allons pas détailler les différentes structures cérébrales une à une en décrivant leur rôle respectif, mais nous allons reprendre les processus comportementaux évoqués dans la partie 1.1, et essayer de voir à quelles structures ils ont globalement été rattachés. Nous commencerons par évoquer les corrélats neurophysiologiques des processus permettant la prise de décision dirigée vers un but et permettant de réduire l'ambiguïté *via* l'apprentissage, avant d'aborder les régions et mécanismes cérébraux qui entrent en jeu lorsque l'environnement devient, en plus, stochastique et/ou volatil. L'idée est de justifier pourquoi nous avons décidé, dans ce projet de thèse, d'enregistrer le signal en provenance du cortex frontal, en incluant le cortex cingulaire médian, mais aussi le cortex pariétal.

Mais avant de commencer, une petite mise au point concernant les dénominations des régions cérébrales, qui varient beaucoup selon les époques et les domaines, est nécessaire. Toute région corticale citée dans cette partie (en particulier pour les cortex préfrontal latéral, médian et orbital) sera nommée à partir des délimitations cyto-architectoniques des aires de Petrides et Pandya (reprises dans (Petrides et al., 2012)), auquel le lecteur pourra se référer dans la figure 1.2. Dans cette partie, nous évoquerons également le rôle crucial des ganglions de la base, constitué d'un réseau complexe de structures, connectées entre elles et avec les régions du cortex préfrontal, comme illustré dans la figure 1.3.

1.2.1 Mécanismes communs à la prise de décision basée sur la valeur

Comme nous l'avons vu dans la partie comportement, les comportements dirigés vers un but sont orientés en fonction de différents calculs de la valeur estimée des différentes options. Ainsi, la sélection d'action basée sur la récompense commence par le calcul de la valeur des options, des actions et des stimuli, représentée de manière hétérogène dans le cerveau. La représentation des valeurs dans le cerveau est un point intéressant car cela nécessite de réfléchir à ce qui est ou doit être représenté exactement : la valeur de l'action choisie, la valeur globale de toutes les actions, la

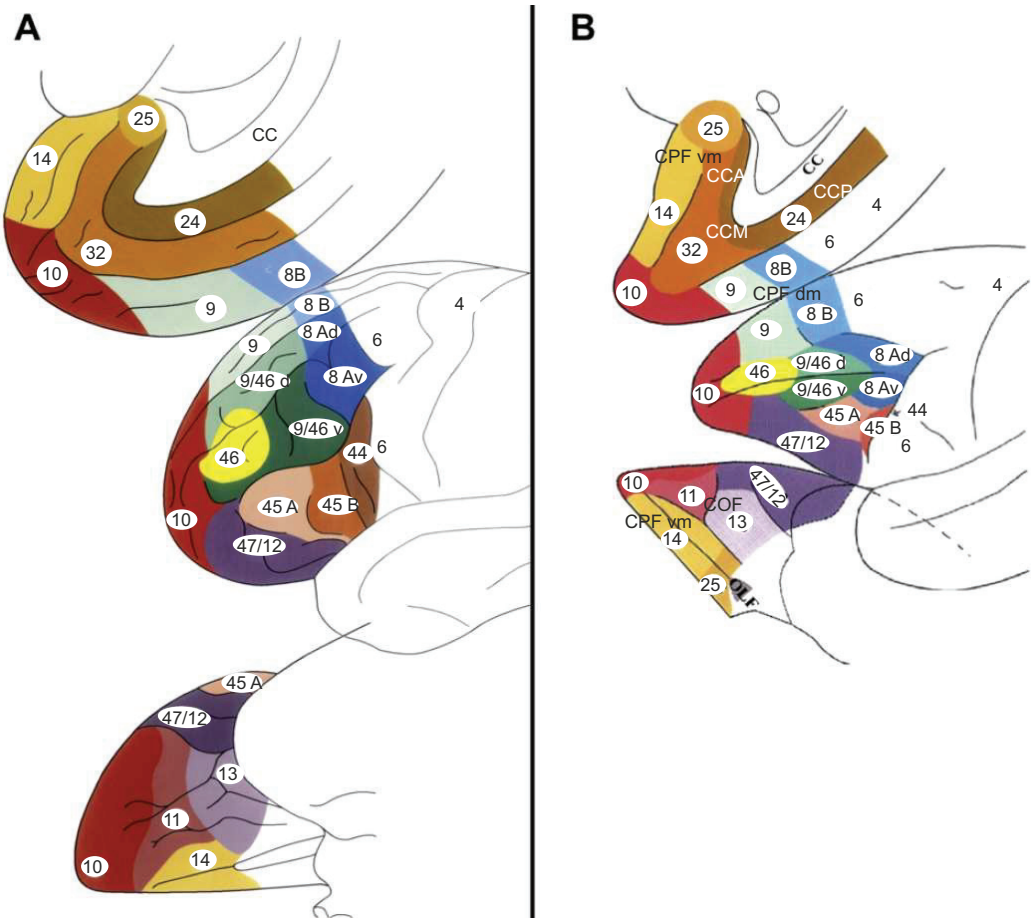

Figure 1.2 – Cartes cyto-architectoniques de la surface des lobes frontaux latéral (haut), médian (milieu) et orbital (bas) chez le cerveau de l'Homme (A) et du singe macaque (B), d'après les délimitations de (Petrides et al., 2012). Abréviations : CPF : cortex préfrontal (dm : dorso-médian ; vm : ventro-médian), CCX : cortex cingulaire (avec X= A : antérieur, M : médian, et P : postérieur), COF : cortex orbito-frontal et CC : corps calleux. Le cortex préfrontal est représenté par les aires colorées (excluant donc les aires 6 et 4). Le CCM a historiquement été appelé CCA puis CCA dorsal, mais nous ferons référence à cette région sous le terme CCM (Procyk et al., 2014). Nous appellerons CPF dorso-médian (CPFdm) le regroupement des aires 9 et 8, et CPF dorso-latéral (CFP dl), le regroupement des aires 46 et 8. Le COF fait référence aux aires 11 et 13, et le CPFvm à l'aire 14.

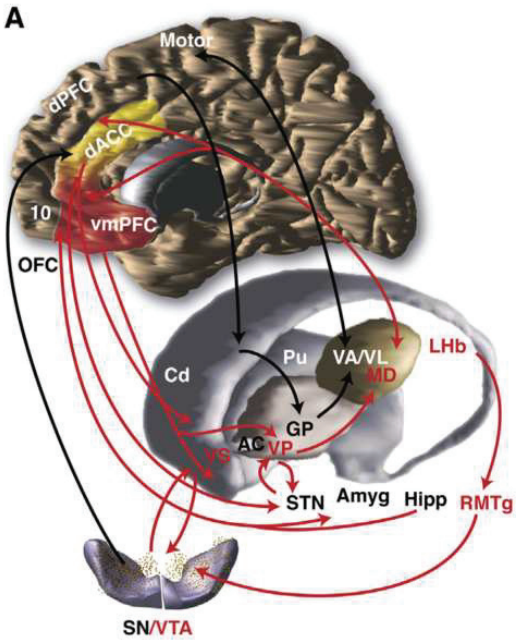

Figure 1.3 – Schéma illustrant les connections entre les structures des ganglions de la base et celles du cortex préfrontal, représentant le circuit de la récompense, d'après (Haber and Behrens, 2014). AC, commissure antérieure ; Amyg, amygdale ; Cd, noyau caudé ; dACC, cortex cingulaire médian ; dPFC, cortex préfrontal dorso-médian ; GP, globus pallidus ; Hipp, hippocampe ; LHb, habenula latérale ; OFC, cortex orbito-frontal ; Pu, putamen ; RMTg, noyau tegmental rostro-médian ; SN, substance noire pars compacta ; STN, noyau sous-thalamique ; VA/VL/MD, noyaux ventral antérieur/ latéral ventral / dorsal médian du thalamus ; vmPFC, cortex préfrontal ventro-médian ; VS, striatum ventral ; VP, pallidum ventral ; VTA, aire tegmentale ventrale.

valeur des actions non choisies... Ensuite, particulièrement en contexte d'incertitude, la valeur "objective" d'une action n'est pas toujours disponible pour le sujet qui ne peut alors que coder sa valeur "subjective". On a vu précédemment que les sujets basaient leurs décisions non pas sur les valeurs objectives des options, mais sur des valeurs subjectives qui intègrent un ensemble de paramètres, comme les coûts des actions, préférences, gains attendus et risques potentiels, etc. Qu'en est-il dans le cerveau ?

Corrélats de la valeur des stimuli

Pour commencer, nous décrirons quelques travaux qui ont identifié des représentations de la valeur des stimuli prédictifs des récompenses. Une première structure importante pour ce rôle semble être le cortex orbito-frontal (COF). En effet, des études montrent que léser cette structure affecte la sensibilité à la dévalua-

tion des renforcements (Gallagher et al., 1999; Izquierdo et al., 2004). Ensuite, des expériences d'électrophysiologie ont montré que la valeur subjective des stimuli représentant des différentes options proposées et choisies dans une tâche de choix économique était reflétée dans l'activité des neurones du COF (Padoa-Schioppa and Assad, 2006). Une autre étude a montré que les neurones du COF représentaient les différentes récompenses de la tâche en cours, en modulant même leur activité en fonction du rang de préférence de l'animal (Hikosaka and Watanabe, 2000). Ce codage est adaptatif, c'est-à-dire qu'il s'adapte à l'étendue (*range*) des valeurs, permettant d'encoder les valeurs au travers de divers contextes présentant des étendues et types de valeurs différents. Cependant, il existe des situations où il peut être utile de ne pas juger la valeur en fonction du contexte, afin d'en avoir une représentation en terme absolu. De manière intéressante, ce codage adaptatif de la valeur dans le COF ne concerne pas tous les neurones : la majorité encode en fait la valeur mais de manière fixe (Kobayashi et al., 2010)). Les neurones de l'amygdale jouent également un rôle dans l'encodage des valeurs des stimuli puisque leur activité suit les changements de valeur des stimuli lors d'une tâche de *reversal* (renversement des associations auparavant correctes) (Paton et al., 2006).

Corrélats de la valeur de l'action

Cependant, pour obtenir les récompenses, il est souvent nécessaire de réaliser une action, et la sélection de cette action est basée en partie sur la valeur qu'on peut espérer obtenir en la choisissant. Le cortex intrapariétal latéral (LIP) a été impliqué dans la représentation de la valeur attendue d'une action (*expected value*) (Platt and Glimcher, 1999). Dans une expérience chez le singe, Platt et Glimcher ont montré que l'activité des neurones du LIP encodait la récompense à laquelle le singe pouvait s'attendre après la réalisation d'une action (saccade), dans un protocole où magnitude et probabilité de récompense étaient variées indépendamment. Une étude suivante, utilisant une tâche avec récompense différée, a montré que l'activité des neurones de LIP évoluait de façon à représenter de manière initiale dans l'essai, la valeur subjective pour ensuite représenter, en complément, la probabilité de réaliser l'action pour l'option donnée (Louie and Glimcher, 2010).

D'autres régions semblent également porter des informations sur les récompenses attendues après la réalisation d'actions spécifiques lors d'une décision. Deux régions ont fait l'objet d'une attention particulière : le striatum et notamment sa partie dorso-*médiane*, et le cortex cingulaire. Cela a été mis en évidence notamment grâce à des expériences de lésion. En effet, des lésions du striatum dorso-médian chez le rat provoquent une insensibilité à la dévaluation des renforcements dans une tâche de conditionnement instrumental (Yin et al., 2005). Le cortex pré-limbique, projetant très fortement sur le striatum dorso-médian, aurait un rôle dans l'*acquisition* de ces réponses dirigées vers un but, puisque des lésions de cette région provoquent les mêmes déficits que ceux obtenus après dévaluation, mais seulement pendant les premières étapes de l'acquisition (Killcross and Coutureau, 2003). Le noyau sous-thalamique semble également participer à l'évaluation de la valeur des actions choisies car son activité après un choix est modulée différemment selon que l'animal (ici, un singe) a reçu sa récompense préférée ou non après un choix entre plusieurs réponses (Espinosa-Parrilla et al., 2015).

Chez le singe, Shima et Tanji ont montré le rôle des aires cingulaires motrices dans la sélection appropriée des actions guidées par la récompense, grâce à des enregistrements électrophysiologiques et des manipulations pharmacologiques (Shima and Tanji, 1998). Une autre découverte clé a été de montrer une sélectivité du cortex cingulaire médian (CCM) pour les représentations liées à l'*action* par rapport aux représentations liées au stimulus. Dans un protocole de Go/No-Go instruits par des indices visuels, les neurones du cortex préfrontal médian (CPFm) encodaient préférentiellement la réponse en elle-même, la possibilité de récompense, ou l'interaction entre ces deux facteurs, mais, en comparaison, très peu de neurones de cette zone encodaient l'identité du stimulus ou l'interaction entre le stimulus et la récompense. Ce patron d'activité était en fait renversé dans le cortex préfrontal dorso-latéral (CPFdl) (Matsumoto et al., 2003). Par la suite, des travaux manipulant probabilité et magnitude des récompenses dans des tâches de décision basées sur des actions, ont permis de montrer que les activités des neurones du CCM corrélaient avec la valeur attendue de la récompense et l'intégration de cette valeur au cours de la tâche (Amiez et al., 2006; Kennerley et al., 2009). Le CCM permet également d'intégrer

l'effort physique dans la valeur des actions (Skvortsova et al., 2014), ce qui participe à son rôle d'intégrateur des coûts et bénéfices des options choisies permettant de déterminer la plus optimale. Les études de lésion contrastent vraiment le rôle du COF et du CCM dans les représentations des valeurs des stimuli (donc des objets, ou des biens) et des actions respectivement (Rudebeck et al., 2008), ce qui correspond bien à leur connectivité anatomique respective (le COF recevant plus d'inputs sensoriels des cortex auditifs, visuels et autres cortex sensoriels, alors que le CCM est fortement connecté aux cortex moteur et prémoteur (Haber and Behrens, 2014).

Ainsi, il existe un réseau étendu de structures corticales et sous-corticales impliquées dans la signalisation du degré d'attente d'une récompense et dans l'attribution de valeurs lors des choix guidés par la valeur, avec des régions plutôt impliquées dans l'encodage des valeurs des actions, et d'autres des valeurs des stimuli.

Représentation des variables qui influencent la valeur des résultats des actions

Il est possible de trouver dans presque toutes les régions du cerveau des neurones dont l'activité est modulée par toute variable décisionnelle utilisée par l'expérimentateur pour manipuler la valeur attendue du résultat d'une action. Mais le CCM, le CPFdl et le COF ressortent particulièrement ; qu'il s'agisse de variables indiquant la magnitude ou la probabilité de la récompense (Amiez et al., 2006), les préférences (Hikosaka and Watanabe, 2000), l'effort requis pour obtenir la récompense (Kennerley et al., 2009), la confiance dans l'action choisie (Kepecs et al., 2008)... Ces multiples représentations rendent la compréhension et la hiérarchisation des différentes aires très complexe. Une solution consiste à enregistrer différentes régions en simultanée. Par exemple, Kennerley et collaborateurs ont enregistré en parallèle dans le COF, le CCM et le CPFl en simultané ce qui permet de comparer les régions pendant que l'animal est dans le même état (Kennerley et al., 2009). Les auteurs ont trouvé que *les trois régions* encodaient toutes les variables de la tâche. Ils ont néanmoins mis en évidence une particularité du CCM : comparé aux autres régions, ses neurones représentaient plus la *combinaison* de plusieurs variables. Cela suggère un rôle dans le multiplexage des représentations des valeurs.

L'apprentissage des valeurs dans le cerveau

Si les valeurs des stimuli et des actions se retrouvent associées à leur renforcement, c'est à la suite d'un processus d'apprentissage. Nous allons détailler dans cette partie les mécanismes cérébraux permettant d'apprendre les valeurs des actions et stimuli, en commençant par le rôle de la dopamine et d'autres neurotransmetteurs, comme substrats pour l'apprentissage associatif, puis en étudiant leur influence sur d'autres régions.

La dopamine représente un signal de récompense renforçant les stimuli et les actions précédant sa libération (initialement suggéré par les expériences d'autostimulation par Olds et Milner (Olds and Milner, 1954) ; puis par de nombreuses études comportementales et de manipulation pharmacologique ou lésionnelle du système dopaminergique (Brozoski et al., 1979; Sawaguchi and Goldman-Rakic, 1991; Simon et al., 1980). Une distinction a d'abord été faite entre la libération de dopamine dans le striatum dorsal jouant plutôt un rôle dans l'apprentissage sensorimoteur et dans le striatum ventral, jouant un rôle dans l'apprentissage des associations stimulus-récompense (White, 1989). Puis, il a été montré que les réponses dopaminergiques aux récompenses dépendent de manière cruciale de leur *imprédictibilité* (Apicella et al., 1992). Un rôle de la libération dopaminergique a ensuite été suggéré, non pas dans le signalement de la récompense en tant que tel, mais dans l'erreur de prédiction de la récompense (Schultz et al., 1997; Schultz, 1998). L'erreur de prédiction de la récompense n'est pas le seul fait de la dopamine, puisqu'une étude a également montré qu'une partie des interneurones cholinergiques du striatum (les neurones TANs *striatal tonically active neurons*) encodait un signal similaire à l'erreur de prédiction de la récompense, négative ou positive, en augmentant et diminuant son activité en réponse à la présence ou l'omission d'une récompense (Apicella et al., 2009). Une différence entre les neurones dopaminergiques et les TANS serait que l'activité des TANs serait modulée par le contexte, contrairement aux neurones dopaminergiques (Apicella, 2007). Varazzani et coll. ont comparé l'implication respective des systèmes dopaminergique et noradrénergique dans des décisions impliquant un effort et ont montré des rôles complémentaires (Varazzani et al., 2015). Les neurones dopaminergiques encodent les valeurs des récompenses mais ces va-

leurs sont modulées par le coût de l'effort à fournir, permettant ainsi de déterminer si l'action en vaut la peine. Une fois cette estimation réalisée, les neurones noradrénergiques prennent le relais et prépare le système à l'effort à venir en modulant leur activité en fonction de la quantité d'effort à fournir.

Ainsi, le signal d'erreur de prédiction dopaminergique constitue une information précieuse représentant la nécessité d'adapter le comportement en cours ou non. Pour être utile, ce signal doit être utilisé par les régions impliquées dans l'apprentissage, et dans l'adaptation du comportement. Nous allons maintenant détailler le rôle de deux structures dans l'apprentissage associatif : le CCM et le COF, et nous verrons comment le CCM en particulier intègre le signal dopaminergique pour l'apprentissage.

Lier les stimuli au renforcement

Il a tout d'abord été reporté que des lésions du COF chez le singe ne provoquent aucun déficit pour l'apprentissage des associations stimulus-résultat sauf en situation de *reversals* conduisant à des persévérations (Izquierdo et al., 2004). Originellement, ces déficits ont été interprétés notamment comme une insensibilité à l'absence de récompense ou à la persévération du choix. Puis Walton et coll. ont de nouveau testé l'effet de cette lésion sur la capacité de *reversal* mais dans un contexte de tâche de *reversal probabiliste* (et non déterministe) (Walton et al., 2010). Dans cette étude, les singes lésés étaient tout aussi capables de réaliser les *reversals* que les contrôles, montrant même un comportement de changement de réponse (*switching*) plus élevé. Le déficit n'apparaissait que lorsque les *reversals* étaient réalisés dans des environnements où les niveaux de stochasticité étaient très variables. Walton et coll. ont utilisé les idées de Thorndike (1933) sur la "diffusion de l'effet" ('*Spread of effect*') pour montrer que les singes lésés fondaient en fait leur choix sur l'*historique* des choix précédents les plus récompensant, plutôt que sur les résultats récents des actions. Ces déficits n'apparaissaient que dans un contexte de *reversals* où la stochasticité était très changeante. En effet, quand l'environnement change tout le temps, il est très difficile de déterminer le meilleur choix dans l'historique des choix précédents. Ainsi, les lésions du COF conduisent à un déficit pour assigner la

valeur appropriée au stimulus responsable de la récompense (lien *causal* entre un stimulus et sa récompense). Un point particulièrement intéressant de cette étude est qu'elle suggère la présence de plusieurs systèmes d'apprentissage, portés par des structures différentes : un premier système supporté par le COF permettant l'apprentissage des contingences entre les stimuli et leurs conséquences ; et au moins un deuxième système, non COF dépendant, moins précis temporellement, basé sur les choix précédents.

Ces résultats vont dans le sens de l'étude de Tanaka et collaborateurs, qui ont contrasté l'encodage des actions permettant d'obtenir des récompenses *immédiates* par rapport à des récompenses *futures* (Tanaka et al., 2004). Ils ont montré que si des sujets apprennent à choisir les actions sur la base des récompenses immédiates, une activation significative est observée en IRMf dans le COF latéral et dans le striatum. Quand les sujets apprennent à agir de telle sorte à obtenir de grandes récompenses dans le *futur*, tout en s'exposant à de petites pertes immédiates, les cortex préfrontal dorsolatéral et inférieur pariétal, ainsi que le noyau raphé dorsal et le cervelet s'activent. Des analyses de régression en utilisant un modèle d'apprentissage par renforcement, calculant les estimations des erreurs de prédiction et récompenses prédites futures des sujets, révèlent en fait que l'activation des régions ventro-antérieures de l'insula et du striatum reflète la prédiction des récompenses immédiates, alors que les régions dorso-postérieures celle des récompenses futures. Ainsi, ces différentes régions seraient impliquées dans la prédiction des récompenses mais à des échelles de temps différentes.

Mais les études de lésion ou d'IRMf n'indiquent pas si le rôle de ces structures est plus dans l'encodage de la valeur plutôt que dans l'apprentissage de cette valeur. Les études d'électrophysiologie le permettent en distinguant temporellement les réponses neuronales des décisions et des récompenses. Dans le COF, au moment du *feedback*, on peut trouver à la fois des neurones codant pour les récompenses, et des neurones codant pour les choix réalisés précédemment (Tsujimoto et al., 2009). La présence de ces deux informations au même moment est un pré-requis pour un système d'apprentissage. Dans le COF, l'encodage des réponses réalisées se fait de manière indépendante du fait qu'elles aient été récompensées ou non. Cela suggère

dans ce cas que le COF représente les réponses choisies de manière indépendante de leur résultat, ce qui ne va pas dans le sens d'une fonction d'apprentissage des valeurs des actions, mais plutôt de représentation des choix appropriés. Au contraire du COF, les neurones du CPFdl encodent mieux les réponses choisies quand elles étaient récompensées, suggérant cette fois un rôle du CPFdl dans l'apprentissage des valeurs des actions (Tsujimoto et al., 2009).

Lier les renforcements aux actions

Contrairement au COF, le CCM est concerné par l'intégration au cours du temps de l'information liée à la récompense permettant la prise de décision basée sur les associations entre les actions et les récompenses. Les neurones du CCM encodent les récompenses reçues dans l'essai en cours mais de manière modulée par les récompenses reçue précédemment (Seo and Lee, 2007). Des injections de muscimole (agent inhibiteur, car agoniste du GABA) dans le CCM induisent des déficits pour trouver la meilleure option dans une tâche d'apprentissage probabiliste et des déficits pour changer de réponse en réaction à un changement de récompense (Amiez et al., 2006; Shima and Tanji, 1998). Des études ont cherché à caractériser plus précisément la nature de ces déficits. Au contraire des singes avec une lésion du COF (qui basaient leur choix sur l'historique des renforcements passés), des singes avec une lésion du CCM basaient leur choix sur le résultat le plus récent seulement (Kennerley et al., 2006). Ainsi, le COF permet la construction d'un lien causal entre un stimulus et sa récompense immédiate, alors que le CCM permet d'intégrer l'historique des choix précédents. Dans une autre étude, l'activité des neurones du CCM en réaction aux *feedback* positifs diminuait au fur et à mesure de l'apprentissage des actions à réaliser, en parallèle de la diminution de l'erreur de prédiction sur les actions à réaliser, suggérant un encodage des erreurs de prédiction concernant les valeurs des actions (Matsumoto et al., 2007). Ainsi, un rôle fondamental du CCM, au moment du *feedback*, serait la mise à jour des valeurs des actions, grâce à un système d'erreur de prédiction construit à partir des choix précédents (Rushworth et al., 2004). Nous verrons que cette fonction est particulièrement pertinente dans un contexte volatil et qu'elle est permise par un deuxième rôle important du CCM dans la détection des

évènements surprenants pertinents pour l'adaptation (voir partie sur la volatilité).

La sélection des options

La prise de décision en faveur de l'une ou l'autre option peut être formalisée par une accumulation d'évidence pour chacune de ces options (Bogacz et al., 2006). Cette accumulation d'évidence prendrait en compte la représentation interne de la valeur, mais également les coûts associés. De nombreux modèles mathématiques ont été développés sur ce principe, en particulier en ce qui concerne les décisions perceptuelles, dans des paradigmes où il est facile de contrôler la quantité d'information accumulée pour chaque option. Le *drift diffusion model* (DDM) suggère que la décision se fait en faveur de l'option dont la variable de décision, qui augmente *via* le taux de *drift*, excède un certain seuil de décision (Ratcliff and Rouder, 1998). Kiani, Hanks et Shadlen ont montré cette propriété dans les neurones de LIP en utilisant une tâche d'accumulation d'évidence en faisant varier la cohérence de points mouvants et en demandant à des singes de faire une saccade dans la direction supposée des points (Kiani et al., 2008). Quelques expériences ont montré que ce type de modèle peut s'appliquer pour les décisions non perceptuelles. Par exemple, Sigman et Dehaene ont utilisé une tâche très simple où il faut comparer deux nombres et dire lequel est le plus grand (Sigman and Dehaene, 2005). Dans cette étude, un DDM capture bien l'augmentation des temps de réaction et le taux d'erreur quand la différence entre les deux nombres se réduit.

Mais lorsqu'on parle de choix d'une "option", dans la plupart des décisions, il s'agit en fait de la sélection d'une action parmi plusieurs possibles. Plusieurs modèles de la façon dont le cerveau décide entre les actions ont été proposés (Cisek, 2012). Deux modèles s'opposent notamment sur un point : Est-ce que le cerveau décide entre les représentations abstraites (par exemple, en comparant les valeurs de chaque offre) et ensuite prépare le plan d'action approprié ou est-ce que les plans d'actions sont déjà représentés en compétition dans le cerveau, et que cette compétition est biaisée par une variété de facteurs, comme, par exemple, la valeur subjective de chaque offre ?

Les données expérimentales, si elles ne permettent pas parfaitement de trancher,

apportent quelques éléments de réponse. Selon le premier modèle, la planification motrice n'interviendrait qu'une fois les décisions prises. Cependant, plusieurs études ont montré que les neurones des régions sensorimotrices représentent les cibles potentielles et les actions bien avant que l'animal ait décidé entre elles (Cisek and Kalaska, 2005; Klaes et al., 2011). De plus, il est difficile de comprendre comment le cerveau serait capable de calculer les coûts des actions sans encoder de représentation de ces actions potentielles. Or, les choix sont influencés par les coûts des actions. En effet, à valeur égale, des sujets humains vont préférer les actions plus faciles à réaliser en terme biomécaniques (Cos et al., 2011). Ensuite, les activités enregistrées dans les régions sensorimotrices sont modulées par les variables décisionnelles. Par exemple, l'activité neuronale liée à une action est plus forte si l'action représentée a plus de probabilité de se réaliser ou si elle conduit à plus de récompense (Michelet et al., 2010). Ces modulations ont été observées dans les cortex pariétaux, frontaux, et même dans le cortex moteur primaire chez l'Homme. Mais tous ces éléments ne prouvent pas que les actions soient préparées puis en compétition pour la sélection. En effet, ces éléments pourraient servir à calculer les coûts des actions et préparer le système sensorimoteur pour les mouvements les plus probables sans pour autant être impliqués de manière causale dans la sélection finale. Des expériences d'intervention sur le système sensorimoteur ont permis de questionner son implication causale dans les décisions. Par exemple, des micro-stimulations perturbant le colliculus supérieur chez des singes réalisant une tâche de recherche visuelle provoquent des déficits de sélection des saccades dans la bonne zone (une tâche contrôle confirmant que les déficits sont non moteurs) (McPeek and Keller, 2004). Ainsi, plusieurs éléments supportent l'idée d'une compétition biaisée entre les représentations sensorimotrices (Cisek, 2012).

Il faut faire la distinction entre les mécanismes par lesquels on se biaise pour un choix plutôt qu'un autre, et ceux par lesquels on sélectionne un de ces choix. Comme proposé par les modèles d'accumulation pour les décisions perceptuelles, un modèle proposé par Roitman et Shadlen suggère que lorsque le cerveau décide entre des actions, l'engagement envers une de ces actions est réalisé lorsque l'activité dans les régions sensorimotrices atteint un seuil (par exemple, dans les neurones de LIP

(Roitman and Shadlen, 2002)). Les décisions plus abstraites pourraient suivre les mêmes règles. Par exemple, le CCM est impliqué dans la décision d'explorer versus exploiter et il a été montré que les neurones du CCM atteignent un seuil juste avant que le singe décide de partir en exploration (Hayden et al., 2011).

Le modèle du consensus distribué (Cisek, 2012) propose que la compétition entre les actions se réalise à deux niveaux interconnectés : un premier niveau où l'activité représente les mouvements spécifiques en compétition les uns avec les autres (*à droite ou à gauche ?*), et un second niveau, plus haut, représentant les choix dans l'espace abstrait des buts (*la pomme ou l'orange ?*), en compétition eux aussi, et présents dans les parties plus antérieures du cortex frontal, fortement connecté avec les parties sensorimotrices. Cela rejoint le concept du contrôle cognitif, dont nous allons décrire les corrélats neurophysiologiques dans un des paragraphes suivants.

Le suivi et l'évaluation (monitoring) des performances

Le suivi des performances concerne plusieurs types de "performances" : de la détection d'erreurs exécutives motrices, d'inattention jusqu'à la détection d'erreurs liées à des changements externes ou encore la détection des *feedback* (positifs ou négatifs). L'ensemble de ces évènements engendre une déflection dans le signal EEG (appelé potentiel relié aux erreurs (*error related negativity ERN*) et potentiel relié aux *feedback* (*feedback related negativity FRN*)) : elle consiste en une négativité fronto-centrale (vers 200ms *post-feedback*), suivie immédiatement d'une positivité fronto-centrale, puis un peu plus tard, d'une positivité pariétale, appelées respectivement N2, P3a et P3b (Falkenstein et al., 1991; Gehring et al., 1993; Miltner et al., 1997). Les deux activités fronto-centrales semblent provenir de la même source dans le CCM (Dehaene et al., 1994; Hauser et al., 2014), ou de l'aire motrice supplémentaire (Bonini et al., 2014).

Dans la théorie d'apprentissage par renforcement et du potentiel lié à l'erreur (*RL-ERN theory*) proposée par Holroyd et Coles, le potentiel évoqué aux erreurs refléterait l'influence sur l'activité des neurones du CCM des signaux d'erreur de prédiction de la récompense reçus *via* les connections entre le CCM et le système dopaminergique (Holroyd and Coles, 2002). La diminution de l'activité dopaminer-

gique en réaction à une erreur de prédiction négative (Schultz, 1998) provoquerait une inhibition de l'activité des neurones du CCM, qui serait responsable de la déflection négative enregistrée à la surface du cortex correspondant à la FRN. Holroyd et Coles ont proposé que le CCM utiliserait ces signaux dopaminergiques d'erreur de prédiction de la récompense pour sélectionner les meilleurs plans d'actions permettant l'adaptation comportementale. Une autre hypothèse (notamment proposée dans le modèle PRO (*predicted response outcome*), (Alexander and Brown, 2011)) propose que ces potentiels reflètent plutôt la surprise (c'est-à-dire l'erreur de prédiction absolue ou non signée (positive et négative)). En effet, des études plus récentes sur la FRN ont montré que l'amplitude de la FRN était équivalente pour des évènements de valence opposée mais avec le même degré d'inattendu (Ferdinand et al., 2012; Hauser et al., 2014; Holroyd and Krigolson, 2007; Oliveira et al., 2007; Walsh and Anderson, 2011). Les données chez le singe contribuent aux deux hypothèses puisqu'on trouve des neurones du CCM encodant à la fois l'erreur de prédiction positive, négative, mais aussi non orientée (*unsigned*) (Matsumoto et al., 2007; Quilodran et al., 2008).

En fait, peu d'études ont étudié directement le rôle de la dopamine dans le CCM. L'étude de Wilkinson a montré, grâce à une microdialyse in vivo chez le rat, qu'une décharge de dopamine corticale était libérée dans le CCM lors de l'association entre un stimulus et un choc électrique (qui doit sûrement être considéré comme un évènement surprenant) (Wilkinson et al., 1998). Une étude de Vezoli et Procyk a montré, de manière indirecte, l'influence du système dopaminergique sur la FRN, car ce potentiel est modulé par des injections systémiques d'un antagoniste dopaminergique, l'halopéridol (Vezoli and Procyk, 2009). Des manipulations locales du taux de dopamine dans les structures de l'apprentissage restent encore à réaliser afin de mieux comprendre comment son signal est utilisé. Une expérience importante serait de combiner l'enregistrement de ces potentiels évoqués par différents types de *feedback* à des perturbations pharmacologiques dopaminergiques *locales* dans le CCM, afin de tester directement le rôle de la dopamine dans le CCM dans la génération et la modulation de ces potentiels. Il s'agit en fait d'un des buts à long terme de ce projet de thèse, que nous sommes en train de développer (voir partie "Programme

expérimental de la thèse"). Cette expérience (si nous parvenons à la mener à son terme !) apportera des données précieuses pour comprendre le rôle de la dopamine dans la fonction du CCM.

Un débat actuel dans la littérature sur la FRN consiste à savoir si ce potentiel est uniquement lié à la détection des *feedback* surprenants, c'est-à-dire au signalement d'un besoin d'adapter le comportement ; ou si ce potentiel reflète aussi ce qu'il est nécessaire de faire pour s'adapter en conséquent. Les études actuelles apportent des éléments dans un sens puis dans l'autre. Nous détaillerons ces études dans l'introduction du premier papier d'électrophysiologie de cette thèse, mais à titre d'exemple, des études ont rapporté des changements de l'amplitude de la FRN qui étaient prédictifs des adaptations comportementales, comme de choisir la même cible ou non par rapport à celle juste choisie (Cohen et al., 2007) ; ou du niveau d'apprentissage dans la tâche (Frank et al., 2005). Mais d'autres études ont mis en évidence des situations où l'amplitude de la FRN était complètement décorrélée des adaptations comportementales par les sujets (Chase et al., 2010; Mars et al., 2004; Walsh and Anderson, 2011). Dans ce travail de thèse, nous essayerons d'apporter des éléments pour faire avancer le débat en regardant si les potentiels évoqués aux *feedback* que nous enregistrons chez le singe sont prédictifs des changements comportementaux (voir première étude d'électrophysiologie).

Un autre débat concerne le rôle de la P3 par rapport à la FRN. Ces deux signaux sont difficiles à dissocier car ils se chevauchent partiellement dans le temps. Les générateurs de la P3 pariétale seraient distribués dans le cortex, autour de la jonction temporo-pariétale (Polich, 2007). Les composantes précoces fronto-centrales reflèteraient des signaux d'alarme indiquant l'imminence de la nécessité d'adapter le comportement (Ullsperger et al., 2014). L'évidence s'accumulerait pour donner la composante pariétale, semblant représenter le résultat d'un processus de décision. Lorsque ce signal excèderait un certain seuil, il provoquerait la redirection de l'attention et la conscience de la décision de s'adapter (Steinhauser and Yeung, 2012). Une hypothèse relie la P3 à la fonction du locus coeruleus et aux effets de la norépinephrine dans le cortex dans l'éveil et dans la balance entre exploration et exploitation (Nieuwenhuis et al., 2005). Le fait que le taux d'apprentissage influence la fiabilité

et le contenu informationnel des *feedback* dans les environnements incertains (un haut taux d'apprentissage donnant plus de poids aux *feedback* récents) est reflété dans ces activités. Dans une tâche de prise de décision basée sur l'apprentissage des probabilités de récompense associées à des stimuli, une plus grande P3b au moment du *feedback* a été associée à de hauts taux d'apprentissage et prédisait les changements de comportement (Fischer and Ullsperger, 2013). Ainsi, la P3 a elle aussi été reliée au signalement du besoin de s'adapter, ainsi qu'à l'adaptation comportementale elle-même. Cela suggère que plus d'études sont nécessaires afin de distinguer les rôles respectifs de la FRN et la P3 dans l'adaptation.

Dans tous les cas, il est intéressant de noter que ce système de détection des *feedback* dans le CCM semble fortement *incarné* (*embodied*) (Procyk et al., 2014). Le CCM possède des cartes somato-motrices reflétant l'activité des membres utilisés (œil, main, pied, langue...) et il a été montré que les activités reliées au *feedback* sont localisées dans les régions du CCM correspondant à la modalité du *feedback* reçu (Amiez and Petrides, 2014). Ainsi, les *feedback* visuels sont retrouvés dans les régions somato-motrices cingulaires liées à l'œil, les *feedback* "jus de fruit" dans les régions de la langue, etc... Ces résultats mettent en évidence le fait que ces mécanismes d'adaptation se sont développés en relation avec le monde physique. Une nouvelle question liée au monde moderne concerne les situations où les *feedback* donnés sont abstraits comme le pouvoir, ou les promesses d'argent dans les études chez l'Homme, et il reste à déterminer si leur analyse est réalisée *via* une généralisation des mécanismes décrits dans le CCM, ou dans des régions plus impliquées dans le traitement des informations abstraites.

Ainsi, une des grandes questions est de comprendre comment l'évaluation des performances permet l'adaptation comportementale. Cette question est le point central des études sur le contrôle cognitif, qui cherchent à comprendre comment ce lien est réalisé.

Le contrôle cognitif

La théorie de la boucle du contrôle cognitif postule que la sélection des actions appropriée est influencée par un mécanisme de contrôle cognitif à la suite de l'étape

d'évaluation des performances (Fuster, 2001) (ce qui rejoindrait l'idée du consensus distribué de Cisek). Le contrôle cognitif fait référence à un panel de mécanismes computationnels permettant : "*[the] active maintenance of task-relevant context and top-down biasing of local competitive interactions that occur during processing*" (Braver et al., 2002). Le CPF s'est vu attribuer un rôle particulier dans le concept de contrôle cognitif en permettant le maintien actif des patrons d'activité représentant les buts ainsi que les moyens pour les atteindre (Miller and Cohen, 2001). D'après Miller et Cohen, toutes les options sont associées à des patrons d'activité dans le cortex préfrontal, soutenues par des connections qui sont renforcées à chaque fois qu'un comportement réussit. Avec le temps, et de nombreuses répétitions, ces représentations préfrontales se complexifient en se combinant avec d'autres et s'associent à des évènements et contingences particulières (pouvant alors être considérés comme des *task-sets*), conduisant à l'apprentissage des actions requises.

Le CPF possède les caractéristiques principales d'un système dont le rôle serait d'exercer le contrôle cognitif. Une première propriété importante d'un tel système est la capacité à abriter les représentations appropriées. Pour cela, il doit avoir accès à toutes les informations disponibles par les autres régions, ce qui nécessite une forte multimodalité et une capacité d'intégration. Le CPF remplit ce critère car il est fortement connecté avec les divers systèmes sensoriels, moteurs, et limbiques, et les aires qui le composent sont aussi fortement connectées entre elles, ce qui fait de lui un nœud central important (Markov et al., 2013). De plus, comme évoqué précédemment, le cortex préfrontal regroupe des régions impliquées dans l'apprentissage des associations (Tsujimoto et al., 2009).

Le contrôle cognitif requiert aussi une maintenance du but à atteindre, de manière robuste jusqu'à ce qu'il soit achevé, en résistant aux interférences. Mais en même temps, cela requiert d'être assez flexible pour le remettre en question s'il devient non pertinent. L'implication du CPFdl dans cette fonction a été montrée grâce à des tâches comportant une période de délai entre un indice présenté indiquant le comportement à réaliser et l'exécution de la réponse. Dans ce genre de tâches, il est nécessaire de maintenir toutes les informations importantes pendant le délai pour obtenir de bonnes performances. Pendant cette période, le CPF montre une activité

soutenue représentant les informations sur le stimulus (di Pellegrino and Wise, 1991), les actions à venir (Asaad et al., 1998; Funahashi et al., 1991), les récompenses espérées (Watanabe, 1996) et des informations plus complexes comme la position séquentielle d'un stimuli au sein d'une série ordonnée (Barone and Joseph, 1989) ou l'association spécifique entre un stimulus et la réponse correspondante (Asaad et al., 1998). La partie dorsolatérale est spécifiquement impliquée dans cette fonction. En effet, des expériences de lésions des CPFdl et CPFdm ont montré que seul le CPFdl est critique pour la bonne réalisation de tâches de mémoire de travail, mais seulement pour les versions spatiales (Levy and Goldman-Rakic, 1999).

D'autres aires présentent cependant ce type d'activité soutenue, comme le cortex visuel dans lequel un stimulus bref peut évoquer une activité persistante (Fuster and Jervey, 1981). Mais ces activités, à la différence de celles du CPF, ne résistent pas aux interférences (comme lorsque la période de délai est remplie de distracteurs (Miller et al., 1996)). Il est également nécessaire d'avoir une représentation particulière des éléments pertinents pour le comportement. Par exemple, les études de Watanabe (Watanabe, 1990, 1992) dans lesquelles des singes sont entrainés à reconnaitre certains stimuli visuels et auditifs comme pertinents pour obtenir une récompense montrent que les neurones du CPF latéral encodent seulement les stimuli quand ils sont pertinents, c'est-à-dire seulement dans les situations où ils étaient prédictifs d'une récompense.

Ensuite, une caractéristique fondamentale d'un système permettant la réalisation du contrôle cognitif consiste à envoyer des signaux qui agissent sur les autres structures. Le CPF communiquerait, par des voies dites *feedback*, des signaux qui peuvent biaiser les systèmes sensoriels, ce qui appuierait son rôle pour diriger l'attention (Knight, 1984). Des signaux envoyés au système prémoteur seraient responsables de la sélection des réponses (Fuster, 2001). Le CPF envoie également des signaux top-down au cortex temporal inférieur pour le rappel des mémoires visuelles qui y sont stockées (Tomita et al., 1999). Cavanagh et collaborateurs proposent un rôle du noyau sous-thalamique (STN) pour l'implémentation du contrôle cognitif après une erreur (Cavanagh et al., 2014). L'activité des potentiels de champs locaux du STN a en effet été reliée avec une augmentation du ralentissement des

temps de réaction après une erreur (*post-error slowing*). Les auteurs proposent que le STN agisse comme un frein sur l'exécution motrice en réaction à une erreur ou en situation de conflit, laissant plus de temps pour la prochaine action. Au cœur du contrôle, les interactions entre le CCM et le CPFdl ont été proposées pour permettre l'intégration de l'évaluation des performances par le CCM pour ajuster le comportement en conséquent (Johnston et al., 2007; Rothé et al., 2011). Dans une étude récente, Alexander et Brown propose un nouveau modèle (le modèle hiérarchique de représentation de l'erreur) pour expliquer ces interactions. Ils proposent que les divers signaux d'erreurs générés par le CCM en réponse aux évènements surprenant soient utilisés pour entrainer le CPFdl à la création de représentations des erreurs attendues, afin de les associer aux stimuli pertinents pour la tâche. Ces représentations seraient maintenues dans le CPFdl et seraient utilisées en retour pour moduler l'activité prédictive dans le CCM afin de générer de meilleures estimations des résultats des actions. Un modèle proposé par Koechlin et collaborateur propose une organisation des différentes régions du CPF selon un axe antéro-postérieur supportant différents niveaux d'abstraction du contrôle cognitif (Koechlin et al., 2003). Nous détaillerons ce modèle après avoir évoqué les corrélats neurophysiologiques des différentes sources d'incertitude, stochasticité et volatilité, car les plus hauts niveaux proposés par ce modèle seraient engagés lors de forte demandes cognitives, conditions qui surviennent en particulier lorsque l'environnement est changeant, et nécessite un comportement flexible.

Si ces études ont permis l'identification des régions clés pour l'implémentation du contrôle cognitif, il reste encore à comprendre les *mécanismes* permettant au contrôle cognitif de s'exercer, et, en particulier, la façon dont ces régions clés s'échangent l'information, et la façon dont ceci est modulé en fonction des demandes de la tâche. C'est pourquoi l'intérêt de la communauté s'intéressant à la prise de décision a commencé à se porter sur les oscillations cérébrales, qui ont été proposées comme mécanisme plausible de communication "on-off", rapide et flexible, entre les régions (Bastos et al., 2015). Un rôle des oscillations thêta (mais aussi beta, comme nous le verrons au paragraphe suivant) a été proposé pour l'implémentation du contrôle cognitif *via* la communication entre le CPF et les autres régions (Cavanagh et al.,

2014). Les oscillations thêta ont été impliquées dans les mécanismes de cohérence entre potentiels d'actions et potentiels de champs locaux (par exemple, dans le CCM du rat et du singe, une augmentation de la puissance thêta est associée à un couplage augmenté entre les potentiels d'action et la phase du cycle thêta (Narayanan et al., 2013; Womelsdorf et al., 2010)). Ce couplage de phase du thêta medio-frontal pourrait agir comme organisateur des processus neuronaux pendant les "points de décision", comme lorsque l'information pertinente pour le choix est intégrée afin de permettre la sélection de l'action (Benchenane et al., 2010; Womelsdorf et al., 2010). En effet, le couplage en thêta entre différentes régions est augmenté dans les conditions nécessitant une augmentation du contrôle. Par exemple, des augmentations de la synchronisation de phase entre le CPFdl, le CCM et les aires sensorimotrices ont été observées après une erreur (van de Vijver et al., 2011). Une étude de *task-switching* chez le singe, dans laquelle les auteurs ont séparé les essais en deux catégories (les essais dits "contrôlés" versus "automatiques") selon les temps de réaction, a permis de mettre en évidence un réseau préfrontal-pariétal dont la cohérence (bidirectionnelle) dans la bande de fréquence thêta était accrue lors des essais contrôlés par rapport aux essais automatiques, durant la phase préparatoire de l'essai (Phillips et al., 2014). Une autre étude a montré qu'une synchronisation thêta se mettait en place des aires motrices vers les aires medio-frontales immédiatement après le *feedback*, puis des aires medio-frontales vers les aires frontales un peu plus tard, et ce, de manière plus forte chez les sujets ayant le mieux réussi apprendre la tâche (Luft et al., 2013). Ces études soutiennent fortement la proposition d'un rôle coordinateur et de transmission d'information pour les oscillations thêta permettant l'exercice du contrôle cognitif.

Cependant, les oscillations beta ont aussi été associées à diverses fonctions "top-down", comme, par exemple, dans le système moteur (Pfurtscheller and Lopes da Silva, 1999), mais aussi au contrôle cognitif. Par exemple, notre équipe a montré une augmentation de la puissance beta durant la période de préparation de l'essai pour les phases du problème demandant le plus de contrôle cognitif, mais aussi avec le temps passé sur la tâche (Stoll et al., 2015). Nous avons donc proposé un rôle de ces oscillations pour refléter le besoin de contrôle et le niveau d'effort

attentionnel. Une augmentation de la synchronisation beta entre les aires frontales et pariétales a également été montrée juste avant la détection réussie d'un changement (Micheli et al., 2015), ou lors des décisions à libre choix, par rapport aux décisions instruites (Pesaran et al., 2008). Ainsi, si les oscillations semblent jouer un rôle crucial dans ces processus, celui-ci, et notamment la spécificité des différentes bandes de fréquence, n'est pas encore très clair. Des études comparant les comportements de ces deux types d'oscillations, ainsi que leurs interactions, au cours de tâches de décision nécessitant du contrôle, sont encore nécessaire afin de comprendre leur rôle respectif, et le fonctionnement des régions clés, en tant que réseau. Comme nous le verrons, ce type d'analyses est au cœur des objectifs de ce travail de thèse.

Maintenant que nous avons eu un aperçu des mécanismes neurophysiologiques permettant d'apprendre de l'environnement, nous allons voir comment ces mécanismes s'adaptent lorsque l'environnement devient plus complexe à comprendre. Nous commencerons par étudier les mécanismes permettant de gérer l'incertitude en situation de risque.

1.2.2 Corrélats neurophysiologiques du risque et représentations de la stochasticité

Un comportement adapté nécessite de prendre en compte le niveau de risque d'un choix. Or, comme nous l'avons vu dans la partie sur les processus comportementaux, les choix des humains et animaux sont fortement influencés par leur estimation des probabilités de chaque option. On peut trouver des corrélats neuronaux de certaines prédictions de la *Prospect Theory* (Kahneman and Tversky, 1979) dans les environnements de risque, testés grâce à des paradigmes manipulant la variance de la distribution des probabilités de la récompense. Par exemple, Tom et coll. se sont intéressés aux corrélats neuronaux de l'aversion aux pertes (Tom et al., 2007). Dans un paradigme de décisions prises à partir d'instructions, les auteurs proposaient un choix entre deux options avec une probabilité égale de gagner ou de perdre, mais dont la magnitude changeait, afin d'identifier le degré d'aversion aux pertes chez les sujets. Ils ont d'abord montré que le même réseau cérébral était activé pour les gains et les pertes, mais dans des directions opposées. Ce réseau incluait les régions dé-

crites précédemment pour leur rôle dans l'encodage des valeurs (CPFvm, striatum, COF) mais de manière intéressante, le degré d'encodage du ratio gain/perte était prédictif de ce même ratio dans le comportement. L'aversion comportementale aux pertes était donc corrélée à l'aversion "neurale" aux pertes, suggérant un encodage de la valeur subjective de la décision. Une autre étude a montré une autre prédiction de la *Prospect Theory* : l'effet de cadrage (*framing effect*) (Martino et al., 2009). Les sujets devaient prendre une décision entre un choix risqué et un choix sûr mais l'amplitude des pertes et des gains était déterminée par les montants que le sujet avait déjà perdu ou gagné. Dans ce contexte, l'activité de l'amygdale était représentative du fait que les décisions étaient "cadrées" ou pas.

Des corrélats neuronaux du risque ont également été retrouvés à l'échelle d'enregistrements unitaires chez l'animal, en cherchant des activités modulées de la même façon que la fonction mathématique du risque (forme caractéristique en U inversé). Fiorillo et coll. ont montré une double réponse des neurones dopaminergiques dans une tâche Pavlovienne, sans choix, de perception des probabilités (Fiorillo et al., 2003). Dans cette expérience, des stimuli visuels sont présentés et indiquent la probabilité de recevoir une récompense de magnitude fixe (les probabilités variant de 0 à 1 par pas de 0.25). Tout d'abord, les neurones dopaminergiques montrent une réponse transitoire, phasique, en réponse au stimulus prédictif de la récompense, qui augmente avec la probabilité de récompense. Si la magnitude de la récompense est variée, ces neurones encodent la valeur attendue moyenne de la récompense (Tobler et al., 2005). C'est également le cas dans le striatum, en particulier pour les neurones cholinergiques TANS (Apicella). L'activité BOLD du striatum reflète en fait à la fois la magnitude de la recompense, sa probabilité, et le produit des deux, à savoir, la valeur attendue (Tobler et al., 2007). Ces deux types de neurones encodent également l'erreur de prédiction de la récompense (Schultz et al., 1997; Apicella et al., 2011). Une hypothèse serait que les neurones TANS encodent les erreurs engageant les récompenses obtenues dans un contexte d'habitude, alors que les neurones dopaminergiques détecteraient les erreurs lors de l'acquisition de nouvelles associations entre les stimuli et leurs conséquences (Apicella et al., 2009). Il a également été montré que l'erreur de prédiction de la récompense en provenance du striatum conduisait

à une modulation de l'activité du cortex prémoteur en réaction aux évènements visuels surprenants, ce qui va dans le sens du rôle du striatum dans le task-switching et la sélection des actions motrices pertinentes (den Ouden et al., 2010). Il n'est pas clair si l'activité du striatum représente le risque, puisque les études se contredisent. Preuschoff et collaborateurs ont trouvé que l'activité BOLD dans le striatum corrélait avec le niveau de risque (Preuschoff et al., 2006), mais Tobler et collaborateurs n'ont pas répliqué ce résultat, en montrant que l'activité dans le striatum reflétait seulement la valeur attendue (Tobler et al., 2009). Ils expliquent notamment la différence entre les deux études par le fait que dans leur étude, ils ont régressé leur modèle sur des activations phasiques, qui ne corrèlent pas avec le risque. En regardant les activations toniques, ils ont trouvé qu'elles reflétaient bien le risque mais les auteurs argumentent que ces activations, plus prolongées ; sont difficiles à décorréler de l'attitude des sujets face au risque. Concernant les neurones dopaminergiques, en supplément, au moins un tiers des neurones enregistrés dans l'étude de Fiorillo montrent également un autre type de réponse : une activation additionnelle, plus lente et plus soutenue, pendant l'intervalle entre le stimulus et la récompense. Cette activation est plus forte pour les probabilités de 0.5 et moins forte pour les probabilités plus faibles et plus élevées, correspondant à la forme en U inversé du risque. Ainsi, les neurones dopaminergiques encodent à deux moments différents deux types d'information : l'information concernant les prédictions et erreurs de prédiction de la valeur de la récompense, et l'information concernant le niveau de risque lié à ces récompenses. Cette dernière fonction pourrait fournir des informations sur le degré de risque dans la distribution des récompenses aux structures chargées de gérer le risque en lui-même et de prendre des décisions en conséquent. Cela pourrait également être un moyen de moduler les signaux de la valeur de la récompense par le niveau de risque, afin de moduler l'utilité attendue (la valeur subjective) de la récompense pour les individus très sensibles au risque. Des signaux liés au niveau de risque sont également retrouvés au niveau cortical, modulant l'activité liée aux mouvements ; comme par exemple dans le cortex cingulaire postérieur (CCP) (McCoy and Platt, 2005).

Chez l'Homme, des expériences d'IRMf ont utilisé le même type de protocole

pour distinguer risque et valeur de la récompense attendue ((Preuschoff et al., 2006) (tâche de cartes), (Tobler et al., 2007)). Ces études ont montré les implications du putamen, striatum ventral, globus pallidus, du CCM et mésencéphale dans l'encodage de la valeur de la récompense attendue (probabilités de récompense). Les activations liées au risque ont notamment montré les implications du striatum ventral, du mésencéphale et du thalamus. L'activation de l'insula antérieure co-variait avec la différence entre le risque estimé (information fournie par la tâche) et sa prédiction (erreur de prédiction du risque) (Preuschoff et al., 2006).

Tobler et coll. ont également montré des activations du COF latéral qui augmentaient en parallèle avec le risque (Tobler et al., 2007). De manière intéressante, le gyrus frontal supérieur antérieur a montré un signal de risque décroissant seulement chez les personnes aversives au risque, et le gyrus frontal inférieur caudé un signal de risque croissant uniquement chez les personnes ayant une inclination pour le risque. Cela suggère que les signaux de risque ne sont pas les mêmes entre les différents individus et varient en fonction de l'attitude des individus vis-à-vis du risque, ce qui pourrait expliquer les différent comportements face prises de décision en situation de risque.

Les divers systèmes de neurotransmetteurs jouent également un rôle dans l'estimation du niveau d'incertitude environnementale. L'acétylcholine (Ach) aurait un rôle crucial dans l'incertitude attendue (provoquée par un environnement probabiliste stable). Cette idée a émergé grâce à des expériences où un indice prédisait correctement l'association stimulus-réponse-conséquence dans 80% des essais. La fiabilité de cet indice était constante tout du long de la session et consistait donc en une mesure de la stochasticité de la tâche. Les niveaux d'ACh étaient inversement corrélés au niveau d'estimation de la validité de l'indice (Witte et al., 1997). Cela suggère que l'ACh correspond à une forme d'incertitude qui peut être apprise à travers l'expérience des associations stimulus-réponse-conséquence passées (Yu and Dayan, 2005). Des études suggèrent que l'ACh augmente de manière soutenue en fonction de l'incertitude attendue de l'environnement quand l'attention doit être maintenue (Dalley et al., 2001). Cela implique que la capacité de prédire les associations d'un environnement nécessite un mécanisme soutenu temporellement pour

estimer l'incertitude.

Ces études montrent que plusieurs structures et systèmes sont impliqués dans la détection et l'estimation du risque. Une question émanant de ces études est de savoir si l'estimation du risque et l'apprentissage pour diminuer l'ambiguïté sont supportés par des mécanismes différents. En effet, la prise de décision en contexte d'ambiguïté (incertitude associée à des probabilités inconnues) pourrait être considérée comme un cas spécial, plus complexe, de prise de décision en contexte de risque (incertitude associées à des probabilités connues). Cependant, plusieurs études montrent des réseaux différents activés lors de deux types de contextes, ce qui supporte l'idée qu'il s'agit de deux mécanismes différents. Tout d'abord, l'étude en IRMf de Hsu et coll. utilise un protocole permettant de dissocier l'ambiguïté du risque en contrastant des options contenant différentes quantités d'informations (Hsu et al., 2005). Par exemple, dans la situation dite du "plateau de cartes", le sujet doit faire un choix entre un pari risqué où les probabilités de gagner sont connues et un pari ambigu où seulement une partie des probabilités est connue (et des options contrôles "sûres" permettant le contraste). Le signal BOLD a montré une activation plus forte pour les paris ambigus par rapport aux paris risqués dans le COF, l'amygdale, et le CPFdm. En utilisant un protocole similaire, Huettel et collaborateurs ont montré que les préférences des individus pour le risque et l'ambiguïté dans une tâche de prise de décision permettaient de prédire les activations de différentes régions : l'activité BOLD dans le CPFl était notamment corrélée avec les préférences en terme d'ambiguïté alors que celle du cortex pariétal postérieur était modulée par les préférences en terme de risque (Huettel et al., 2006).

Ces études montrent l'implication d'un réseau étendu pour l'estimation de l'incertitude liée au risque, dont les niveaux d'activation pourraient expliquer les différences individuelles de comportement face au risque, (Huettel et al., 2006) et qui varient selon le type de décision (Huettel et al., 2005). Dans la tâche comportementale utilisée pour ce travail de thèse, l'incertitude dépend à la fois de l'apprentissage au cours d'un problème, mais aussi du niveau de "bruit" des *feedback* (*feedback* probabilistes), ce qui ne permettra pas d'identifier des réseaux spécifiques pour l'une ou l'autre forme d'incertitude, mais qui permettra d'obtenir un comportement qui

doit faire face à la combinaison des deux.

1.2.3 Corrélats neurophysiologiques des processus développés pour s'adapter à la volatilité

La volatilité de l'environnement nécessite une grande flexibilité comportementale chez les individus. Comme nous l'avons abordé dans la partie sur le comportement, le concept de *task-sets* a été proposé comme moyen de réagir de manière flexible à un changement. De nombreuses études ont voulu mettre en évidence des corrélats neurophysiologiques sous-jacents aux différents sous-processus supportant le concept de *task-sets*.

Corrélats neuronaux des *Task-sets*

Les corrélats neuronaux des *task-sets* peuvent être recherchés en combinant les divers paradigmes comportementaux décrits précédemment avec des techniques d'enregistrements cérébraux. De nombreux éléments de la tâche vont corréler avec le signal cérébral. Mais si ces différents éléments interagissent ou se combinent pour former un *task-set* cohérent, il est peut-être possible de trouver des corrélats du *task-set* lui-même, en supplément de ses éléments constitutifs. Par exemple, la représentation d'un *task-set* serait indépendante des changements dans les stimuli si la tâche reste la même, alors que les représentations perceptuelles de ces stimuli changeraient. Si un tel niveau de représentation du *task-set* existe, une façon de comprendre son rôle serait à travers l'étude de comment et à quel moment il change et s'adapte, en particulier quand la tâche elle-même change.

Ainsi, nous commencerons notre revue des corrélats neuronaux des *task-sets* par les études cherchant à établir un rôle spécifique de régions particulières pour le *task-switching*. En fait, ces structures varient beaucoup en fonction des tâches étudiées, ce qui indique qu'elles travaillent probablement au sein d'un réseau, plutôt que d'avoir un rôle spécifique dans un processus particulier. Dosenbach et collaborateurs ont mis en évidence des régions montrant l'activité d'un réseau commun à différents types de signaux liés aux *task-sets* testés dans 10 tâches différentes (Dosenbach et al., 2006). Ces régions incluent le CCM /CPFm et l'insula antérieure bilatérale/l'operculum

frontal et ont été considérés comme le "cœur du système de *task-set*". Une idée parcimonieuse serait que le *task-set* puisse être représenté comme l'interaction entre les différentes représentations des éléments de la tâche, plutôt que dans des représentations concrètes séparées. Cela implique que les méthodes d'enregistrement du signal les plus pertinentes pour aborder cette question sont celles permettant d'étudier des assemblées de neurones, et plus particulièrement leurs interactions (ce qui pousse vers la réalisation d'études sur les oscillations, plutôt sur des neurones isolés par exemple). Néanmoins, il est important d'identifier les éléments principaux de ce réseau, et la façon avec laquelle ils corrèlent à la tâche, au *task-set*, et entre eux, afin de comprendre leur fonctionnement en réseau.

Corrélats neuronaux des sous-processus des *Task-sets*

La représentation des règles

Parmi les composants les plus fondamentaux dans un *task-set* figurent les règles, requises pour exécuter la tâche. Ainsi, la première étape pour comprendre les *task-sets* est de comprendre où et comment les règles sont représentées dans le cerveau. Une structure clé semble être le CPF. En effet, des patients avec des lésions du cortex préfrontal latéral montrent souvent une "inertie cognitive", qui a notamment été associée avec des difficultés à générer des règles (Levy and Dubois, 2006). Des neurones représentant les règles ont été reportés dans de nombreuses sous-régions du CPF (Hoshi et al., 2000; Ott et al., 2014; Sakagami and Tsutsui, 1999). Cependant, des activités liées aux règles, du moins à des règles simples, peuvent être trouvées dans de nombreuses autres régions cérébrales. Par exemple, Wallis & Miller ont montré que la simple règle "identique versus différent" était encodée de manière plus forte dans le cortex prémoteur que dans le préfrontal (Wallis and Miller, 2003). Une explication pourrait être que les réponses demandées dans cette tâche différaient beaucoup par leurs exigences motrices. De manière similaire, des neurones représentant les règles ont été trouvés dans différentes régions en fonction du domaine sensoriel qui était requis pour réaliser la tâche. Par exemple, des neurones spécifiques d'une règle "*location-match versus shape-match*" ont été retrouvés de manière respective dans la partie la plus postérieure et une partie un peu moins postérieure du

sillon principal (Hoshi et al., 2000). Ces découvertes semblent indiquer qu'il n'existe pas de "région représentative des règles", et comme les règles sont un élément clé du *task-set*, cela soutient l'idée d'une représentation du *task-set* au niveau d'un réseau.

Bien que ces différents neurones montrent une activité corrélée avec les règles comportementales, cela ne signifie pas forcément que ce soit ce qu'ils représentent. Leur activité pourrait en effet être modulée de manière indépendante (ou de manière corrélée) avec des mesures comme l'attention. Ainsi, les variations de localisations anatomiques décrites précédemment pourraient être liées à des variations du niveau d'attention requis par les différents domaines sensoriels. Un mécanisme de modulation pourrait bien être les oscillations cérébrales, car elles ont par exemple étaient montrées comme modulant la direction de l'attention visuelle sur des stimuli (Bosman et al., 2012).

Si les règles peuvent se baser sur quelque chose de très concret, s'appliquant par exemple à un domaine sensoriel particulier, ou à des stimuli particuliers ; elles peuvent également se baser sur des notions abstraites, représentant les stratégies comportementales générales. On peut alors se demander si les mêmes régions sont impliquées. Des activations en IRMf dans le CPFvl ont été observées lorsqu'on demande à des sujets de détecter des similarités entre des images, alors que des activations du CPFdl sont observées dans des conditions où il est demandé de trouver les images partageant un même concept abstrait (Garcin et al., 2012). Des neurones représentant des règles abstraites, c'est-à-dire indépendants de domaines sensoriels spécifiques, sont également présents dans le cerveau et encodent différents types d'opérations ou de stratégies cognitives. Ce type de neurones a été observé dans plusieurs régions. Par exemple, des neurones codant la règle "*match versus non match*" ont été retrouvés dans les cortex CPFdl, CPFvl et COF, ainsi que dans le cortex prémoteur (Asaad et al., 2000; Wallis et al., 2001; White and Wise, 1999). D'autres études montrent également un encodage de règles abstraites antagonistes au sein d'une seule et même région (par exemple, les stratégies "*win-stay*" versus "*lose-shift*" dans la même région du CPFdl (Genovesio et al., 2005)). L'ensemble de ces études suggère, comme nous l'avons mentionné précédemment, l'existence d'un réseau distribué pour la représentation des règles, fortement dépendant de la

tâche employée. Si tel est le cas, comprendre la communication au sein de ce réseau pourrait être la clé pour en inférer ces propriétés.

Dans cette direction, une étude chez le singe, utilisant une tâche avec deux règles en alternance (couleur versus orientation) a montré que les règles étaient également représentées à l'échelle des oscillations des potentiels de champs locaux (PCL) dans le CPFdl (Buschman et al., 2012). En effet, la synchronisation des PCL permet de former des ensembles dont le degré de synchronie dans les fréquences beta permet de distinguer les règles. Les auteurs ont également montré que des neurones règles-spécifiques se synchronisaient aux ensembles des PCL représentant la règle en cours. Ce mécanisme de changements tâche-spécifiques de la synchronie oscillatoire pourrait être une des façons par lesquelles se ferait la sélection des neurones spécifiques à la tâche pertinente. Cette étude laisse cependant des questions ouvertes, notamment concernant la façon dont le changement de règle (ou de la tâche) est généré, en d'autres mots, comment le changement entre les réseaux s'opère. Une étude chez le rat a montré que, au moment où le comportement de l'animal révèle qu'il a détecté un changement de règle, des ensembles neuronaux dans le CPF médian passaient (dans la plupart des cas) brusquement d'un état représentant la règle connue à un autre état représentant la nouvelle règle (Durstewitz et al., 2010). Les auteurs suggèrent que l'apprentissage des règles serait un processus de décision basé sur l'accumulation d'évidence qui serait accompagné de brusques moments d' "*insight*". Il reste également à déterminer comment ces réseaux oscillatoires locaux interagissent avec les signaux oscillatoires en provenance des autres régions en lien avec la tâche, permettant le bon usage de ces règles.

Maintenant que nous avons une petite idée de comment le cerveau représente les règles, il faut nous pencher sur la façon dont ces règles et leur utilisation pourraient être liées ensemble en un *task-set*.

Maintenance des différents *Task-sets* durant l'exécution d'une tâche

Exécuter une tâche de manière optimale requière la capacité de maintenir en mémoire et d'utiliser le *task-set* en cours. Cela signifie l'application de règles précédemment apprises, et il est possible que la maintenance de ces règles soit le fait du

CPF *via* les patrons d'oscillations locaux (Buschman et al., 2012). De plus, il est possible que d'autres régions aient une influence plus générale dans la maintenance du *task-set*. Par exemple, le CCM et l'insula antérieure/operculum frontal montrent une activation maintenue durant l'intervalle entre l'indice de changement et les cibles quand il est demandé aux sujets d'alterner entre différentes tâches (Dosenbach et al., 2006).

Le CPF pourrait également être impliqué dans ce processus car il montre une activation spécifique en IRMf durant les tâches mixes en comparaison aux tâches simples, ce qui suggèrerait un rôle de cette région pour garder plusieurs *task-sets* actifs (Braver et al., 2003). Le CPF est impliqué dans la représentation de la tâche mais n'est pas crucial pour induire l'activation des aires postérieures tâche-spécifiques, comme démontré par le fait que les patients avec une lésion du CPF montrent une activité pré-tâche soutenue dans ces régions postérieures. Le rôle du CPF pourrait être de permettre une interaction inter-régionale plus efficace entre les aires postérieures pertinentes pour la tâche.

Compétition et alternance (switching) entre les différents *Task-sets*

Il est possible que les êtres humains soient capables de garder plus d'un *task-set* en mémoire de telle sorte à pouvoir changer de manière optimale quand cela est nécessaire. Collins & Koechlin proposent que les humains peuvent maintenir jusqu'à 3 ou 4 *task-sets* en même temps (Collins and Koechlin, 2012). L'alternance entre deux *task-sets* semble impliquer de nombreuses structures. Yeung & Cohen ont montré dans une étude en IRMf chez l'Homme que, durant un changement entre une catégorisation basée sur les mots et celle basée sur les visages, il y avait une activité augmentée dans les cortex CPFl et pariétal (Yeung et al., 2006). Mais des neurones reliés au changement (*switch*) ont également été trouvés dans la région arquée inférieure chez le singe (Kamigaki et al., 2012). Dans cette étude consistant en une tâche d'alternation entre des règles de correspondance (*matching*) basées sur la forme ou sur la couleur, cette région a montré une activité sélective au moment où les singes avaient à changer de règle. Et en effet, des injections de muscimole dans cette zone ont conduit à des altérations doses-dépendantes des performances

d'alternance.

Une façon d'identifier des corrélats de la compétition entre des *task-sets* est de rechercher une activité reliée à un *task-set*, qui aurait une influence sur un autre *task-set*. Comme décrit dans la partie comportementale, des coûts de mixage sont observés lorsqu'on doit gérer deux tâches qui s'alternent. Il y aurait deux explications pour cela. La première idée suggère que les sujets retourneraient à un état neutre à la fin de chaque essai afin de choisir la tâche correcte, même dans les essais de répétition. Le CPFl pourrait être impliqué dans ce phénomène car il est activé de la même façon entre les essais de changement et de répétition dans une tâche de *task-switching* (Johnston et al., 2007). Une seconde possibilité pourrait être que les deux *task-sets* sont maintenus et analysés ensemble, en parallèle, à chaque essai. Dans ce cas, les coûts de mixage correspondraient à la compétition continue entre les deux *task-sets* au moment de la sélection de la réponse (Gilbert and Shallice, 2002). Le cortex prémoteur pourrait être impliqué dans ce choix car différents sets moteurs pourraient être représentés dans cette région (étude de modélisation (Cisek, 2006)). Un autre bon exemple nous est fourni par l'étude de l'inertie d'un *task-set*. Des corrélats neuronaux de l'inertie des *task-sets* ont été trouvés par Yeung & Cohen dans une étude en IRMf incluant deux tâches (catégoriser des mots ou des visages) qui activent des régions séparées (Yeung et al., 2006). Dans ces régions, l'activité sélective de la tâche non pertinente permettait de prédire les coûts comportementaux associés au changement d'une tâche à l'autre. Cette activité pourrait représenter le *task-set* non acteur, en compétition avec le *task-set* en cours. De plus, différentes régions ont été associées à l'inhibition du *task-set*, comme le CPFvl, puisqu'il était actif à la fois au cours des essais de changement et au cours des essais *No Go* dans une tâche de *Go-No Go* dans une étude en imagerie chez l'Homme (Allport et al., 1994). Chez le singe, des études unitaires suggèrent un rôle de la pré-SMA dans l'inhibition des *task-sets* (Isoda and Hikosaka, 2007).

Le *task-switching* pourrait aussi être relié à l'idée de conflit entre 2 *task-sets* en compétition. Dans une tâche où un stimulus peut être associé à deux réponses différentes, le conflit émergerait à deux étapes : au moment de la représentation du stimulus, ou au moment de la sélection de la réponse. Le conflit durant l'étape de

sélection de la réponse a été associé à une activité augmentée dans l'aire motrice pré-supplémentaire, et possiblement dans le CCM chez l'Homme (bien que ce point soit encore débattu, étant donné que les résultats d'imagerie et d'enregistrements unitaires ne se recoupent pas) (Crone et al., 2006; Liston et al., 2006). Le conflit à l'étape de représentation du stimulus a été associé à une activation augmentée du CPFdl pour les stimuli non-pertinents (Liston et al., 2006). Le CCM a également été associé à l'idée de "conscience" du conflit entre les règles d'un *task-set* et la réponse, à l'étape de sélection de la réponse (Dehaene et al., 2003).

Processus préparatoires

Des changements de tâches peuvent être connus en avance ou planifiés dans les cas où ils sont eux même initiés par le sujet, ou quand ils sont signalés ou prédictibles. Dans ces situations, des processus préparatoires sont nécessaires afin de sélectionner le *task-set* de manière optimale. Des analyses en IRM ont montré que les variations dans les patrons d'interactions inter-régionales changent en fonction de la règle. Par exemple, la force de la connectivité fonctionnelle entre des sous-régions du CPF varie différemment en fonction de l'opération qui va être effectuée, reflétant potentiellement le processus préparatoire (Sakai and Passingham, 2003). Le CCM pourrait également jouer un rôle puisqu'il montre une augmentation d'activité dans les tâches comparant une sélection volontaire versus instruite de la tâche à effectuer (Forstmann et al., 2006). On pourrait postuler qu'un changement de tâche volontaire demanderait un plus grand degré de préparation. Dans ce sens, Weissmann et collaborateurs ont montré une activité augmentée dans CCM /CPFm en réponse à un indice signalant une tâche plus difficile, requérant plus de contrôle. Une interprétation proposée serait le rôle du CCM pour diriger l'attention vers les stimuli pertinents pour la tâche en cours (Weissman et al., 2005). De plus, Rushworth et coll. ont montré que le CCM et le CPFm sont importants dans le maintien des associations entre les actions et leurs résultats, et l'implémentation des *task-sets* (Rushworth et al., 2004).

Sohn & Carter ont proposé 2 types de préparation pour le changement : endogène ou exogène (comme décrit dans la partie comportementale de cette partie) (Sohn

et al., 2000). Dans leur étude en IRMf, les sujets humains devaient alterner entre 2 tâches. Ils ont proposé qu'une préparation endogène est portée par le CPFl et le cortex pariétal car, au cours de la préparation, l'augmentation d'activité dans ces régions est supérieure quand les sujets ont la suspicion qu'un switch va arriver, par rapport à quand ils ne l'ont pas. Les auteurs argumentent également que, en contraste, la préparation exogène se déroule dans le CPFdm et dans le pariétal postérieur car, au cours de la préparation, lorsque les sujets ne s'attendent pas à un changement, l'activité dans ces aires est supérieure pour les essais de changement comparés aux essais de répétition.

Une organisation antéro-postérieure ?

On peut résumer les rôles généraux de ces structures sur la base de ces études. Il a été proposé un gradient antéro-postérieur du CPF antérieur au cortex prémoteur pour représenter les éléments du *task-set*, du plus abstrait au plus concret (Badre and D'Esposito, 2009; Koechlin et al., 2003). Le CPF antérieur semble avoir un rôle supérieur pour maintenir actifs plusieurs *task-sets*. Les régions de la surface latérale du cortex préfrontal ont des rôles plus spécifiques, comme la représentation générale de la règle autour du sillon principal, l'inhibition des *task-sets* non-pertinents par le CPFvl et la représentation des règles abstraites par le CPFdl. Plus postérieurement, les sets moteurs et les règles concrètes sont représentés dans les aires prémotrices. Et empiétant sur le tout, le CCM serait impliqué dans la détection des évènements pertinents et la maintenance des associations entre les actions et leurs résultats.

Koechlin et collaborateurs ont proposé une structure théorique de cette idée d'axe antéro-postérieur pour la sélection des actions au sein du CPFl, dans ce qu'ils ont appelé "le modèle en cascade du contrôle cognitif" (Koechlin et al., 2003). Dans ce modèle, le type d'information représenté est très abstrait dans la partie la plus rostrale du préfrontal latéral, et devient progressivement de plus en plus concret lorsqu'on se rapproche des aires prémotrices. Ils attribuent un rôle du CPF fronto-polaire dans le "*branching control*", une fonction permettant de gérer plusieurs tâches à la fois en maintenant dans un état "suspendu" les éléments non pertinents pour le moment, mais potentiellement pertinents plus tard. Ensuite, le CPFl antérieur serait

impliqué dans le contrôle épisodique, permettant de prendre en compte les évènements passés pour orienter le comportement en cours. De manière plus concrète, le CPFdl aurait un rôle dans le *contrôle contextuel*, permettant de prendre en compte les éléments pertinents à partir du contexte en cours. Enfin, à la base de la hiérarchie temporelle, le cortex prémoteur intègrerait toutes les informations de la précédente cascade avec l'information concernant le stimulus lui-même afin de faire le choix final pour orienter l'action, dans ce qu'ils appellent le *contrôle sensori-moteur*. Une étude réalisée sur des patients avec des lésions spécifiques des régions mentionnées dans ce modèle a confirmé leur rôle critique dans les rôles proposés par le modèle cascade, qui était principalement basé sur des résultats de neuroimagerie (Azuar et al., 2014). Cette étude a montré une organisation asymétrique de ces régions : les fonctions nécessitant le plus haut niveau de contrôle cognitif, c'est-à-dire celles portées par les régions les plus antérieures, dépendent de l'intégrité des régions les plus postérieures. Au contraire, les tâches les plus simples peuvent être réalisées par les régions postérieures seules. Ainsi, les plus hauts niveaux de contrôle ne sont engagés que lorsque les plus bas niveaux de contrôle ne peuvent pas réaliser la sélection des actions. Une étude récente a justement proposé un mécanisme électrophysiologique *via* les oscillations cérébrales expliquant comment ces différentes régions étaient engagées ou non en fonction du degré d'abstraction requis par la tâche (Voytek et al., 2015). Les auteurs ont analysé des enregistrements d'électrocorticographie chez des sujets humains réalisant une tâche au cours de laquelle s'alternaient différentes règles exigeant différents niveaux d'abstraction. Ils ont montré que les modulations de l'amplitude des oscillations dans la bande du haut gamma différaient en fonction du degré d'abstraction requis pour générer une réponse. Ces modulations de l'amplitude du gamma étaient organisées, en essai par essai, par un couplage de phase avec les oscillations thêta (couplage phase-amplitude, CPA). Les auteurs ont observé ce CPA au sein des régions du CPF et entre le CPF et le cortex moteur/prémoteur pendant les périodes pertinentes pour la tâche. De manière cruciale, le couplage des régions frontales en direction des régions motrices était d'autant plus fort que le degré d'abstraction requis était élevé ; ce qui montre que la communication des régions frontales vers les régions motrices est nécessaire lorsque ce sont des règles

abstraites qui régissent les actions à réaliser. Les auteurs proposent le CPA comme mécanisme indexant le transfert des informations pertinentes dans le cortex frontal. Ces résultats soutiennent la proposition que les oscillations jouent un rôle crucial pour la coordination des comportements complexes (Lisman and Jensen, 2013).

Les changements non signalés : résoudre le dilemme exploration/exploitation

Dans un environnement volatil où les changements sont non signalés, il est nécessaire de posséder un système permettant d'estimer s'il faut persévérer dans le mode de réponse utilisé ou s'il faut partir en exploration. Nous avons vu, dans la partie sur le comportement, l'importance capitale de la détection de la surprise pour l'apprentissage. Comme nous l'avons vu, le CCM a été impliqué dans la détection et l'attribution de valeur aux évènements inattendus mais pertinents (Vezoli and Procyk, 2009). Ces évènements peuvent être la présence ou l'absence d'un renforcement (signalant éventuellement la nécessité de changer de stratégie) ou un échec dans la production d'une action (signalant la nécessité de réallouer de l'attention dans la réalisation de l'acte moteur par exemple). Ainsi, alors que, dans le système dopaminergique, les même cellules encodent les erreurs de prédiction de la récompense positives et négatives en augmentant ou diminuant respectivement leur taux de décharge (Schultz et al., 1997) ; dans le CCM, cet encodage est réalisé par différentes populations de neurones, et consiste dans tous les cas en une augmentation du taux de décharge (Matsumoto et al., 2007; Quilodran et al., 2008; Sallet et al., 2007). Ces neurones sont également capables de distinguer les erreurs de choix ou d'exécution (Quilodran et al., 2008). Cela suggère que l'erreur de prédiction dopaminergique peut être utilisée directement pour adapter les valeurs des actions, alors que les signaux du CCM représentent un niveau d'abstraction de l'information plus élevé, comme la catégorisation des *feedback*.

Le CCM pourrait avoir un rôle de *régulateur global* du comportement permis grâce à son activité d'intégration du suivi des performances et de la tâche. En effet, des études ont montré une modulation de l'activité du CCM en fonction de l'état d'exploration ou d'exploitation (Procyk et al., 2000; Quilodran et al., 2008), ou entre des périodes volatiles ou stables (Behrens et al., 2007). Les modulations

de l'activité du CCM en exploration et exploitation ont été modélisées grâce à des modèles d'apprentissage par renforcement *via* la régulation d'un paramètre appelé taux d'exploration (Khamassi et al., 2013). Dans une étude chez le rat, l'activité neuronale populationnelle dans le CCM a montré que le réseau basculait entre des états complètement indépendants selon que de la nourriture était introduite ou non dans l'environnement (Caracheo et al., 2013). Les auteurs ont montré que des mesures de l'«entropie" neurale diminuaient fortement lorsque les rats passaient de l'exploration de l'environnement à l'exploitation d'une source de nourriture fiable. Ces résultats rejoignent ceux de Karlson et coll. , qui ont aussi montré chez le rat que l'activité dans les neurones du CCM change de manière abrupte en devenant plus instable lorsque l'animal choisit de basculer en exploration (Karlsson et al., 2012). Les auteurs ont interprété ce changement comme un reset du réseau signalant un état d'incertitude élevé au démarrage de l'exploration, et qui se stabiliserait progressivement en s'approchant d'un état d'exploitation.

Le fait de détecter correctement un changement dépend de la capacité à estimer en direct le niveau de la volatilité environnementale. Behrens et collaborateurs ont utilisé un modèle Bayésien pour suivre les estimations de la volatilité par les sujets et ont montré que le signal BOLD dans le CCM corrélait avec ces estimations pendant la phase de monitoring de l'essai, c'est-à-dire au moment du *feedback* (Behrens et al., 2007). Les niveaux de l'activité du CCM reflétaient en fait la saillance de chaque nouvelle information à prédire les *feedback* futurs. De manière intéressante, ces variations permettaient de prédire les différences de taux d'apprentissage entre les sujets. Ces signaux influençant le taux d'apprentissage ont été identifiés de manière indépendante des signaux représentant l'erreur de prédiction. Cela suggère que les variations dans l'activité du CCM reflètent l'adaptation flexible du paramètre d'apprentissage par renforcement qu'est le *taux d'apprentissage* en fonction des demandes de la tâche. Ainsi, les résultats des études ayant rapporté que l'activité du CCM encode les erreur de prédiction de la récompense sont peut être la conséquence de cette fonction de *metalearning* (Matsumoto et al., 2007; Quilodran et al., 2008).

Un mécanisme permettant l'implémentation des estimations de la volatilité par le CCM pourrait se baser sur le système noradrénergique émanant du locus coe-

ruleus, car il a été aussi montré comme ayant un rôle crucial dans pour basculer entre exploration et exploitation. Selon Aston-Jones et Cohen, les changements associés dans le type d'activité des neurones de cette région (phasique pour l'exploitation, tonique pour l'exploration) pourraient être provoqués par des changements au sein du CCM (ce qui est supporté par de fortes projections cingulaires vers cette structure (Aston-Jones and Cohen, 2005). Cette proposition d'un rôle de la noradrénaline pour la balance exploration/exploitation fait écho aux études montrant son rôle dans le signalement de l'incertitude inattendue (Bouret and Sara, 2005; Yu and Dayan, 2005). En effet, le système noradrénergique préfrontal serait engagé dans les situations impliquant un changement des contingences stimulus-réponse-conséquence (Dalley et al., 2001). Des expériences de Preuschoff et coll. ont utilisé la dilatation pupillaire (la taille de la pupille corrèle avec la noradrénaline, à la fois chez l'Homme et l'animal) pour montrer que la noradrénaline signale le degré d'incertitude inattendue (Preuschoff et al., 2011). La taille de la pupille serait en effet étroitement corrélée avec l'incertitude inattendue, et ce, de manière indépendante de l'incertitude attendue. La noradrénaline faciliterait donc les *Shifts* attentionnels et cognitifs pour l'adaptation du comportement, en signalant quand les attentes sur le monde nécessitent d'être révisées. Preuschoff et coll. ont fait la distinction entre les changements de contingences rares (auxquels elle réfère comme à l'incertitude inattendue) ou fréquents (volatilité). Une activité de type phasique serait impliquée dans le premier cas alors que la volatilité serait signalée de manière tonique par de hauts niveaux de noradrénaline (Yu, 2007).

Représentations de la fiabilité des stratégies

Enfin, lorsque le monde est très changeant, et que cette volatilité est difficile à évaluer, il est nécessaire d'évaluer la fiabilité de la stratégie utilisée en permanence et de la comparer avec celle d'autres stratégies potentielles, au cas où certaines deviendraient plus avantageuses. Donoso et collaborateurs ont utilisé la tâche de manipulation de *task-sets* et le modèle PROBE présentés dans le papier d'Anne Collins (Collins and Koechlin, 2012) pour trouver des corrélats de la fiabilité des différentes stratégies dans le signal BOLD (Donoso et al., 2014). Ils ont trouvé que

le cortex préfrontal comportait deux réseaux dévoués à ces évaluations. Le signal en provenance d'un réseau de régions allant du cortex préfrontal ventromédian au dorsomédian corrélait avec la fiabilité de la stratégie en cours, permettant d'arbitrer entre l'ajustement de cette stratégie et l'exploration de nouvelles stratégies. Un deuxième réseau impliquant les régions du cortex préfrontal polaire à latéral corrélait quant à lui avec la fiabilité de deux ou trois stratégies alternatives, permettant l'arbitration entre l'exploitation de ces alternatives et l'exploration de toutes nouvelles stratégies. Une question à présent serait de comprendre comment ces signaux de fiabilité modulent le réseau et comment ces stratégies alternatives sont récupérées ou mises de côté en fonction, et il semble de nouveau pertinent de proposer d'aller chercher du côté des oscillations pour comprendre les mécanismes sous-jacents.

En résumé, au cours de cette première partie, nous avons détaillé les processus permettant d'apprendre de l'environnement incertain. Nous avons également détaillé les régions qui semblent permettre un tel apprentissage et l'estimation des différentes sources d'incertitude. Ainsi, toutes ces recherches ont permis l'identification d'un réseau étendu et complexe de régions, corticales et sous-corticales, semblant supporter ces processus. Cependant, ces études ne sont pas suffisantes pour qui veut vraiment comprendre comment le cerveau permet aux organismes d'apprendre de l'environnement. Un aspect reste à creuser, maintenant que ce réseau de structures a été identifié : il s'agit de mieux comprendre comment ce réseau fonctionne...en tant que réseau. En effet, toutes ces différentes informations représentées dans le cerveau - qu'il s'agisse de la représentation des stimuli, des actions, et de leur lien avec les récompenses, en passant par les stratégies adoptées, les stratégies alternatives, et le niveau de fiabilité de ces stratégies, jusqu'aux estimations des différents niveaux d'incertitude de l'environnement - toutes ces représentations doivent nécessairement être partagées, liées et s'influencer au sein d'un réseau dynamique entre les différentes régions identifiées, afin de permettre un comportement flexible. Ainsi, un nouveau défi de la recherche en neuroscience sur la prise de décision doit être de comprendre le fonctionnement en réseau de ces régions. C'est le but à long terme de mon projet de thèse, que je décrirai dans la partie "Questions de thèse".

Mais auparavant, il est nécessaire de pointer un deuxième point sur lequel ces études ne se sont pas penchées. Tous les processus et mécanismes que nous avons décrits jusqu'à présent sont plus ou moins rapides selon les espèces et les individus, et sont globalement efficaces et adaptés à l'environnement... sauf dans certaines situations où ils sont sous-optimaux, et biaisés dans des directions qui ne semblent pas logiques. Une possibilité serait que ces mécanismes sont adaptés pour l'environnement dans lequel ils se sont développés, ce qui expliquerait pourquoi dans certains cas ils sont adaptés, et dans d'autres, ils sont contre-productifs. Plusieurs études font pencher la balance en faveur de cette proposition. En effet, les caractéristiques de l'environnement dans lequel un organisme grandit ou évolue (Biernaskie et al., 2009; Tebbich and Teschke, 2014), ou les a priori d'un organisme sur son environnement (Hertwig et al., 2004; Payzan-LeNestour and Bossaerts, 2011), influencent fortement la façon dont cet organisme prend des décisions. Dans la deuxième partie de cette introduction, nous nous pencherons donc sur les mécanismes qui sont susceptibles d'influencer (voire d'améliorer) les mécanismes développés par les organismes pour apprendre de leur environnement incertain.

Chapitre 2

Apprendre à apprendre, ou comment raffiner et généraliser les processus pour apprendre de l'environnement

Dans cette deuxième partie, nous nous concentrerons sur le processus d'*apprendre à apprendre* en proposant l'idée qu'il est à l'origine de la mise en place et de la mise au point des mécanismes d'apprentissage utilisés pour faire face à l'environnement incertain, décrits dans la première partie. Nous essaierons de tisser des liens entre cette littérature, celle sur l'apprentissage latent et celle opposant les apprentissages dits *model based* et *model free*.

2.1 Processus comportementaux

2.1.1 *Learning-set*

Comme nous l'avons abordé dans la première partie, même des organismes très simples sont capables d'apprendre à partir de leur expérience en renforçant les associations impliquant les stimuli ou actions désirés (Dickinson, 1985). Cependant, si ces processus d'apprentissage par renforcement conduisent à un comportement

adapté à l'environnement en cours, celui-ci n'est pas forcément flexible, ni très efficace. L'apprentissage par renforcement est un processus lent : une action ou un stimulus est associé à une récompense au cours d'un certain nombre d'essais. Cependant, il est possible d'observer, notamment chez les primates, un processus beaucoup plus rapide, en particulier lorsqu'ils engagent leur contrôle cognitif. Un enfant de 3 ans ou un singe macaque naïf mettent longtemps à apprendre à choisir entre un objet récompensé A et un objet non-récompensé B, mais une personne adulte n'a besoin que de 2 ou 3 essais pour faire l'association (Ochoki et al., 1975). La raison de cette différence est que cette personne a *appris à apprendre de manière efficace*.

Au cours de cette partie, nous développerons des arguments pour montrer que le processus d'apprendre à apprendre ne consiste pas en une simple amélioration du taux d'apprentissage avec l'expérience. Nous montrerons qu'il s'agit d'une stratégie implémentée, qui permet d'acquérir de l'*efficacité* et de la *flexibilité* et qui repose de manière importante sur l'utilisation de la mémoire prospective.

Apprentissage intra- et inter-problème

Le processus d'*apprendre à apprendre* est un processus d'amélioration de l'apprentissage *inter*-problème car il s'acquière au cours d'un grand nombre de problèmes rencontrés dans une tâche, et conduit à l'acquisition d'un *learning-set*. Ce processus a longtemps été considéré comme une version plus évoluée de l'apprentissage classique (*intra*-problème), jusqu'à ce que Harlow propose une façon de le quantifier de manière expérimentale (Harlow, 1949) et que des études suggèrent qu'il s'agit de processus distincts, portés par des régions cérébrales différentes. La Figure 2.1 montre la différence entre les processus d'apprentissage intra- et inter-problème, dans une expérience menée chez le singe (Harlow, 1949). L'apprentissage des premiers problèmes de discrimination (1 à 8) est lent, et la courbe d'apprentissage *intra-problème* est typique d'un apprentissage par essai-erreur. Aucune amélioration n'est observée sur les premiers essais puis la courbe montre une accélération positive. Au fur et mesure que les animaux rencontrent des problèmes supplémentaires, un gain de performance progressif est observable sur les premiers essais de chaque problème, et la forme de la courbe devient plus sigmoïde, montrant une accélération négative.

Après une centaines de problèmes rencontrés, la courbe d'apprentissage montre une cassure entre l'essai 1 et l'essai 2 faisant passer les performances de la chance à la perfection ou quasi-perfection en un essai (Harlow, 1949). Ainsi, le processus d'apprendre à apprendre ou apprentissage *inter-problème* ou le *learning-set* permet de rendre les décisions plus efficaces. Nous définirons un *learning-set* comme l'amélioration progressive des performances jusqu'à une asymptote où elles deviennent stables à un niveau optimal ou quasi-optimal. Nous verrons à présent que cette amélioration des performances n'est pas due à une amélioration du processus d'apprentissage intra-problème.

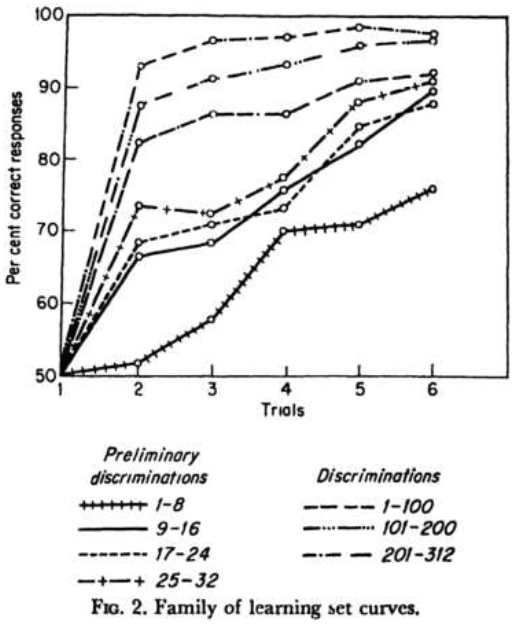

FIG. 2. Family of learning set curves.

Figure 2.1 – Courbes d'apprentissage de discrimination au cours des blocs successifs de problèmes, chez le singe macaque. Cette figure met en évidence la différence entre l'apprentissage intra-problème (courbe des problèmes 1 à 8) et inter-problème (autres courbes). D'après (Harlow, 1949).

Eléments en faveur d'une différence fondamentale entre apprentissages intra- et inter-problème

Certains peuvent s'interroger sur l'existence d'une réelle différence entre l'apprentissage intra-problème et l'apprentissage inter-problème. Un premier élément en faveur d'une différence entre les deux processus est un argument développemental. En effet, cette capacité à améliorer l'apprentissage au fur à mesure des problèmes

n'émerge qu'après un certain âge, à la fois chez l'homme et le singe. Des singes rhésus peuvent résoudre des problèmes individuels de discrimination dès l'âge de 60 jours, en 50 à 100 essais. Mais ces mêmes animaux sont complètement incapables de former un *learning-set* de discrimination. Cette capacité ne commence à se développer qu'à partir de 150 jours et n'est vraiment efficace qu'à partir de la deuxième ou troisième année (Harlow, 1959).

Le deuxième élément en faveur d'une différence entre apprentissage intra- et inter-problème est que certains animaux présentent des difficultés à réaliser le deuxième (Figure 2.2). Harlow rattache ces différences inter-spécifiques à la position phylogénique, où seules les espèces les plus "évoluées" auraient cette capacité (Harlow, 1949). Une étude chez le chat a montré qu'ils parvenaient à acquérir un *learning-set*, de discriminations visuelles mais les courbes d'apprentissage atteignaient un niveau de performance beaucoup plus faible que chez le singe (Warren and Baron, 1952). Des études chez le rat montrent qu'ils atteignent rarement les mêmes performances que les primates pour des discriminations visuelles (Koronakos and Arnold, 1957; Bailey and Thomas, 2001). Mais ces conclusions sont peut-être à considérer avec précaution. Dans l'étude de Koronakos et Arnold, les rats font preuve d'un apprentissage inter-problème limité (Koronakos and Arnold, 1957). Dans cette étude, 20 rats ont été entrainés sur une série de problèmes de discrimination visuelle avec un critère de performance de 80% (ou un maximum de 80 essais si le critère n'était pas atteint). Seulement 5 rats ont réussi à atteindre le critère sur tous les problèmes et leur courbe de performance s'est améliorée atteignant un maximum de 70% de réponses correctes sur les essais 1 à 20 du problème 6. Il faut noter cependant que dans cette étude, le nombre de problèmes donnés aux animaux était petit, et n'a peut-être pas été suffisant pour permettre la formation d'un *learning-set*. Une autre étude montre que les rats peuvent acquérir des *learning-set* sur certaines tâches dont une tâche de discriminations visuelles avec des *reversals* mais les performances n'excèdent jamais 80% (Bailey and Thomas, 2001). Cependant, Warren remarque qu'une comparaison inter-spécifique n'est vraiment pertinente que si les tests sont adaptés aux espèces en question, en particulier pour ce qui concerne les modalités sensorielles testées (Warren, 1965). Ainsi, les rats ont de bien meilleures performances dans une tâche

d'acquisition d'un *learning-set* lorsqu'elle est basée sur des discriminations *olfactives* que visuelles, en parvenant même aux performances parfaites sans erreur dès le 2ème ou 3ème essai (Slotnick et al., 2000; Slotnick and Katz, 1974). Par contre, à notre connaissance, il n'existe pas d'étude chez le rat montrant un transfert réussi à une nouvelle tâche (par exemple de *reversals*) après acquisition d'un *learning-set*. Ainsi, si les rats possèdent la capacité de d'améliorer l'efficacité de leur apprentissage au cours des problèmes, il leur manque peut-être la flexibilité conférée par le processus d'apprendre à apprendre.

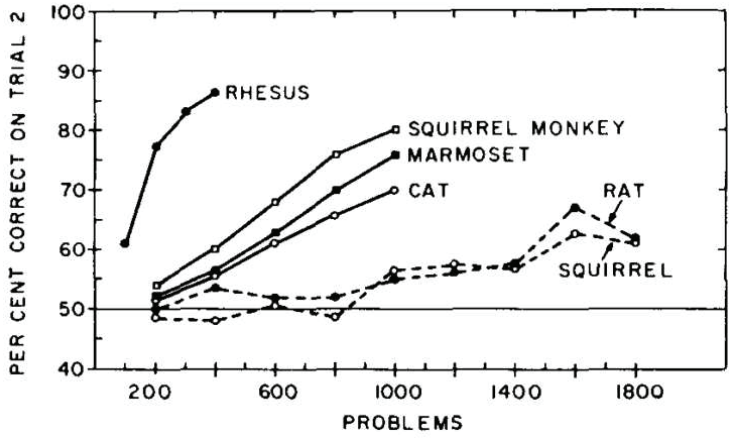

Figure 2.2 – Différences inter-spécifiques de performance à l'essai 2 au cours des problèmes de discrimination visuelle chez les primates, carnivores et rongeurs, d'après (Warren, 1965).

Le dernier élément suggérant que le processus d'apprendre à apprendre ne consiste pas en une simple amélioration du taux d'apprentissage avec l'expérience vient de l'expérience de Murray et Gaffan, qui suggère que le processus d'association reste stable au cours de l'apprentissage de problèmes de discrimination (Murray and Gaffan, 2006). En effet, des singes, chez lesquels la capacité d'apprendre à apprendre a été bloquée comportementalement en espaçant les essais d'un même problème par un intervalle de 24h, montrent une courbe d'apprentissage intra-problème stable même après une centaine de problèmes rencontrés (Figure 2.3). De plus, des études de lésions menées par la même équipe à Oxford montrent également que léser le cortex préfrontal conduit à des déficits d'acquisition du *learning-set* tout en laissant indemne l'apprentissage intra-problème, ce qui suggère qu'ils reposent sur des mécanismes séparés (Browning et al., 2007). L'apprentissage intra-problème, ou par

essai-erreur, a fait l'objet de nombreuses investigations, à la fois pour en comprendre les mécanismes computationnels, mais aussi pour en trouver les corrélats neurophysiologiques, et commence à être bien compris. Mais, si l'apprentissage inter-problème ne consiste pas en l'amélioration du processus intra-problème, en quoi consiste-t-il exactement ? Nous verrons qu'il s'agit, au moins dans le cas des *learning-set* de discrimination, de la capacité à implémenter une stratégie, et de manière cruciale, en se basant sur la mémoire prospective.

Learning-set et transfert

Un premier élément spécifique au processus d'apprendre à apprendre est que des sujets ayant appris à apprendre cessent de résoudre des problèmes sur la base de l'identité des stimuli individuels, utilisant à la place une *règle comportementale* qui peut être appliquée à tout un panel de stimuli différents (Wilson et al., 2010). Cette *capacité d'abstraction* est couramment utilisée chez des sujets les plus sophistiqués faisant preuve de contrôle cognitif pour réaliser des choix adaptés. Cette capacité est à l'origine d'un élément critique de processus d'apprendre à apprendre : la maitrise facilitée de nouvelles tâches. Cela signifie que ce processus permet un certain niveau de *transfert* possible entre différentes tâches. Une des rares études à avoir suggéré cet effet est celle de Schrier, où il montre que des macaques ayant acquis un *learning-set* pour une tâche de discrimination sérielle montrent des coûts de transfert significativement réduits pour une tâche de *reversal* d'associations stimulus-récompense, par rapport à des singes sans *learning-set* (Schrier, 1966). Il est intéressant de noter que ces tâches semblent à première vue assez différentes. En effet, dans le premier cas, les singes doivent apprendre à discriminer successivement une grande série d'objets, alors que dans la deuxième tâche, ils doivent apprendre, oublier et ré-apprendre de manière répétée un set limité d'objets. En fait, ce qui lie ces deux tâches est que la même règle comportementale peut être appliquée dans les deux cas : une règle de *win-stay lose-shift*. Cette règle a pu être acquise au cours de l'apprentissage de la première tâche et peut également s'appliquer à la deuxième (Wilson et al., 2010). Dans cet exemple, les tâches partagent suffisamment de similarités pour qu'une simple règle d'abstraction permette le transfert de l'une à l'autre. En l'occurrence,

l'étude de Jang suggère que l'acquisition d'un *learning-set*, dans le cas de l'apprentissage des *reversals* chez le macaque, consiste en l'acquisition d'une règle *lose-shift* (plutôt qu'une règle win-stay qui est déjà utilisée de manière forte au début de l'apprentissage et se maintient au cours de l'acquisition) (Jang et al., 2015). Nous n'argumentons pas qu'un *learning-set* facilite le transfert entre n'importe quels types de tâches, en particulier si elles ne partagent pas des règles ou pré-requis similaires. L'idée défendue ici est plutôt que le *learning-set* représente une des formes d'abstraction les plus précoces et simples, constituant ainsi une étape d'apprentissage importante pour la construction cognitive abstraite.

Learning-set et mémoire prospective

Cette stratégie se base de manière cruciale sur la mémoire *prospective*. C'est-à-dire que cette règle permet de prendre une décision à propos d'un stimulus ou d'une situation spécifique *en avance*. Un singe naïf devant discriminer un objet récompensé A d'un objet non récompensé B aura besoin de sélectionner les deux options pour apprendre sur chacune et changer leur valeur de renforcement respective. Mais un singe avec un *learning-set* pour cette tâche saura déjà quoi faire à propos de l'objet B après avoir choisi l'objet A, et ce, en un essai seulement (d'où son aspect prospectif). Un contrôle cognitif efficace permettant l'application d'un *task-set* approprié à un contexte donné requiert précisément cet aspect prospectif. Il n'est en effet pas très utile de posséder une règle dans le contexte K s'il est nécessaire de tester tous les éléments de ce contexte pour confirmer que cette règle est bien applicable dans ce contexte.

L'importance de la mémoire prospective dans la formation d'un *learning-set* a été démontrée par Murray et Gaffan (Murray and Gaffan, 2006). Ils démontrent en effet que l'usage de la mémoire prospective est une condition *sine qua non* à la formation d'un *learning-set*. Un entrainement classique sur une tâche de discrimination chez le singe consiste à faire résoudre un par un, de manière successive, les problèmes à l'animal, en les faisant apprendre jusqu'au critère. Dans leur étude, Murray et Gaffan ont empêché l'utilisation de la mémoire prospective en séparant chaque essai d'un même problème d'un intervalle de 24h. Les animaux devaient toujours apprendre les

différents problèmes au critère, sauf qu'au lieu d'être présentés de manière sérielle, ils étaient présentés par groupe de 10, à raison d'un essai par jour pour chaque problème. Les résultats sont présentés dans la Figure 2.3 en comparaison avec ceux de Harlow de 1949. De manière intéressante, cette manipulation n'a pas empêché les singes d'apprendre la tâche sur les 100 premiers problèmes, et ce, de manière strictement identique à un apprentissage classique sériel (comme dans l'étude de (Harlow, 1949)). La différence apparait en fait au-delà de ces 100 premiers problèmes, lorsqu'un entrainement classique conduit normalement à l'apparition du *learning-set*, observable par une forte amélioration des performances dès le 2ème essai de chaque problème. Les singes de l'étude de Murray et Gaffan ne montrent pas cette amélioration, leur courbe restant identique à celle des 100 premiers problèmes. Ces singes sont donc capables d'apprendre, mais sont incapables de former un *learning-set*. Cette étude consiste en une preuve supplémentaire de la distinction entre les processus d'apprentissage intra- et inter-problème. Il est important de noter que : 1) La différence n'est pas due à une difficulté plus grande de la tâche (en effet, les singes apprennent tout aussi bien les 100 premiers problèmes). L'acquisition d'un *learning-set* ne consiste donc pas à une amélioration de la capacité d'association. 2) Dans les deux cas, les singes reçoivent le même entrainement (ils passent le même nombre d'essais totaux sur la tâche). Cela montre que l'acquisition d'un *learning-set* ne consiste pas juste à "apprendre plus". La différence réside dans la façon dont cet apprentissage est réalisé, et l'élément crucial ici est l'utilisation de la mémoire prospective.

En résumé, le développement d'un *learning-set* permet non seulement de rendre les choix plus *efficaces*, mais aussi plus *flexibles*, et surtout *généralisables* via l'‘acquisition d'une stratégie prospective. Ces éléments sont les conditions *sine qua non* de l'expression du contrôle cognitif. C'est pourquoi, ils permettent de suggérer que le processus d'*apprendre à apprendre* est à l'origine de la mise au point et de l'optimisation des processus du contrôle cognitif.

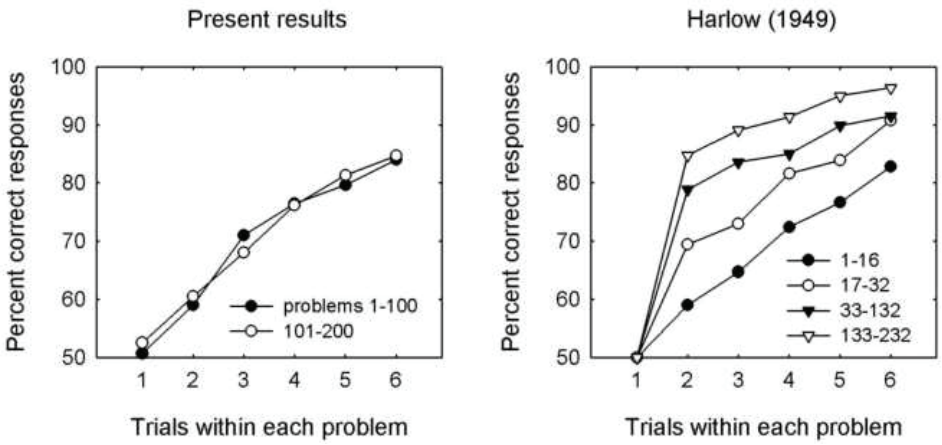

Figure 2.3 – Comparaison des courbes d'apprentissage de discrimination concurrente dans l'étude de (Murray and Gaffan, 2006) à gauche et (Harlow, 1949) à droite. On n'observe aucune acquisition d'un *learning-set* au cours des problèmes dans la première figure. La différence entre les deux protocoles est l'intervalle temporel entre deux essais : de 24h pour l'étude de Murray et Gaffan et immédiat dans l'étude originale de Harlow. D'après Murray et Gaffan, l'intervalle de 24h empêche la formation de la mémoire prospective, cruciale à la formation d'un *learning-set*.

Learning-set dans un environnement incertain

Nous avons vu dans la première partie (voir partie 1) que différents processus liés à l'expression du contrôle cognitif sont utilisés pour prendre des décisions dans des environnements nécessitant de gérer plusieurs sources d'incertitude. Nous avons vu également que ces comportements étaient complexes et couteux à réaliser, soumis à des biais, et qu'il existe une variabilité entre les espèces, mais aussi au sein d'une même espèce, entre les individus. Une hypothèse qui émerge de ces conclusions et des recherches sur le *learning-set* est que cette variabilité pourrait être liée aux caractéristiques de l'environnement dans lequel les individus évoluent et à la manière dont cela influencerait leur façon d'*apprendre à apprendre*. A notre connaissance, il n'existe pas vraiment d'étude qui se soit penchée sur la question mais quelques données existantes vont dans le sens que, dans la nature, évoluer dans un environnement volatil favorise la flexibilité comportementale. Par exemple, une étude chez le pinson a montré que les individus vivant dans une zone avec une grande variabilité dans la disponibilité des ressources de nourriture ont de meilleures performances dans des tâches de *reversal* que d'autres individus vivant dans une zone plus stable (Tebbich and Teschke, 2014). De la même façon, le fait que les chimpanzés seraient moins aversifs au risque que les autres primates non humains serait lié au degré d'incerti-

tude dans leur environnement naturel (Heilbronner and Hayden, 2015). Des études chez le rat, mais aussi le bourdon, ont montré que leur capacité à détecter un changement dans le taux de récompense dépendait de leur expérience concernant les changements de l'environnement acquise lors des entrainements précédents (Biernaskie et al., 2009; Gallistel et al., 2001). Ensuite, comme nous l'avons vu dans la première partie, les *a priori* sur la volatilité de l'environnement, qui sont possiblement acquis lors des premières expériences dans cet environnement, influencent également fortement la façon de réagir à un changement. Par exemple, dans l'expérience de Payzan-Lenestour et Bossaerts, les sujets étaient incapables de détecter les changements lorsque l'éventualité de leur existence ne leur était pas suggérée dans des consignes (Payzan-LeNestour and Bossaerts, 2011). De la même façon, le degré d'aversion au risque dépend de la façon dont on rencontre les évènements stochastiques, comme l'a montré l'expérience de Hertwig et coll. comparant décisions instruites et décisions à partir de l'expérience (Hertwig et al., 2004). Toutes ces études contribuent à l'idée que la façon dont on apprend à l'origine influence de manière très forte la façon de se comporter par la suite. Comprendre les mécanismes d'acquisition des *learning-set* dans les environnements incertains permettrait donc d'apporter des éléments pour comprendre l'origine des comportements actuels face au risque et à la volatilité. Plusieurs études ont montré que l'acquisition de *learning-set* dans des environnements incertains est possible (acquisition d'un *learning-set* en contexte de volatilité : études de *learning-set* avec des *reversals*, et en contexte de stochasticité : apprentissage des tâches probabilistes). Cependant, il n'existe pas vraiment d'étude qui ont cherché à comprendre les mécanismes permettant l'acquisition d'un *learning-set* dans de telles conditions, et encore moins lorsque les différents types d'incertitude sont entremêlés (c'est-à-dire dans des tâches d'apprentissage avec des *reversals* probabilistes). Nous verrons que ces questions constituent une grande partie de mon travail de thèse.

Agir de manière efficace et flexible dans un environnement volatil, comme c'est le cas après l'acquisition d'un *learning-set* pour une tâche de *reversals*, nécessite de maitriser l'idée que l'environnement peut changer, pour réagir de manière appropriée. Le fait que l'environnement soit changeant peut être considéré comme une

information cachée ou latente sur la *structure* de l'environnement, car le sujet n'en est pas forcément informé et doit l'apprendre par lui-même. De même, au cours des expériences de transfert entre deux tâches après acquisition d'un *learning-set*, il a été suggéré que les caractéristiques importantes de la première tâche sont rendues abstraites, ce qui a pour conséquence de faciliter l'apprentissage des nouvelles tâches. D'un point de vue computationnel, cette façon d'extraire des invariants peut être considérée comme une façon d'apprendre la structure sous-jacente (Braun et al., 2010). Ainsi, dans la partie suivante, nous allons essayer de mettre en lien la littérature du *learning-set* avec la littérature sur l'apprentissage de la structure latente.

2.1.2 L'apprentissage de la structure

L'apprentissage dit latent fait d'abord référence de manière historique aux expériences de Tolman d'apprentissage en absence de récompense chez le rat (Tolman and Honzik, 1930). Dans cette expérience, l'apprentissage spatial dans un labyrinthe de 3 groupes de rats est comparé pendant 17 jours. Les rats du premier groupe reçoivent une récompense dès qu'ils atteignent la sortie du labyrinthe. Les rats du second groupe suivent le même entrainement à l'exception qu'ils ne reçoivent pas de récompense à la sortie du labyrinthe pendant les 10 premiers jours, puis reçoivent une récompense au cours des 7 jours suivants. Le troisième groupe ne reçoit jamais de récompense. Les rats du groupe 1 apprennent plus vite que les autres pendant les 10 premiers jours. Cependant, les rats du 2$^{\text{ème}}$ groupe rattrapent leur retard et finissent avec de meilleures performances que le premier groupe à partir du jour où la récompense a été introduite. Tolman a interprété ces résultats comme le signe qu'un apprentissage s'est tout de même déroulé dans le 2$^{\text{ème}}$ groupe pendant la première période, malgré l'absence de renforcement. Ces idées vont à l'encontre des théories comportementales majoritaires de l'époque qui considéraient que l'apprentissage d'une association entre un stimulus et une réponse ne s'effectue qu'en présence d'un renforcement, même si d'autres chercheurs de l'époque, comme Guthrie, avaient proposé d'autres théories allant plus dans le sens de Tolman (par exemple, que la contiguïté entre un stimulus et une réponse était responsable de l'association, plutôt que le renforcement qui en résultait) (revue (Jensen, 2006)). Ces théories

alternatives proposaient que la *réponse* elle-même soit suffisante pour provoquer un changement de comportement en changeant l'orientation de l'organisme vis-à-vis du stimulus (comme par exemple, entrer dans un cul-de-sac d'un labyrinthe). Ce type d'apprentissage peut être aussi un argument pour le rôle des renforcements négatifs dans l'apprentissage (un cul-de-sac pouvant être considéré comme une punition). Ces résultats peuvent également être interprétés comme l'apprentissage de la *structure* de l'environnement.

L'apprentissage de la structure et des causes latentes

Les êtres humains montrent une forte tendance à trouver de l'ordre dans le chaos (Sun et al., 2015). En effet, lorsqu'on leur demande de produire des séquences aléatoires, ils en sont généralement incapables : ils utilisent les alternances en excès et évitent les répétitions. Sun et collègues ont développé un modèle permettant de reproduire les choix des sujets dans une tâche de production de séquences aléatoires. Ils ont montré que ce modèle développe une représentation biaisée en faveur de l'alternance, en étant très sensible à la structure statistique émergeant naturellement des séquences aléatoires. Ce modèle montre que notre biais comportemental face à ce qui est aléatoire constitue en fait un mécanisme d'apprentissage efficace de la *structure statistique* de l'environnement intégrée dans les séquences aléatoires (Sun et al., 2015). Les études de Collins et Frank montrent que les humains utilisent les structures par défaut pour apprendre une tâche, même si cela n'apporte aucun avantage comportemental (Collins et al., 2014; Collins and Frank, 2013). Ils montrent des transferts facilités ou plus difficiles à une nouvelle tâche selon la structure apprise par les sujets lors d'une première tâche. Dans une étude suivante, ils montrent également que cette capacité à créer des règles hiérarchiques généralisées était présente chez l'enfant de 8 mois, suggérant que le cerveau humain est prédisposé à extraire de l'information de l'environnement bruité (Werchan et al., 2015).

Cette idée d'une recherche des structures cachées comme mécanisme d'apprentissage efficace est défendue par Gershman et Niv, qui argumentent que les théories d'apprentissage par renforcement proposent un mécanisme trop lent pour expliquer la façon dont les individus font des choix dans le monde réel (Gershman et al.,

2010). A la place, ils proposent donc que la capacité des animaux et des humains à apprendre rapidement de nouveaux problèmes repose sur leur capacité à tirer avantage du haut degré de structure présent dans les tâches naturelles. Par exemple, dans le cas de la perception, les situations où l'évidence sensorielle disponible est faible poussent le cerveau à faire des inférences sur la structure cachée du monde.

Un des problèmes rencontrés pour inférer une structure à partir des observations est que, aux yeux du sujet, il existe un large éventail (possiblement infini) de structures pouvant expliquer l'environnement. Dans ce cas, il est nécessaire d'identifier la *structure causale* la plus pertinente. Gershman et Niv proposent un cadre normatif pour expliquer comment le cerveau résout le problème de l'apprentissage de la structure, et la façon dont cela peut alimenter les mécanismes d'apprentissage par renforcement (Gershman et al., 2010). Ils utilisent un modèle Bayésien pour expliquer comment mettre à jour les croyances probabilistes sur les structures causales, en réaction à l'arrivée de nouvelles informations. Ce modèle leur permet de réinterpréter les résultats des expériences classiques de conditionnement. Par exemple, les théories d'association classiques ont du mal à expliquer le fait qu'une association originelle entre un son et un choc ne soit jamais complètement désapprise après une étape d'extinction. En effet, il suffit de remettre l'animal dans le contexte original de l'acquisition pour que la réponse de crainte en réaction au son réapparaisse, suggérant qu'il prédit de nouveau l'occurrence d'un choc (ce qu'on appelle l'effet de renouvellement ou *renewal*) (Bouton, 2004). Selon ces chercheurs, ce résultat a du sens si on considère que les réponses conditionnées de l'animal reflètent ses inférences sur les structures latentes de l'environnement, à chaque étape de l'expérience. Ainsi, l'animal considèrerait que les sons et les chocs sont générés par une même cause latente durant la phase d'acquisition, et que le profil de sons non associés au choc est généré par une autre cause latente pendant la phase d'extinction. Par conséquent, le fait de replacer l'animal dans son contexte d'acquisition va le conduire à considérer que la cause présente lors de l'acquisition est de nouveau d'actualité, et à prédire de nouveau des chocs. Redish et collaborateurs proposent une théorie computationnelle pour expliquer ces effets de renouvellement (Redish et al., 2007). Ils proposent un mécanisme permettant de "diviser en états" (*state splitting*), c'est-à-

dire de créer de nouveaux états lorsque les statistiques perçues changent de manière radicale. L'idée serait que la création de nouveaux états est déclenchée lorsque le panel des états à disposition se montre inadéquat pour obtenir un comportement adapté. Ce mécanisme rappelle le modèle d'évaluation et création de *task-sets* de Collins et Koechlin, dans lequel un nouveau *task-set* est créé si aucun des *task-sets* en stock n'est suffisamment fiable (Collins and Koechlin, 2012).

Braun et collaborateurs proposent de voir l'acquisition d'un *learning-set* comme l'apprentissage de la structure de la tâche (Braun et al., 2010). Par exemple, l'apprentissage de la structure a été proposé comme une explication au fait que les individus construisent des généralisations basées sur une quantité limitée d'information (Kemp et al., 2004). Kemp et collaborateurs proposent un modèle computationnel Bayésien d'apprentissage *causal* pour expliquer comment apprendre à apprendre permet d'acquérir des connaissances abstraites qui peuvent être pertinentes pour d'autres tâches (Kemp et al., 2010). Leur modèle consiste en plusieurs représentations à différents niveaux d'abstraction. Le niveau le plus haut permet d'organiser les objets en catégories et de spécifier la puissance causale et les caractéristiques principales de ces catégories. Pour tester leur modèle, ils ont étudié l'apprentissage causal en un essai (*one shot learning*) chez des sujets humains et ont montré que le niveau supérieur du modèle permet d'expliquer comment les sujets font des inférences sur un nouvel objet avec très peu de données disponibles. Cette étude renforce l'idée qu'un *learning-set* représente une forme d'abstraction facilitant l'apprentissage.

2.1.3 *Model-free* vs *model-based*

Un autre domaine, celui faisant la distinction entre les apprentissages dits *model-based* et *model-free*, s'intéresse aux changements d'apprentissage qui surviennent entre le début et la fin de l'entrainement. Dans cette partie, nous essaierons de voir dans quelles mesures on peut relier cette littérature à celle du *learning-set*.

Les modèles d'apprentissage par renforcement dits *model-based* incluent une notion d'état interne, un modèle de la structure de l'environnement, qui est une représentation explicite du monde et des règles qui le régissent, sur lequel le sujet peut baser son apprentissage. Contrairement à l'apprentissage par renforcement dit *model-*

free, où il est nécessaire d'essayer toutes les actions possibles, l'apprentissage *model-based* utilise les représentations internes des différents états qu'il s'est construit. Traditionnellement, les apprentissages par renforcement *model-free* et *model-based* ont été différenciés en fonction de leur sensibilité à la dévaluation (Daw et al., 2005). Du fait que les animaux deviennent de moins en moins sensibles à la dévaluation avec l'entrainement, il a été proposé que le surentrainement consiste en un passage d'un système *model-based* à un système *model-free* (Daw et al., 2005; Wan Lee et al., 2014). Daw et collaborateurs proposent qu'au début de l'apprentissage, l'animal utilise un "système de recherche en arbre", utilisant des prédictions à court terme sur les conséquences immédiates de chaque action. Cette méthode est couteuse en mémoire, en temps et prompt aux erreurs. Mais comme les prédictions sont construites en direct, elle permet une certaine flexibilité pour réagir rapidement à un changement des circonstances, comme dans le cas où le renforcement est dévalué. C'est la caractéristique d'un comportement dirigé vers un but. Au contraire, avec le surentrainement, l'animal se construit des représentations plus stables des valeurs de chaque action. C'est une méthode dite de "*caching*", dans laquelle chaque action est associée à une future valeur attendue, qui n'est pas modifiée par les résultats en cours. C'est la caractéristique d'un comportement de type habitude. La transition d'un système à l'autre ne serait pas irréversible : les individus alterneraient probablement entre les deux systèmes en permanence. En effet, l'étude de Daw et collaborateurs montre que la transition entre les deux systèmes au cours de l'entrainement dépend notamment de la complexité du choix (Daw et al., 2005). Par exemple, lorsque l'animal doit choisir parmi deux actions pour obtenir deux récompenses différentes par rapport à une tâche plus simple (une action pour une récompense), les actions restent sensibles à la dévaluation, même après un surentrainement. Les auteurs proposent en fait que les deux systèmes sont utilisés alternativement, selon leur pertinence relative face aux diverses situations. Les auteurs ont développé un modèle pour expliquer comment l'arbitrage entre les deux est réalisé. Ils suggèrent un rôle de l'incertitude dans les valeurs produites par chacun de ces modèles, dans cet arbitrage. Selon ce modèle, les actions choisies sont celles avec la plus haute valeur. Les deux systèmes d'apprentissage par renforcement peuvent produire différentes estimations de

ces valeurs, dont le degré de précision varie selon les situations. L'arbitrage entre ces valeurs serait basé sur leur degré d'incertitude relative. Les deux systèmes sont en permanence incertains sur ces estimations, en particulier à cause de l'ambiguïté (ils commencent naïfs) et dans le cas des environnements volatils. Daw et coll. ont conçus un modèle Bayésien permettant de quantifier l'incertitude de chaque système. Le système *model-based* (de recherche en arbre) utilise son expérience dans la tâche pour estimer les transitions entre états et les récompenses (autrement dit, pour reconstruire les branches de l'arbre). Le système *model-free* (apprentissage par différence temporelle) estime les valeurs à long terme, directement de l'expérience, sans construire "d'arbre" mais en utilisant une méthode de *bootstrapping*, un processus auto-soutenu. Les résultats des simulations avec ces modèles montrent que quand le système de recherche en arbre domine, les choix des actions sont sensibles à la dévaluation, contrairement aux situations où le système de caching domine, reproduisant les données comportementales. Une étude de Wan Lee et collaborateurs, dont nous parlerons dans la partie sur les corrélats neurophysiologiques, propose les bases neurales d'un tel système d'arbitrage (Wan Lee et al., 2014). On peut se demander si l'acquisition d'un *learning-set* ne consisterait pas en un développement de cette capacité à alterner de manière efficace entre les systèmes *model-based* et *model-free*. En tout cas, cette capacité devrait être fortement liée à l'utilisation adaptée du contrôle cognitif. C'est ce que montre l'étude d'Otto et collaborateurs (Otto et al., 2014). Ces chercheurs proposent un rôle du système cognitif en tant que processus par lequel un système peut dominer l'autre. Pour cela, ils ont testé des sujets dans une tâche de contrôle cognitif requérant l'utilisation d'information contextuelle reliée au but, afin de surmonter des réponses de types habitude. En parallèle, ils ont testé les sujets dans une tâche de choix séquentielle. Les différences individuelles dans la tâche de contrôle cognitif étaient prédictives d'un comportement *model-based* dans la tâche de séquence. Ce lien comportemental entre contrôle cognitif et apprentissage par renforcement *model-based* suggère qu'ils sont portés par des mécanismes communs. Nous avons vu précédemment les liens forts entre *learning-set* et contrôle cognitif en proposant que *l'acquisition d'un learning-set peut être vue comme la mise au point du contrôle cognitif*. Cela nous permet d'émettre

l'éventualité d'un rôle du processus d'apprendre à apprendre dans le développement du système *model-based*, et dans le contrôle de l'arbitration entre les deux systèmes. Des recherches supplémentaires sont nécessaires, par exemple, en étudiant le développement des deux systèmes, *model-based* et *model-free*, lors de l'acquisition d'un *learning-set*.

2.2 Corrélats neurophysiologiques

Dans cette partie, nous allons décrire des études qui ont cherché les corrélats neurophysiologiques des 3 processus que nous venons de décrire : le processus d'apprendre à apprendre, l'apprentissage de la structure latente et l'apprentissage *model-based* versus *model-free*. L'idée est de voir si ces processus ont des bases physiologiques communes, entre eux, mais aussi avec les processus du contrôle cognitif évoqués dans la partie 1. Cela soutiendrait l'hypothèse que les premiers participent à la mise en place et la mise au point du second.

2.2.1 *Learning-set*

La formation d'un *learning-set* peut s'effectuer pour des tâches très différentes. Cela suggère une fonction supérieure, permettant de mettre en relations d'autres fonctions. Si tel est le cas, cette fonction serait potentiellement plus supportée par un réseau que par une seule région. Dans cette partie, nous détaillerons des études qui suggèrent que cette fonction repose de manière cruciale sur l'intégrité du cortex préfrontal dans son ensemble.

L'étude de Jang et collaborateurs semble suggérer que différentes régions cérébrales sont impliquées chacune à leur façon dans le processus d'acquisition d'un *learning-set*. Grâce à une méta-analyse regroupant les résultats de plusieurs études chez le singe, ces chercheurs ont montré que des lésions de différentes régions cérébrales provoquent des déficits différents dans l'acquisition d'un *learning-set* pour des *reversals* (Jang et al., 2015). Ils distinguent 3 types d'effets sur selon les lésions : soit un retard, soit aucun effet, soit une accélération dans l'acquisition du *learning-set* par rapport aux performances de groupes contrôles non lésés. Les lésions par aspi-

ration touchant les cortex orbitofrontal et rhinal ainsi que les lésions excitotoxiques de l'hippocampe ont conduit à des déficits dans l'acquisition du *learning-set*. La lésion excitotoxique du COF médian (aire 14) n'a pas eu d'effet, alors que les lésions de l'amygdale, ainsi que des aires du COF fortement connectées à l'amygdale, et certaines parties du CPFm ont conduit à une amélioration des performances sur les premiers *reversals*. Les analyses Bayésiennes conduites par les auteurs suggèrent que les différentes lésions provoquent des déficits différents car elles résultent en différents *niveaux initiaux des a priori* de l'animal quant au niveau de la volatilité environnementale, et non à cause de différences dans l'acquisition elle-même. Les auteurs proposent des hypothèses pour expliquer le fait que certaines lésions conduisent, de manière étonnante, à une *amélioration* de l'acquisition. La première hypothèse est que les singes avec de telles lésions auraient des *a priori* plus forts sur le fait que l'environnement est volatil, comme proposé par le modèle Bayésien. Cela pourrait être qu'ils ont un *a priori* fort sur le fait que les associations entre les stimuli et les récompenses ne sont pas stables au cours du temps. Une autre hypothèse est que ces singes formeraient des associations *plus faibles* entre les stimuli et les récompenses. Cette interprétation n'est pas soutenue par le fait que ces singes montrent les plus fortes stratégies de *win-stay lose-shift*. Mais cette interprétation est confortée par le fait que des singes avec des lésions de l'amygdale présentent un comportement d'extinction plus rapide que des singes contrôles (Izquierdo and Murray, 2005). Il est aussi possible qu'une capacité plus faible à former des associations stimulus-récompense conduise à des *a priori* plus fort sur l'instabilité environnementale. Il faut néanmoins relever un point important concernant les conclusions de ces auteurs sur le fait que les différences d'acquisition des *learning-set* sont dues à des différences dans les *a priori* sur la volatilité et pas à des différences dans l'acquisition elle-même : aucune des lésions ne concerne le CPF dans son ensemble.

Wilson et collaborateurs ont pourtant proposé un rôle particulier pour le CPF dans l'acquisition du *learning-set* (Wilson et al., 2010). Ce rôle impliquerait le CPF dans son ensemble, et ne consisterait pas en la somme des fonctions spécifiques de chacune des sous-régions préfrontales. Les données allant dans ce sens proviennent notamment d'études de lésions chez le singe. Des singes avec des lésions de sous-

régions du CPF , comme le COF ou le CPFvl, montrent des déficits dans une tâche de mémoire épisodique (*object-in-place scene-learning task*) (Baxter et al., 2008, 2007), mais l'ampleur de ces déficits est bien moindre comparée à celle provoquée par une lésion du CPF entier (Browning et al., 2005; Wilson et al., 2010). Des singes avec de telles lésions ne sont plus capables d'apprendre des discriminations simples, même après une centaine d'essais d'entrainement. Le fait que cette lésion provoque des déficits comportementaux si importants argumente pour un rôle crucial du CPF. Cependant, elle ne permet pas de révéler les mécanismes sous-jacents. Une façon plus fine d'investiguer le rôle du CPF est de le priver spécifiquement de ses interactions avec les autres régions. Par exemple, des lésions de déconnection entre le CPF et le cortex temporal peuvent être réalisées en supprimant le CPF dans un hémisphère, et le cortex temporal inférieur dans l'autre hémisphère. Les voies d'interaction cortico-corticales entre ces deux régions étant de manière prédominante ipsilatérales, ce type de lésion croisée permet d'abolir uniquement la communication entre les deux régions en préservant leur rôle individuel respectif. Les déficits liés à cette lésion sont nombreux mais un type de tâche est préservé : l'apprentissage des associations entre objets et récompenses lors d'une tâche d'apprentissage de discrimination de 10 paires concurrentes (Parker and Gaffan, 1998). Pour comprendre les limites de cette lésion, Browning et coll. ont réalisé la lésion décrite ci-dessus, mais après que les singes aient acquis un fort *learning-set* pour la tâche d'apprentissage de discrimination (Browning et al., 2007). De manière intéressante, après lésion, les singes sont toujours capables de faire la tâche, mais avec un taux d'apprentissage similaire à celui qu'ils avaient au tout début de l'apprentissage, avant la formation du *learning-set*. Ainsi, la communication entre les cortex frontal et temporal n'est pas impliquée dans l'apprentissage intra-problème, mais est cruciale pour l'apprentissage inter-problème. Ces données apportent un argument supplémentaire à l'idée que ces deux types d'apprentissage sont fondamentalement différents et portés par différentes régions. Reliant ce travail à l'étude de Murray et Gaffan (Murray and Gaffan, 2006), Browing et coll. proposent que cette lésion abolit la mémoire prospective, c'est-à-dire la représentation apprise lors d'un essai de quel sera le choix correct à l'essai suivant (le processus de représentation

d'évènements temporellement distants). Ces études mettent en avant le rôle crucial du CPF dans son ensemble dans l'*organisation d'évènements temporels complexes* (Wilson et al., 2010). Cela a été spécifiquement testé avec l'étude de Browning et Gaffan (Browning and Gaffan, 2008). La tâche donnée aux singes consiste en une discrimination concurrente, mais au lieu de 2 objets à discriminer, il s'agit d'une séquence de 2 fois 2 objets. Cela permet de créer des évènements temporellement complexes. La tâche contrôle est identique à la première à l'exception qu'un délai vide (sans objet) remplace la période temporelle de présentation des $2^{\text{ème}}$ objets. Les singes avec une lésion déconnectant le CPF du temporal inférieur montrent des déficits pour se rappeler la séquence de 2 objets, mais pas de déficits pour la tâche contrôle. De manière intéressante, ce déficit émerge parce que les singes contrôles montrent plus de facilité à réaliser la tâche de séquence par rapport à la tâche avec le délai vide, contrairement aux singes lésés. Cela montre que la lésion n'empêche pas l'apprentissage d'une telle tâche, mais empêche l'amélioration de l'apprentissage conférée par la capacité à gérer les évènements temporels complexes.

Toutes ces études suggèrent un rôle global du CPF dans l'organisation des évènements temporels complexes, qui serait cruciale pour l'acquisition d'un *learning-set*. Si l'intégrité du CPF est requise pour cette fonction, cela suggère que cette fonction est portée par l'existence d'un fonctionnement en réseau entre ses différentes sous-régions, comme l'étude de Browning et collaborateurs, s'intéressant au rôle spécifique de la communication entre le CPF et le cortex temporal, a commencé à le montrer (Browning et al., 2007). Une question qui se pose alors, à laquelle ces études ne permettent pas de répondre, concerne la dynamique de ce réseau pendant la mise en place du *learning-set*. Comment se développe la communication entre les régions du CPF lors du processus d'apprendre à apprendre ? Est-ce que le réseau final correspond à celui mis en place par un individu étant capable d'exercer un bon contrôle cognitif ? Ces questions mériteraient d'être posées lors de recherches futures.

2.2.2 Apprentissage latent

Si des études comportementales ont démontré l'utilisation de structures pour l'apprentissage (Collins and Frank, 2013), il reste à trouver les corrélats neurophy-

siologiques correspondant. Plusieurs études proposent l'implication de différentes régions dans l'apprentissage des structures, ce qui suggère que cela doit dépendre des modalités sur lesquelles la structure est basée. Par exemple, l'hippocampe semble être impliqué dans l'apprentissage des structures latentes lorsque des rats sont testés dans des expériences de conditionnement associés à différents contextes (Ji and Maren, 2005). En effet, dans ces conditions, des lésions de l'hippocampe avant entrainement éliminent l'effet de renouvellement, ce qui pourrait s'expliquer comme un déficit de l'animal à inférer l'existence de nouvelles causes latentes. D'après Gershman et Niv, les lésions de l'hippocampe conduisent l'animal à n'attribuer qu'une seule cause latente à toutes ses observations.

Dans d'autres tâches, le CPF est impliqué. Hampton et collaborateurs ont montré en scannant des sujets réalisant une tâche d'apprentissage de *reversal* probabiliste, que l'activité BOLD du CPFvm corrélait plus avec un modèle computationnel exploitant la structure de la tâche qu'avec un modèle de simple apprentissage par renforcement (Hampton et al., 2006). Collins et collaborateurs ont utilisé une tâche où les sujets apprennent spontanément en utilisant une structure pour les règles (Collins et al., 2014). Ils ont montré que des potentiels évoqués enregistrés au niveau du CPFl sont organisés en fonction des structures choisies par les participants. Ces marqueurs étaient également prédictifs de la capacité des participants à généraliser la structure des règles à de nouveaux contextes.

Redish et collaborateurs proposent un mécanisme pour créer des structures : un mécanisme de division en différents "états" lorsqu'aucun n'est pertinent pour le comportement en cours, via un rôle de la dopamine (Redish et al., 2007). Leur proposition est que les erreurs de prédiction négatives tonique résulteraient en une plus grande probabilité de création d'un nouvel état.

Ces études montrent une fois encore le rôle crucial du CPF, mais aussi de la dopamine, qui étaient également impliqués dans les mécanismes supportant le contrôle cognitif. D'autres structures, comme l'hippocampe, semblent également jouer un rôle, mais plutôt dans les situations où il s'agit plutôt d'une structure concrète. Est-ce que le CPF aurait un rôle dans l'apprentissage de la structure "abstraite" de l'environnement (comme sa structure statistique) et d'autres structures, comme

l'hippocampe, dans l'apprentissage de sa structure physique ?

2.2.3 *Model-free* vs *model-based*

Les approches *model-free*, comme l'apprentissage par différence temporelle, apportent une interprétation forte à l'activité des neurones dopaminergiques et de leurs projections striatales dorsolatérales (Schultz et al., 1997). L'autre classe de modèle dit *model-based* a plutôt été associée au réseau du CPF au sens large (mais pouvant inclure des régions plus médiales du striatum). Différentes lésions cérébrales perturbent l'un ou l'autre des deux systèmes. Les lésions du cortex infralimbique chez le rat conduisent à une sensibilité à la dévaluation, quel que soit le niveau d'entrainement, suggérant un rôle de cette structure dans le système *model-free* (Coutureau and Killcross, 2003). A l'inverse, d'autres lésions, présentes au sein d'un large réseau, affectent le système *model-based*. Ces régions incluent le cortex médial prélimbique (Balleine and Dickinson, 1998; Coutureau and Killcross, 2003; Killcross and Coutureau, 2003), les régions du striatum dorsomédial (Yin et al., 2005), de l'amygdale basolatérale (Blundell et al., 2003) associées aux régions préfrontales (expériences chez le rat), et le COF (expérience chez le singe) (Izquierdo et al., 2004). En effet, des lésions de ces régions éliminent la sensibilité à la dévaluation, même en condition de faible entrainement. Les résultats des lésions indiquent que chaque système peut se substituer à l'autre, même dans des circonstances où il n'aurait normalement pas eu le dessus, comme suggéré par le modèle de Daw et collaborateurs (Daw et al., 2005).

Wan Lee et collaborateurs ont cherché les corrélats neurophysiologiques d'un mécanisme d'arbitrage entre les deux systèmes qui avait été proposé de manière théorique par Daw et collaborateurs (Daw et al., 2005) et qui serait basé sur la fiabilité respective des prédictions produites par chaque système (Wan Lee et al., 2014). Ils ont montré que le signal BOLD en provenance des CPFl inférieur et fronto-polaire encode ces valeurs de fiabilité et le résultat de leur comparaison. Les auteurs ont trouvé des interactions fonctionnelles entre les régions faisant l'arbitrage et celles supportant le système d'évaluation *model-free* (à savoir le putamen et le cortex moteur supplémentaire) mais pas avec celles supportant le système *model-*

based. Cela suggère une asymétrie dans le fonctionnement de "l'arbitre" : au lieu de moduler l'un ou l'autre des systèmes, il contrôlerait le système *model-free* de manière sélective (qui peut être alors considéré comme un système par défaut). Si l'on veut faire le lien avec la littérature sur le *learning-set*, ces études renforcent peut être l'idée que l'acquisition d'un *learning-set* consiste en l'acquisition de la capacité de basculer en mode *model-based* pour apprendre de manière efficace et flexible. Des études suivant l'évolution du signal dans les régions soutenant les deux modes de fonctionnement, et surtout l'évolution de leurs échanges dynamiques lors de l'acquisition d'un *learning-set*, permettraient de répondre à ces questions.

Chapitre 3

Questions de thèse

Comme nous l'avons développé au fil de l'introduction, apprendre de l'environnement fait appel à un large panel de processus dont l'efficacité peut être améliorée au cours de l'acquisition d'un *learning-set*. Cette étape où l'on apprend à apprendre est cruciale pour la flexibilité comportementale, car elle influence la généralisation et donc le transfert à d'autres tâches. Etudier la façon dont le processus d'apprendre à apprendre conduit en la mise en place et la mise au point de des processus d'apprentissage de l'environnement est un moyen pertinent de mieux les comprendre. Cependant, ces mécanismes sont encore assez peu connus, et la plupart des études ont été effectuées dans des contextes déterministes. Il a pourtant été montré que l'environnement dans lequel on apprend à apprendre une tâche, et notamment son degré d'incertitude, influence de manière cruciale le niveau de flexibilité comportementale ultérieure (Biernaskie et al., 2009; Tebbich and Teschke, 2014). Ainsi, en contexte expérimental, si l'on veut vraiment comprendre les processus comportementaux et mécanismes physiologiques qui permettent de prendre des décisions dans un environnement incertain, il est nécessaire d'entrainer les sujets dans un tel environnement dès le départ, afin de maitriser l'origine de leurs comportements. Le but final de ce travail de thèse étant d'étudier la prise de décision dans un environnement où il est nécessaire de gérer différentes sources d'incertitude, nous avons donc délibérément choisi d'entrainer nos singes dans un tel contexte, ce qui est rarement fait dans notre domaine. Apprendre à apprendre dans un contexte mêlant plusieurs

sources d'incertitude oblige le sujet à apprendre à gérer l'ambiguïté décroissante liée au fait qu'il apprenne à réaliser la tâche petit à petit et à apprendre à estimer les contributions relatives du risque et de la volatilité. La première grande question à laquelle ce travail de thèse tentera de répondre est donc la suivante :

"Comment se développent les réponses liées à l'incertitude de l'environnement au cours de l'acquisition d'un learning-set dans un environnement incertain ?"

Le deuxième aspect de ce travail de thèse concerne les corrélats neurophysiologiques de la prise de décision dans un environnement combinant risque et volatilité, nécessitant d'alterner en permanence entre des comportements d'exploration et d'exploitation. Si l'implication de certaines régions a déjà été mise en évidence dans ce genre de processus, les mécanismes neurophysiologiques permettant la réalisation de ces processus sont encore peu connus. L'étude du signal continu enregistré en électroencéphalographie, électrocorticographie, ou dans les potentiels de champs locaux permet une première approche de ces mécanismes, via l'étude des potentiels évoqués et des oscillations. Les potentiels évoqués représentent la réponse calée (time-locked) sur un stimulus, et les oscillations, la réponse induite. Des variations de l'amplitude de ces potentiels, ou de la puissance des oscillations renseignent sur les modulations à l'échelle de la population neuronale située sous l'électrode. Par exemple, une plus forte puissance des oscillations indique une augmentation de la synchronisation de la population. De manière intéressante, les différentes bandes de fréquence observables dans le signal continu ont été associées à différents mécanismes cognitifs.

De plus, la façon dont les régions interagissent au sein d'un réseau dynamique pour intégrer l'information pour permettre l'adaptation du comportement est encore largement méconnue. Une façon de mettre en évidence ces échanges est l'étude des relations dynamiques entre régions. Une théorie actuelle, proposée par Pascal Fries, suggère un rôle des oscillations cérébrales comme mécanisme de coordination des échanges d'information entre les régions (Bastos et al., 2015). Les oscillations cérébrales sont créées par les variations des potentiels excitateurs et inhibiteurs pré-

synaptiques et leurs variations influencent directement le seuil d'excitabilité des neurones, en d'autres mots, leur capacité à réagir à l'information reçue. Ainsi, l'échange entre deux régions serait le plus efficace lorsque leur oscillations seraient synchronisées, ce que Fries appelle la "communication par la cohérence". Les analyses des modulations du signal continu, qu'il s'agisse des potentiels évoqués, des variations de puissance des oscillations, ou des relations entre les oscillations sont un outil précieux pour comprendre les mécanismes cérébraux sous-jacents aux processus cognitifs. Peu d'études existent pour l'instant, en particulier en ce qui concerne les domaines cognitifs, et notamment, aucune étude n'a encore essayé de faire ce genre d'analyses pour comprendre les mécanismes de prise de décision dans les environnements complexes impliquant de réagir à différentes formes d'incertitude. C'est le deuxième but de mon projet de thèse. Pour cela, nous avons choisi d'étudier les modulations du signal continu en provenance des aires préfrontale et pariétale, dont nous avons vu l'implication dans les mécanismes de l'adaptation. Ainsi, la deuxième grande question de ce travail de thèse est la suivante :

"Quelle est l'implication respective des réponses oscillatoires de surface en provenance des cortex préfrontal et pariétal dans les comportements d'exploration et exploitation dans un environnement incertain ?"

Dans le chapitre suivant, nous présenterons le programme expérimental que nous avons suivi pour ce travail de thèse afin de répondre à ces deux grandes questions. Le chapitre 5 présentera les résultats d'une première étude conçue pour répondre à la première question. Le chapitre 6 et 7 proposeront des résultats préliminaires de deux études mises en place pour répondre à la deuxième question.

Chapitre 4

Programme expérimental de la thèse

Afin de répondre à ces deux grandes questions, ce travail de thèse s'est déroulé en deux étapes. La première étape a consisté à entrainer des singes pour leur procurer un *learning-set* dans une tâche nécessitant de la flexibilité comportementale dans un environnement incertain. La deuxième étape a consisté en la conception d'un implant permettant l'enregistrement chronique de la dynamique oscillatoire des cortex préfrontal et pariétal. Dans cette partie, je vais détailler les réflexions que nous avons développées et les difficultés que nous avons rencontrées.

Développement d'une nouvelle tâche et d'une stratégie d'entrainement

Un des points d'intérêt majeurs du travail de l'équipe d'Emmanuel Procyk concerne les corrélats comportementaux et neurophysiologiques de l'adaptation du comportement et du contrôle cognitif. Lorsque je suis arrivée dans l'équipe, la tâche comportementale courramment utilisée était la tâche de résolution de problèmes (*Problem Solving Task*, PST). Dans cette tâche, le singe doit trouver le stimulus récompensé dans un set de 4 stimuli affichés à l'écran. Cette tâche se résout par essai-erreur, jusqu'à ce que le singe découvre le bon stimulus (période de recherche), et soit autorisé à répéter 3 fois de suite le choix correct (période de répétition). Ensuite, un signal de changement indique qu'un nouveau stimulus sera récompensé et que l'animal doit repartir en recherche. L'intérêt majeur de cette tâche est qu'elle

permet de contraster de manière contrôlée des périodes nécessitant différents degrés de contrôle cognitif : élevé lors de la recherche, plus faible lors de la répétition. Cette tâche a permis à l'équipe de publier plusieurs travaux quant aux rôles complémentaires du CCM et du CPFdl dans la détection des évènements pertinents pour l'adaptation et le contrôle cognitif (Quilodran et al., 2008; Rothé et al., 2011; Stoll et al., 2015; Vezoli and Procyk, 2009). Cependant, cette tâche comporte certaines limites. La première est qu'elle se déroule dans un environnement complètement déterministe. La deuxième est que la période d'exploitation (répétition) est limitée et le retour à l'exploration est forcé et signalé. Or, comme nous l'avons vu dans cette introduction, le monde réel est stochastique et volatil, et un des défis auxquels sont confrontés les animaux en permanence est de savoir appréhender l'incertitude que cela provoque, et de gérer la balance entre exploration et exploitation.

Adaptation d'une nouvelle tâche pour le singe

Nous avons donc décidé d'essayer une nouvelle tâche combinant ces aspects et avons commencé une collaboration avec l'équipe d'Etienne Koechlin. Nous avons adapté, pour le singe, leur tâche de manipulation de *task-sets*, originellement conçue pour l'homme (Collins and Koechlin, 2012). La tâche consiste à trouver, de manière concurrente, deux associations stimulus-cible parmi un panel de deux stimuli et 3 cibles (par exemple, le stimulus A est associé à la cible 1 et le stimulus B à la cible 3) (Figure 4.1). La tâche est probabiliste car dans 10% des essais, un *feedback* inverse est donné. Ces essais sont appelés essais "*Trap*" et consistent en un *feedback* négatif sans récompense après une bonne réponse, et un *feedback* positif avec une récompense après une mauvaise réponse. De plus, la tâche est réalisée dans un contexte volatil car les deux associations correctes changent au bout d'un moment. La tâche originale de Collins et Koechlin est plus complexe dans le sens où les sujets doivent découvrir les associations correctes entre 3 stimuli et 4 cibles (alors que notre version ne comporte que 2 stimuli et 3 cibles). Cependant, la structure est similaire : les *feedback* sont probabilistes et les contingences changent régulièrement.

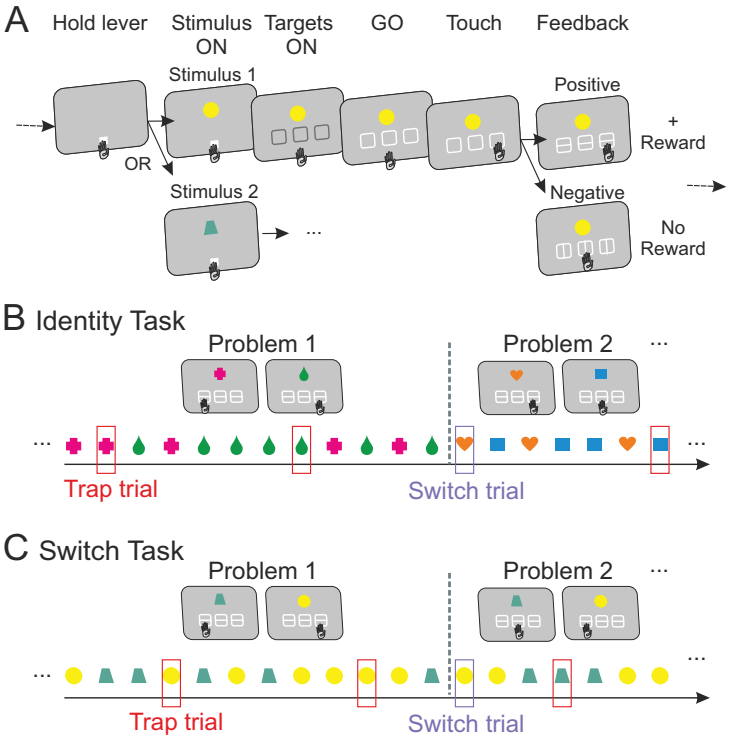

Figure 4.1 – Schéma des tâches utilisées. A. Structure d'un essai (commun aux deux tâches). Le singe doit toucher un "levier" sur un écran tactile pour démarrer l'essai. Il sélectionne ensuite une des 3 cibles en réponse à un stimulus central. Le singe doit apprendre les cibles associées à 2 stimuli de manière concurrente. Les *feedback* visuels positifs ou négatifs consistent en des barres horizontale et verticales sur les cibles. Un *feedback* positif est suivi d'une récompense de jus de fruit. B. La tâche d'identité. Chaque problème comprends 2 "mappings", consistant chacun en l'association entre un des 2 stimuli et une des 3 cibles. Un stimulus à la fois est présenté à l'écran. Lors des essais dits "*Trap*", un *feedback* trompeur est donné, consistant en un *feedback* positif suivi d'une récompense après un choix incorrect ; et un *feedback* négatif sans récompense après un choix correct. Les essais *Trap* apparaissent de manière pseudo-aléatoire avec une probabilité de 10%. Après qu'un critère de performance (de 17 réponses correctes sur 20) ait été atteint, le problème change (correspondant à l'essai dit "Switch"), et deux nouveaux mappings avec 2 nouveaux stimuli sont sélectionnés de manière aléatoire. C. La tâche de Switch. Cette tâche est identique à la tâche d'identité, à l'exception que les stimuli restent les mêmes après qu'un problème change. Seuls les associations entre les stimuli et les réponses changent. Ainsi, les Switch entre problèmes ne sont pas détectables visuellement.

Stratégie d'apprentissage

Réussir à apprendre cette tâche complexe à des singes n'était pas un pari facile. Nous avons donc développé une stratégie d'apprentissage, basée sur la littérature du *learning-set*. Nous avons élaboré notre stratégie en nous inspirant des expériences de transfert entre deux tâches, après acquisition d'un *learning-set*, et notamment d'une l'étude de Schrier montrant un très bon transfert entre une tâche de discrimination et une tâche de *reversal* chez le singe (Schrier, 1966). Nous avons choisi d'entrainer les singes sur une version simplifiée de la tâche, mais possédant la même structure que la tâche finale. L'idée était de permettre l'acquisition d'un *learning-set* apportant à la fois des informations pertinentes sur la structure de la tâche et une stratégie efficace pour réaliser la tâche, afin de permettre un transfert facilité à la tâche finale. La tâche d'entrainement (Figure 4.1.B) est une version simplifiée de la tâche finale (Figure 4.1.C) car les changements de contingence sont signalés visuellement par un changement des stimuli (deux nouveaux stimuli pour chaque nouveau problème) alors que la tâche finale consiste en des changements non signalés (les stimuli restent les mêmes au cours des différents problèmes). La Figure 4.2 décrit les différentes phases d'entrainement au cours des 3 premières années de cette thèse, pour chaque singe.

L'acquisition d'un *learning-set* est traditionnellement réalisée en changeant le problème une fois que l'animal a atteint un certain critère de performance. Cela permet de s'assurer d'un certain niveau de performance dans la tâche. Ainsi, dans notre version de la tâche, les contingences ne changent qu'une fois que l'animal atteint un critère de performance (de 17 essais corrects sur 20 essais successifs). Le changement au critère diffère de la tâche utilisée chez l'homme dans laquelle le changement survient après un nombre aléatoire d'essais. Ce choix de l'apprentissage au critère n'a été réalisé que pour l'étape d'entrainement pour l'acquisition du *learning-set*, ainsi que pour les premiers problèmes lors du transfert à la nouvelle tâche. Pour les étapes d'enregistrements cérébraux, le critère a été enlevé.

Cette tâche requiert un comportement complexe et flexible d'apprentissage de l'environnement, nécessitant de gérer plusieurs sources d'incertitude. Cela va donc nous permettre de répondre à un ensemble de questions que voici.

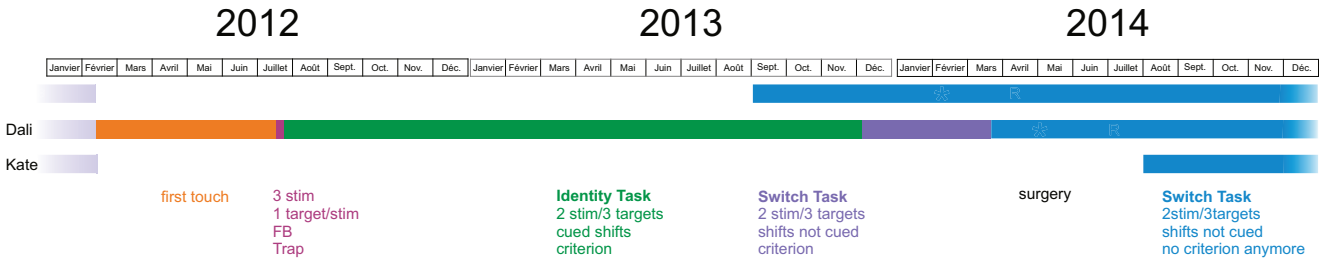

Figure 4.2 – Schéma des phases d'entrainement au cours des années 2012 à 2014. Après une phase d'habituation à l'expérimentateur, à la sortie en canne, à la chaise de transport puis au box expérimental (mauve), les singes ont d'abord appris à toucher des cibles à l'écran pour obtenir du jus de fruit ("first touch", orange). Les singes ont ensuite appris à toucher une cible, mais pas le stimulus associé (prune). Les *feedback* visuels, avec les essais *Trap*, ont été introduits à ce moment-là. Après quelques jours, les singes ont commencé la vraie tâche : la tâche d'identité (vert). Le changement d'un problème à l'autre était signalé par un changement de l'identité des deux stimuli, et intervenait lorsque le singe atteignait le critère de performance. Une fois que les performances se sont stabilisées à un haut niveau (*learning-set* acquis), le singe a été transféré sur la tâche de Switch (violet). Le singe Kate a passé plus de temps sur la tâche d'identité parce qu'elle réalisait moins d'essais par jour (et non à cause de performances moins bonnes). L'étape suivante a consisté à enlever le critère de performance, les changements de problème intervenant après un nombre aléatoire d'essais, devenant ainsi indépendant des performances (bleu). Les étoiles rouges indiquent le jour de l'implantation des singes Pippa et Dali, et le R dans un rond jaune, les premiers enregistrements du signal.

— Est-ce que les singes sont capables d'acquérir un *learning-set* pour une tâche de discrimination concurrente combinant stochasticité et volatilité ?

— Comment se développe leur réponse aux *feedback* incertains au cours de l'acquisition du *learning-set* ?

— Est-ce que les singes sont capables de combiner deux associations, faisant ainsi preuve de l'utilisation d'un *task-set* ?

— Est-ce que l'acquisition du *learning-set* pour la tâche simplifiée permet un transfert sans coût à la tâche finale ?

— Comment les singes résolvent-ils cette tâche : en particulier, comment distinguent-ils les situations où les *feedback* surprenants sont dus aux essais *Trap* (stochasticité) de celles où ils sont dus à un changement des contingences (volatilité) ?

Conception et réalisation d'un nouveau type d'implant chronique

La deuxième étape de ce projet de thèse a été le développement d'un implant permettant le suivi de l'activité corticale oscillatoire pendant toute la session de travail de l'animal, et de manière reproductible d'une session à l'autre sur plusieurs mois, voire plusieurs années. Dans cette partie, nous décrirons les étapes de la conception

à l'implantation, que nous avons d'abord réalisées sur un singe pilote, puis sur deux singes entrainés à la tâche.

Localisation des électrodes

Nous avons décidé de cibler les cortex préfrontal et pariétal afin d'étudier les dynamiques des interactions entre ces deux régions au cours de la tâche (Figure 4.3). L'implant est constitué de 46 électrodes placées sur la surface de ces cortex. Les positions centrales F21, F4, F24, F27, F7 et F10 sont situées à la verticale du cortex cingulaire médian, afin d'enregistrer les potentiels évoqués aux *feedback*.

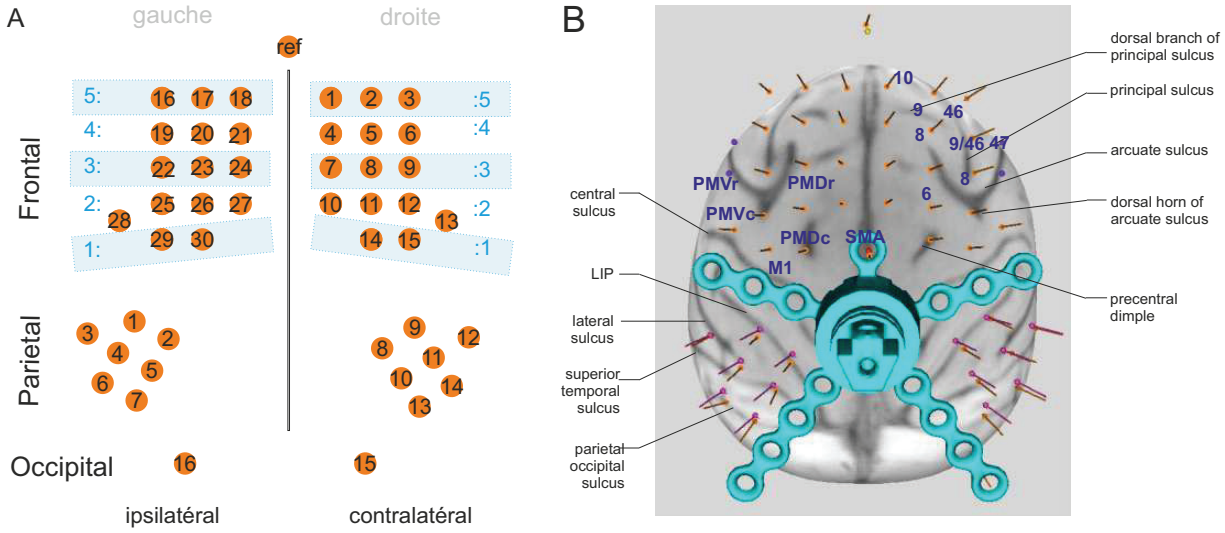

Figure 4.3 – Localisation des 46 électrodes. A. Numérotation des électrodes. B. Reconstruction 3D du cerveau et localisation des électrodes. Les localisations des aires ont été réalisées d'après les travaux de Petrides et Pandya (1194). PMV : cortex premoteur ventral, PMD : cortex premoteur dorsal, M1 : cortex moteur, SMA : aire motrice supplémentaire.

Conception de l'implant

Ce genre d'implant chronique avait été réalisé sur deux singes (Stoll et al., 2015). Cependant, le défi consistait à concevoir un nouvel implant moins invasif que les précédents qui consistaient en un gros bloc de ciment recouvrant toute la surface supérieure du crâne. Nous avons travaillé avec l'entreprise Gray Matters pour concevoir ce nouvel implant.

L'idée du nouvel implant a été de ne pas cimenter les fils des électrodes sur le

crâne mais de les laisser courir sous la peau. Afin d'obtenir un implant le plus petit possible, nous avons aussi utilisé de nouveaux micro-connecteurs capables de s'insérer dans la barre de tête. La Figure 4.4 différentes vues de l'implant ainsi que son positionnement sur le crâne. La barre de tête est pourvue de 5 pattes, que nous avons incurvées pour qu'elles s'ajustent à la forme du crâne, et qui sont maintenues par des vis dans le crâne. Une pièce supplémentaire consiste en une plateforme conçue pour porter les deux micro-connecteurs. Les électrodes sont insérées dans le crâne et reposent sur la dure-mère. Les fils des électrodes serpentent le long du crâne pour se rejoindre dans les fentes situées à la base de la barre de tête, remonter le long d'un conduit interne et sortir sur la plateforme supportant les connecteurs auxquels ils sont soudés. L'implant a également été conçu pour laisser de l'espace pour l'implantation ultérieure de chambres latérales, permettant d'avancer des canules d'injection jusqu'au cortex cingulaire.

Le métal de la barre de tête est du tungstène, métal biocompatible. Les électrodes sont des fils en chlorure argent enrobés de téflon (*Phymep*) ; soudés à une extrémité aux têtes des électrodes (en chlorure d'argent, *Science Product*) et à l'autre extrémité au micro-connecteur (*Omnetics*). L'étape des micro-soudures a été délicate et fastidieuse. Afin de renforcer et d'isoler les soudures, nous les avons recouvertes d'un verni isolant (*Acrylic protective lacquer*, Electrolube). La stérilisation classique étant impossible due aux soudures et à la fragilité des micro-connecteurs, nous avons contacté le service central de stérilisation des hôpitaux de Lyon pour réaliser une stérilisation à froid au gaz (peroxyde d'hydrogène). Les parties de l'implant en contact avec les fils des électrodes (fentes à la base, conduit interne et plateforme) ont été également isolées avec du vernis isolant. Les fils et les connecteurs ont ensuite été cimentés (ciment dentaire, *SuperBond C&B*) afin de renforcer le montage et d'empêcher tout mouvement.

Stratégie d'implantation chirurgicale

Comme le but était de laisser courir les fils des électrodes sur le crâne, en les laissant sous la peau, nous avons réfléchi à une stratégie d'incision provoquant le moins

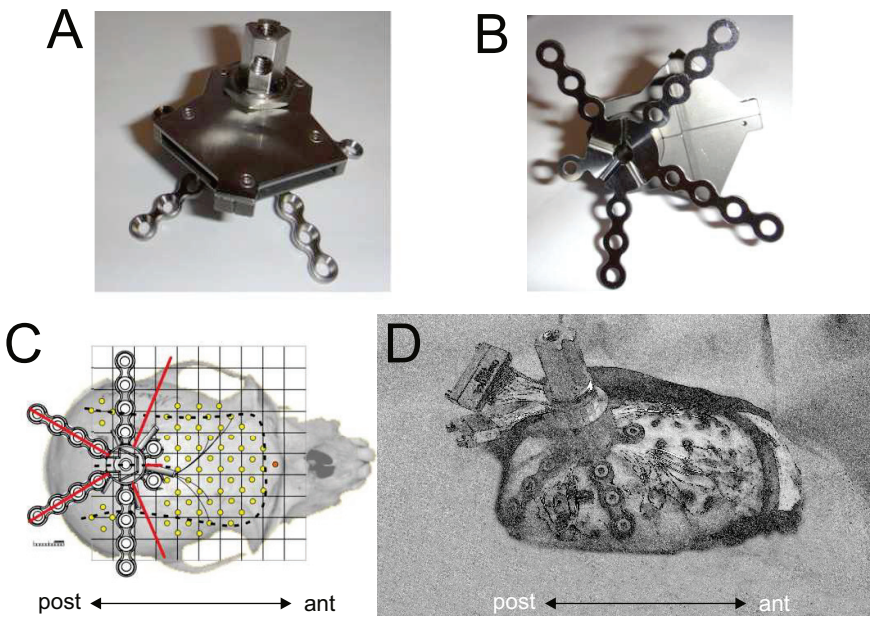

Figure 4.4 – Implant et positionnement sur le crâne. A. Vue postérieure de l'implant, montrant la plateforme conçue pour porter les micro-connecteurs. B. Vue de dessous de l'implant, montrant le conduit intérieur et les 4 fentes conçus pour laisser passer les fils. C. Vue schématique de l'implant sur le crâne montrant la position des pattes. D. Photo de l'implant sur le crâne, montrant les électrodes en place.

de dégâts et une meilleure cicatrisation possible du tissu recouvrant les électrodes. Pour cela, au lieu du plan d'ouverture antéro-postérieur utilisé traditionnellement, nous avons réalisé une incision en forme de U allant "d'une oreille à l'autre" en passant par l'avant, permettant de réaliser un "rabat" de peau. Nous avons ensuite désolidarisé la peau et les muscles du crâne, que nous avons soigneusement enrobés de compresses imprégnées de saline en permanence afin de les préserver le plus possible. A la fin de l'opération, les muscles ont été replacés sur le crâne et suturés de manière à se maintenir dans leur position originelle. Le rabat de peau a été ensuite percé en son centre pour permettre le passage de la barre de tête. Le défi est de réaliser le plus petit trou possible afin que la peau soit positionnée le plus proche de la barre pour permettre une meilleure repousse des tissus.

Pour localiser les positions des électrodes sur le crâne, nous avons utilisé le logiciel de neuro-navigation Brainsight (Rogue Research). Quelques jours avant la chirurgie, le singe est légèrement anesthésié pour permettre l'enregistrement de points de repère sur la peau de la tête de l'animal, placé dans un cadre stéréotaxique. Ces points sont

utilisés pour faire correspondre précisément la position réelle de l'animal avec le scan de son cerveau, et permettre la géo-localisation des électrodes le jour de la chirurgie. Une fois les électrodes placées et fixées avec du ciment au niveau de leur entrée dans le crâne, leur position réelle est ré-enregistrée dans Brainsight.

Conception et réalisation d'une chambre d'injection

Une dernière étape de ce projet, que nous n'avons pas encore évoquée, est en cours de réalisation. Nous voulons cibler le CCM à l'aide de canules d'injection. L'idée est d'obtenir des données causales quant au rôle de cette structure (et de certains neuromodulateurs agissant dans cette structure) sur le comportement d'adaptation et dans la dynamique oscillatoire entre les cortex préfrontal et pariétal. Pour cela, nous avons conçu une chambre supposée s'insérer latéralement, en dessous des positions des électrodes préfrontales latérales (Figure 4.5). Nous avons également conçu un adaptateur permettant de s'insérer sur la chambre et de permettre la fixation de micro-descendeur afin de descendre les canules d'injection jusqu'à leur cible. L'implantation du singe pilote est prévue pour la janvier 2016.

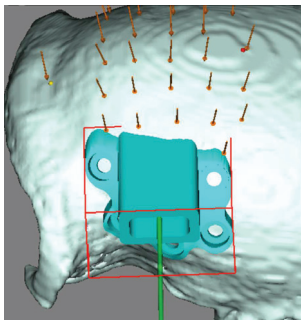

Figure 4.5 – Représentation 3D de la chambre d'injection, sur une reconstitution d'un crâne de singe (réalisée à l'aide de Brainsight)

Chapitre 5

Etude Comportementale :

Apprendre à apprendre dans un environnement incertain

Dans cette étude, nous nous sommes intéressés aux processus permettant de prendre des décisions dans un environnement incertain, et en particulier, au développement de ces processus au cours de l'acquisition d'un learning set. Nous montrons, qu'au fur et à mesure que les animaux améliorent leurs performances dans la tâche, ils développent une stratégie exploratoire en réaction aux feedback incertains crées par l'environnement stochastique et volatil. Ces animaux montrent ensuite une excellente capacité de transfert à une version plus complexe de la tâche. Ce travail a été accepté à la publication par Learning & Memory.

Learning To Learn About Uncertain Feedback

Maïlys C.M. Faraut[1,2,3], Emmanuel Procyk[1,2],
and Charles R.E. Wilson[1,2,3]

[1]INSERM U846, Stem Cell and Brain Research Institute, Bron 69500, France
[2]Université de Lyon, Lyon 1, UMR S-846, Lyon 69003, France
[3]Corresponding authors: mailys.faraut@inserm.fr & charles.wilson@inserm.fr

Keywords: feedback, adaptation, flexibility, learning, monkey

Abstract

Unexpected outcomes can reflect noise in the environment or a change in the current rules. We should ignore noise but shift strategy after rule changes. How we learn to do this is unclear, but one possibility is that it relies on learning to learn in uncertain environments. We propose that acquisition of latent task structure during learning to learn, even when not necessary, is crucial. We report results consistent with this hypothesis. Macaque monkeys acquired adaptive responses to feedback whilst learning to learn serial stimulus-response associations with probabilistic feedback. Monkeys learned well, decreasing their errors to criterion, but they also developed an apparently non-adaptive reactivity to unexpected stochastic feedback, even though that unexpected feedback never predicted problem switch. This surprising learning trajectory permitted the same monkeys, naïve to re-learning about previously learned stimuli, to transfer to a task of stimulus-response remapping at immediately asymptotic levels. Our results suggest that learning new problems in a stochastic environment promotes the acquisition of performance rules from latent task structure, providing behavioural flexibility. Learning to learn in a probabilistic and volatile environment thus appears to induce latent learning that may be beneficial to flexible cognition.

Introduction

Our environment is both noisy and volatile, and we track the resulting uncertainty to guide our choices (Behrens et al., 2007). Monkeys and humans successfully make choices and switch strategies in tasks with both features (Walton et al., 2010; Donoso et al., 2014; Pleskac, 2008; Rudebeck et al., 2008; McGuire et al., 2014; Collins and Koechlin, 2012). In these studies, increasingly sophisticated models, often derived from Bayesian principles, explain behaviour and its neural correlates in well-trained subjects. But subjects are trained before data are acquired, and the training process, during which subjects learn to learn about the different causes of unexpected feedback, is surprisingly understudied. In these tasks, an unexpected negative feedback might be due to the noise, or more formally the stochasticity of an event or outcome. Such feedback should be ignored for optimal performance. Unexpected feedback may also be due to a change in the settings or rules of our volatile environment. This should induce a switch of choice strategy. Although we don't know how we learn to learn about these features, learning to learn appears to provide tuning of cognitive control processes required for efficient adaptation (Collins and Frank, 2013).

The way in which we learn is likely to modify our response to stochastic elements of the environment. This is especially true in many naturalistic settings, such as foraging. By learning to be more efficient we may make the environment more changeable by depleting resources quicker and so increasing the need to switch strategies - this is a common problem for optimal choice in foraging situations (McNamara and Houston, 1985; Ollason, 1980). Hence improved learning can increase volatility, and in the context of increased volatility, stochastic outcomes become less easy to detect and ignore (Payzan-LeNestour and Bossaerts, 2011).

This predicts that the response to stochasticity should modify with learning to learn. The strategies employed by human subjects are shaped by a priori information about types of environmental uncertainty in the task (Payzan-LeNestour and Bossaerts, 2011; Hertwig et al., 2004). Across nature, learning in stochastic (as opposed to deterministic) environments leads to greater behavioural flexibility as measured by remapping tasks (Tebbich and Teschke, 2014; Gallistel et al., 2001; Biernaskie et al., 2009). So in these cases, subjects learn to learn about uncertain tasks, and in doing so acquire abstract information about latent task structure, even when not necessary (Collins and Frank, 2013; Gershman et al., 2010; Kornell et al., 2007). Deep neural networks can now acquire such latent structure and generalize it across a range of tasks (Mnih et al., 2015), suggesting that this acquisition process could be crucial to behavioural flexibility. We sought to show that learning to learn in a stochastic and volatile environment specifically drives the acquisition of latent information. Importantly, this la-

tent information should impact performance and strategy (Collins and Frank, 2013) in ways that identify the nature of the learning to learn process.

Learning to learn was established in deterministic tasks in primates by the seminal work of Harlow on 'learning sets' (Harlow, 1949). Monkeys acquire a learning set that provides optimal performance on tasks of simple associative learning, as well as reversal learning (Murray and Gaffan, 2006). Learning set is posited as a memory dependent performance rule, for example 'win-stay lose-shift' based on previous feedback (Murray and Gaffan, 2006). But as soon as feedback is determined by any probabilistic rule, optimal performance would require choice driven by more than the single previous outcome - it is unclear whether simple 'win-stay lose-shift' is then adapted or modified in order to track a longer feedback history.

Monkeys with learning set flexibly transfer between new learning and remapping learning (Schrier, 1966). This flexibility derives from 'win-stay lose-shift', as in deterministic tasks the rule applies equally to new and remapped associations. Hence monkeys naïve to remapping or reversal learning are able to remap their knowledge very efficiently. These deterministic tasks are a special case, but if such benefits of learning to learn are generalizable to more realistic stochastic settings, transfer effects to remapping tasks should also be observable in such settings.

In this study, therefore, we follow the evolution of responses to stochasticity across learning to learn in a paradigm in which the level of volatility is driven by changing performance. We tracked the progress of monkeys as they acquired new problems in a stochastic environment. Each time monkeys successfully learned a problem, a new problem was presented for learning, imposing volatility in addition to the stochastic nature of the task. We followed how the monkeys learned to learn over a large number of problems, and showed that as volatility increased with learning, monkeys acquired an exploratory response to the stochastic environment, even though the task did not require this. We then tested whether learning to learn in a volatile environment permitted flexible choosing, as learning set does in deterministic settings. Monkeys repeatedly remapped what they learned. Despite their naivety to remapping, the monkeys were immediately able to do so optimally.

Results

Adaptive responses to standard and unexpected feedback differ through learning.

Our study proceeded in two major steps. In the first, monkeys performed an "Identity Task" (IT, **Figure 1.B**), in which they learned pairs of object-response associations in a probabilistic environment on a 90/10 schedule - that is 10% of trials received 'Trap' feedback where the opposing feedback was

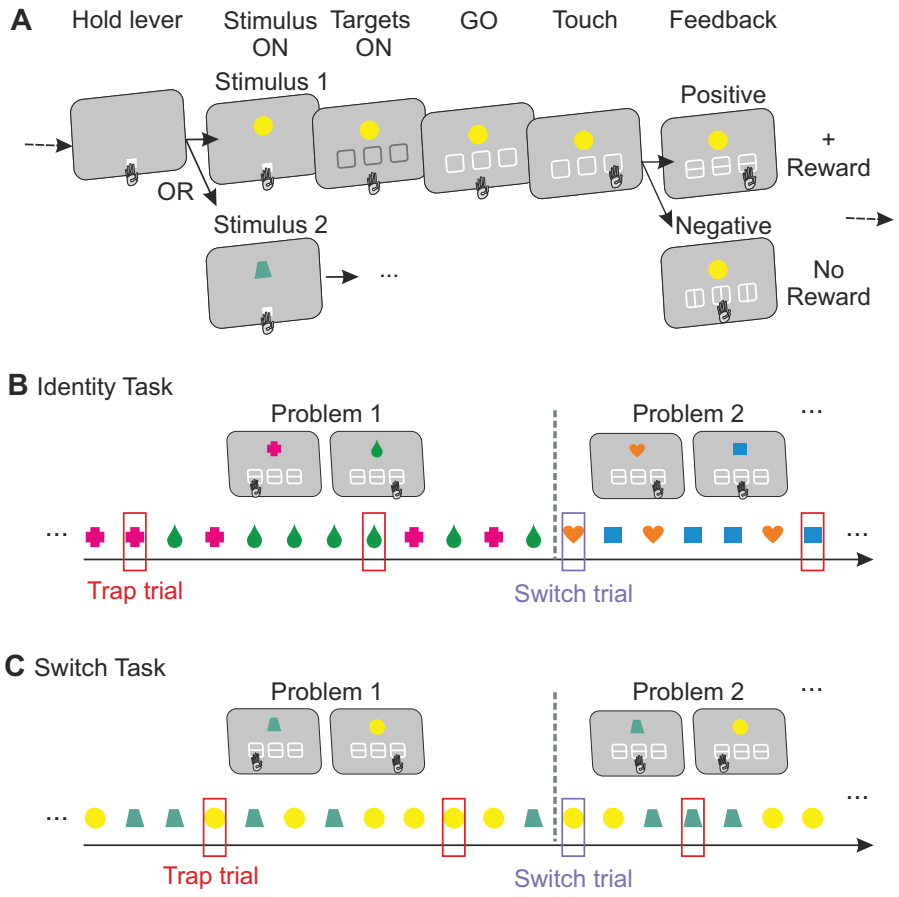

Figure 1: Task design. A. Structure of a single trial - common to both tasks. The monkey holds a touch screen "lever" to initiate the task, and then selects one of 3 targets in response to a central stimulus. Monkeys learn about 2 stimuli concurrently. Positive or negative visual feedback is horizontal or vertical bars respectively. Positive feedback was followed by the delivery of juice reward. **B. Identity Task.** Each problem comprised two mappings, between each of the two stimuli and a single target. One stimulus at a time was randomly presented. On Trap trials, misleading feedback was given: positive feedback with juice reward after an incorrect choice; negative feedback with no reward after a correct choice. Trap trials occurred pseudo-randomly with p=0.1. After a performance criterion (17/20 correct responses), the problem changed (Switch trial), and 2 new mappings with 2 new stimuli were randomly selected. **C. Switch Task.** Identical to the Identity Task, except that stimuli remained the same when the problem changed. Only the mappings between stimuli and responses changed. Thus, Switches between problems were not visually detectable.

given compared to the rule currently in force. We refer to this probabilistic feedback as the stochasticity of the environment. A pair of these associations formed a problem. Monkeys serially learned a long sequence of new problems, changing problem after learning the current one to a performance criterion. Each new problem was incidentally signalled by a change in the identity of the stimuli, the first trial of the new problem being a Switch trial. We measured the learning of the monkeys on individual problems by the number of errors to reach the criterion. But in addition to this learning, monkeys also 'learned to learn', by reducing the number of errors to criterion over many problems (**Figure 2.A**), and learning to rapidly adapt to new stimuli after a Switch (**Figure S1**). This process eventually stabilized to consistently less than 50 errors to criterion (vertical dotted lines). We call this the "stabilized period" even if some learning still occurred, as shown by the continuous improvement of percentage correct per problem (**Figure 2.B**).

Reward maximization in this task would require the monkeys to learn the rule, ignore the Trap trials, and switch rule only in the presence of new stimuli. Whilst theoretically optimal performance like this is obtainable in deterministic tasks, achieving optimality in a stochastic environment is costly. In order to understand how the monkeys adapted to feedback noise over learning, we studied their response to the uninformative Trap trials.

Trap reactivity refers to the change in performance after a Trap feedback compared to performance before it (example in **Figure 4.A**). Initially monkeys appeared to take very little notice of Trap feedback (**Figure 2.D**), showing Trap reactivity around 0. This potentially maximized rewards, as Trap trials are uninformative about the rule and unpredictable. Trap reactivity then increased significantly over problems in all monkeys (**Figure 2.D**, glm, interaction Trials x Learning_Bins, $p<0.001$). This increase was strongly correlated with performance (correlation between Trap reactivity and percentage of correct responses per learning bin, $R=0.7119$, $p<0.001$, **Figure 2.E**), a counter-intuitive result. Optimal responding demands that monkeys decrease Trap reactivity as they improve at the task, given that Trap feedback is never predictive of a switch. But importantly improving performance is also increasing the level of volatility (**Figure 2.C**), as monkeys reach criterion and switch quicker. This is a naturalistic situation - in foraging increased efficiency of foraging also requires increased shifting between patches. In IT, by analogy, learning increases the ratio of switches to Traps, making Traps harder to distinguish from switches. In this context an increase in Trap reactivity is less counter-intuitive (correlation between Trap reactivity and switch:Trap ratio per learning bin, $R=0.6808$, $p<0.001$, **Figure 2.F**).

Trap reactivity increased even when the monkeys were performing each problem well, when only exploitation periods were consid-

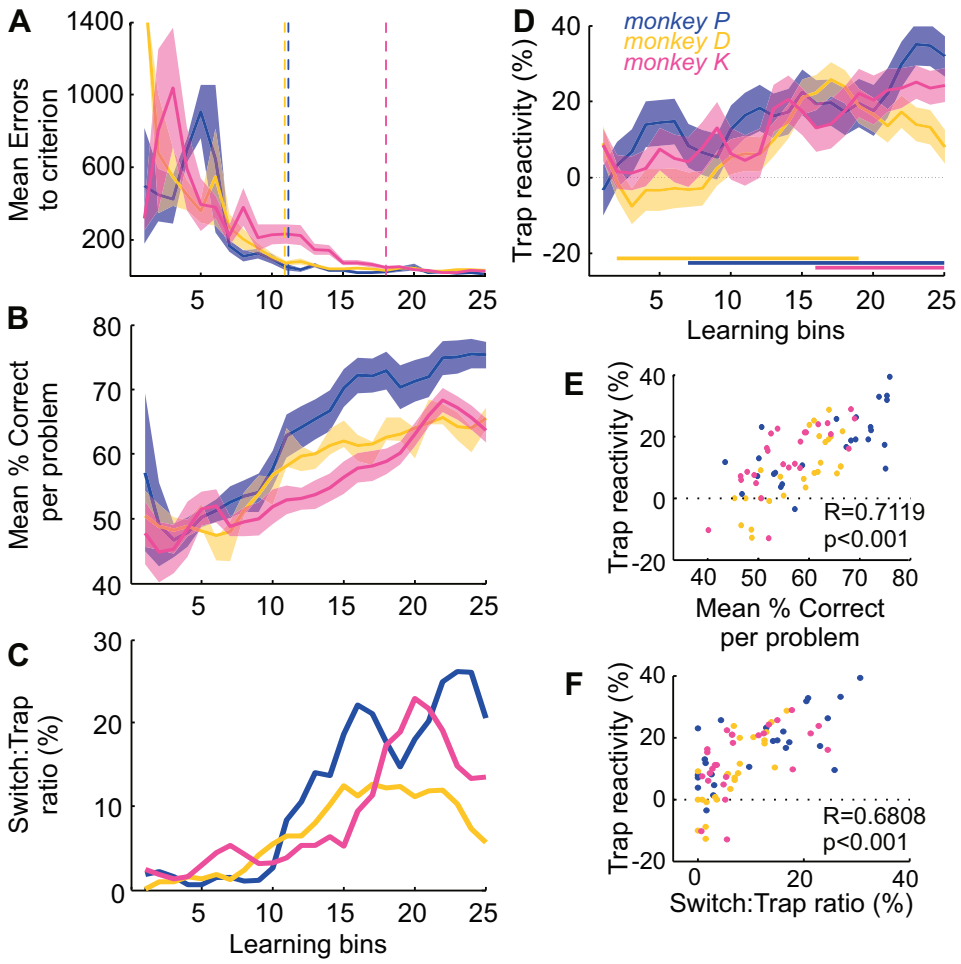

Figure 2: Acquisition of the Identity task. **A.** Mean errors to criterion across learning bins, showing significant learning. Vertical do ed lines indicate the stabilization point of the curve for each monkey (*see Methods*). Shaded area is sem. **B.** Mean percentage of correct responses to criterion for problems of each learning bin. **C.** Ratio of Switch over Trap trials (taken as an index of the volatility), for each learning bin. **D.** Value of the Trap reactivity across learning bins. Trap reactivity is the difference in performance before and after the Trap trial. It increases from an initial value around zero, especially later in learning (horizontal bars). **E.** Correlation between Trap reactivity and percentage correct. **F.** Correlation between Trap reactivity and the switch:Trap ratio. In all figures: Monkey P, blue; monkey D, yellow and monkey K, pink.

ered (glm on exploitation trials only, interaction Trials x Learning_Bins, p<0.001) (**Figure 3.A**). This effect was robust to variation in the criterion used for selecting exploration/exploitation periods (**Figure S2**). Importantly, the increase in Trap reactivity was not driven by an increase of general performance with learning, but by a decrease of performance on trial Trap+1 **Figure 3.B**. Trap reactivity also increased during exploration periods, although less strongly (glm on exploration trials only, interaction Trials x Learning_Bins, p<0.01) (**Figure 3.C**). Further, the increase in Trap reactivity could not be accounted for by the increase in correct trials with learning. Even when we considered only the Trap trials in the form of surprising negative feedback after a correct choice ("Negative Trap"), the effect persisted (Separate glms for positive and negative Trap feedback. Interaction Trials x Learning_Bins, p<0.01 and p<0.001 respectively) (**Figure 3.E & F**). This increasing reactivity to unexpected feedback with learning was specific to Trap feedback. We compared Trap reactivity across learning bins with Switch reactivity (taken here as the mean performance on the trials following a switch). While performance after a Trap decreased with learning (accounting for the increase of Trap reactivity) (Separate glms for positive and negative Trap feedback on performance at Trap+1, factor Learning_Bins, p<0.001 in both cases), performance on the trials following a Switch trial increased after a negative feedback and remained high and stable across learning after a positive feedback (Separate glms for negative and positive Switch feedback on performance at Switch+1,+2 and+3, factor Learning_Bins, p<0.01 & p>0.05) (**Figure 3.G & H**). This shows that monkeys are increasing the volatility in part by becoming efficient at switching, and it is specifically whilst they do this that they also increase their Trap reactivity; they learn to switch well, but they also learn to switch to unexpected outcomes. It should be noted that Switch reactivity here is performance on the 3 trials after a Switch, whereas Trap reactivity remains performance at Trap+1. The Switch here provides new stimuli with no reward history (unlike the Trap), and so an appropriate response to it requires integrating feedback from more than one trial. A difference between the effects on performance of Trap and Switch is also obtained, however, if we check this result only on trial Switch+1 (**Figure S3**. Separate glms for positive and negative Trap feedback on performance at Trap+1, factor Learning_Bins, p>0.05 in both cases).

In the stabilized period at the end of the IT, Trap reactivity was highly significant and persistent (reverse Helmert contrast comparing performance at trial Trap+1 vs previous trials, p<0.001, for the 3 monkeys) (**Figure 4.A**). Performance was also lower 2 trials after a Trap (reverse Helmert contrast, trial Trap+2 vs previous trials, p<0.001, for the 3 monkeys) but then returned to initial pre-trap levels (reverse Helmert contrast, trial Trap+3

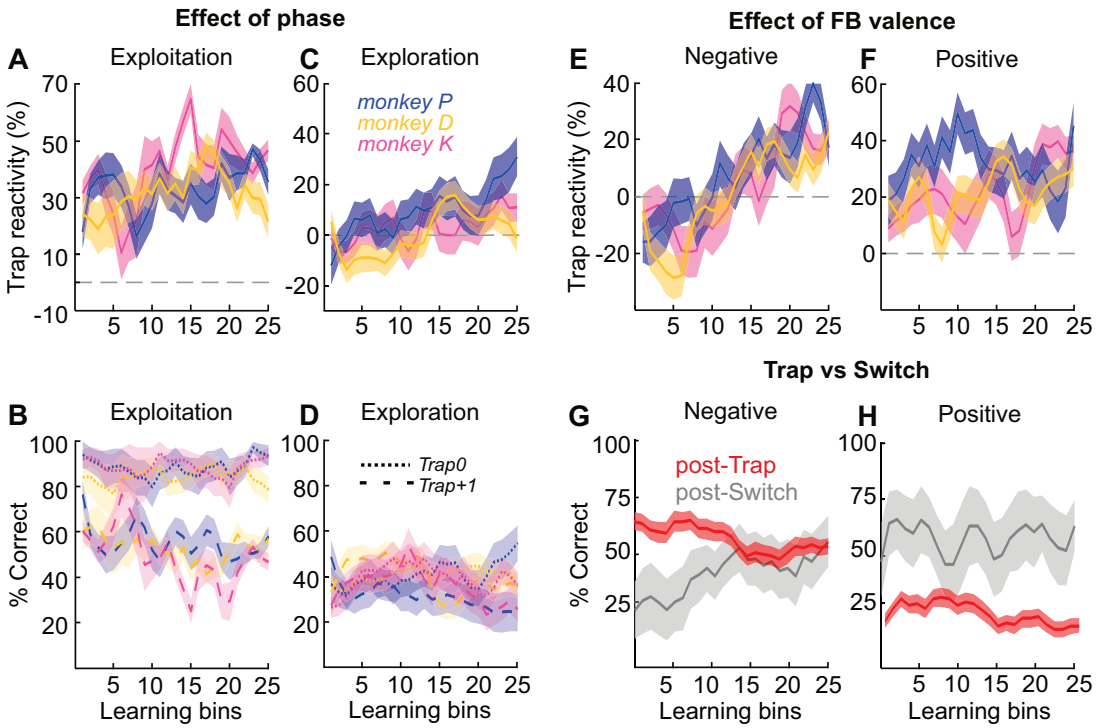

Figure 3: Modulations of Trap reactivity and performance with learning set on IT. A to D. Effect of phase on Trap reactivity across learning: **A**: in exploitation and **C** in exploration. Effect of phase on performance at the trial of the Trap (Trap0, dotted line) and at the next trial (Trap+1, dashed line): **B**: in exploitation and **D**: in exploration. **E & F.** Effect of feedback valence on Trap reactivity across learning: **E**: after a negative Trap feedback (received after a correct response) and **F**: after a positive Trap feedback (after an incorrect response). **G & H.** Comparison of performance at trial Trap+1 (red) and Switch+1 to Switch+3 trials (gray) after a **G**: negative or **H**: positive feedback (concatenated data from the 3 monkeys). In all figures: Monkey P, blue; monkey D, yellow and monkey K, pink.

vs previous trials, ns, for the 3 monkeys) (**Figure 4.A**). The Trap reactivity acquired by the monkeys had a number of properties. Trap reactivity was stimulus specific – that is specific to the stimulus associated with the Trap feedback (glm, interaction Trials x Stimulus_Similarity, $p<0.001$) (**Figure 4.B**, top). Trap reactivity showed a stronger value in response to a positive Trap (after an incorrect response) compared to a negative Trap (glm, interaction Trials x FB_Valence, $p<0.001$) (**Figure 4.B**, middle). Trap reactivity was also stronger in exploitation than exploration (glm, interaction Trials x Phase, $p<0.001$) (**Figure 4.B**, bottom), demonstrating that monkeys were not shedding their Trap reactivity each time they mastered a problem. This shows that Trap reactivity is a strategy adapted to the overall environmental volatility in the whole task, and not simply a signal of exploration on an individual problem.

Monkeys apply different choice strategies following normal or Trap feedback, further supporting this argument. In deterministic studies of learning set, a Win-Stay Lose-Shift (WSLS) structure is latent within the task. Adopting a WSLS strategy provides an optimal performance rule, and this is posited as the rule acquired in learning set (Murray and Gaffan, 2006; Wilson et al., 2010). WSLS, however, becomes non-optimal in non-deterministic designs - in our case because Trap feedback should be ignored. Nevertheless, our monkeys acquired a significant WSLS strategy for normal feedback (binomial test, $p<0.05$), but significantly more so than for Trap feedback, which approached chance levels of WSLS (glm on win-stay lose-shift values, main effect of Normal_or_Trap, $p<0.001$). Indeed the monkeys showed significantly greater acquisition of WSLS for normal feedback (same glm, interaction Normal_or_Trap x Learning_Bins, $p<0.05$) (**Figure 4.C**). This shows that in the main our monkeys learn the stochastic task as other monkeys have learned the deterministic one (Harlow, 1949; Murray and Gaffan, 2006; Izquierdo et al., 2004) - using normal feedback to apply WSLS. But in addition the monkeys clearly learned to differentiate unexpected feedback (Trap and Switch trials), and applied a different strategy to that class of feedback. Specifically, on Trap trials monkeys are not using WSLS (**Figure 4.C**, low WSLS after Trap), but they are significantly changing response (**Figure 4.A**, Trap reactivity). Their strategy after Trap feedback therefore appears to be more random than these alternatives - and could thereby be interpreted as exploratory.

Different responses to feedback and Trap reactivity were also reflected in reaction times (RTs, **Figure 4.D**). RT decreased in the trial following positive feedback (two-sample Kolmogorov-Smirnov test, $p<0.05$, for both (normal or Trap) cases, for monkey D & K. Monkey P's data could not be analyzed due to technical problems); and increased after negative feedback (two-sample Kolmogorov-Smirnov test, $p<0.05$). These post-correct

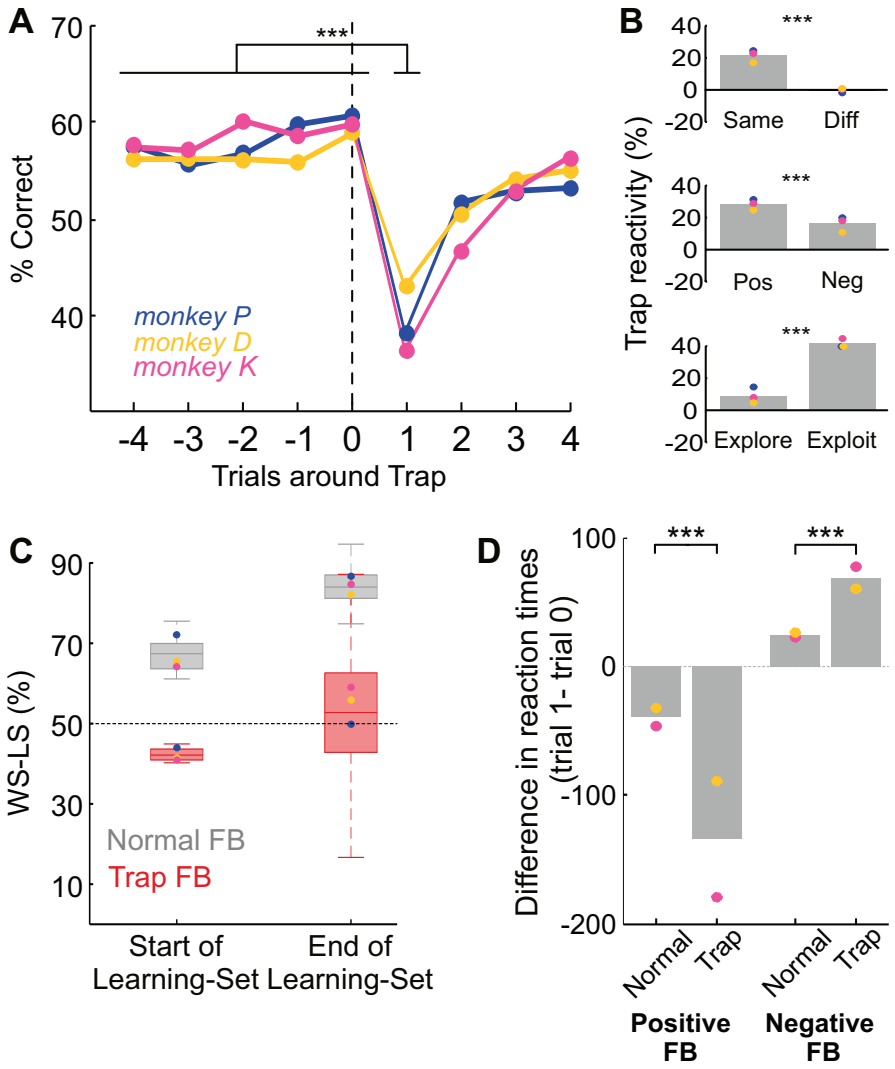

Figure 4: Trap reactivity on IT. A. Percentage correct around Trap trial (trial 0) for all monkeys. Trap reactivity is the drop of performance between trials 0 and 1. **B.** (top)Trap reactivity is modulated by the stimulus of the trial, specifically being present only when the subsequent trial is on the same stimulus as the Trap trial. (middle) Trap reactivity is greater for 'Positive' Traps (positive feedback after incorrect choice) than 'Negative' Traps (negative feedback after correct choice). (bottom) Trap reactivity is greater during exploitation compared to exploration period (See Experimental Procedure). **C.** Proportion of trials upon which Win-Stay Lose-Shift (WS-LS) strategy is applied after feedback in the Identity Task, split between normal and Trap feedback. 50% represents a random (not WS-LS) strategy. Boxplots represent group data, circles each monkey's mean. WS-LS is significant and increasing across learning for normal feedback, but absent for Trap feedback. **D.** Effect of feedback on reaction times (RTs). Plot shows the difference (trial 1 – trial 0) in RTs before and after the feedback in question. Full circles indicate a significant ($p<0.001$) difference between trials 0 and 1 in every case. Feedback on trial 0 can be positive or negative, and that feedback can either be Trap feedback or « normal ». Left panel shows post-correct speeding, which is further increased by Trap feedback. Right panel shows post-error slowing, again increased by Trap feedback. **In all figures**: Monkey P, blue; monkey D, yellow and monkey K, pink. Gray bars represent averages of all monkeys. Stars indicate significant difference between conditions. ***: $p<0.001$; *: $p<0.05$; ns: non-significant

speeding and post-error slowing effects were accentuated significantly after a Trap trials (multi-factor ANOVA on RT differences, factor Normal_or_Trap: $p<0.001$, in both negative and positive cases).

Over the course of learning to learn about a task in a probabilistic environment, monkeys both adapted their responses to the task contingencies, but also modified their response to unexpected feedback, even though such a response was not the most rewarding strategy. Instead of learning to ignore Trap feedback, they became reactive to all unexpected feedback, Trap or Switch. This result suggests a fundamental influence of a probabilistic learning environment on the way in which animals learn about tasks. The explorative value of unexpected feedback, we propose, drives a performance rule which, although not necessary for the initial task (Collins and Frank, 2013), is nevertheless important in promoting generalization of the learning.

Adaptive responses to feedback promote flexible decisions.

In a second step, we sought to test how having learned to learn with both stochasticity and volatility would serve the monkeys when moving to a task of higher complexity with less information.

To test this we transferred the monkeys to the Switch Task (ST). Here the identity of stimuli was fixed each day, and did not change between problems. Only the rule that associated stimuli to responses changed (**Figure 1.C**), requiring monkeys to remap what they knew about the current objects and the responses without any cue to the change of rule. Reversal learning is a specific form of remapping task - here remapping was more complicated than simple reversal given there were 3 options and 2 objects. It should be stressed that these monkeys, naïve to cognitive testing at the start of the experiment, had never re-learned a new rule for the same stimulus. We sought to compare monkeys' performance when starting ST with their initial performance on IT. Monkeys made many errors when first exposed to IT (**Figure 2.A**). Monkeys performing classical reversal learning for the first time (Harlow, 1949; Izquierdo et al., 2004) show high error rates, just as when they start to learn simple discrimination problems. But importantly monkeys' initial error rates in reversal learning are lower if they have previously acquired a deterministic learning set (Schrier, 1966). ST and IT maintained the same parameters in terms of volatility and stochasticity of the environment, potentially promoting good performance after transfer. Monkeys worked to the same performance criterion, and worked on a 90/10 schedule, receiving 10% Trap trials.

Monkeys maintained low and stable errors to criterion when starting the ST compared to their performance at the end of the IT. That is, their transfer from new problem learning to remapping learning was perfect, regardless of the fact that these monkeys had never remapped a response to a stimulus in their

lives (no significant difference between errors to criterion at the end of IT and the start of ST, Kruskal Wallis test, H=0.89, 1 d.f., p=0.34) (**Figure 5.A**). This performance did not improve further, so the monkeys started this task with immediately asymptotic learning (linear regression over bins of problems, p>0.05 for the 3 monkeys).). We propose that this high level of performance was reached because a learning to learn process on IT prepared monkeys to transfer to the new and more complex task.

We cannot specifically attribute high-level performance on ST after transfer to the learning of IT in a stochastic and volatile setting. This is because we do not have a control group that acquired a deterministic version of IT and then transferred to the same stochastic ST as described here. It is very unlikely, given the rich literature of training on remapping tests, that monkeys with such a training regime would also show good and asymptotic transfer to ST, but without such a control group we cannot make that claim. We can nevertheless draw two conclusions. First, monkeys with our stochastic training regime are capable of asymptotic task transfer. Second, there is at least some evidence to support an assertion that this is because of the stochastic nature of IT. Specifically, a number of results support the assertion that Trap reactivity has driven efficient transfer. Trap reactivity on the ST closely matched that of the IT, even though unexpected feedback in the ST could be a signal of either a Trap or a change in rule. As such it is important for the monkeys to discriminate Trap from Switch trials. Trap reactivity was still present (reverse Helmert contrast, on performance at trial Trap+1 vs previous trials, p<0.001, for monkeys P and D) (**Figure S4.A**) and maintained at the level of the IT (glm, no significant interaction Trials x Task, for the 50 last problems of IT versus 50 first problems of ST (**Figure 5.B**), with no change over problems (glm, no significant interaction Trials x Learning_Bins). Trap reactivity had the same properties: a stronger effect on trials with the same stimulus as the Trap trial (glm, interaction Trials x Stimulus_Similarity, p<0.001) (**Figure S4.B**); and a stronger effect after positive compared to negative Trap feedback (glm, interaction Trials x FB_Valence, p<0.001) (**Figure S4.C**). Trap reactivity was still significantly stronger in exploitation trials than exploration (glm, interaction Trials x Phase, p<0.001) (**Figure S4.D**). We also observed post-correct speeding and post-error-slowing (two-sample Kolmogorov-Smirnov test, p<0.05, for both (normal or Trap) cases, for each monkey, except monkey K showing no significant post-negative feedback slowing) (**Figure S4.E**). In terms of RTs, monkeys reacted differently to Trap compared to normal feedback only when the feedback was positive (multifactor Anova on reaction time differences, factor Normal_or_Trap: p<0.05 and non-significant, for positive and negative cases respectively).

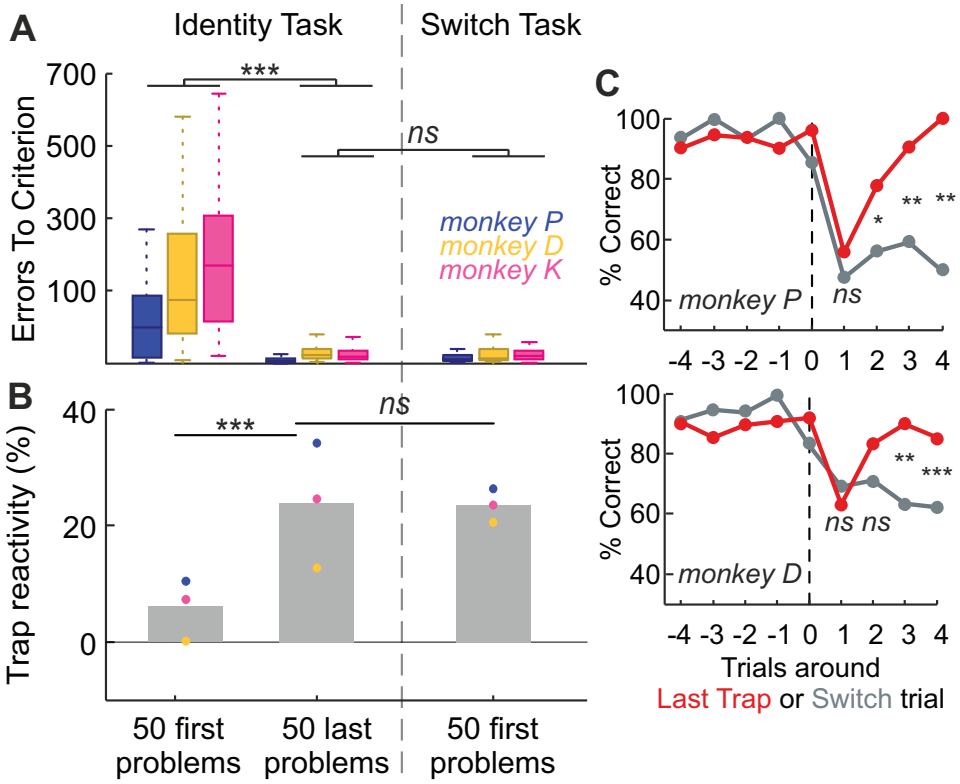

Figure 5: Switch task. A. Excellent transfer from IT to the ST. Errors-to-criterion for the 50 first problems and the 50 last problems of the IT; and for the 50 first problems of the ST. **B.** Trap reactivity is unchanged by task transfer, having increased during learning, suggesting that Trap reactivity acquired in IT is adaptive to ST. Gray bars represent averages of all monkeys. **C.** Contrast of behaviour around Last Trap trials and Switch trials. Monkey P (top) shows performance that distinguishes the two forms of unexpected feedback from trial 2 – the earliest possible moment. Monkey D's performance (bottom) distinguishes at trial 3. In this case only, performance after Switch is calculated on the basis of the rules of the previous problem, to provide a comparable score between Switch and Trap. P-values compare performance at LastTrap vs Switch. *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; ns: non-significant.

Trap reactivity is an adapted strategy for using unexpected feedback to differentiate Switch from Trap trials. The only way to distinguish Trap from Switch trials in ST is to maintain a record of the feedback history a number of trials after unexpected feedback. Exploratory responses during this period will make adaptation to Switch even more efficient. Monkeys were immediately capable of doing this in ST, suggesting they had learned to learn in this fashion during IT. Monkeys very quickly discriminated the two situations (at trial+2 for monkey P, and trial +3 for monkey D; glm, Performance at Switch vs Trap trial +3, $p<0.001$ for monkeys P & D, Monkey K was excluded from analysis due to an insufficient number of trials), indicating that they were able to efficiently integrate the feedback history (**Figure 5.C**). In fact, Traps and Switches were theoretically dissociable af-

ter 2 trials, when the continued presence of the unexpected feedback can first be assessed, given that Trap trials never occur on 2 successive trials. As such, from the start of ST, monkeys were distinguishing volatility from stochasticity optimally or near optimally, despite the fact that the ST contained far fewer cues to aid the monkeys. The pattern of target selection after a Last Trap or a Switch did differ between the two tasks (**Figure S4.F**). This shows that monkeys were indeed sensitive to the lack of switch cue, but adapted their behaviour so rapidly that there were no significant differences in errors to criterion. There were no differences in reaction times in trials following a Last Trap compared to a Switch (Two-sample Kolmogorov-Smirnov test comparing distributions for Last Trap vs Switch trials, non-significant).

Discussion

3 monkeys acquired a probabilistic task with signalled rule switches. In doing so they increased their response to misleading information provided by Trap feedback. Hence monkeys' performance was good and stable, but they were not maximizing their rewards for this specific task. This form of responding appeared, however, to be adapted to the changing volatility over learning and continued stochasticity of the reward environment, suggesting that learning to learn lead to choice that was driven by a process that takes these latent variables into account. Monkeys that learned in this way transferred without cost to a more complex version of the task with remapping of previously learned associations and un-signalled rule switches. Good transfer is therefore possible even from tasks that provide stochastic feedback.

The data from the IT shows that over the course of learning to learn about a task in a probabilistic environment, monkeys will both adapt their responses to the task contingencies, but also modify their response to unexpected feedback, even in cases where such a response is not necessarily the most rewarding strategy. This result demonstrates the fundamental influence of a probabilistic learning environment on the way in which animals learn about tasks. Trap reactivity continued to increase even when errors-to-criterion was relatively stabilized. Stabilization of learning to learn is therefore an important concern in training animals for neuroscience experiments, where we wish to separate learning effects from elements of the acquired task (Costa et al., 2014). This consideration is also important in situations where learning to learn might be used in wider applications (e.g. (Bavelier et al., 2012). If latent information is being acquired, for example about the structure of the task (Collins and Frank, 2013), classical measures of learning might not capture this ongoing process, introducing a risk of cutting short the learning to learn process before the attendant advantages can be obtained.

What is the specific process occurring as the monkeys learn to learn? From a mod-

elling perspective, adapting responses to Trap feedback could be akin to a matching of the learning rate for unexpected feedback to the volatility of the environment, a process reflected in decision making after learning in humans (Behrens et al., 2007; Payzan-LeNestour and Bossaerts, 2011; Courville et al., 2006). But the emergence of a significant difference in response strategy after different forms of feedback (**Figure 4.C**) is striking, suggesting that monkeys are genuinely learning to detect unexpected outcomes and explore after them. It is unclear whether a simple modulation of model learning rate could account for such a categorical change in strategy, but this provides evidence for at least two levels of information acquisition during learning to learn. First, as in deterministic tasks, monkeys are increasing their proportion of WSLS on normal feedback trials. Second, by acquiring the information about the statistics of unexpected outcomes, something that can only be learned by integrating across many trials, monkeys learn to maintain an exploratory strategy to these trials. What is particularly striking in the results from IT is that monkeys are not obliged to acquire this exploratory strategy on this task - there is a clear signal to explore in the change of stimuli - yet they nevertheless do.

It is of note that monkeys' final level of Trap reactivity in IT represents the final Trap/Switch ratio in the task. The fact that Trap reactivity is present in exploitation, increases with experience, and correlates with percentage correct reinforces the idea that it is a learning-driven strategy. The exploratory value of unexpected feedback, we propose, drives a latent performance rule which, although not necessary for the initial task (cf (Collins and Frank, 2013), but yet is robustly acquired, and might potentially be important in promoting generalization of the learning. Whether this acts as an account of the role of probabilistic environments in other species remains an open question (Tebbich and Teschke, 2014; Gallistel et al., 2001; Biernaskie et al., 2009).

Generalization of learning was expressed in the transfer from new learning (IT) to remapping (ST). Whilst there is some evidence for this process being driven by the acquired Trap reactivity, the aim of this study was to follow learning in a stochastic environment, and so we do not make the specific claim that good transfer was because learning was in a stochastic as opposed to deterministic environment.

Nevertheless, our findings do show very clearly that remapping can be performed at asymptotic levels without prior remapping experience. Classically such remapping, and the special case of reversal learning, has been associated with a process of cognitive inhibition, in which subjects make large numbers of errors on initial remapped problems, and subsequently acquire the ability to inhibit efficiently their previous learning. Here the monkeys performed their first ever remapping problems just as well as they had been per-

forming new discriminations, a remarkable result considering the difficulty usually induced by initial remapping (Harlow, 1949; Izquierdo et al., 2004). The result argues either that monkeys do not need inhibition to complete the task, or that in learning to learn the IT they acquired the ability to inhibit. Monkeys never have cause to unlearn or ignore any of their previous stimulus learning during the IT, as stimuli are never repeated, and so it is unlikely that they have learned to inhibit specific stimulus-response associations. There remains the possibility that simply inhibiting prior responses (and not their associations) is sufficient. But rather, because monkeys treat unexpected feedback in the same manner in IT and ST, we hypothesise that this reactivity promotes rapid differentiation of Trap feedback from Switch feedback in ST, the crucial prerequisite for efficient performance. A deterministic to stochastic transfer experiment would confirm this hypothesis, contributing to a growing body of evidence that questions the importance of the cognitive process of inhibition (Stuss and Alexander, 2007).

Accounts of learning set in deterministic tasks have shown that instead of inhibiting, monkeys are applying the prospective memory-dependent WSLS performance rule as a result of their learning set (Murray and Gaffan, 2006; Wilson et al., 2010). This rule explains the capacity to remap with only prior experience of serial discriminations (Schrier, 1966), as the rule applies equally to both tasks. Both of these functions - learning set and inhibition - have been closely associated with frontal cortex and damage to it (Miller and Cummings, 2007; Browning et al., 2007), but when the two explanations were explicitly contrasted, the learning set mechanism was clearly predictive of performance after lesions in monkeys (Wilson and Gaffan, 2008), again calling into question the cognitive inhibition process.

Our data link into this work on learning set in that both processes require the integration of temporally discontinuous information into coherent structures of action, a process strongly associated with prefrontal cortex (Fuster, 2008; Browning and Gaffan, 2008; Wilson et al., 2010). In the case of deterministic learning set, these structures need to link two consecutive trials in order to apply the performance rule. In our data, the capacity to link a longer series of outcomes over time, and to extract from those outcomes a performance rule that generalizes to all unexpected feedback, is crucial to adapting responses to volatility (Behrens et al., 2007). This process is also likely to be dependent on frontal cortex mechanisms, and more specifically the mid cingulate cortex (Kennerley et al., 2006; Quilodran et al., 2008). Our study has shown the complex time-course of this process, laying the groundwork for longitudinal electrophysiological investigation of the physiological mechanisms. Learning to learn in a probabilistic environment therefore drives formation of these extended temporal structures, inducing latent learning effects.

Materials and Methods

Subjects and materials

Three rhesus monkeys (Macaca mulatta), two females and one male, weighing 7 kg, 8 kg and 8.5 kg (monkeys P, K & D respectively) were used in this study. Ethical permission was provided by the local ethical committee "Comité d'Éthique Lyonnais pour les Neurosciences Expérimentales", CELYNE, C2EA #42, under reference C2EA42-11-11-0402-004. Animal care was in accordance with European Community Council Directive (2010) (Ministère de l'Agriculture et de la Forêt) and all procedures were designed with reference to the recommendations of the Weatherall report, "The use of non-human primates in research". Laboratory authorization was provided by the "Préfet de la Région Rhône-Alpes" and the "Directeur départemental de la protection des populations" under Permit Number: #A690290402.

Monkeys were trained to perform the task seated in a primate chair (Crist Instrument Co., USA) in front of a tangent touch-screen monitor (Microtouch System, Methuen, USA). An open-window in front of the chair allowed them to use their preferred hand to interact with the screen (all 3 monkeys were left-handed). Presentation of visual stimuli and recording of touch positions and accuracy was carried out by the EventIDE software (Okazolab Ltd, www.okazolab.com).

Behavioral tasks

Principle of the task

The task is an adaptation for monkeys of the task described for human subjects in (Collins and Koechlin, 2012). Across successive trials, a problem consisted in the monkey concurrently finding, by trial and error, the correct mappings between stimuli and targets, within a set of 2 stimuli (stimulus 1 and 2) and 3 targets (target A, B and C) (**Figure1**). For example, a problem would consist in concurrently finding the two associations: "stimulus 1 with target A" and "stimulus 2 with target C". Monkeys learned problems to a behavioral criterion. The task contained stochasticity, in the form of unreliable feedback, and volatility, in the form of switches between problems.

Monkeys initially learned a version of the task in which the volatility was made evident by changes in stimulus (**Figure 1.B**). During this version, we tracked the monkeys' learning about stochasticity and volatility of the environment. We then studied how this learning was applied in a second version of the task in which volatility was unsignalled (**Figure 1.C**).

Procedure of the task

▷ *Trial procedure*

The structure of a single trial and a single problem was always the same, regardless of the form of the task. Monkeys initiated each

trial by touching and holding a lever item, represented by a white square at the bottom of the screen (**Figure ??.A**). A fixation point (FP) appeared. After a delay period, a stimulus was displayed at the top of the screen (Stim ON signal), and was followed after a delay by the appearance in the middle of the screen of three targets (Targets ON signal). Stimuli consisted of square bitmap images of either an abstract picture or a photograph, of size 65x65mm. Targets were three empty grey squares, of the same size as the stimulus. After a further delay all targets turned white, providing the GO signal following which monkeys were permitted to make their choice by touching a target. Monkeys maintained touch on the chosen target for a fixed amount of time in order to receive visual feedback on that choice. Feedback consisted of horizontal (positive) or vertical (negative) bars within each of the three targets. A positive feedback was followed by the delivery of about 1.8 mL of 50% apple juice. After the completion of a trial, a new stimulus was picked within the set of 2 stimuli and monkeys were allowed to begin a new trial. Timing for each event gradually increased across learning to progressively train monkeys to hold their hand on the screen without moving after each action.

▷ *Problem procedure*

A problem consisted of the monkeys learning about 2 stimuli concurrently. For a given trial, one of the 2 stimuli was pseudo-randomly selected (50% of each stimulus over 10 consecutive trials). The two concurrent stimuli were never associated to the same target. Hence, there were 6 possible mappings of the 2 stimuli and the 3 targets. Each mapping was randomly selected (with the constraint that the 2 mappings of a problem had to be different from each other), so that the 2 mappings of a problem could never be predicted nor learned. The only way to find them was to proceed by trial and error based on feedback provided after each choice.

After reaching a performance criterion (defined as a total of 17 correct responses out of 20 successive trials), the problem changed and 2 new mappings were randomly selected. We refer to this change of problem as a 'Switch'. Switches only occurred after the performance criterion was reached and after a correct response. These Switches provide the volatility in the environment of the task.

In addition to and separate from this volatility, a stochastic reward environment was created by providing misleading feedback (called 'Trap feedback') on 10% of trials. Trap trials occurred pseudo-randomly once every 10 trials, with the constraint that there were at least 2 consecutive normal trials between each Trap trial. Trap feedback was the inverse of that determined by the current mapping - as such Trap feedback after a correct response consisted of negative feedback (see below) and no reward; Trap feedback after an incorrect response consisted of positive feedback and a reward.

▷ *Task Version*

We trained the monkeys in two successive

steps: 1) the Identity Task (IT) (**Figure 1.B**) and 2) the Switch task (ST) (**Figure 1.C**). The 2 tasks were strictly identical at the level of individual trials and at the level of the problems, and both tasks contained 10% Trap trials, thus monkeys learned each task directly in a probabilistic environment.

The single but crucial difference between the two tasks was the nature of the Switch between problems. In the initial IT, when the problem switched, both the identity of the stimuli and the responses were changed. That is, after a problem Switch, monkeys learned about new objects and new rules. Stimuli were always novel to the monkey in a new problem. By contrast, in the ST, monkeys worked on the same pair of stimuli throughout the session. As such only the responses were changed - the stimuli remain the same across problems, and so the monkeys were learning about new rules for the same objects after a problem Switch. Thus, Switches between problems were visually detectable in the Identity Task whereas the only way to detect a Switch in the Switch task was by trial and error using feedback on subsequent trials.

▷ *Task motivation*

In order to motivate and maintain performance at a stable level throughout each daily session, animals were asked to complete a fixed number of problems each day (number varying throughout learning between 100 and 350 trials). Upon successfully completing this number of problems, monkeys received a large reward bonus (50 ml of fruit juice, calculated based on the effectiveness in motivating the monkey).

Behavioral and statistical analyses

Principles of analyses

The major behavioral measure in these tasks was errors to criterion, the number of errors made by the monkey to reach the performance criterion of 17 correct answers out of 20 successive trials. In addition we studied the mean percentage of correct responses on specific subclasses of trials. In particular, we focused on trials around the Trap and Switch by aligning on these events and calculating percentage correct for the surrounding trials. Trap or Switch trials were referred to as trial Trap0 and Switch0 respectively.

Behavioral analyses focused on two major questions: first the learning of volatility, in the form of improvement in performance across problems and hence the formation of a learning set. Second the learning of stochasticity, in the form of changes in response to the Trap trials, which provided unreliable feedback on 10% of trials.

Analysis of volatility: learning set and errors to criterion

We analyzed the data from IT both during acquisition of the learning set, and during the subsequent stable performance. Data were analyzed up until the endpoint where the monkey had completed 400 problems with less than 50 errors to criterion. Monkeys did not

learn at the same rate, but to render performance equivalent in terms of learning progression, we separated the data into 25 bins, referred to as learning bins. Therefore the bins for each monkey did not contain exactly the same number of problems, but after the 25 bins all 3 monkeys had reached the same behaviorally defined level of stable performance (**Figure 2.A**).

We then split these data into an acquisition phase and a stable phase. The acquisition phase was completed when the learning set had been acquired to the point of stabilization of the errors to criterion per problem i.e. when the learning curve became flat. As a marker of this transition we determined the 'stability point' (dashed lines on Fig. 2.A). This was determined for each monkey by a sliding linear regression (window of 40 points) on the errors to criterion curve in order to detect when the slope would become non-significant at $p<0.05$. Data after the stability point were deemed to be in the stable period.

Analysis of stochasticity: Trap trials and logistic regression

▷ *Effect of Trap feedback on performance*

In order to initially test the effect of a Trap feedback on performance, we used Reverse Helmert coding. This compares each level of a categorical variable to the mean of the previous levels, by using a specific contrasts matrix within a generalized linear model of the binomial family. We compared performance at the trial following a Trap (trial Trap+1) with the mean performance of the trials before and including the Trap (trials Trap-3, -2, -1 and 0). In order to test for the subsequent recovery of performance, we compared performance at trial Trap+2 and Trap+3 with the mean performance of trials Trap-3, -2 and-1.

▷ *Modulation of Trap reactivity*

We observed modulations of performance after Trap trials (hereafter named Trap reactivity). Trap reactivity was modulated by a number of different factors, showing the different influences on the learning of stochasticity. To evaluate these modulations, trial-by-trial performance was fitted with logistic regressions. Models were of the form: $Y_i = \beta X_i$ where X corresponds to fixed-effects design matrix. We also measured the score of win-stay lose-shift strategy after a Trap or a normal feedback (with a score of 1 for a change or a maintenance of previous response after an incorrect or a correct feedback respectively and a score of 0 for the reverse pattern of responses) and evaluated modulations of this strategy using the same binomial models as for the performance. We also observed modulation of the counts of different targets types selected after a Trap or a Switch trial. We fitted these counts with a glm using a Poisson regression for the model 'Target' (see above). All models were applied to the two tasks, except the 'Trap or Switch' model that was applied on ST data only. All statistical procedures were performed using R (R Development Core Team 2008, R foundation for Sta-

tistical computing) and the relevant packages (MASS, car).

Different combinations of the following factors were included as explanatory variables to calibrate the different models: (1) 'Monkey' (3 levels: monkey P, K or D); (2) 'Trials' (trial-1 pre-Trap or trial 0 Trap; and trial+1 post-Trap) referring to the trial before and the trial after a Trap trial; (3) 'Learning_bins' corresponding to 25 bins of trials in the IT. The size of the groups differed for each monkey; (4) 'Trap_Valence' (positive or negative) referring to the valence of the Trap feedback (positive after an incorrect response, or negative after a correct response); (5) 'Stimulus_Similarity' (same or different), referring to the fact that the considered trial tested the same stimulus or not as the Trap trial; (6) 'Phase' (exploration or exploitation), referring to the phase within the problem. We used the following criterion: 'exploration' trials were trials associated to a performance of no more than 3/5 correct over a sliding window of 5 trials, whereas 'exploitation' trials were those with a performance of 4/5 or more; (7) 'Trap_or_Switch', referring to the identity of trials being either around a Trap or around a Switch trial. Here, only the last Trap trial before each Switch trial was considered. Similarly a factor 'Normal_or_Trap' was used for distinguishing the effects of normal versus Trap feedback; and (8) 'Target_type' ('Good', 'Second' or 'Exploratory'), referring to the type of target selected. "Good" indicates the correct target for the stimulus in the current trial; "Second" indicates the other correct target of the problem, which is incorrect in the current trial; and "Exploratory" indicates the third target, which is never correct in the current problem).

We tested the data using 5 different models to understand the different influences on the learning of volatility:

'Learning-Set' model. This model tested whether Trap reactivity was modulated across learning. It included the factors 'Monkey', 'Trials' and 'Learning_bins', and was applied selectively on trials that had the same stimulus as the Trap trial. A possible confounding factor of a learning effect on Trap reactivity was the valence of the Trap feedback. At the beginning of learning, monkeys made more errors and thus received positive feedback on Trap trials more often than at the end of learning. The effect of learning on Trap reactivity could partially be the consequence of this unequal number of positive versus negative Trap feedback, and on the unequal relevance of each feedback valence. To account for this possibility, we applied the 'Learning-Set' model on a subset of data with only positive Trap trials and on another subset with only negative Trap trials. An influence of learning on Trap feedback reactivity would be represented as a significant 'Trials x Learning_bins' interaction. Significant interactions 'Trials x Learning_bins' in both models would indicate that the effect of learning is not independent from the valence of the Trap feedback.

"Trap reactivity modulation" model. This model tested the influence of the behavioral context on established Trap reactivity, and contained the factors 'Monkey', 'Trials', 'Trap_Valence' and 'Phase'. This model was tested on the stable period of performance after the stability point. Similarly, we tested in a separate model this influence of the factor 'Stimulus_Similarity'.

"Trap or Switch" model. This model tested how fast monkeys were able to differentiate between a Trap trial and a Switch trial on the ST, when there was no stimulus change to signal the difference. It included the factors 'Monkey', 'Trials' and 'Trap_or_Switch'. To render the trials included as equivalent as possible, we included only data around each Switch and the Trap that immediately preceded it ('Last Trap'). We also selected only trials with the same stimulus as the Trap or Switch, and only in the stable period. For procedural reasons unrelated to the current experiment, Monkey K provided limited data on the ST. There were thus insufficient trials for to power this analysis, and that monkey's data were excluded.

"WSLS" model. This model tested modulations of the win-stay lose-shift strategy after a normal compared to a Trap trial, as a function of learning bins. We thus used the factors 'Normal_or_Trap' and 'Learning_Bins'.

"Target" model. This model tested how fast monkeys were able to differentiate between a Trap trial and a Switch trial, in terms of proportions of targets selected by monkeys. We thus compared counts of each type of target selected after a Trap or Switch. We included the factors 'Target_type', 'Trap_or_Switch' and 'Monkey'.

Model Selection

Models were selected using a standard procedure of constructing the model starting with all possible interactions between the included factors as described above. In a stepwise manner we evaluated the contribution of each level of fixed effect. We used the drop1 function, repeatedly testing the effect of dropping the highest-order interaction fixed-effect term on the fit (Zuur et al., 2009). Models were selected using AIC, and changes in AIC between models were tested using a chi-square test ($P < 0.05$). The principle of model selection was identical for all models. It should be noted that the factor Monkey was included in these models, accounting for individual differences between monkeys and improving fit.

Reaction times

Reaction times were calculated as the time between the GO signal and the lever release (in order to further select a target on screen). Measures beyond 2 seconds were not included in the analysis. Due to a technical fault in the software, reaction time measurements for monkey P were inaccurate during the first task (IT), and were excluded from analyses. This fault was corrected for the second task (ST).

Acknowledgements

This work was supported by Agence Nationale de la Recherche, Fondation Neurodis (C.R.E.W.), Fondation pour la Recherche Médicale (M.C.M.F.), and by the labex CORTEX ANR-11-LABX-0042. C.R.E.W. is funded by a Marie Curie Intra-European Fellowship (PIEF-GA-2010-273790). M.C.M.F. is funded by Ministère de l'enseignement et de la recherche. EP is funded by Centre National de la Recherche Scientifique. We thank K. Knoblauch for assistance with statistical methods, F.Stoll for advice, M. Valdebenito, M. Seon, and B. Beneyton for animal care and C. Nay for administrative support. Conflict of Interest: None declared.

Author Contributions

M.C.M.F, EP & C.R.E.W designed the research; M.C.M.F & C.R.E.W performed the research and analyzed the data; M.C.M.F, EP & C.R.E.W wrote the paper.

References

Bavelier, D, Shawn Green, C, Pouget, A, and Schrater, P. Brain plasticity through the life span: Learning to learn and action video games, 2012.

Behrens, Timothy E J, Woolrich, Mark W, Walton, Mark E, and Rushworth, Matthew F S. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9): 1214–1221, 2007.

Biernaskie, Jay M., Walker, Steven C., and Gegear, Robert J. Bumblebees learn to forage like Bayesians. *The American Naturalist*, 174(3):413–423, 2009.

Browning, Philip G F and Gaffan, David. Prefrontal cortex function in the representation of temporally complex events. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(15):3934–3940, 2008.

Browning, Philip G F, Easton, Alexander, and Gaffan, David. Frontal-temporal disconnection abolishes object discrimination learning set in macaque monkeys. *Cerebral Cortex*, 17 (4):859–864, apr 2007.

Collins, Anne and Koechlin, Etienne. Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10(3):e1001293, 2012.

Collins, Anne G E and Frank, Michael J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190–229, 2013.

Costa, Vincent D, Tran, Valery L, Turchi, Janita, and Averbeck, Bruno B. Dopamine modulates novelty seeking behavior during decision making. *Behavioral neuroscience*, 128(5):556–66, 2014.

Courville, Aaron C., Daw, Nathaniel D., and Touretzky, David S. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

Donoso, Maël, Collins, Anne G E, and Koechlin, Etienne. Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science (New York, N.Y.)*, 344(6191):1481–6, 2014.

Fuster, J M. *The Prefrontal Cortex*, volume 1. 2008.

Gallistel, C R, Mark, T a, King, a P, and Latham, P E. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *Journal of experimental psychology. Animal behavior processes*, 27(4):354–372, 2001.

Gershman, Samuel J, Blei, David M, and Niv, Yael. Context, learning, and extinction. *Psychological review*, 117(1):197–209, 2010.

Harlow, H F. The formation of learning sets. *Psychological review*, 56(1):51–65, jan 1949.

Hertwig, Ralph, Barron, Greg, Weber, Elke U., and Erev, Ido. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539, 2004.

Izquierdo, Alicia, Suda, Robin K, and Murray, Elisabeth a. 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):7540–7548, 2004.

Kennerley, Steven W S.W., Walton, Mark E M.E., Behrens, Timothy E J T.E.J., Buckley, M.J. Mark J, and Rushworth, M.F.S. Matthew F S. Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9(7):940–947, jul 2006.

Kornell, Nate, Son, Lisa K., and Terrace, Herbert S. Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18(1):64–71, 2007.

McGuire, Joseph T, Nassar, Matthew R, Gold, Joshua I, and Kable, Joseph W. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, 84(4): 870–881, 2014.

McNamara, John M. and Houston, Alasdair I. Optimal foraging and learning. *Journal of Theoretical Biology*, 117(2):231–249, 1985.

Miller, Bruce L and Cummings, Jeffrey L. The human frontal lobes: Functions and disorders (2nd ed.). In *The human frontal lobes: Functions and disorders (2nd ed)*, pages xx, 666. 2007.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei a, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, Petersen, Stig, Beattie, Charles, Sadik, Amir, Antonoglou, Ioannis, King, Helen, Kumaran, Dharshan, Wierstra, Daan, Legg, Shane, and Hassabis, Demis. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Murray, Elisabeth a and Gaffan, David. Prospective memory in the formation of learning sets by rhesus monkeys (Macaca mulatta). *Journal of experimental psychology. Animal behavior processes*, 32(1):87–90, 2006.

Ollason, J G. Learning to forage–optimally? *Theoretical population biology*, 18(1):44–56, 1980.

Payzan-LeNestour, Elise and Bossaerts, Peter. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, 7(1):e1001048, 2011.

Pleskac, Timothy J. Decision making and learning while taking sequential risks. *Journal of experimental psychology. Learning, memory, and cognition*, 34(1):167–185, 2008.

Quilodran, René, Rothé, Marie, and Procyk, Emmanuel. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2):314–325, 2008.

Rudebeck, Peter H, Behrens, Timothy E, Kennerley, Steven W, Baxter, Mark G, Buckley, Mark J, Walton, Mark E, and Rushworth, Matthew F S. Frontal cortex subregions play distinct roles in choices between actions and stimuli. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(51):13775–13785, 2008.

Schrier, Allan M. Transfer by macaque monkeys between learnin-set and repeated -reversal tasks. *Perceptual and Motor Skills*, (23):787–792, 1966.

Stuss, Donald T and Alexander, Michael P. Is there a dysexecutive syndrome? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1481): 901–915, 2007.

Tebbich, Sabine and Teschke, Irmgard. Coping with Uncertainty: Woodpecker Finches (Cactospiza pallida) from an Unpredictable Habitat Are More Flexible than Birds from a Stable Habitat. *PLoS ONE*, 9(3):e91718, 2014.

Walton, Mark E., Behrens, Timothy E J, Buckley, Mark J., Rudebeck, Peter H., and Rushworth, Matthew F S. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, 65 (6):927–939, 2010.

Wilson, Charles R E and Gaffan, David. Prefrontal-inferotemporal interaction is not always necessary for reversal learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(21):5529–38, may 2008.

Wilson, Charles R E, Gaffan, David, Browning, Philip G F, and Baxter, Mark G. Functional localization within the prefrontal cortex: Missing the forest for the trees? *Trends in neurosciences*, 33(12):533–40, dec 2010.

Zuur, Alain F., Ieno, Elena N., Walker, Neil J., Saveliev, Anatoly a, Smith, Graham M., and Ebooks Corporation. *Mixed Effects Models and Extensions in Ecology with R*. 2009.

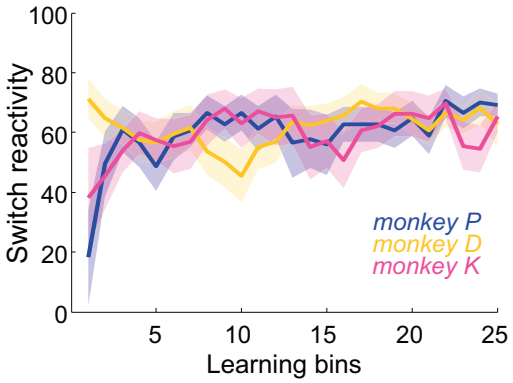

Figure S1: Switch reactivity across learning bins in the IT. The Switch reactivity score can be thought of as the size of the drop in performance between trials Switch-2 and Switch 0 (noting that Switch-1 is by definition 100%). Monkeys rapidly learn to adapt to Switches in this task, in which they are signaled by new stimuli. Shaded areas represent sem. Monkey P, blue; monkey D, yellow and monkey K, pink.

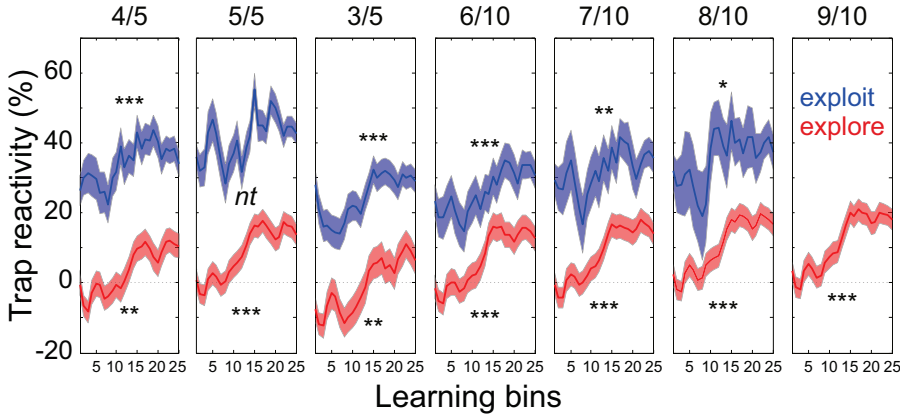

Figure S2: Effects of modulation of the exploration/exploitation criterion on Trap reactivity over learning bins in the IT. The first panel shows the criterion that is currently used in this paper (4/5 : exploitation when 4 correct trials out of 5). Other panels show the effects of modulation of the number of correct trials used for the criterion (either 5/5 or 3/5) and modulations for another sliding window (6/10 to 9/10). For visibility, data are concatenated accross monkeys. Stars indicate significant increase of Trap reactivity with learning bins (glm, Trials x Learning_bins, ***: $p<0.001$; **: $p<0.01$; *: $p<0.05$; nt: not tested).Statistics are missing forTrap reactivity in the 5/5 panel because all performance are at 100% in exploitation for the 5/5 criterion preventing from using the test. Data are not shown for exploitation with the 9/10 criterion because of an insufficient number of trials. Red: exploration (trials with performance <criterion), blue : exploitation (trials with performance >criterion).

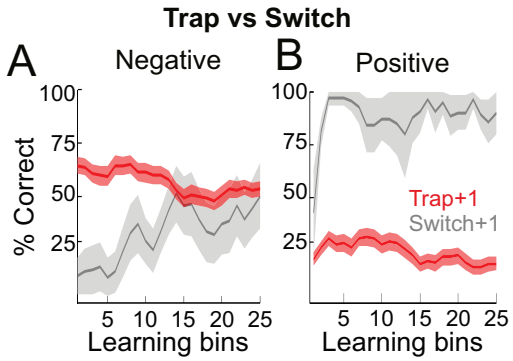

Figure S3: Comparison of performance at trial Trap+1 (red) and Switch+1 (gray) after a **A:** negative or **B:** positive feedback (concatenated data from the 3 monkeys).

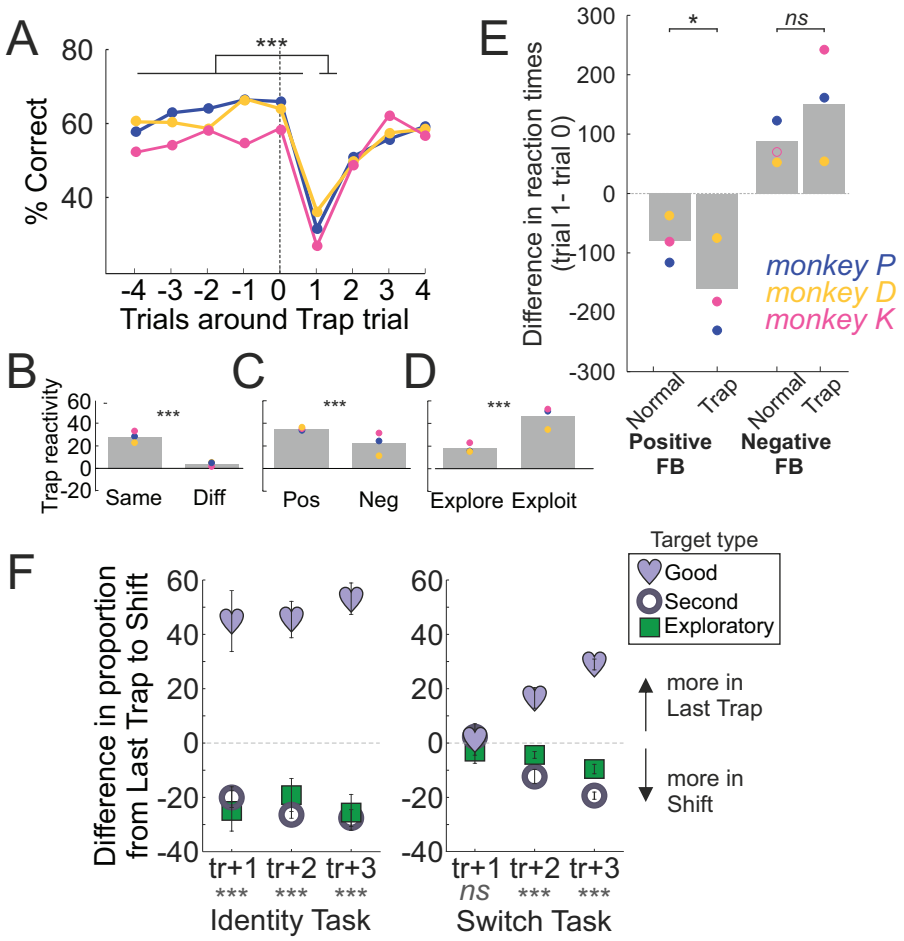

Figure S4: Related to Figure 5. A. Trap reactivity is maintained in ST: Percentage correct calculated for trials around Trap trial (trial 0). This Trap reactivity maintains similar properties to the IT as follows: **B.** Trap reactivity present for trials of the same stimulus to the Trap, **C.** Trap reactivity greater for Positive than Negative Traps, and **D.** Trap reactivity greater in exploitation. **E.** Effect of feedback on reaction times, with exactly the same calculations and conventions as **Figure 4 C.** Post-correct speeding is present and greater after Trap trials. Post-error slowing is present in all but one case (open circle: non-significant difference between trials 0 and 1; closed circles – significant difference with $p<0.001$). Post- error slowing is not significantly increased for Trap trials in this task. **F. Difference in proportion of chosen target following Last Trap versus Switch, for the 3 following trials, for the 2 tasks**. The 3 'Target types' refer to the rules of the problem of the Last Trap. "Good" indicates the correct target for the given stimulus; "Second" indicates the other correct target of the mapping, that is incorrect for the given stimulus; and "Exploratory" indicates the third target, that is never correct in the mapping in which the Last Trap occurred. The pa ern of target selection after a Last Trap or a Switch did differ between the two tasks. Monkeys immediately differentiated Traps from Switches in the IT, even one trial after unexpected feedback, by using the cue of new stimuli. In the ST, however, this differentiation was absent in trial+1, as should be expected given the Switch is not cued. But after trial+2, monkeys had sufficient information to distinguish Trap from Switch trials, and quickly made the distinction. As such, monkeys were indeed sensitive to the lack of switch cue, but adapted their behaviour so rapidly that there were no significant differences in errors to criterion. P-values correspond to the interaction Proportions x LastTrapOrSwitch. Bars represent sem across monkeys. ***: $p<0.001$; *: $p<0.05$; ns: non-significant.

Chapitre 6

Etude des Potentiels Evoqués :

Modulations des potentiels évoqués au feedback fronto-pariétaux dans un environnement incertain

Dans cette étude, nous nous sommes intéressés aux corrélats électrophysiologiques des processus qui se sont mis en place suite à l'acquisition du learning set. En particulier, nous avons étudié les modulations des potentiels évoqués au feedback, enregistrés en surface au niveau des aires frontales et pariétales, en fonction des paramètres de la tâche.

Feedback Related Potentials encode the feedback unexpectedness and subsequent behavioral adaptation in an uncertain environment in monkeys

Maïlys C.M. Faraut[1,2,3], Charles R.E. Wilson[1,2,3]
and Emmanuel Procyk[1,2]

[1]INSERM U846, Stem Cell and Brain Research Institute, Bron 69500, France
[2]Université de Lyon, Lyon 1, UMR S-846, Lyon 69003, France
[3]Corresponding authors: mailys.faraut@inserm.fr & charles.wilson@inserm.fr

Keywords: feedback, adaptation, flexibility, learning, monkey

Abstract

Adaptive behavior relies on the interplay between accurate performance monitoring and relevant subsequent adjustments of strategies. Feedback Related Potentials (FRPs), which can be recorded over the frontal midline of the scalp, are thought to emerge from activity in the mid-cingulate cortex. They have been associated with the detection of unexpected events that are relevant for adaptation. However, a current debate concerns whether modulations of FRPs amplitude are predictive of subsequent behavioral adjustments. In this study, we recorded FRPs from one monkey performing a probabilistic reversal task in which the stochasticity of rewards combined to the volatility of the environment require a steady reactive and flexible behavior. We show that FRPs were modulated by feedback valence and the degree of unexpectedness of the feedback. Interestingly, FRPs were also predictive of the strategy (shift or stay) selected in the following trial. We discuss whether it is possible to dissociate a surprise effect from the implementation of behavioral adaptation in these protocols.

Introduction

Our environment is complex : it is changing, stochastic and sometimes both; which brings uncertainty to nearly every decision we make (Huettel et al., 2005). In order to reduce uncertainty about consequences of their actions, organisms have developed strategies to learn from the environment. These strategies rely on feedback detection and evaluation, enabling subsequent behavioral adjustments (Miller and Cohen, 2001). Investigations of the neural correlates of these processes have revealed a major role of the midcingulate cortex (MCC) for performance monitoring, in interaction with the dorsolateral prefrontal cortex (DLPFC) for implementing control (Rothé et al., 2011; Voloh et al., 2015). The MCC is thought to be the source of the Feedback Related Negativity (FRN), a negative deflection recorded at the surface of the brain over fronto-central electrodes after feedback onset (Hauser et al., 2014; Miltner et al., 1997). The most recent accounts of the FRN suggest that its amplitude is modulated by the unexpectedness or degree of surprise of an outcome, suggesting that it represents a kind of unsigned reward prediction error (Ferdinand et al., 2012; Hauser et al., 2014; Holroyd and Krigolson, 2007; Oliveira et al., 2007; Walsh and Anderson, 2011). Thus, the FRP was suggested to be involved in the detection of feedback relevant for adaptation (Vezoli and Procyk, 2009).

A current behavioral theory postulates that learning can be driven by the surprisingness of an event, regardless of its valence (Courville et al., 2006). According to the reinforcement learning theory of the error related negativity (RL-ERN), as learning changes the degree of unexpectedness of outcomes, it should be associated to modulations of the ERN (Holroyd and Coles, 2002). This has been shown in many studies in which a transfer from the time of the feedback to the time of the response with learning was evident in the anti-correlated changes between the FRN and the ERN (Groen et al., 2007; Pietschmann et al., 2008), but also in studies using high versus low probability outcomes to show modulations of the FRN depending on the degree of unexpectedness of outcomes (Cohen et al., 2007; Walsh and Anderson, 2011).

But an ongoing debate concerns whether the FRN is actually being used for adapting subsequent behavior, in addition to signaling the prior outcome. It is conceivable that these are separate processes – there are many steps in the process between signaling a need to adapt behavior and subsequently successfully doing so. Further, given that the FRN is a summated surface signal, it remains possible that the subtleties of encoding subsequent adaptation are not detectable within it. Nevertheless, some studies have reported FRN changes associated to subsequent behavioral adjustments. For example, the FRN amplitude was found to be predictive of adjustments in future behavior by discriminating whether subjects will choose the same or opposing target on the following trial (Cohen

et al., 2007). The FRN has also been shown to predict inter-individual differences in learning from feedback as its magnitude was larger in negative learners (individuals who learned more from the negative feedback) than in positive learners (Frank et al., 2005).

However, many studies also found no relationship between FRN modulations and related behavioral adaptation. In a study by Chase and colleagues using a probabilistic reversal task, subjects were given the particular instruction not to shift until they were certain that a shift really occurred (Chase et al., 2010). FRN amplitude was related to the degree of expectedness of the feedback, by being more negative for surprising negative feedback (due to stochastic feedback or Shift) than for expected negative feedback (when subjects were voluntary perseverating to confirm that a shift really occurred). Thus, FRN amplitude was bigger when subjects didn't change, and smaller after subjects decided to change, suggesting that amplitude was not related to subsequent behavioral changes. Similarly, Mars and colleagues compared different types of feedback leading to different levels of behavioral adjustments but found no relationship with FRN amplitude (Mars et al., 2004). In a study using a probabilistic learning task, FRN magnitude was first shown to increase in parallel to learning. But in a second condition no learning was required as participants were given instructions about the reward probabilities of the different stimuli in advance, and indeed performance was immediately at asymptote. Nevertheless the magnitude of the FRN came to discriminate the different probabilities only after several trials, showing a supposed learning effect despite no behavioral evidence of learning, and therefore demonstrating a complete dissociation between performance and FRN levels (Walsh and Anderson, 2011). Thus, it is still unclear whether the FRN is actually limited to the detection of unexpected feedback relevant for adaptation, signaling the need for adaptation or whether it is also used for behavioral adaptation.

One confounding element of many studies is that the need for adaptation is often related to losses, and larger losses are themselves associated with larger FRN. Hence, a design that can fully respond to this question requires that incorrect feedback does not always signal the need for a change of behavior. To some extent this was the case in the study by Chase et al., but in that case it was achieved by overcoming natural response tendencies with specific instructions. Probabilistic reversal tasks do provide the sorts of trials required, at least to some extent. Probabilistic feedback means that there will be unexpected negative (and unexpected positive) outcomes, which the subject should have learned to ignore – and that should therefore not trigger adaptation. Thus, each feedback valence has potentially differing implications for adaptation, depending on the learning and task context. In these tasks, the challenge is to be able to distinguish irrelevant from relevant feedback because they should trigger dif-

ferent levels of behavioral adjustment.

In this study, we recorded FRPs from one monkey performing a probabilistic reversal task in which he had to flexibly adapt to the stochastic and volatile environment. We show that FRPs were modulated by feedback valence and degree of unexpectedness, but were also predictive of the upcoming strategy that the monkey will use at the following trial.

Results

Behavior

Monkey D adapted flexibly to the changing and stochastic environment created by the probabilistic reversal task (**Figure 1**) with a mean accuracy per problem of 65.82% +/- 2.57 (**Figure 2.A**) for a mean of 3.8 switches per session. Reaction times were not different after a negative compared to a positive feedback (two-sample Kolmogorov-Smirnov test, $p<0.05$) (**Figure 2.B**). During the progressive solving of a problem, his behavior alternated between periods of exploration and exploitation, as well as periods of re-exploration, for a given stimulus (see Methods) (**Figure 2.C&D**). As learning progressed, he also shifted more his response compared to the previous trial with the same stimulus (**Figure 2.E**). Monkey D was reactive to Trap trials as shown by a decrease of performance after a Trap trial (**Figure 2.F**, see Etude Comportementale). This change in behavior was characterized by the use of different strategies following normal and Trap trials: for most trials, he used a Win-Stay Lose-Shift strategy, whereas for Trap trials, he used a particular exploratory strategy (**Figure 2.G**). Thus, monkeys didn't learn to ignore Trap trials as they shifted their choice compared to their pre-Trap strategy, probably because of the presence of the uncued shifts requiring a constant reactivity to unexpected feedback.

Trial-to-trial and within problem modulations of feedback related potentials

Monkey D was implanted with an ECoG implant covering the surface of the frontal and parietal cortices (**Figure 3**). This technique enabled us to investigate modulations of evoked-related potentials by the various factors of the task. We focused our analysis on the feedback period of the task, a crucial period for performance monitoring.

We first confirmed that Feedback related potentials (FRPs), in particular at the time-window usually described for the FRN (taken here as: 0.20 and 0.35 sec post-feedback) encoded feedback valence. As expected, the FRP significantly discriminated positive from negative feedback in all frontal and parietal electrodes, by showing a more negative amplitude after negative compared to positive feedback (permutation test based on unpaired t-test, $p<0.05$ for all electrodes) (**Figure 4**). We then looked at whether this effect would appear at different times depending on the recorded cortical region, but no significant difference in the latency of the effect was observed between frontal (mean latency: 0.18

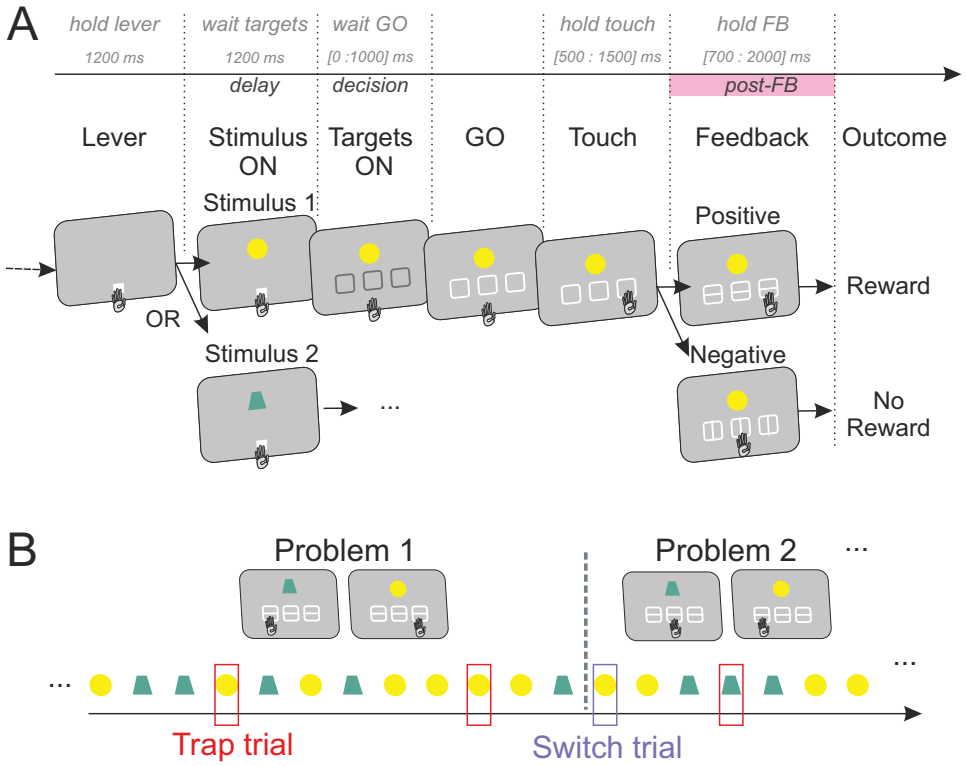

Figure 1: Probabilistic Reversal Task. A. Structure of a single trial. After holding a touch screen 'lever', one of 2 stimuli appears, followed by the onset of three targets. After a GO signal, the monkey selects one target by touching it. Visual feedback is represented by horizontal or vertical bars superimposed on all three targets for positive or negative feedback respectively. Positive feedback was followed by the delivery of juice reward. **B.** Example of a series of trials. Each problem constituted two mappings between the two stimuli and one of the responses. One stimulus at a time was randomly presented. On Trap trials, a misleading feedback was given to the monkey: a positive feedback with a juice reward after an incorrect choice; or a negative feedback with no reward after a correct choice. One Trap trial occurred pseudo-randomly within each set of 10 trials. After a random number of trials (between 60 and 85 trials), the problem changed (Switch trial), and 2 new mappings with the same 2 stimuli were randomly selected. Thus, Switches between problems were not predictable.

sec +/- 0.01) and parietal (mean latency: 0.15 sec +/- 0.01) electrodes (unpaired t-test, $p<0.05$) (**Figure 4.D**).

Beyond the encoding of valence, the FRP has also been suggested to be an index of the level of unexpectedness of an outcome, which should change as learning progresses. We thus checked whether a feedback of a given valence was encoded differently in the FRP depending on the phase of learning within a problem (**Figure 5 & 6**). We considered the exploration and re-exploration periods as phases with ongoing learning whereas the exploitation period as the phase with the most advanced learning (or requiring less cognitive control) (see Methods). The expectedness effect should have differing impact on the neural response to correct and incorrect feedback depending on these periods.

As learning progresses, positive feedbacks

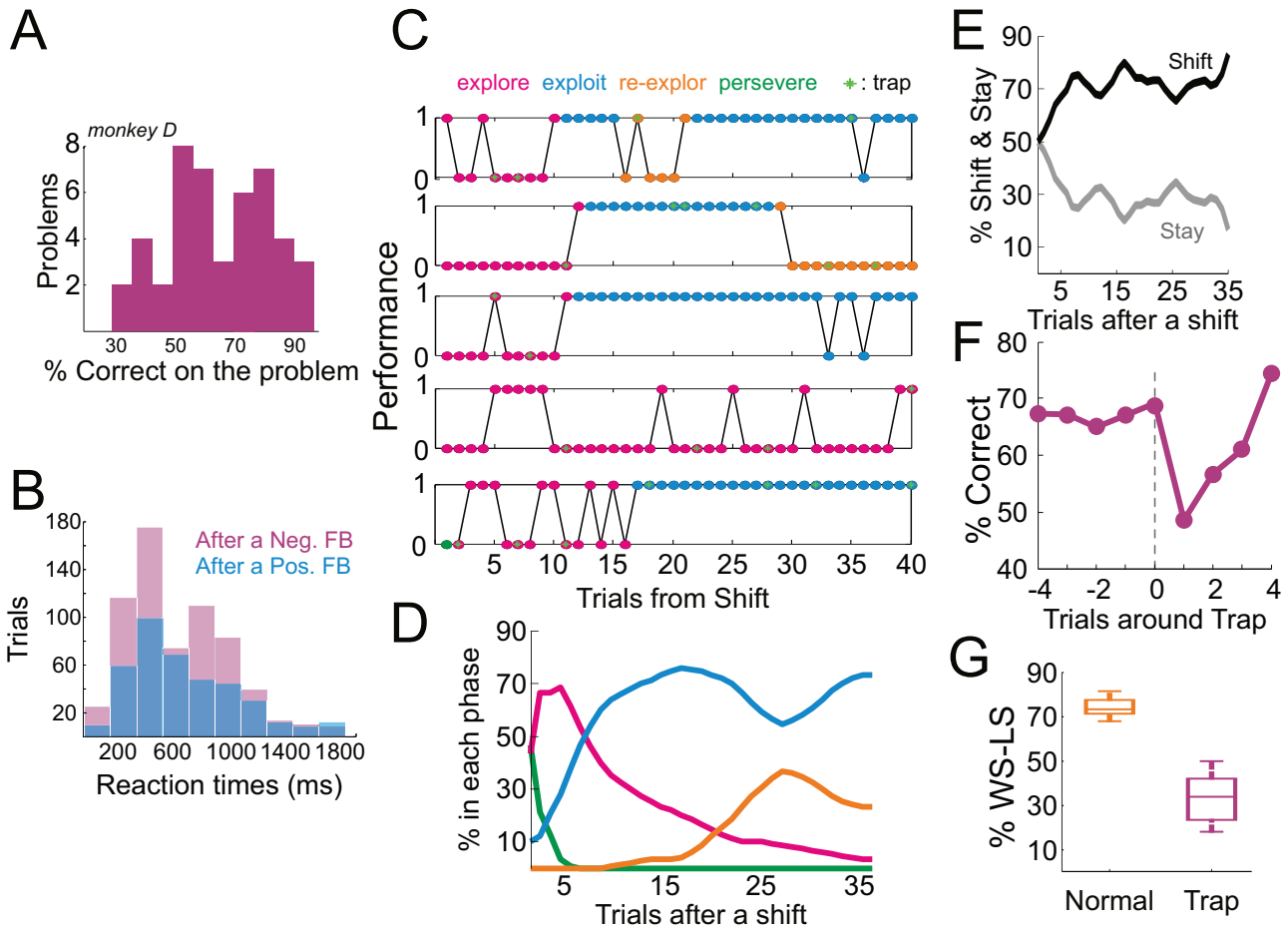

Figure 2: Behavior of monkey D in the task on 12 recordings sessions. **A.** Percentage correct per problem. **B.** Reaction times depending on whether the monkey received a positive (pink) or negative (blue) feedback on the previous trial. **C to G**: Analyses carried out on trials on a single stimulus, to show treatment of individual stimuli (note that monkeys in fact learned concurrently about 2 stimuli). **C.** Examples of performance on series of trials after a switch for trials of a specific stimulus and illustration of the "Phase" criterion: exploration (red), exploitation (blue), re-exploration (orange), perseveration (green) trials. Trap trials are also shown (green stars). **D.** Post-Shift evolution of the percentage in each phase of trials of a specific stimulus (same color code as in C). **E.** Post-Shift evolution of the percentage of trials leading to either a shift (black) or a stay (gray) at the next trial with the same stimulus. **F.** Performance around Trap trials (referred as trial 0) on trials with the same stimulus. **G.** Percentage of Win-Stay Lose-Shift strategy after a Normal or a Trap trial on trials with the same stimulus.

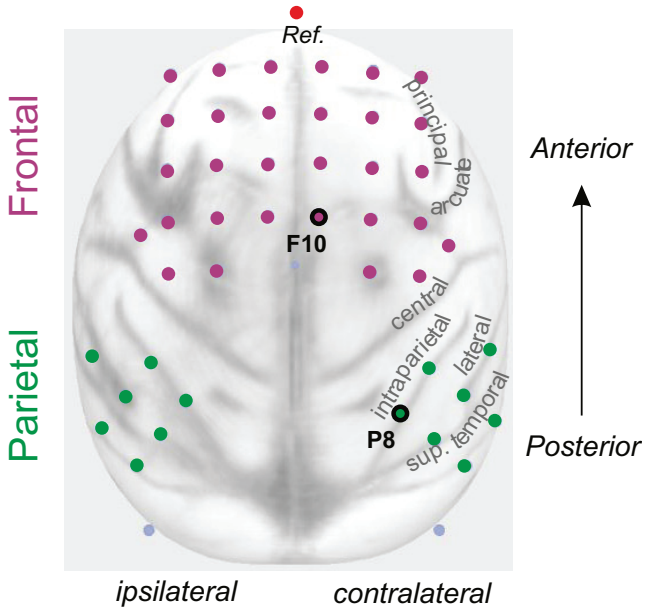

Figure 3: Location of intra-cranially implanted ECoG electrodes with the names of major sulci in the underlying cortex. Positions of example electrodes (F10 over frontal and P8 over parietal cortex) are also indicated (black circles). The reference electrode (Ref) is located on the skull, at the apex of frontal regions.

become more frequent and thus, a positive feedback received at the end of learning should be less surprising than at the beginning. Thus, if more negative FRP magnitude reflects more unexpected feedback, we should expect a more negative FRP (that we will call FRPp) in exploration or re-exploration compared to exploitation for positive feedback. The FRPp peak was not different in the exploration compared to the exploitation phase (permutation test based on unpaired t-test between explore and exploit conditions, $p>0.05$ for all electrodes) (**Figure 5.A & B**), hence failing to meet with the prediction. However, the FRPp was significantly more negative in re-exploration compared to exploitation (permutation test, $p<0.05$ for the electrodes indicated on the **Figure 5.C**, 3rd panel). Re-exploration is a complicated part of the data to interpret. As can be seen in **Figure 2.C**, re-exploration seems to occur when the monkey anticipates a switch and responds as if the switch has already occurred, a behaviour potentially driven by the Trap feedback. If the monkey does believe that a switch has occurred, positive feedback again becomes unexpected, and so a more negative FRPp to positive feedback is to be expected. Re-exploration, however, is open to alternative interpretations.

We predicted an opposite pattern for negative feedback. Their occurrence decreases with learning, such as a negative feedback received at the end of learning should be more surprising. Thus, the FRN peak should be the most negative in exploitation. We observed such an evolution of the FRN after negative feedback (FRPn) with learning phases

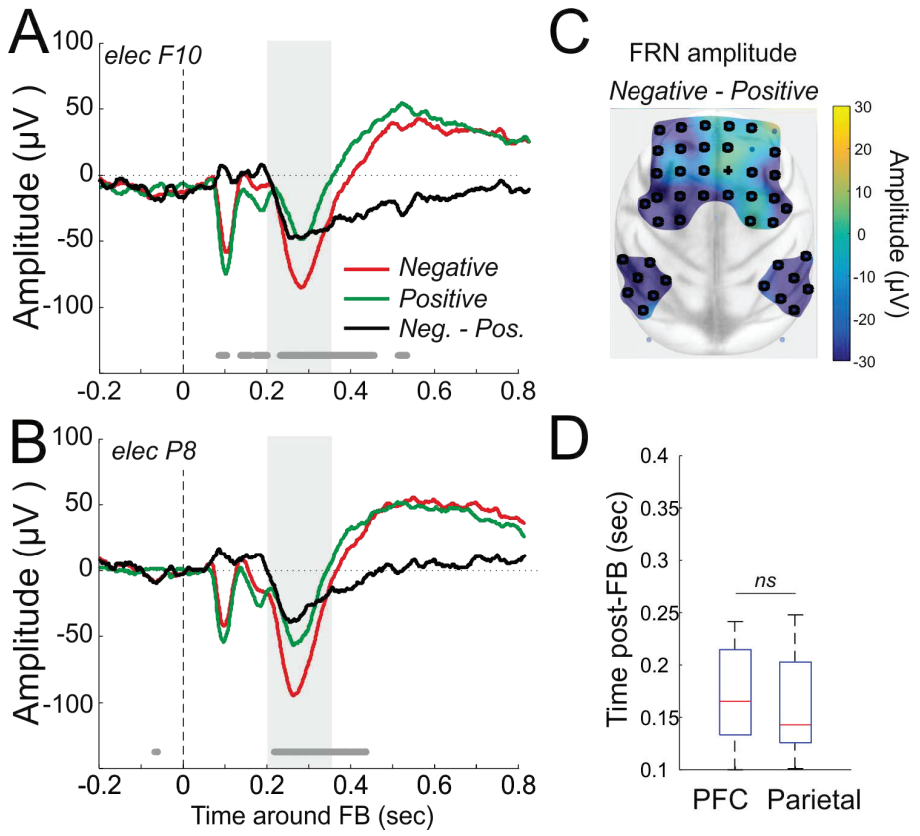

Figure 4: Effect of feedback valence on Feedback related potentials, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), for negative (red) and positive (green) feedback. The black line represents the Negative – Positive feedback difference. Gray rectangle represents the time-window considered for the FRN (0.2 to 0.35 sec post-feedback). Horizontal bars at the bottom of the plot represent a significant difference between the conditions (permutation test, $p<0.05$). **C.** Cartography of the mean 'Negative-Positive' amplitude difference over the FRN time window. Empty circles represent a significant negative difference (permutation test, $p<0.05$). **D.** Box-plot comparing the repartition of the latencies of the effect between frontal and parietal electrodes (non-paired t-test). ns:non-significant.

(**Figure 6**). The FRPn amplitude was more negative in exploitation compared to exploration and re-explorations phases, especially in the contra-lateral electrodes (permutation test, $p<0.05$ for the electrodes indicated on the **Figure 6.C**, 1st and 3rd panels).

We thus showed, in accordance with the literature, that the FRP magnitude was modulated by feedback valence but also by feedback expectedness, whatever the valence, as shown by modulations across learning phases (summary of the effects in **Figure 7**, left panel). However, one major debate in the FRP literature concerns whether FRPs, which carry information about feedback, can also be associated with subsequent behavioral adaptation based on this information. To test this, we looked at whether FRP amplitude modulations were predictive of the strategy that the monkey was to apply in the next trial (labeled

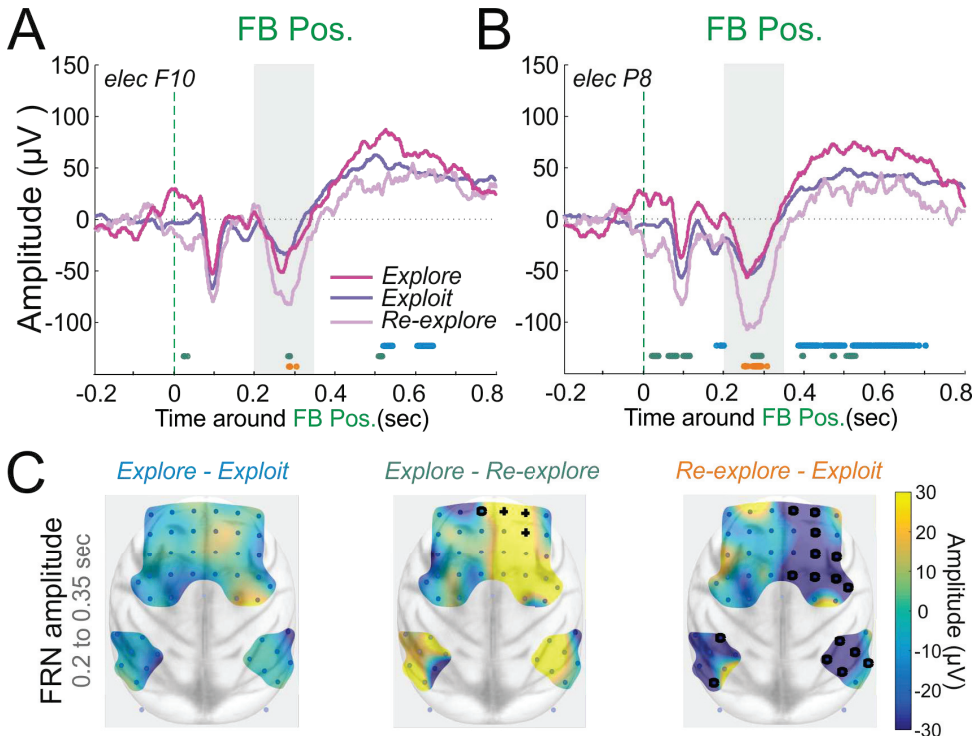

Figure 5: Effect of intra-problem learning phase on Feedback related potentials after a positive feedback, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), in exploration (dark pink), exploitation (purple), and re-exploration (mauve) trials. Horizontal bars at the bottom of the plot represent a significant difference between the conditions: explore vs exploit (blue), explore vs re-explore (green) and re-explore vs exploit (orange) (permutation test, $p<0.05$). **C.** Cartography of the mean amplitude difference over the FRN time window. Crosses and empty circles represent a significant positive and negative difference respectively (permutation test, $p<0.05$).

as 'stay' vs 'shift', referring to whether the monkey kept or changed his response on the next trial with the same stimulus). FRPs significantly discriminated between those conditions, in most frontal and parietal electrodes, by showing a more negative amplitude for the 'Shift' compared to the 'Stay' condition (permutation test based on unpaired t-test between shift and stay conditions, $p<0.05$ for the indicated electrodes) (**Figure 8**). No significant difference was observed between frontal (mean latency: 0.16 sec +/-0.01) and parietal (mean latency: 0.14 sec +/-0.01) electrodes (unpaired t-test, $p<0.05$) (**Figure 8.D**).

These results suggest that a more negative peak of the FRP reflects a change of strategy at the next trial, as it is associated to either a negative feedback, signaling the need for a change or in trials leading to a subsequent shift. However, given that shift trials mostly occur after negative feedback (as monkeys use a WSLS strategy, **Figure 2.G**), the negative deflection could be driven by the negative valence of the feedback, and not by the need to change behavior. One way to address this

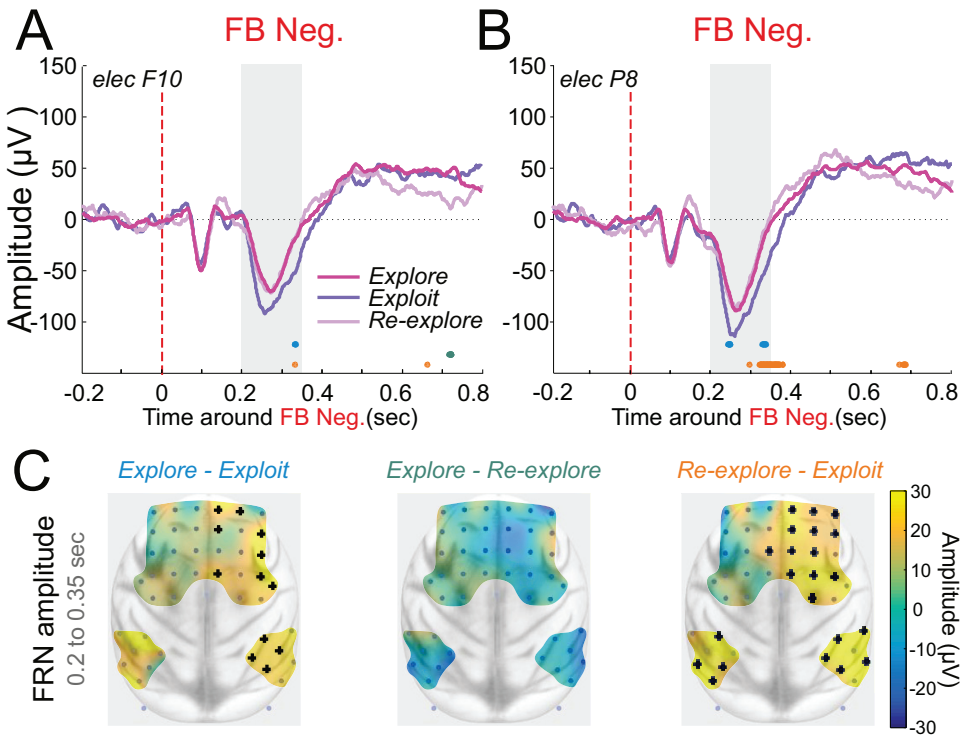

Figure 6: Effect of intra-problem learning phase on Feedback related potentials after a negative feedback, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), in exploration (dark pink), exploitation (purple), and re-exploration (mauve) trials. Horizontal bars at the bottom of the plot represent a significant difference between the conditions: explore vs exploit (blue), explore vs re-explore (green) and re-explore vs exploit (orange) (permutation test, $p<0.05$). **C.** Cartography of the mean amplitude difference over the FRN time window. Crosses and empty circles represent a significant positive and negative difference respectively (permutation test, $p<0.05$).

confound is to look at the amplitude of the FRN after positive feedback that was followed by a shift. If the negative deflection reflects the subsequent behavioral change, the FRPp should be higher in shift compared to stay trials; whereas, if it is reflecting negative feedback valence, we should not see any difference. The FRPp magnitude was more negative in the 'shift' compared to the 'stay' condition (permutation test based on unpaired t-test between shift and stay conditions, $p<0.05$ for the indicated electrodes) (**Figure 9.A, B & C**) suggesting an encoding of the subsequent behavioral adaptation in the FRPp. There was no difference between 'shift' and 'stay' conditions after a negative feedback (permutation test, $p>0.05$ for all electrodes) (**Figure 9.D, E & F**) which suggests a strong encoding of the negative valence whatever the subsequent decision, in the case of negative feedback. One might argue that negative feedback is such a strong signal to adapt behavior that the large negative deflection is masking any potential effect of that actual subsequent choice of the monkey. A summary of these effects is shown in **Figure 7** (right panel).

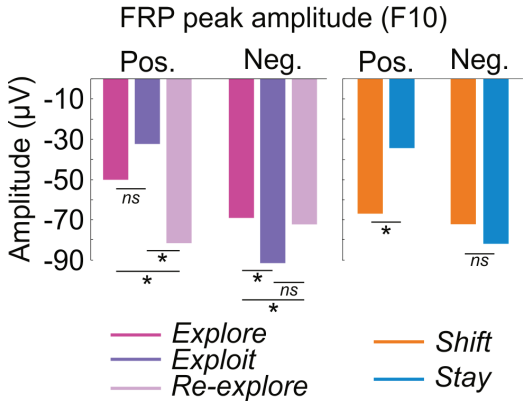

Figure 7: Summary of the effects on the FRP of feedback valence depending on the phase and on the upcoming strategy, shown here as the maximum peak amplitude at the electrode F10, during the 0.2 to 0.35 sec post-feedback time-window. The statistics correspond to the previously described permutation tests. *: p<0.05.ns: non-significant.

However, another hypothesis for the most negative deflection after positive feedback leading to a shift compared to a stay trial could be that those trials were mostly unexpected trials (as unexpected trials were related to more negative peaks, as shown in **Figure 5**). On 184 trials with a positive feedback followed by a shift, 53% were in exploration, 23% in exploitation and 22% in re-exploration. Thus, the majority of trials were from a learning phase during which positive feedback were the most unexpected (exploration). Consequently, we cannot rule out the fact that bigger negative deflections followed by a shift are more reflecting the unexpectedness of the feedback than the need for subsequent behavioral adaptation. A way to do it would be to compare the FRP amplitude after a positive feedback leading to a shift in exploration compared to exploitation. If the more negative FRP amplitude reflects the unexpectedness, it should be more negative in exploration compared to exploitation. If it reflects the need of behavioral change, no difference should be observed. No difference was observed on the contralateral electrodes (but a significant difference was present in ipsilateral electrodes, permutation test, p<0.05 for the indicated electrodes) (**Figure 10**), which makes the conclusions hard to draw.

Our last test to disambiguate the effect of unexpectedness and of subsequent adaptation was to look at whether the FRP was differently modulated after Trap and No Trap trials in exploitation. Indeed, in exploitation, given that the correct association is known, Trap feedback should be surprising but should be ignored and not lead to subsequent adaptation. We thus compared Trap and No Trap trials with the same feedback valence, during the exploitation period, that is when the monkey should be able to differentiate the two. We only compared negative Trap versus No Trap trials because there weren't enough pos-

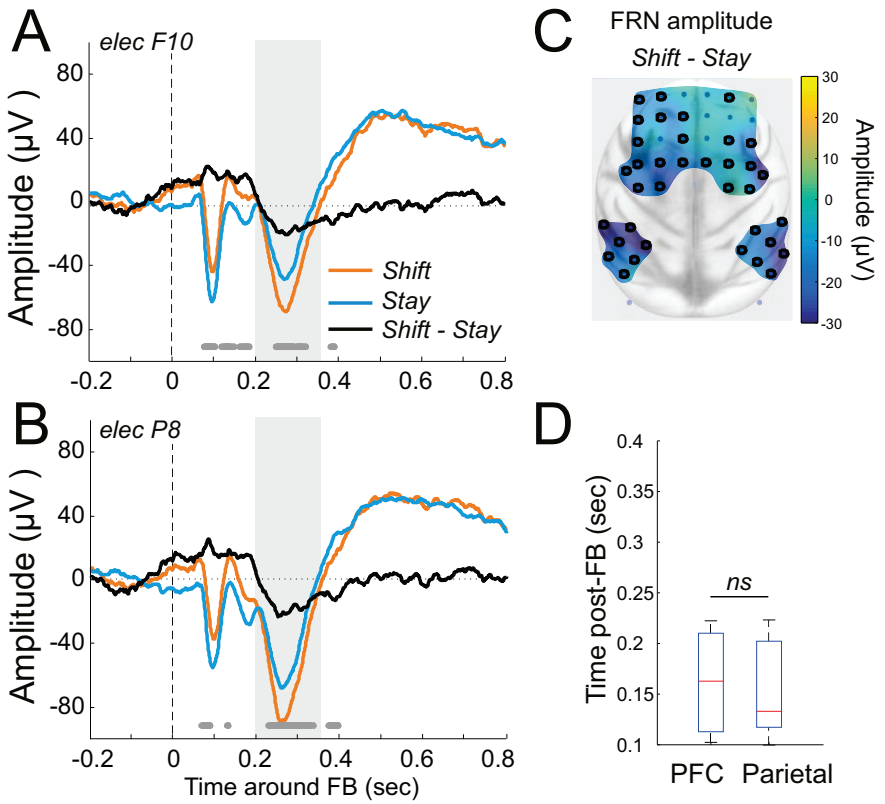

Figure 8: Effect of the upcoming strategy (shift vs stay) on Feedback related potentials, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), for Shift (orange) and Stay (blue) trials. The black line represents the 'Shift-Stay' difference. Horizontal bars at the bottom of the plot represent a significant difference between the conditions (permutation test, $p<0.05$). **C.** Cartography of the mean 'Shift-Stay' amplitude difference over the time window of the FRN. Black crosses and Empty circles represent a significant positive and negative difference respectively (permutation test, $p<0.05$). **D.** Box-plot comparing the repartition of the latencies of the effect between frontal and parietal electrodes (non-paired t-test). *: $p<0.05$.

itive Trap trials in exploitation. The FRP peak was not different after a negative Trap trial compared to a negative No Trap trial in exploitation (**Figure 11**). This result is not unexpected because the monkey didn't ignore a Trap trial, but mostly shifts strategy after it (on 113 Trap trials with a negative feedback in exploitation, only 6 lead to a stay strategy at the next trial) (**Figure 2.F**).

Discussion

In a task where monkeys flexibly adapted to their changing and stochastic environment by dynamically alternating between periods of exploration and exploitation, we confirmed that the FRP reflected feedback valence and showed that it was modulated by the learning phase, suggesting an encoding of the feedback unexpectedness, whatever the valence. The FRP amplitude was also predictive of

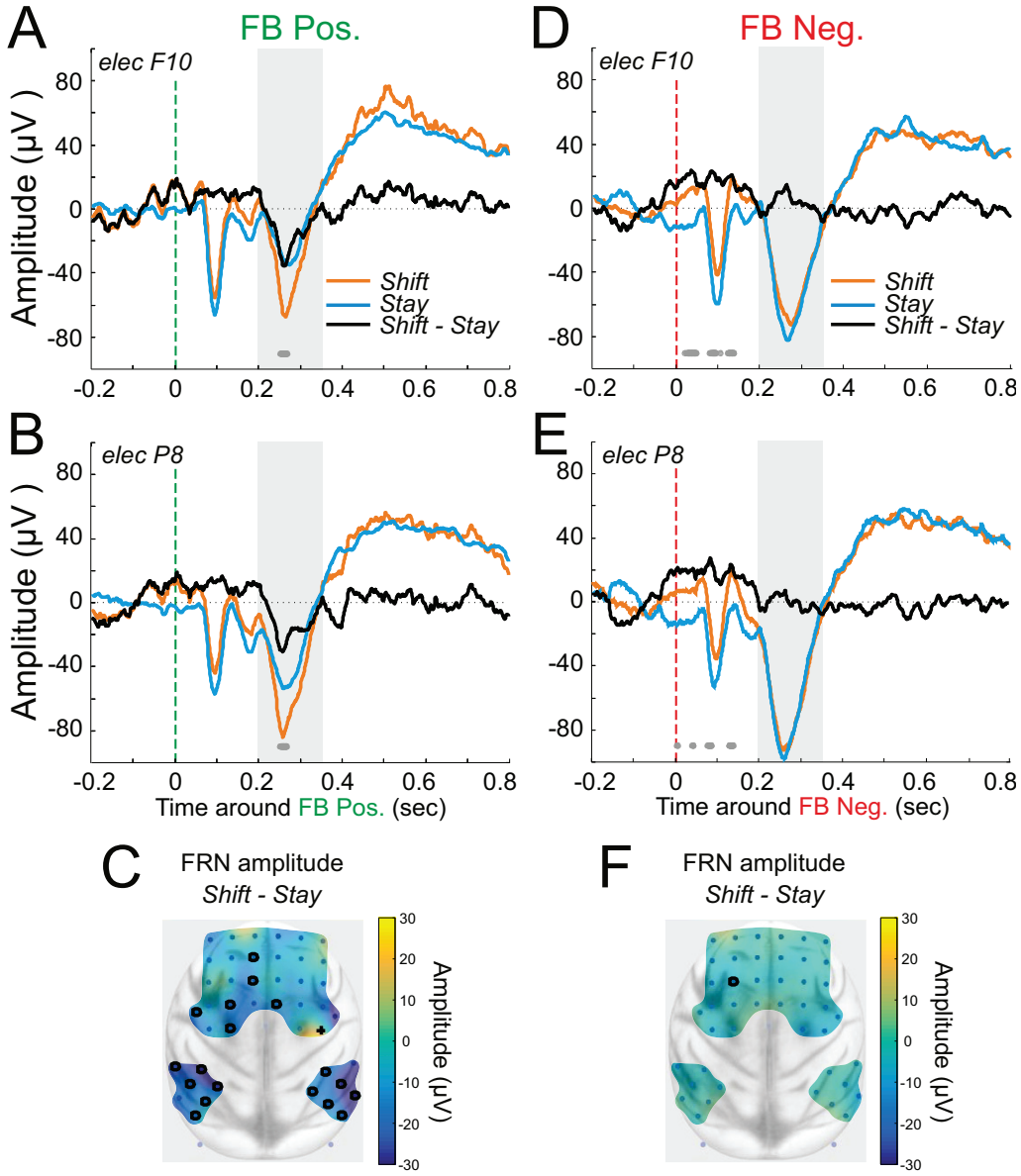

Figure 9: Effect of the upcoming strategy (shift vs stay) on Feedback Related Potentials depending on feedback valence. A, B, D, E: Example of FRPs over a frontal (elec F10, **A & D**) and a parietal electrode (elec P8, **B & E**) following a positive (left panel) or negative (right panel) feedback that will be followed by a shift (orange) or stay (blue) trial. The black line represents the 'Shift-Stay' difference. Horizontal bars at the bottom of the plot represent a significant difference between the conditions (permutation test, $p<0.05$). **C & F.** Cartographies of the mean 'Shift-Stay' amplitude difference over the time window of the FRN. Black crosses and Empty circles represent a significant positive and negative difference respectively (permutation test, $p<0.05$).

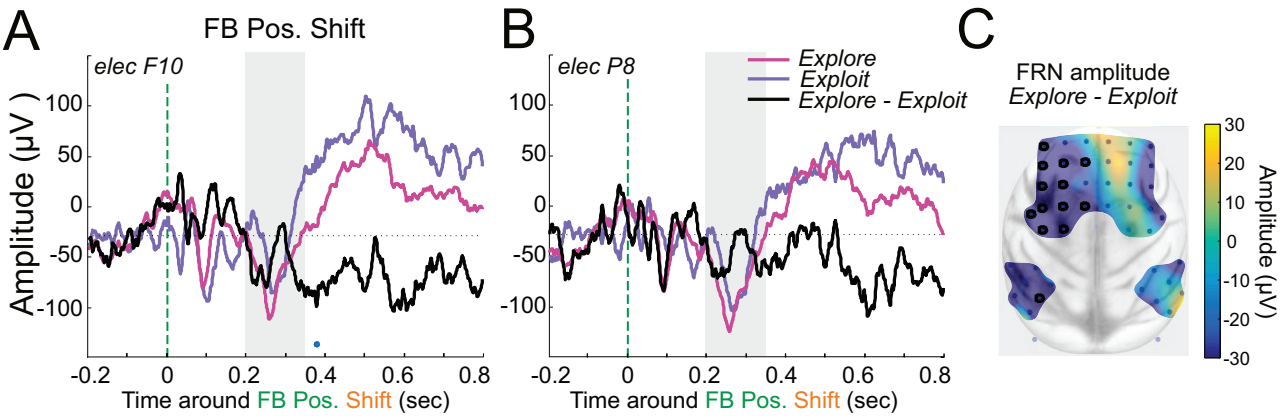

Figure 10: Effect of intra-problem learning phase on Feedback related potentials after a positive feedback followed by a shift, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), in exploration (dark pink) and exploitation (purple). Horizontal bars at the bottom of the plot represent a significant 'explore – exploit' difference (permutation test, $p<0.05$). **C.** Cartography of the mean amplitude difference over the FRN time window. Crosses and empty circles represent a significant positive and negative difference respectively (permutation test, $p<0.05$).

the strategy that the monkey will apply at the next trial after a positive but not a negative feedback. This suggests an encoding in the FRP of subsequent behavioral adjustment in reaction to a feedback, but in an asymmetric manner between positive and negative feedback. Hence we show that the FRP signal demonstrates a number of behavioral elements when studied in a complex learning task, supporting a number of results from the human cognitive literature.

Holroyd & Coles proposed a theory linking the ERN, involved in error detection, with the activity of the mesencephalic dopaminergic system (Holroyd and Coles, 2002). The dopaminergic system has been suggested to reflect prediction errors, because a phasic decrease in the dopaminergic activity is observed after the occurrence of an event that is worse than expected (Schultz et al., 1997). Holroyd & Coles propose that the FRN reflects the transmission of a reinforcement learning signal to the MCC, with the idea that the dopaminergic decrease after a negative feedback would disinhibit MCC neurons. Supports for this theory come from studies showing that the FRN amplitude is sensitive to dopaminergic pharmacology (Vezoli and Procyk, 2009; Zirnheld et al., 2004). The MCC would use this prediction error signal to choose between action plans, leading to adapted behavior (Holroyd and Coles, 2002). Our data showing a more negative FRP amplitude after negative compared to positive feedback and a bigger amplitude in the phase where negative feedback should be the more unexpected (i.e. in exploitation) support the idea that the FRP amplitude is larger after errors, and fits with the hypothesis postulating a link between FRN after negative feedback

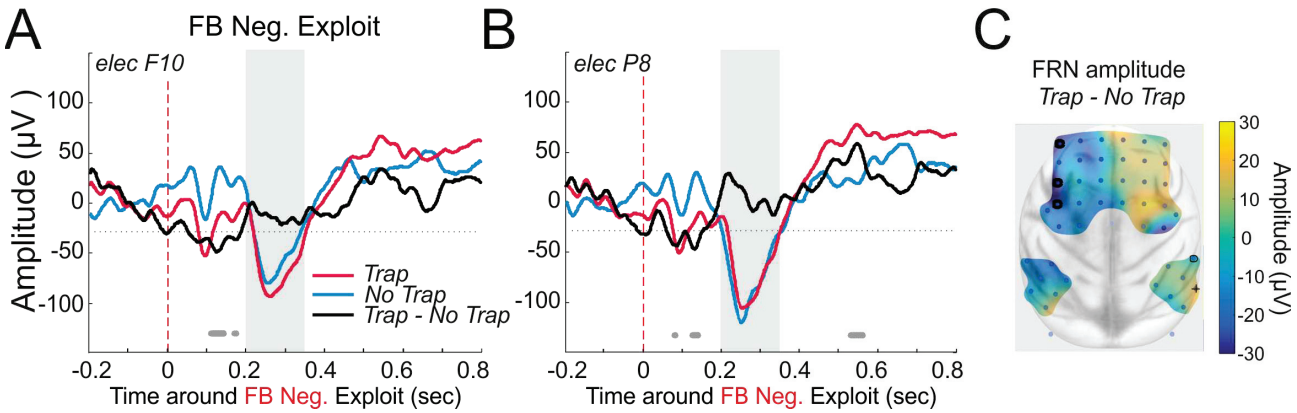

Figure 11: Effect of Trap trials on Feedback Related Potentials after a negative feedback in exploitation, over a frontal (elec F10, **A**) and parietal electrode (elec P8, **B**), for Trap (red) and No Trap (blue) negative feedback. The black line represents the 'Trap – No Trap' difference. Horizontal bars at the bottom of the plot represent a significant difference between the conditions (permutation test, $p<0.05$). **C.** Cartography of the mean 'Trap – No Trap' amplitude difference over the time window of the FRN. Note that the difference is never significant over this time window (permutation test, $p>0.05$).

and dopaminergic negative prediction errors.

However, our data concerning the FRP after positive feedback are more difficult to interpret in this framework. Indeed, the FRP amplitude is also negative after positive feedback, which doesn't fit with the idea that MCC neurons are disinhibited only after events that are worse than expected. Moreover, the relationship between the amplitude of the FRP after positive feedback and unexpectedness of the feedback is difficult to interpret as the FRP was similar in exploration and exploitation, but more negative in re-exploration compared to exploitation. An interpretation could be that positive feedback are never really surprising, except in re-exploration, when the monkey believes that a shift occurred and doesn't expect them anymore on the responses that triggered them. Indeed, in re-exploration, the monkey might think he knows what's going on, which is actually not the case, and that's why he is surprised by feedback; whereas in exploration the monkey might know that he doesn't know what's going on making him having fewer expectations.

Our results also show FRP modulations depending on the chosen subsequent strategy, suggesting a role in either signaling the need for a change, or in implementing this change. However, our design cannot disambiguate between the effect of the unexpectedness and the need for a change. One might have anticipated that our task would induce unexpected feedback that were not associated to subsequent behavioral adjustments, in that the monkeys might have learned to ignore the Trap feedback once they were in exploitation, modulating the level of expectedness to these feedback. In this protocol, however, monkeys did not learn like this. Rather, they reacted to Trap feedback, even in exploitation, so we

were unable to disambiguate unexpectedness from subsequent adaptation. Monkeys appear to have learned like this because of the volatile environment in which they acquired the task, and latterly because of the repeated occurrence of uncued shifts, both of which might have pushed monkeys toward being reactive to any unexpected feedback. This interpretation is supported by the data from the acquisition of the task by these monkeys, in which, rather than coming to ignore Trap feedback, they increase their response to it (see our previous paper). During the training of the task, shifts occurred only after a performance criterion was reached. The consequence was that the frequency of the shifts increased as monkeys got better at the task. Interestingly, their reactivity to Trap trial, that was absent at the start of learning, increased with learning, in parallel to the increase of volatility, to become stable once the learning and thus the volatility remained at a high and stable level. Thus, the high level of volatility in our task must be responsible of the high reactivity of the monkeys to Trap trials. As a consequence, our protocol didn't enable us to get surprising feedback which didn't lead to behavioral change, so that we can't differentiate the effects of surprise from the effects of behavioral adjustment in the FRP.

This interpretation therefore leads to the suggestion that it remains possible to induce the monkeys to ignore Trap feedback, if they were to work on the same probabilistic task without the shifts. However, a critic of this approach is that if probabilistic feedback don't trigger adaptation anymore, it might be that, in a sense, they have stopped being unexpected. This is why probabilistic contexts are said to trigger expected uncertainty (Payzan-LeNestour and Bossaerts, 2011). Accordingly, it was shown that random variations of reinforcement, if stable, are not surprising anymore and thus don't trigger learning (Courville et al., 2006; Gallistel and Gibbon, 2000).

Moreover, one can argue that trying to dissociate the contributions of unexpectedness and need for a behavioral shift is not particularly relevant to understand how the brain really works. Indeed, these are in practice rarely dissociated in real life and the brain might have evolved to treat them as the same thing. Thus, by creating situations that could not, or rarely, happen in real life, it is certainly possible to find some correlates in the brain, but this does not necessarily mean that the correlates reveal a process of normal functioning. The fact that we can design a task aiming at disambiguating certain confounded factors doesn't mean that we necessarily should do it. When designing a task to find the corresponding neural correlates, the crucial question should be "what does the brain need to know?" and not "How would the brain encode this?", which pushes toward more ecological designs (Tom Schonberg et al., 2011). In this context, the question would be: "how can the brain know when to keep or change the ongoing behavior?" Our design enables to

test this question by using unexpected events leading or not to a rule change (Trap and shift) and phases carrying different levels of evidence about what should be done (exploration, exploitation, re-exploration).

The fact that FRPn was not modulated by Trap compared to No Trap trials in exploitation suggests that the monkey is considering any negative feedback in exploitation as a shift. This shows that we succeeded inducing a flexible behavior in reaction to the uncertainty triggered by the combined stochasticity and volatility of the environment. An interesting further analysis will be to look at the trials following a Trap in exploitation, and to compare those trials with the ones consecutive to a shift. We might find correlates of the decision concerning whether it is a Trap or a shift.

Another aspect of discussion around these data is the specific contribution of the P3 to our signal. The P3 is a positive amplitude ERP component that can be observed after the presentation of a stimulus, in a large area around the midline and with peak latency between 300 and 600 ms (Sutton et al., 1965). The P3 is important to take into account as it overlaps with the FRN leading to distortion of its amplitude. Methodological approaches have been proposed to differentiate the two, such as looking at the difference wave (e.g. negative minus positive), or to study the components resulting from an independent component analysis (ICA) (Gentsch et al., 2009). The P3 must be particularly important to take into account when studying the FRN as it has been associated to functional roles very close to the FRN. The P3 amplitude is inversely proportional to the frequency or probability of stimuli, which has been interpreted as a signal of context updating (Donchin, 1981; Donchin and Coles, 1988), with the idea that higher amplitudes are associated to higher efforts to revise the model. Another hypothesis links the P3 to the function of the locus coeruleus and the effects of the norepinephrine in the cortex (Nieuwenhuis et al., 2005). According to this hypothesis, larger P3 should be triggered by high levels of arousals or task relevance, requiring the need of behavioral adaptation. Many studies tried to find dissociated functional roles for the FRN and the P3. For example, Yeung & Sanfey found that the P3 magnitude was sensitive to reward magnitude but not to reward valence, contrary to the FRN (Yeung and Sanfey, 2004). Interestingly, in this study, the authors also showed that the P3 amplitude elicited by unchosen outcomes was correlated with the degree of subsequent behavioral adjustment. Chase et colleagues also showed that the P3 was modulated by behavioral adaptation, contrary to the FRN (Chase et al., 2010). However, other studies showed dissociation between P3 amplitude and subsequent behavior. Inter-individuals differences in learning between negative and positive learners were reflected in the FRN but not in the P3 amplitude (Frank et al., 2005). Our results brought new evidence for

an implication of the FRP in behavioral adaptation, but more analyses are required to exclude any potential contribution of the P3. Thus, it is still unclear whether either the FRN or the P3 encode behavioral adjustment, and further studies should focus on their relative contributions.

Oscillations might be the way to disambiguate between all these effects. One should not forget that the FRP we studied here occurred around 200 ms post-feedback, which might be a bit early for a decision to be taken. Oscillations are later induced phenomenon implying a large network of structures (Tallon-Baudry et al., 1999). Frontal theta oscillations have also been related to cognitive control implementation (Cavanagh et al., 2014) and are thought to share redundant but also independent information with the FRP signal (Hajihosseini and Holroyd, 2013; Munneke et al., 2015).

Material and methods

Subjects and materials

Two rhesus monkeys (Macaca mulatta), one female and one male, weighing 7 kg and 8 kg (monkeys P and D respectively) were used in this study. Ethical permission was provided by the local ethical committee “Comité d’Éthique Lyonnais pour les Neurosciences Expérimentales”, CELYNE, C2EA #42, under reference C2EA42-11-11-0402-004. Animal care was in accordance with European Community Council Directive (2010) (Ministère de l’Agriculture et de la Forêt) and all procedures were designed with reference to the recommendations of the Weatherall report, "The use of non-human primates in research". Laboratory authorization was provided by the "Préfet de la Région Rhône-Alpes" and the "Directeur départemental de la protection des populations" under Permit Number: #A690290402. Monkeys were trained to perform the task seated in a primate chair (Crist Instrument Co., USA) in front of a tangent touch-screen monitor (Microtouch System, Methuen, USA). An open-window in front of the chair allowed them to use their preferred hand to interact with the screen (both monkeys left-handed). Presentation of visual stimuli and recording of touch positions and accuracy was carried out by the EventIDE software (Okazolab Ltd, www.okazolab.com). During the behavioral task, eye movements were monitored using an Iscan infrared system (Iscan, Inc.). Electrophysiological data were recorded using a Blackrock multichannel system (Blackrock).

Behavioral task

Principle of the task

The task is an adaptation for monkeys of the task described for human subjects in (Collins and Koechlin, 2012). Across successive trials, a problem consisted in the monkey concurrently finding, by trial and error, the correct mappings between stimuli and targets, within a set of 2 stimuli (stimulus 1 and 2) and 3 targets (target A, B and C) (**Figure 1**). For

example, a problem would consist in concurrently finding the two associations: "stimulus 1 with target A" and "stimulus 2 with target C". Monkeys learned problems to a behavioral criterion. The task contained stochasticity, in the form of unreliable feedback, and volatility, in the form of switches between problems. Importantly, monkeys were given stochastic feedback from the start of the training (see our Behaviour Paper for details on the training strategies).

Procedure of the task

▷ *Trial procedure*

The structure of a single trial and a single problem was always the same, regardless of the form of the task. Monkeys initiated each trial by touching and holding a lever item, represented by a white square at the bottom of the screen (**Figure 1.A**). A fixation point (FP) appeared. After a delay period of 1200ms (delay 1), a stimulus was displayed at the top of the screen (Stim ON signal), and was followed after a second delay of 1200ms (decision period) by the appearance in the middle of the screen of three targets (Targets ON signal). Stimuli consisted of square bitmap images of either an abstract picture or a photograph, of size 65x65mm. Targets were three empty grey squares, of the same size as the stimulus. After a further delay randomly varying between 0, 700 and 1000ms all targets turned white, providing the GO signal following which monkeys were permitted to make their choice by touching a target. Monkeys maintained touch on the chosen target for a random amount of time from 500 to 1500 ms in order to receive visual feedback on that choice. Feedback consisted of horizontal (positive) or vertical (negative) bars within each of the three targets, that needed to be hold for 700 to 2000ms. A positive feedback was followed by the delivery of about 1.8 mL of 50% apple juice. After the completion of a trial, a new stimulus was picked within the set of 2 stimuli and monkeys were allowed to begin a new trial.

▷ *Problem procedure*

A problem consisted of the monkeys learning about 2 stimuli concurrently. For a given trial, one of the 2 stimuli was pseudo-randomly selected (50% of each stimulus over 10 consecutive trials). The two concurrent stimuli were never associated to the same target. Hence, there were 6 possible mappings of the 2 stimuli and the 3 targets. Each mapping was randomly selected (with the constraint that both associations were changed), so that the 2 mappings of a problem could never be predicted nor learned. The only way to find them was to proceed by trial and error based on feedback provided after each choice. After a random number of trials between 60 and 85 trials, the problem changed and 2 new mappings were randomly selected, without any cue signaling the change. We refer to this change of problem as a 'Switch'. These Switches provide the volatility in the environment of the task. In addition to and separate

from this volatility, a stochastic reward environment was created by providing misleading feedback (called 'Trap feedback') on 10% of trials. Trap trials occurred pseudo-randomly once every 10 trials, with the constraint that there were at least 2 consecutive normal trials between each Trap trial. Trap feedback was the inverse of that determined by the current mapping – as such Trap feedback after a correct response consisted of negative feedback (see below) and no reward; Trap feedback after an incorrect response consisted of positive feedback and a reward.

▷ *Task motivation*

In order to motivate and maintain performance at a stable level throughout each daily session, animals were asked to obtain a fixed number of rewards each day (190 rewards). Upon successfully completing this number of rewards, monkeys received a large reward bonus (50 ml of fruit juice, calculated based on the effectiveness in motivating the monkey).

The analyses that are presented in this paper are restricted to monkey D data. Monkey P was also trained for the same task and recorded the same way, but her data have not been analyzed yet.

Behavioral analyses

For monkey D, 12 recordings sessions were selected for analysis, on the basis of global performance on each day superior or equal to 55% of correct responses with a minimum of 3 shifts. Reaction times were calculated as the time between the GO signal and the lever release (in order to further select a target on screen). Measures beyond 2 seconds were not included in the analysis. We analyzed the data in two ways to look at both trial-to-trial and within-problem modulations of performance. For trial-to-trial analysis, we measured performance before and after Trap trial. We also measured the score of win-stay lose-shift strategy after a Trap or a normal feedback (with a score of 1 for a change or maintenance of previous response after an incorrect or a correct feedback respectively and a score of 0 for the reverse pattern of responses). For within-problem analysis, we classified the trials as a function the "Phase" within a problem, defining 4 levels: exploration, exploitation, re-exploration and perseveration. We used the following criterion: 'exploration' trials were trials that received the same stimulus, that were associated to performance of no more than 3/5 correct over a sliding window of 5 trials and that were never preceded by 'exploitation' trials. 'Exploitation' trials were those with performance of 4/5 or more. 'Re-exploration trials' were trials for which performance went back above 3/5 correct after an exploitation period. 'Perseveration' trials consisted in post-Switch trials that were correct relative to the previous rules. These trials were excluded from 'exploration' trials.

Surgical procedures

Surgical procedures were performed under aseptic conditions. The monkey was sedated on the morning of surgery with both ketamine (10 mg/kg) and xylazine (0.5mg/kg) following pre- anesthetic treatment with glycopyrrolate (0.006 mg/kg). Once sedated, the monkey was given antibiotic (amoxicillin, 8.75 mg/kg) for prophylaxis of infection, and a nonsteroidal anti-inflamma- tory (ketoprofen, 2mg/kg) agent for analgesia. The head was shaved and an intravenous cannula put in place for intraoperative delivery of fluids (sterile saline drip, 5 mL/h/kg). The monkey was intubated, placed on isoflurane anesthesia (0.5–2.75%, to effect, in an O2 and NO2 mix), and then mechanically ventilated. Heating blankets allowed maintenance of normal body temperature during surgery. Heart rate, oxygen saturation of hemoglobin, expired CO2, body temperature, and respiration rate were monitored continuously throughout surgery. Each animal was implanted with a head-holder (Rogue Research) and intracranial electrodes. Both the placement of the electrodes and their depth were determined from separately acquired structural MRI images of each monkey using guidance of the Brainsight neuronavigation system. Holes were drilled through the skull and then stainless steel surgical screws (Synthes) were fixed into the holes, with the aim at each site of advancing the screw through the thickness of the bone to rest on the duramater. Each electrode was connected to a micro-connector (Omnetics). Skin and muscles were repositioned on the skull, above the electrodes wires.

For both monkeys, a grid of 46 electrodes spaced by 7 mm was implanted throughout the frontal, sensorimotor and parietal cortices (**Figure 2**). A supplemental electrode serving as reference was screwed into the bone of the thick brow of the monkey on the midline anterior to the frontal grid.

Electrophysiological Data Processing

All electrodes were referenced to the most frontal reference electrode (**Figure 2**). The signal from each electrode was amplified and filtered (1–250 Hz) and digitized at 781.25 Hz. Data analysis was performed off-line with FieldTrip toolbox (Oostenveld et al. 2011) and homemade Matlab scripts (Matlab, The MathWork, Inc.). Trials longer than 5 sec were excluded from the analysis. Movement artifacts were removed by decomposing ECoG recordings with an independent component analysis, using the logistic infomax algorithm (Bell and Sejnowski 1995).

▷ *Feedback related potentials (FRP)*

For FRP calculation, data were aligned to the feedback signal and averaged over trials. Trials in which monkeys withheld the hand from the screen before 1 sec were removed from the analysis. Data analyzed here correspond to 12 sessions (days) of acquisition on monkey D.

Different combinations of the following fac-

tors were used for comparing FRP amplitudes: (1) FB_valence (positive/negative), corresponding to the valence of the feedback (it refers to the feedback the monkey has just received), (2) Shift/Stay (2 levels: shift and stay). It corresponds to whether the monkey has repeated (stay) or changed (shift) his response compared to the previous trial with to the same stimulus. (3) Phase (exploration/ exploitation/ re-exploration), referring to the phase within the problem, as explained in the "Behavioral analysis" section. (4) Normal_or_Trap (normal/Trap) was used for distinguishing the effects of normal versus Trap feedback.

Permutation tests

In order to detect significant differences in FRPs amplitude between the factors, we used t-tests and permutation tests. We first began by down-sampling the condition with the most trials between the 2 compared conditions. To do this, we selected randomly between the trials of the condition with the most trials in order to obtain the same number of trials as in the condition with the less trials. Then, we did an unpaired t-test over time bins between the 2 conditions; and replicated this operation a 100 times. We then averaged the 100 obtained vectors of t-values. Then, we did a permutation test by randomly selecting trials within the 2 conditions in order to create 2 random sets, which were compared using an unpaired t-test. We replicated this a 1000 times. We then calculated how many times the real t-values were bigger than 2.3 standard deviations from the mean of the distributions of the random t-values, to determine whether there was a significant difference between the 2 conditions at each time bin. This procedure was replicated for the 46 electrodes.

Latencies of effect detection

We analyzed the latency of the effect of factors that were found significant in the models, for both FRP and TF data. A permutation test (with 1000 permutations) was run on the difference of the signal between the 2 levels of each factor, using unpaired samples t-test. Latencies of the effect were taken as the first significant time bin after 0.1ms after the feedback signal.

References

Cavanagh, James F, Sanguinetti, Joseph L, Allen, John J B, Sherman, Scott J, and Frank, Michael J. The Subthalamic Nucleus contributes to post-error slowing. *Journal of Cognitive Neuroscience*, pages 1–8, 2014.

Chase, Henry W, Swainson, Rachel, Durham, Lucy, Benham, Laura, and Cools, Roshan. Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of cognitive neuroscience*, 23(4):936–946, 2010.

Cohen, Jonathan D, McClure, Samuel M, and Yu, Angela J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1481):933–942, 2007.

Collins, Anne and Koechlin, Etienne. Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10(3):e1001293, 2012.

Courville, Aaron C., Daw, Nathaniel D., and Touretzky, David S. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

Donchin, E. Surprise!...Surprise? *Psychophysiology*, 18:493–513, 1981.

Donchin, Emanuel and Coles, Michael G. H. Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, 11(03):357, 1988.

Ferdinand, N. K., Mecklinger, a., Kray, J., and Gehring, W. J. The Processing of Unexpected Positive Response Outcomes in the Mediofrontal Cortex. *Journal of Neuroscience*, 32 (35):12087–12092, 2012.

Frank, Michael J., Woroch, Brion S., and Curran, Tim. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, 47(4):495–501, 2005.

Gallistel, C R and Gibbon, J. Time, rate, and conditioning. *Psychological review*, 107(2):289–344, 2000.

Gentsch, Antje, Ullsperger, Peter, and Ullsperger, Markus. Dissociable medial frontal negativities from a common monitoring system for self- and externally caused failure of goal achievement. *NeuroImage*, 47(4):2023–2030, 2009.

Groen, Yvonne, Wijers, Albertus a, Mulder, Lambertus J M, Minderaa, Ruud B, and Althaus, Monika. Physiological correlates of learning by performance feedback in children: a study of EEG event-related potentials and evoked heart rate. *Biological psychology*, 76(3):174–187, 2007.

Hajihosseini, Azadeh and Holroyd, Clay B. Frontal midline theta and N200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology*, 50(6):550–562, 2013.

Hauser, Tobias U., Iannaccone, Reto, Stämpfli, Philipp, Drechsler, Renate, Brandeis, Daniel, Walitza, Susanne, and Brem, Silvia. The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, 84:159–168, 2014.

Holroyd, Clay B. and Coles, Michael G.H. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109 (4):679–709, 2002.

Holroyd, Clay B. and Krigolson, Olave E. Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, 44(6):913–917, 2007.

Huettel, Scott a, Song, Allen W, and McCarthy, Gregory. Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(13):3304–3311, 2005.

Mars, Rb, De Bruijn, Era, Hulstijn, W, Miltner, Whr, and Coles, Mgh. What if I told you: "You were wrong"? Brain potentials and behavioral adjustments elicited by feedback in a time-estimation task. pages 129–134, 2004.

Miller, Earl K and Cohen, Jonathan D. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24:167–202, 2001.

Miltner, Wolfgang H. R., Braun, Christoph H., and Coles, Michael G. H. Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a “Generic” Neural System for Error Detection, 1997.

Munneke, Gert-Jan, Nap, Tanja S., Schippers, Eveline E., and Cohen, Michael X. A statistical comparison of EEG time- and time–frequency domain representations of error processing. *Brain Research*, 1618:222–230, 2015.

Nieuwenhuis, Sander, Aston-Jones, Gary, and Cohen, Jonathan D. Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, 131(4):510–532, 2005.

Oliveira, Flavio T P, McDonald, John J, and Goodman, David. Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of cognitive neuroscience*, 19(12):1994–2004, 2007.

Payzan-LeNestour, Elise and Bossaerts, Peter. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, 7(1):e1001048, 2011.

Pietschmann, Maria, Simon, Katja, Endrass, Tanja, and Kathmann, Norbert. Changes of performance monitoring with learning in older and younger adults. *Psychophysiology*, 45 (4):559–568, 2008.

Rothé, Marie, Quilodran, René, Sallet, Jérôme, and Procyk, Emmanuel. Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(31):11110–11117, 2011.

Schultz, W, Dayan, P, and Montague, P R. A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275 (5306):1593–9, mar 1997.

Sutton, S, Braren, M, Zubin, J, and John, E R. Evoked-potential correlates of stimulus uncertainty. *Science (New York, N.Y.)*, 150(700):1187–1188, 1965.

Tallon-Baudry, C, Kreiter, a, and Bertrand, O. Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans. *Visual neuroscience*, 16(3):449–459, 1999.

Tom Schonberg, Fox, Craig R., and Poldrack, Russell A. Mind the gap : bridging economic and naturalistic risk-taking with cognitive neuroscience. *Russell The Journal Of The Bertrand Russell Archives*, 15(1):11–19, 2011.

Vezoli, Julien and Procyk, Emmanuel. Frontal feedback-related potentials in nonhuman primates: modulation during learning and under haloperidol. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(50): 15675–83, dec 2009.

Voloh, Benjamin, Valiante, Taufik a., Everling, Stefan, and Womelsdorf, Thilo. Theta–gamma coordination between anterior cingulate and prefrontal cortex indexes correct attention shifts. *Proceedings of the National Academy of Sciences*, (MAY):201500438, 2015.

Walsh, M. M. and Anderson, J. R. Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, 108(47):19048–19053, 2011.

Yeung, Nick and Sanfey, Alan G. Independent coding of reward magnitude and valence in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(28):6258–6264, 2004.

Zirnheld, Patrick J., Carroll, Christine a., Kieffaber, Paul D., O'Donnell, Brian F., Shekhar, Anantha, and Hetrick, William P. Haloperidol Impairs Learning and Error-related Negativity in Humans. *Journal of Cognitive Neuroscience*, 16(6):1098–1112, 2004.

Chapitre 7

Etude des Oscillations :

Etude comparée de la puissance des oscillations beta (10-25 Hz) et thêta (4-8 Hz) fronto-pariétales dans une tâche d'apprentissage dans un environnement incertain

Dans cette étude, nous comparons les oscillations beta et thêta en provenance des aires frontales et pariétales à deux moments clés de la tâche : la période du feedback et la période de décision. Nous mettons en évidence un patron complexe de modulations de la puissance des deux bandes, en fonction des paramètres de la tâche, mais aussi de la période.

Fronto-Parietal Beta (10-25Hz) and Theta (4-8Hz) Activities are Differently Modulated in a Probabilistic Reversal Learning Task

Maïlys C.M. Faraut[1,2,3], Charles R.E. Wilson[1,2,3] and Emmanuel Procyk[1,2]

[1]INSERM U846, Stem Cell and Brain Research Institute, Bron 69500, France
[2]Université de Lyon, Lyon 1, UMR S-846, Lyon 69003, France
[3]Corresponding authors: mailys.faraut@inserm.fr & charles.wilson@inserm.fr

Keywords: feedback, adaptation, flexibility, learning, monkey

Abstract

In our changing and dynamic environment, flexible behavior relies on both performance monitoring and cognitive control. These processes might depend on the dynamic interplay between the mid-cingulate cortex and the lateral prefrontal cortex, for the detection of behaviorally relevant feedback and subsequent control adjustments. However, the mechanisms by which these two regions integrate information from each other in relation with other regions are still unclear. Frontal beta and theta oscillations have both been associated to cognitive control, by showing increased power in situations requiring an increased control. However, few studies have directly compared the two bands to understand their specificities. In this study, we analyzed beta (10-25 Hz) and theta (4-8 HZ) oscillations over frontal and parietal areas in a monkey performing a probabilistic reversal task requiring varying levels of cognitive control. We compared the power of these two bands at two important time points: the feedback period and the decision period. We show both common and separated modulations in relation to cognitive control. In particular, beta power was predictive of the subsequent strategy but at a different periods depending on the feedback valence. Theta power was related to the discrimination of normal versus unreliable negative feedback created by the environmental stochasticity.

Introduction

In our open-ended and constantly changing environment, we need to behave flexibly and remain vigilant for changing contingencies. This flexibility involves monitoring performance and changing levels of cognitive control to produce behavior properly adapted to the current context (Miller and Cohen, 2001). These processes can be measured by tracking the efficacy with which subjects adapt to changing task demands and respond to feedback. A number of neurophysiological markers of cognitive control and performance monitoring emerge from frontal networks, with a key role for the lateral prefrontal cortex (DLPC) and the mid-cingulate cortex (MCC) (Johnston et al., 2007; Quilodran et al., 2008). However, if MCC and DLPF are crucial for flexible behavior, their role is only meaningful when they are considered together as being part of the same dynamic network. The mechanisms by which they dynamically interact as a network and how they connect with other regions such as the sensorimotor and parietal cortical areas are still unclear. These interactions are the focus of ongoing research, which has started to reveal an important role of oscillations for inter-areal communication (Fries, 2005); and in particular a role of beta and theta oscillations in the regulation of cognitive control.

Oscillations in local field potentials (LFPs) in the beta band (20-30Hz) have been linked to many functions, such as the modulation of the motor function (Pfurtscheller and Lopes da Silva, 1999), or as a specific top-down mechanism within the hierarchy of visual areas (Bosman et al., 2012; Bastos et al., 2015). But many studies also revealed a role of beta oscillations in the top-down control of behavior (Buschman and Miller, 2007; Siegel et al., 2012). We have demonstrated a specific role of beta oscillations in the implementation of cognitive control via enhanced power reflecting the need for control and the level of attentional effort (Stoll et al., 2015). Others have demonstrated an important role for increased fronto-parietal beta synchrony, for example before successful change detection (Micheli et al., 2015), or during free compared to instructed decisions (Pesaran et al., 2008). This seems to suggest a role of beta oscillations in the interaction between the nodes of a cognitive control network.

Theta oscillations (4-8 Hz) may also reflect crucial aspects of cognitive control (Cavanagh et al., 2014). High levels of theta power have been associated with conditions of high control (Phillips et al., 2014; van Driel et al., 2015) or with conditions requiring an increase of control, like after an error or a negative feedback (van de Vijver et al., 2011; Cohen et al., 2007); whereas lower levels of theta power were predictive of the commission of an error (Cavanagh et al., 2009). Furthermore, theta oscillations have also been proposed as a potential mechanism for coordinating neuronal ensembles from different regions, in particular during higher cognition, such as cognitive control (Benchenane et al.,

2010; Voloh et al., 2015). For example, theta band phase synchrony between DLPFC, medial frontal and sensorimotor sites was increased after negative feedback (van de Vijver et al., 2011). Another study showed that theta synchronization occurred from left central (motor areas) to mid-frontal areas immediately after feedback, and from mid- frontal to frontal areas later on, and in a stronger way in high compared to low learners (Luft et al., 2013).

Beta and theta oscillations have thus been involved in the same mechanisms, which suggest that they should directly be compared. Separate studies showed that theta power is usually stronger after negative feedback (Cohen et al., 2007) and that Beta power is stronger either after positive feedback (Cohen et al., 2007; HajiHosseini et al., 2012; Hosseini and Holroyd, 2015; van de Vijver et al., 2011) or after negative feedback (Cohen et al., 2009; Leicht et al., 2013). A study by Van de Vijver et colleagues directly compared both bands and suggested that theta oscillations were predictive of learning only from negative feedback, whereas beta oscillations of learning from both negative and positive feedback (van de Vijver et al., 2011). Another study showed that differences in beta and theta band power after feedback distinguished high from low learners in a time estimation task, as high learners had larger mid-frontal theta power and lower sensorimotor beta power after incorrect feedback (Luft et al., 2013). Thus, if it seems that the contributions of both bands can be distinguished, their relative contribution is still ambiguous.

The current study was designed to reveal differences between theta and beta oscillations in two ways. First, the majority of the mentioned studies were based on very simple tasks, requiring a low level of cognitive control, especially in overtrained animals. Crucially, in our task, good performance requires a permanent engagement of cognitive control as the changing and stochastic environment makes routine, or automatic strategy only useful to a certain extent, even in overtrained animals. Second, in contrast to many studies we will compare the modulations of theta and beta oscillations for both feedback and decision periods.

Results

Behavior

Monkey D was implanted with an ECoG implant covering the surface of the frontal and parietal cortices (**Figure 1**). This technique enabled us to investigate the variations of the surface cortical oscillations to the various factors of the task. Behavioral analyses are described in the previous paper (see FRP paper). We focused our analyses on two important time points for performance monitoring and choice respectively: the period of feedback and the period between onset of the stimulus and the targets, the decision period. All recordings were made during performance of the Reversal task detailed in **Figure 2**.

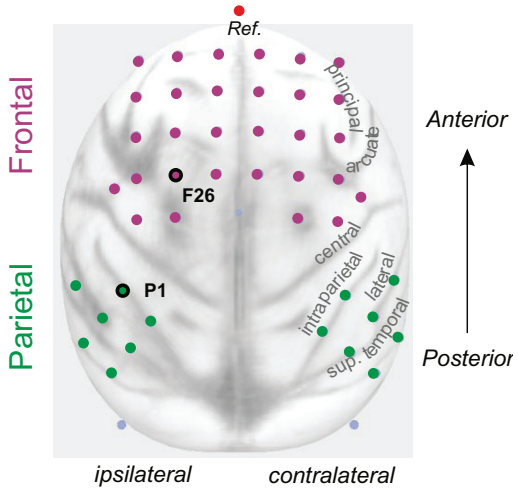

Figure 1: Location of intra-cranially implanted ECoG electrodes with the names of major sulci in the underlying cortex. Positions of example electrodes (F26 over frontal and P1 over parietal cortex) are also indicated (black circles). The reference electrode (Ref) is located on the skull, at the apex of frontal regions.

Feedback period

Trial to trial modulation

Beta (10-25Hz) and theta (6-8Hz) oscillations were observed over prefrontal and parietal cortex after feedback onset (**Figure 3**). We first looked at whether the power of beta and theta bands would reflect the valence of the feedback. Power in both bands was significantly modulated by feedback valence, i.e. with a higher power after a negative compared to a positive feedback (glm, factor FB_valence, $p<0.01$ for both beta and theta models) (**Figure 4. A, B & E, F**). In beta, the effect was present widely over all fronto-parietal electrodes (**Figure 4. A & B**). In theta, by contrast, the effect was restricted to a few posterior electrodes (**Figure 4. E & F**). No significant difference was observed in the latency of the effect in the 2 bands between frontal and parietal electrodes (Beta: mean PFC=0,2873 sec +/-0,0366; mean Parietal= 0,368 sec +/-0,0588; Theta: mean PFC= 0,576 sec +/-0,0383; mean Parietal= 0,5125 sec +/-0,0615; unpaired t-test, $p>0.05$ in both cases) (**Figure 4. C, D & G, H**).

We then looked at whether these oscillations induced by the feedback reflected the strategy (stay or shift) the monkey subsequently chose the next time the same stimulus was presented. Beta, but not Theta, power was significantly, albeit sparsely, modulated by whether or not the monkey repeated his previous response in the next trial (glm, factor Shift/Stay, $p<0.01$), with a higher power after 'Stay' compared to 'Shift' trials (**Figure 5. A&B**). The latency of the significant effect was earlier in frontal compared to parietal electrodes (mean PFC= 0. 2914 sec +/- 0.0265; mean Parietal= 0.3493 sec +/- 0.0671)(unpaired t-test, $p<0.01$ (P=0,0073) , T-value=-2,8187, Df=42, Sd=0,1763) (**Fig-**

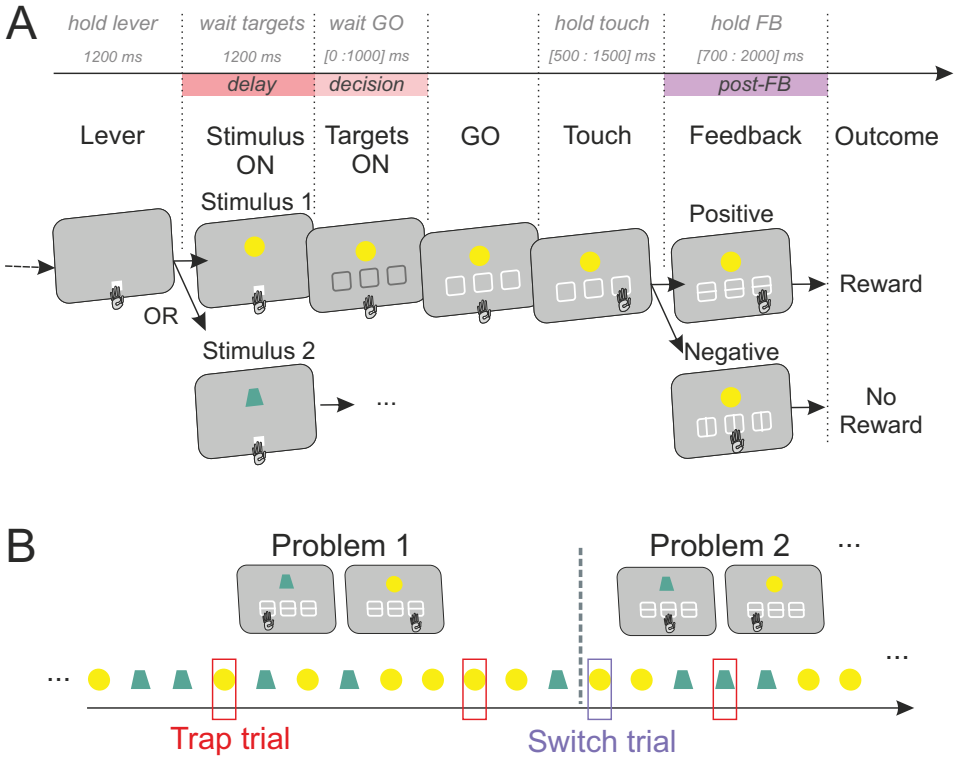

Figure 2: Probabilistic Reversal Task. A. Structure of a single trial. After holding a touch screen 'lever', one of 2 stimuli appears, followed by the onset of three targets. After a GO signal, the monkey selects one target by touching it. Visual feedback is represented by horizontal or vertical bars superimposed on all three targets for positive or negative feedback respectively. Positive feedback was followed by the delivery of juice reward. **B.** Example of a series of trials. Each problem constituted two mappings between the two stimuli and one of the responses. One stimulus at a time was randomly presented. On Trap trials, a misleading feedback was given to the monkey: a positive feedback with a juice reward after an incorrect choice; or a negative feedback with no reward after a correct choice. One Trap trial occurred pseudo-randomly within each set of 10 trials. After a random number of trials (between 60 and 85 trials), the problem changed (Switch trial), and 2 new mappings with the same 2 stimuli were randomly selected. Thus, Switches between problems were not predictable.

ure 5. C&D). Moreover, as there was also a significant interaction between the factor 'Shift/Stay' and the feedback valence (glm, interaction Shift/Stay x FB_valence, $p<0.01$), we ran 2 supplementary models on trials with either only positive or only negative feedback. The factor Shift/Stay was only significant in the model with positive feedback (glm, factor Shift/Stay $p<0.01$ and $p>0.05$ for positive and negative trials respectively), indicating that the Shift/stay effect was driven by positive feedback trials (**Figure 5.B.a**). Thus, Beta power after a positive feedback was higher when the monkey maintained vs. changed his response at the next trial.

Intra-problem modulation

During the solving of problems, monkeys alternated between periods of exploration, ex-

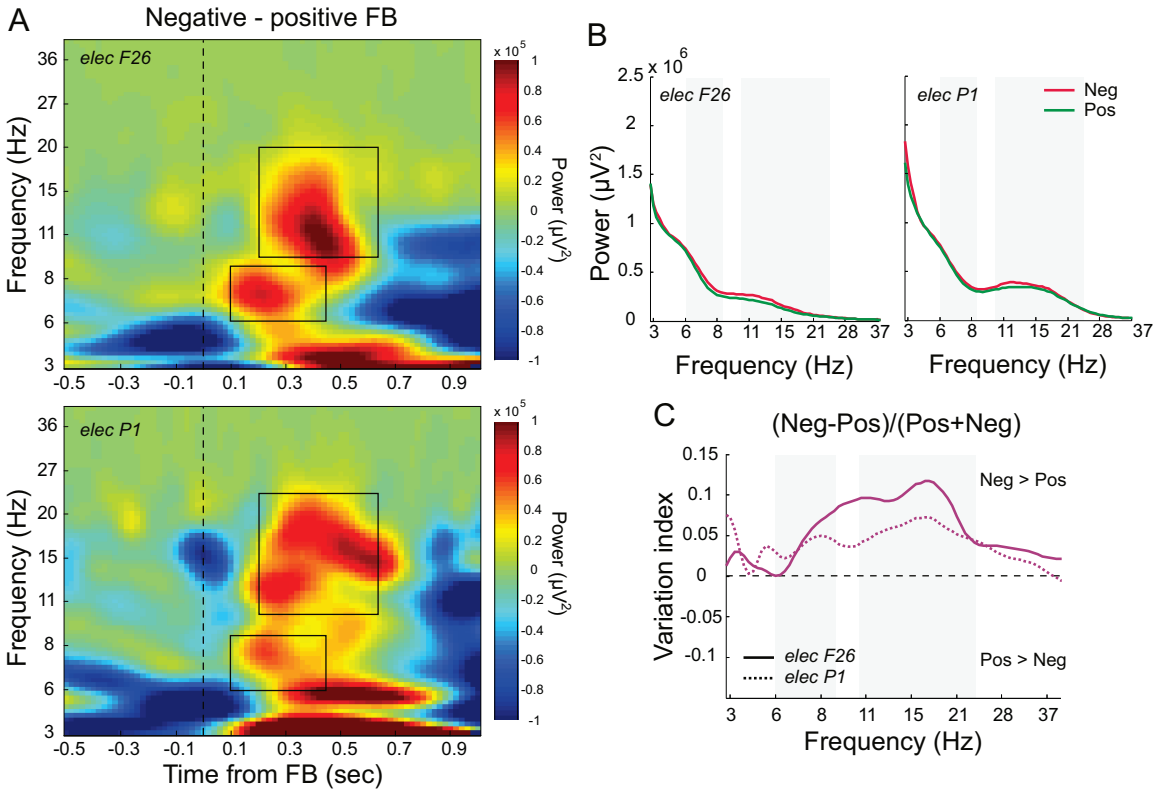

Figure 3: Description of beta and theta oscillations for the period around feedback onset for 2 example electrodes. **A.** Average TF representation for the Negative – Positive feedback contrast for a frontal electrode (elec F26, top figure) and a parietal electrode (elec P8, bottom figure). Black rectangles indicates the time and frequency windows that will be used in the analyses. **B.** Power spectrum densities of the post-feedback time window common to the 2 bands shown in the black rectangles in A, for the example electrodes (elec F26, left; elec P1, right) for positive (green) and negative (red) feedback. **C.** Variation index calculated as (Neg-Pos)/(Pos+Neg)on the power spectrum densities described in B.

ploitation and re-exploration (see the description of behavior in **Figure 2** in the FRP paper). We wondered whether a feedback with the same valence would be encoded differently in beta and theta oscillations depending on those periods, as they might reflect different degrees of learning of the two associations (a positive feedback should be more behaviorally relevant in exploration compared to exploitation for example). Beta power was significantly modulated by the phase within the problem after both positive and negative feedback (2 glm, run on trials with either only positive or negative feedback, factor Phase, $p<0.01$) (**Figure 6 A to C & D to F**). The effects were localized mostly on parietal electrodes, and on some frontal-medial electrodes, and at these sites, beta power was generally higher in exploitation and re-exploration compared to exploration. Theta power was only significantly modulated by the phase after positive feedback (**Figure 6 H to J**). The effects were localized ipsilaterally, and consisted in a higher power for exploitation compared

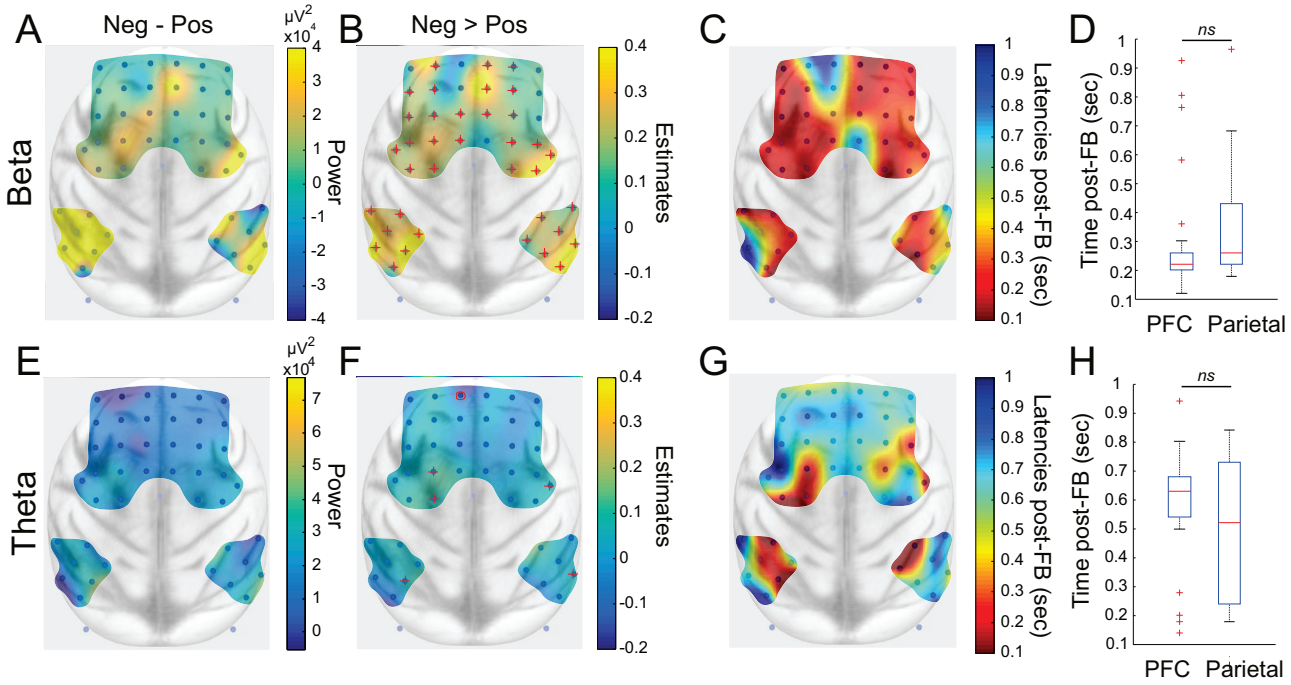

Figure 4: Effects of feedback valence on Beta (top panels) **and Theta** (bottom panels) **oscillations over frontal and parietal electrodes during the feedback period**. Cartographies showing: **A and E**: Beta and Theta oscillatory power contrasts between Negative and Positive feedback. **B and F**: estimates from the model (Negative relative to Positive). Red crosses indicate a significant p-value ($p<0.05$) on the corresponding electrode, with a positive estimate (Negative>Positive), and empty circles a negative estimate (Negative<Positive). **C and G**: Latencies of effect, corresponding to the first time sample after 0.1ms post-FB at which the difference between Negative and Positive becomes significant (permutation test). **D and H**. Boxplot comparing the repartition of the latencies of the effect between frontal and parietal electrodes (non-paired t-test).

to exploration. These results support the idea that the monkey was considering differently a feedback with the same valence depending on his level of learning of the problem.

During the solving of a problem, monkeys should progressively be able to distinguish Trap from normal trials. We tested this by testing whether the signal differed between Trap and normal trials in the exploitation period, during which the monkey was supposed to be able to differentiate them. A significant difference in Theta power (but not in Beta) was observed on the contralateral hemisphere between Trap and normal trials when they consisted in a negative feedback (glm on negative exploitation trials only, factor Normal_or_Trap, $p<0.01$) (**Figure 7**). Interestingly, this difference consisted in higher theta power after the no Trap compared to the Trap trials. Hence, the difference between a negative Trap and normal feedback in exploitation was reflected in oscillatory power.

In summary, Beta and Theta power were modulated differently after feedback. First, if

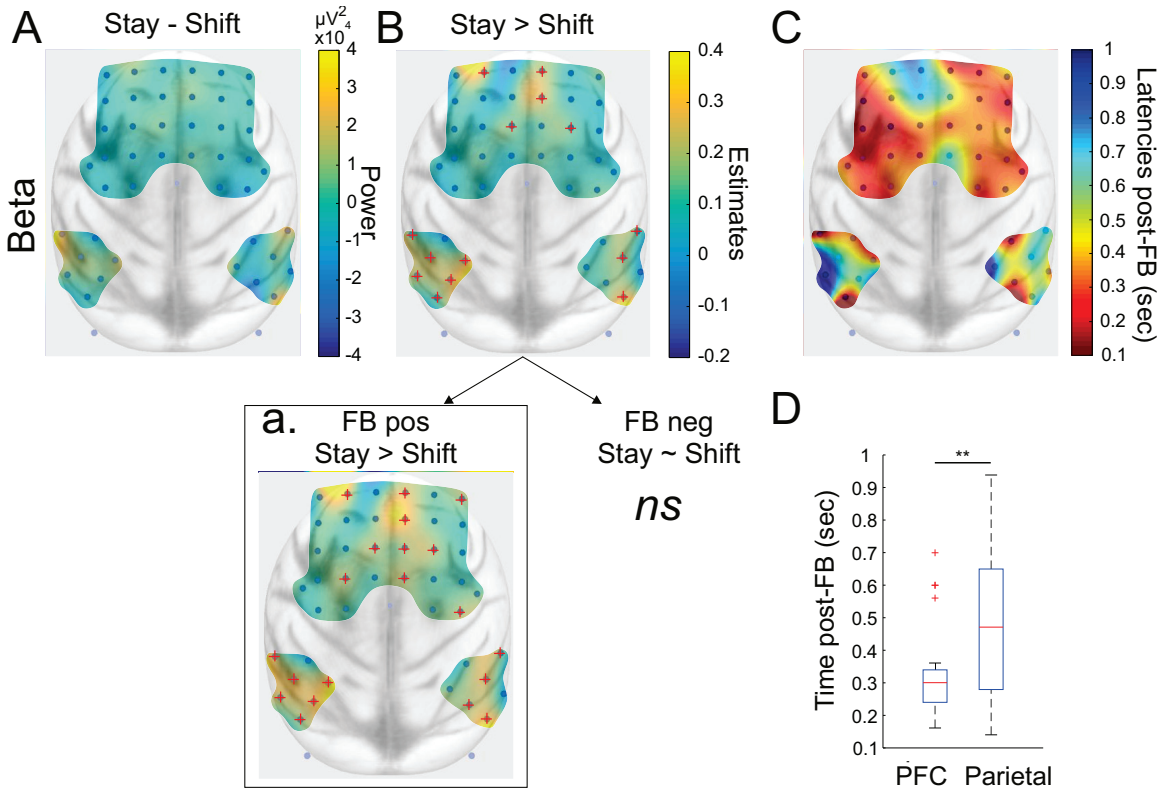

Figure 5: Effects of the upcoming strategy (shift vs stay) on Beta oscillations over frontal and parietal electrodes during the feedback period. Cartographies showing: **A.** Beta oscillatory power contrasts between Shift and Stay trials. **B.** Estimates from the model (Stay relative to Shift). Red crosses indicate a significant p-value ($p<0.05$) on the corresponding electrode, with a positive estimate (Stay>Shift). **B.a** represents the estimates of the same model run on a subset of trials with positive feedback only. **C.** Latencies of effect, corresponding to the first time sample after 0.1sec post-FB at which the difference between Stay and Shift trials becomes significant (permutation test). **D.** Box-plot comparing the repartition of the latencies of the effect between frontal and parietal electrodes (non-paired t-test).**: $p<0.01$.

both bands showed significantly more power after negative compared to positive feedback, the effect was more widespread in beta than theta. Second, only beta power was predictive of the strategy that the monkey would use in the next trial. This effect was significant only after positive feedback and consisted in higher beta power when the monkey maintained rather than changed his response. The effects of feedback valence and of the upcoming strategy on beta power seem somewhat contradictory as higher beta power after a negative feedback suggests that high beta is associated to high control, whereas higher beta power is also found after a positive feedback when monkey will subsequently stay, a situation that presumably demands less control. A possibility could be that high beta levels reflect the fact that an event is relevant for adaptation, being either a negative feedback, or a positive feedback considered as being the correct response. Beta oscillations

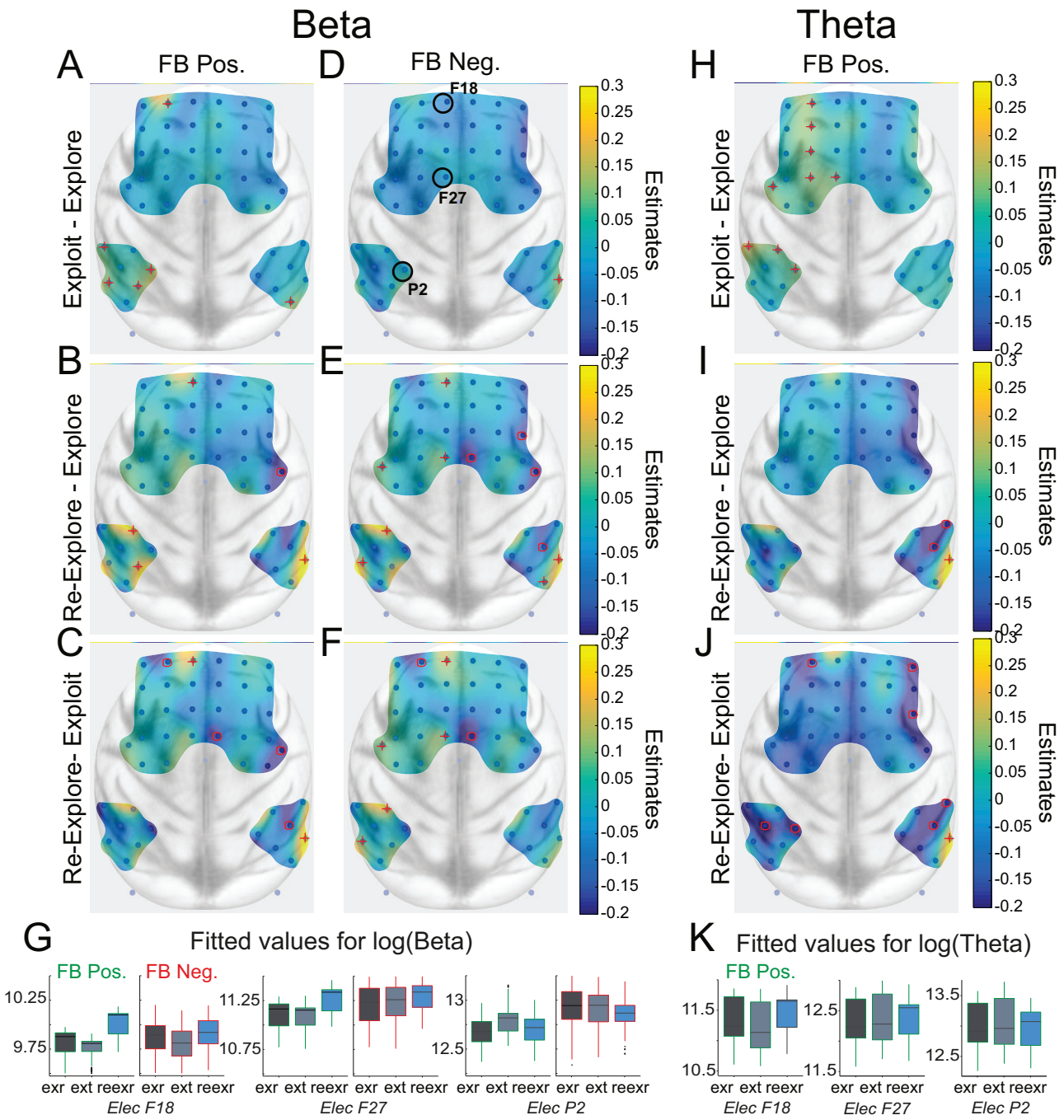

Figure 6: Effects of the factors Phase on the estimates of Beta and Theta oscillations over frontal and parietal electrodes during the feedback period depending on the valence of the current feedback (FB Pos: positive or FB neg: negative). Cartographies showing Beta (**A to F**) and Theta (**H to J**) estimates of the model for the factor Phase for Exploit versus Explore (**A,D,H**) , Re-explore versus Explore (**B,E,I**) and Re-explore versus Exploit (**C,F,J**). Red crosses and circles indicate a significant p-value on the corresponding electrode, with a positive and negative estimate respectively. In Theta, no significant Phase effect was found in the models with negative trials. **G & K**. Fitted values for log(Beta) & log(Theta) extracted from the models; for 3 example electrodes shown in **D.** exr=exploration; ext=exploitation; reexr=re-exploration.

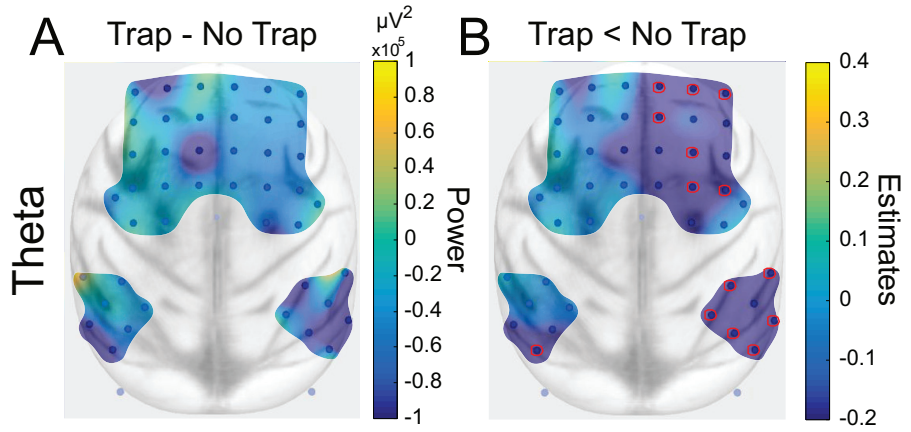

Figure 7: Effect of (negative in exploitation) Trap trials on Theta oscillations over frontal and parietal electrodes during the feedback period. Cartography showing Theta oscillatory power contrasts (**A**) and estimates from the model (Trap relative to NoTrap) (**B**) for the factor Trap_vs_NoTrap for incorrect trials in exploitation only. Red circles indicate a significant p-value ($p<0.05$) on the corresponding electrode, with a negative estimate (Trap<NoTrap).

were also modulated by the phase of learning in a problem, being globally stronger in exploitation and re-exploration after both types of feedback. Theta power was also stronger in exploitation and re-exploration but after positive feedback only; and after normal compared to Trap negative feedback in exploitation.

Decision Period

If the feedback period is an important time for the integration and evaluation of feedback enabling performance monitoring, another crucial period is the decision period, as the monkey has received all information he needs to make the decision. Thus, we focused on oscillations during this period to see whether we could find correlates of the ongoing decision.

Oscillations in the beta (10-17 Hz) and theta (4-6 hz) bands were observed during the decision period (**Figure 8**). We first looked at whether the levels of these oscillations, in particular of theta oscillations, were predictive of the commission of an error in the trial. Indeed, theta power was shown to be lower in the period preceding the commission of an error, which was interpreted as the sign of less control engaged in the trial, consequently leading to an error. Thus, we checked whether theta power in the decision period was predictive of the outcome of the upcoming trial, as being a correct or an incorrect response, but we found no significant effect (glm, factor Current_FB, $p>0.05$). We didn't find any effect in the beta band either (glm, factor Current_FB, $p>0.05$).

Another possibility would be that oscillations during the decision period encode the strategy the monkey will apply in the trial, driven or not by the previous feedback for the same stimulus. Thus, we first checked

whether these oscillations were predictive of the upcoming strategy, and then, whether this was modulated differently depending on the valence of the previous feedback. Power in both beta and theta bands were strongly modulated by whether the monkey shifted or not his response in the current trial compared to previous trial (stronger power for Stay compared to Shift trials. glm, factor Shift/Stay, $p<0.01$ for both bands) (**Figure 9 A to D**). This effect seemed to be driven by negative feedback, as, in both bands, it was significant only on trials following a negative feedback (2 glm on trials with positive or negative feedback, $p>0.05$ & $p<0.01$ respectively) (**Figure 9 B.a, D.a**). The effect was very widespread in theta, and more localized along the frontal midline and parietal cortex in beta.

Only beta oscillations were modulated by the phase, for both positive and negative feedback (**Figure 9**). It seems that the effects are opposite depending on feedback valence as more power is observed in re-exploration compared to exploration consecutive to a positive feedback, and the opposite after a previous negative feedback (**Figure 9.B vs E**).

Thus, to summarize the effects during the decision period, theta (and to a lesser extent, beta) oscillations were predictive of the strategy applied in the trial, but only after negative feedback, by showing a higher power before a "stay" compared to a "shift" trial.

Discussion

One of the aims of this study was to compare the role of theta and beta oscillations in a demanding task as both have been associated with cognitive control. Both types of oscillations were observed in two periods of the task, in the feedback (theta: 6-8Hz and beta: 10-25Hz) and in the decision (theta: 4-6Hz and beta: 10-17Hz) periods.

After feedback, both beta and theta bands were reflective of the feedback valence, showing more power for negative compared to positive feedback. Differences between beta and theta appeared in the effects of the other factors. Indeed, the modulation by feedback valence was changed with the phase within the problem, with more power in exploitation and re-exploration for beta oscillations, and more power in exploration for theta oscillations. We also observed band-specific modulations. The beta, but not the theta, band was predictive of the strategy that the monkey will use at the next trial, but only after positive feedback; by showing higher power for stay compared to shift strategy. Theta oscillations showed a strong encoding of whether an incorrect feedback in exploitation was a normal as opposed to a Trap trial.

In the decision period, both bands were predictive of the upcoming strategy in the trial, by showing more power for stay compared to shift strategy, after negative feedback only. However, only the beta was modulated by the phase, by showing more power in exploitation when the stimulus formerly let to a positive

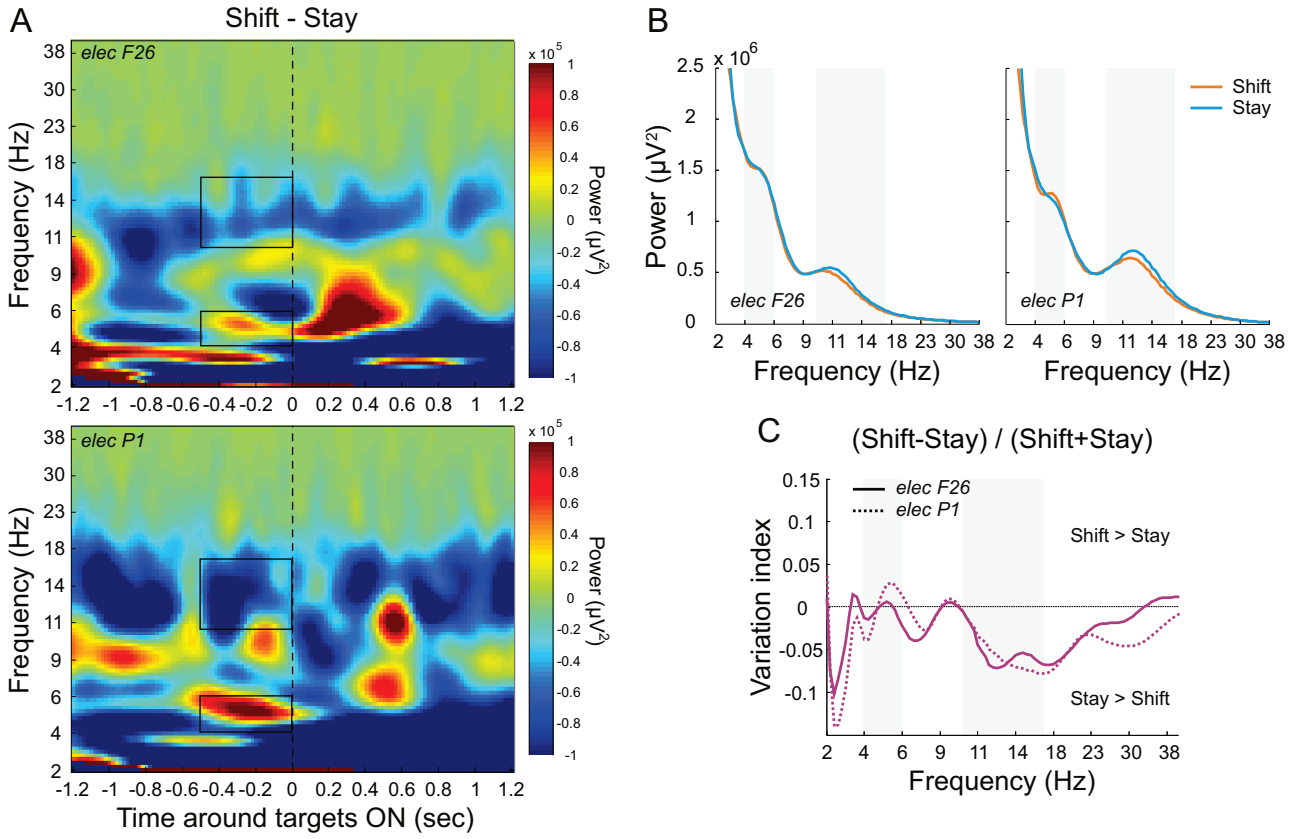

Figure 8: Description of beta and theta oscillations around targets onset showing the decision period for 2 example electrodes. **A.** Average TF representation for the Shift – Stay contrast for a frontal electrode (elec F26, top figure) and a parietal electrode (elec P8, bottom figure). Black rectangles indicates the time and frequency windows that will be used in the analyses. **B.** Power spectrum densities of the post-feedback time window shown in the black rectangles in A, for the example electrodes (elec F26, left; elec P1, right) for shift (orange) and stay (blue) conditions. **C.** Variation index calculated as (Shift-Stay)/(Shift+Stay) on the power spectrum densities described in B.

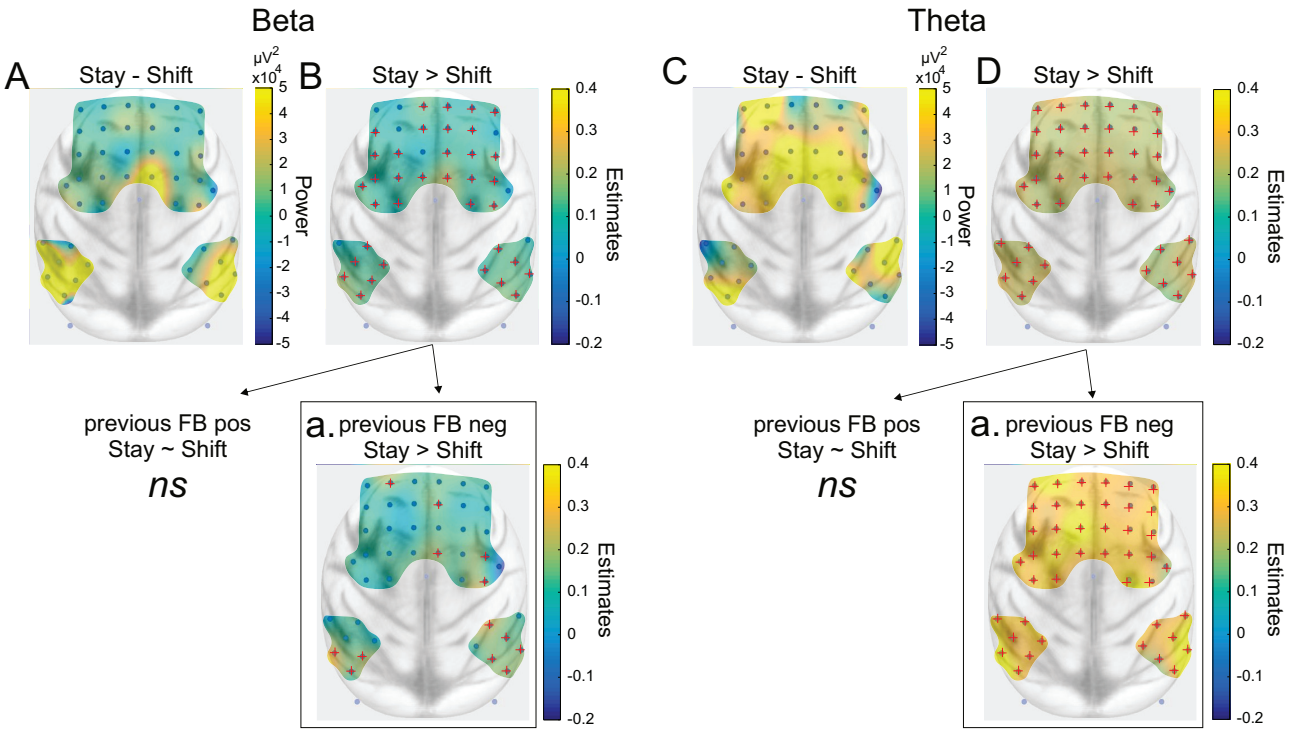

Figure 9: Effects of the upcoming strategy (shift vs stay) on Beta (left panels) **and Theta** (right panels) **oscillations over frontal and parietal electrodes during the decision period**. Cartographies showing: **A.** Beta oscillatory power contrasts between Shift and Stay trials. **B.** Estimates from the model (Stay relative to Shift). Red crosses indicate a significant p-value ($p<0.05$) on the corresponding electrode, with a positive estimate (Stay>Shift). **B.a** represents the estimates of the same model run on a subset of trials with negative feedback. **C & D.** Same cartographies as A & B respectively for Theta oscillations. **D.a** similar to B.a.

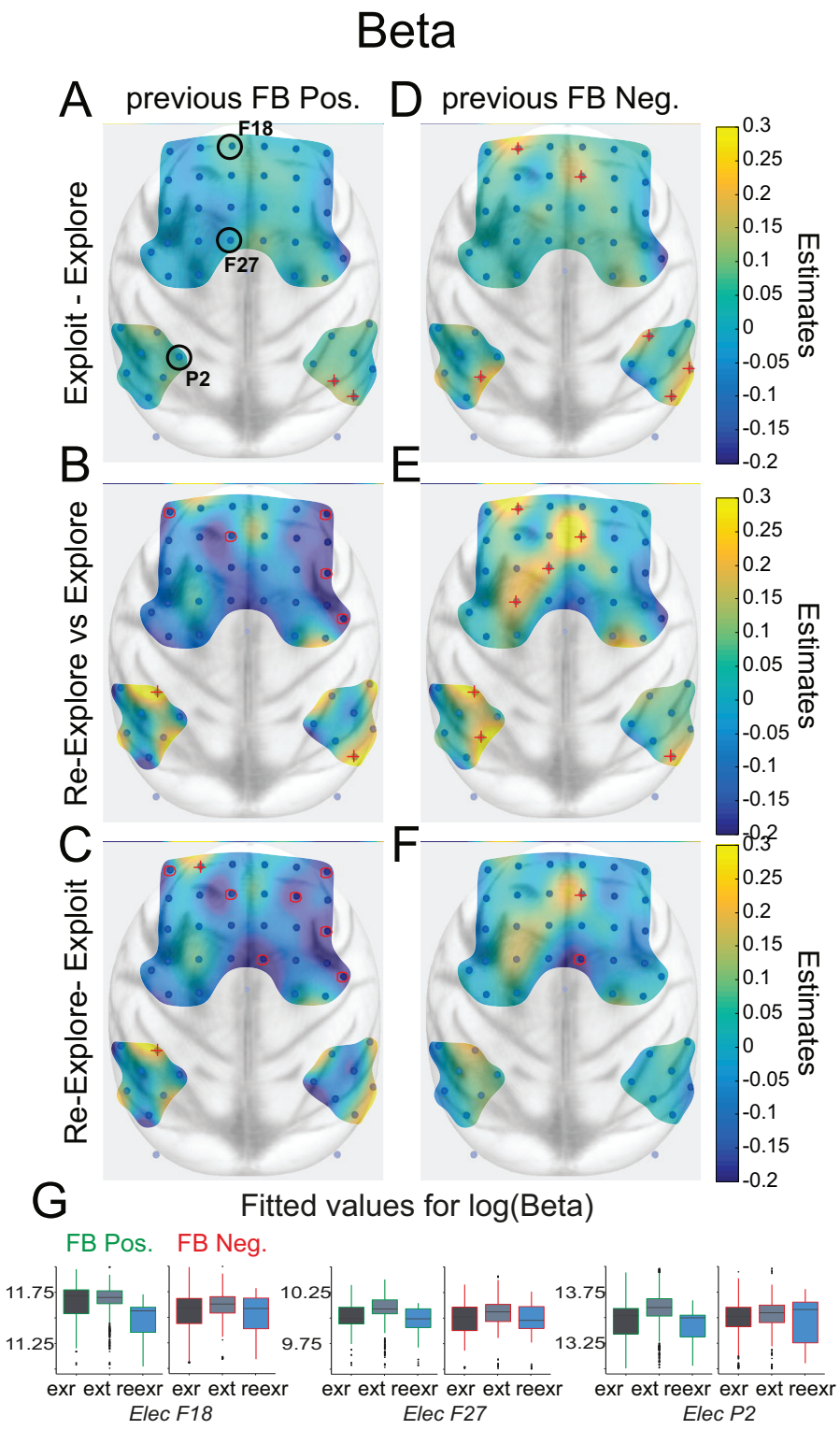

Figure 10: Effects of the factors Phase on the estimates of Beta oscillations over frontal and parietal electrodes during the decision period depending on the valence of the previous feedback (positive (left) or negative (right)). Cartographies showing Beta estimates of the model for the factor Phase for Exploit versus Explore (**A & B**) , Re-explore versus Explore (**C & D**) and Re-explore versus Exploit (**E & F**). Red crosses and circles indicate a significant p-value on the corresponding electrode, with a positive and negative estimate respectively. In Theta, no significant Phase effect was found in the 2 models with either positive or negative trials. G. Fitted values for log(Beta) extracted from the models; for 3 example electrodes shown in A. exr=exploration; ext=exploitation; reexr=re-exploration.

feedback, and more power in exploration following a negative feedback.

The pattern of modulations is complex to interpret. Further analyses and data from the second monkey will help to understand the specific contributions of these two bands in the task. However, some elements can be discussed.

In the literature, "Beta oscillations" have been described at 20-35Hz but also 10-20Hz. This might well explain discrepancies in studies concerning the relationship between beta power and feedback valence, as some studies reported increased beta power after positive feedback (Cohen et al., 2007; Hajihosseini et al., 2012; Hosseini and Holroyd, 2015; van de Vijver et al., 2011) whereas others found increased beta power after negative feedback (Cohen et al., 2009; Leicht et al., 2013). The difference between these results might be related to the exact frequency band referred to as "beta". Indeed, in most studies showing increased beta after positive feedback, "beta" refers to a band above 25 Hz (high beta) (Cunillera et al., 2012; HajiHosseini and Holroyd, 2015; Hosseini and Holroyd, 2015; Leicht et al., 2013; van de Vijver et al., 2011), whereas studies reporting the opposite results refer to a band between 10 and 20 Hz (low beta) (Cavanagh et al., 2012; Cohen et al., 2009; Leicht et al., 2013; Rothé et al., 2011; Stoll et al., 2015). The beta we observe in our study is between 10 and 25Hz, and is stronger after negative feedback, which fits with the "low beta" literature. This suggests that "beta" actually refers to multiple mechanisms, which would explain why it has been associated to so many functions. Our previous study illustrates this well (Stoll et al., 2015). We showed that changes in beta power related to cognitive control demands of the task were specific to one of two beta bands. Both monkeys showed two bands within the beta range, but the specific frequencies differed between the two monkeys. Interestingly, in both monkeys, the effects were always present in the highest band, regardless of the actual frequency.

An interesting result regarding beta power in our study concerns the adaptive strategy in reaction to a feedback. In the feedback period, beta oscillations are predictive of the strategy in the next trial after a positive but not after a negative feedback, whereas in the decision period, the reverse is observed. One interpretation could be that there is an immediate decision concerning the strategy to use after a positive feedback, whereas the decision about what should be done after a negative feedback takes longer to be established and is only present later. Differences of activity between the time of feedback and trial initiation depending on feedback valence have been previously observed in single units and gamma activity in MCC (Quilodran et al., 2008). In this study, increased neuronal activity and gamma power in MCC was observed after negative feedback in the feedback period but not in the delay period. But, as soon as the first correct feedback was obtained, the

activity was significantly reduced at the time of feedback and increased at the delay period. As the task used was deterministic, each feedback was 100% reliable and after the first positive feedback, the information triggered by subsequent positive feedback was not crucial anymore, favoring the transfer from the feedback time to the trial initiation period. In the current task, given that there are 3 options for each stimulus, a positive feedback indicates that the good target for the current stimulus has (likely) been found, whereas a negative feedback only discards one of the three targets. Thus, a positive feedback is immediately informative about the strategy that should be adopted and this information can potentially be encoded immediately after feedback onset. In contrast, reacting to a negative feedback might necessitate remembering what previous choices have already been made for the stimulus, which might explain why the encoding of the strategy occurs later. This activity at the decision time could thus relate the ignition of a working memory or decision process (Stokes, 2015).

In these data we have also revealed theta oscillatory power after the presentation of feedback; and, as described in our previous paper, feedback related potentials (FRP), occurring in the same period of the trial (see FRP paper). An ongoing debate concerns whether FRP and theta reflect the same mechanisms (Hajihosseini and Holroyd, 2013; Luu et al., 2004). A study by Munneke and coll. directly compared both signals at a single trial level and found that the two worked equally well at predicting the single-trial accuracy (Munneke et al., 2015). But they also showed that a model including both types of signal was a better predictor of the accuracy, suggesting that they represent both overlapping and distinct processes. In our study, we found a significant encoding of feedback valence in theta but the effect was not widespread, being significant on just a few separated electrodes, whereas the effect was strongly encoded in the FRP. This difference between theta and the FRP has also been reported by Hosseini and Holroyd as they reported that the FRP amplitude was mainly modulated by feedback valence whereas theta power was mainly sensitive to probabilities (Hajihosseini and Holroyd, 2013). They propose that the FRP is related to reinforcement learning signals whereas theta would reflect the MCC response to unexpected events. Their interpretation fits with data from Cavanagh and coll. showing that theta reflects an unsigned prediction error, indicating the overall degree of surprise (Cavanagh et al., 2012). However, in our data, theta power after feedback is higher in exploration after a positive feedback. Given that a positive feedback should be more surprising in exploration, this result doesn't support the idea that theta encodes surprise.

Another interesting point in our data is that theta power, but not the FRP amplitude, strongly encoded the fact that a negative feedback in exploitation comes from a Trap or a

normal trial. These results would fit with the proposition by Hajihosseini and Holroyd that theta is more related to the encoding of unexpected events than feedback valence. Indeed, here, theta oscillations differentiate between feedback with the same valence but with different degrees of surprise. However, the effect consists in a stronger power for normal compared to Trap trials. Thus, considering this interpretation as true implies that the monkey considers an incorrect feedback after an error in exploitation more surprising than a negative feedback after a correct response (Trap). This is somewhat hard to conceive. Indeed, in exploitation, the monkey should know what the correct responses are. So, the most surprising event should be when the exploited choice is not rewarded, as it is the case after a Trap trial. Here again, our data don't support a role in surprise encoding for theta oscillations.

In conclusion, in our task, beta and theta oscillations showed both common and separated modulations depending on task factors and trial periods. Both seem related to performance monitoring and cognitive control in a different manner. Our data didn't support the proposition that theta encodes surprise. However, theta oscillations after feedback were discriminating Trap from normal negative feedback in exploitation, which might reflect preparation processes for a strategy to react to the uncertain feedback. Beta oscillations seemed to reflect the subsequent strategy to implement after a feedback, but at different times depending on the information that can be extracted from the feedback. These explanations need to be confirmed on a second monkey and with complementary analyses. A closer look at the time-frequency plot reveals complex dynamics, suggesting that the story may be more complicated. For example, regarding electrode P1 in **Figure 3.A**, although we have associated beta to the entire black box, it is possible to interpret the differential activation as two separate sub-bands or activations between 11 and 25 Hz. Moreover, the central frequency of these oscillations seems to change dynamically. Thus, beta oscillations seem to consist of a complex mechanism, and it is possible that averaging their power over a time and frequency window as we did is obscuring part of the relevant information. Finer trial by trial analyses might provide a more pertinent description of the precise time and frequency dynamics of these oscillations. Further analyses are also required to extract the dynamics of these effects among the recorded regions, as cognitive control implementation might well rely on synchrony mechanisms between the different regions (Voloh et al., 2015; Voytek et al., 2015).

Material and methods

Subjects and materials

Two rhesus monkeys (Macaca mulatta), one female and one male, weighing 7 kg and 8 kg (monkeys P and D respectively) were used

in this study. Ethical permission was provided by the local ethical committee "Comité d'Éthique Lyonnais pour les Neurosciences Expérimentales", CELYNE, C2EA #42, under reference C2EA42-11-11-0402-004. Animal care was in accordance with European Community Council Directive (2010) (Ministère de l'Agriculture et de la Forêt) and all procedures were designed with reference to the recommendations of the Weatherall report, "The use of non-human primates in research". Laboratory authorization was provided by the "Préfet de la Région Rhône-Alpes" and the "Directeur départemental de la protection des populations" under Permit Number: #A690290402. Monkeys were trained to perform the task seated in a primate chair (Crist Instrument Co., USA) in front of a tangent touch-screen monitor (Microtouch System, Methuen, USA). An open-window in front of the chair allowed them to use their preferred hand to interact with the screen (both monkeys left-handed). Presentation of visual stimuli and recording of touch positions and accuracy was carried out by the EventIDE software (Okazolab Ltd, www.okazolab.com). During the behavioral task, eye movements were monitored using an Iscan infrared system (Iscan, Inc.). Electrophysiological data were recorded using a Blackrock multichannel system (Blackrock).

Behavioral task

▷ *Principle of the task*

The task is an adaptation for monkeys of the task described for human subjects in (Collins and Koechlin, 2012). Across successive trials, a problem consisted in the monkey concurrently finding, by trial and error, the correct mappings between stimuli and targets, within a set of 2 stimuli (stimulus 1 and 2) and 3 targets (target A, B and C) (**Figure 1**). For example, a problem would consist in concurrently finding the two associations: "stimulus 1 with target A" and "stimulus 2 with target C". Monkeys learned problems to a behavioral criterion. The task contained stochasticity, in the form of unreliable feedback, and volatility, in the form of switches between problems. Importantly, monkeys were given stochastic feedback from the start of the training (see our Behaviour Paper for details on the training strategies).

▷ *Procedure of the task*

Trial procedure

The structure of a single trial and a single problem was always the same, regardless of the form of the task. Monkeys initiated each trial by touching and holding a lever item, represented by a white square at the bottom of the screen (**Figure 1.A**). A fixation point (FP) appeared. After a delay period of 1200ms (delay 1), a stimulus was displayed at the top of the screen (Stim ON signal), and was followed after a second delay of 1200ms (decision period) by the appearance in the middle of the screen of three targets (Targets ON signal). Stimuli consisted of square

bitmap images of either an abstract picture or a photograph, of size 65x65mm. Targets were three empty grey squares, of the same size as the stimulus. After a further delay randomly varying between 0, 700 and 1000ms all targets turned white, providing the GO signal following which monkeys were permitted to make their choice by touching a target. Monkeys maintained touch on the chosen target for a random amont of time from 500 to 1500 ms in order to receive visual feedback on that choice. Feedback consisted of horizontal (positive) or vertical (negative) bars within each of the three targets, that needed to be hold for 700 to 2000ms. A positive feedback was followed by the delivery of about 1.8 mL of 50% apple juice. After the completion of a trial, a new stimulus was picked within the set of 2 stimuli and monkeys were allowed to begin a new trial.

Problem procedure

A problem consisted of the monkeys learning about 2 stimuli concurrently. For a given trial, one of the 2 stimuli was pseudo-randomly selected (50% of each stimulus over 10 consecutive trials). The two concurrent stimuli were never associated to the same target. Hence, there were 6 possible mappings of the 2 stimuli and the 3 targets. Each mapping was randomly selected (with the constraint that both associations were changed), so that the 2 mappings of a problem could never be predicted nor learned. The only way to find them was to proceed by trial and error based on feedback provided after each choice. After a random number of trials between 60 and 85 trials, the problem changed and 2 new mappings were randomly selected, without any cue signaling the change. We refer to this change of problem as a 'Switch'. These Switches provide the volatility in the environment of the task.

In addition to and separate from this volatility, a stochastic reward environment was created by providing misleading feedback (called 'Trap feedback') on 10% of trials. Trap trials occurred pseudo-randomly once every 10 trials, with the constraint that there were at least 2 consecutive normal trials between each Trap trial. Trap feedback was the inverse of that determined by the current mapping – as such Trap feedback after a correct response consisted of negative feedback (see below) and no reward; Trap feedback after an incorrect response consisted of positive feedback and a reward.

Task motivation

In order to motivate and maintain performance at a stable level throughout each daily session, animals were asked to obtain a fixed number of rewards each day (190 rewards). Upon successfully completing this number of rewards, monkeys received a large reward bonus (50 ml of fruit juice, calculated based on the effectiveness in motivating the monkey).

The analyses that are presented in this paper are restricted to monkey D data. Monkey P was also trained for the same task and recorded the same way, but her data have not

been analyzed yet.

▷ *Behavioral analyses*

For monkey D, 12 recordings sessions were selected for analysis, on the basis of global performance on each day superior or equal to 55% of correct responses with a minimum of 3 shifts. Reaction times were calculated as the time between the GO signal and the lever release (in order to further select a target on screen). Measures beyond 2 seconds were not included in the analysis. We analyzed the data in two ways to look at both trial-to-trial and within-problem modulations of performance. For trial-to-trial analysis, we measured performance before and after Trap trial. We also measured the score of win-stay lose-shift strategy after a Trap or a normal feedback (with a score of 1 for a change or maintenance of previous response after an incorrect or a correct feedback respectively and a score of 0 for the reverse pattern of responses). For within-problem analysis, we classified the trials as a function the "Phase" within a problem, defining 4 levels: exploration, exploitation, re-exploration and perseveration. We used the following criterion: 'exploration' trials were trials that received the same stimulus, that were associated to performance of no more than 3/5 correct over a sliding window of 5 trials and that were never preceded by 'exploitation' trials. 'Exploitation' trials were those with performance of 4/5 or more. 'Re-exploration trials' were trials for which performance went back above 3/5 correct after an exploitation period. 'Perseveration' trials consisted in post-Switch trials that were correct relative to the previous rules. These trials were excluded from 'exploration' trials.

▷ *Surgical procedures*

Surgical procedures were performed under aseptic conditions. The monkey was sedated on the morning of surgery with both ketamine (10 mg/kg) and xylazine (0.5mg/kg) following pre- anesthetic treatment with glycopyrrolate (0.006 mg/kg). Once sedated, the monkey was given antibiotic (amoxicillin, 8.75 mg/kg) for prophylaxis of infection, and a nonsteroidal anti-inflamma- tory (ketoprofen, 2mg/kg) agent for analgesia. The head was shaved and an intravenous cannula put in place for intraoperative delivery of fluids (sterile saline drip, 5 mL/h/kg). The monkey was intubated, placed on isoflurane anesthesia (0.5–2.75%, to effect, in an O2 and NO2 mix), and then mechanically ventilated. Heating blankets allowed maintenance of normal body temperature during surgery. Heart rate, oxygen saturation of hemoglobin, expired CO2, body temperature, and respiration rate were monitored continuously throughout surgery. Each animal was implanted with a head-holder (Rogue Research) and intracranial electrodes. Both the placement of the electrodes and their depth were determined from separately acquired structural MRI images of each monkey using guidance of the Brainsight neuronavigation system. Holes were drilled through the skull and then stainless steel surgical screws (Synthes) were fixed into the holes, with the aim

at each site of advancing the screw through the thickness of the bone to rest on the dura-mater. Each electrode was connected to a micro-connector (Omnetics). Skin and muscles were repositioned on the skull, above the electrodes wires.

For both monkeys, a grid of 46 electrodes spaced by 7 mm was implanted throughout the frontal, sensorimotor and parietal cortices (**Figure 1**). A supplemental electrode serving as reference was screwed into the bone of the thick brow of the monkey on the midline anterior to the frontal grid.

▷ *Electrophysiological Data Processing*

All electrodes were referenced to the most frontal reference electrode **Figure 1**. The signal from each electrode was amplified and filtered (1–250 Hz) and digitized at 781.25 Hz. Data analysis was performed off-line with FieldTrip toolbox (Oostenveld et al. 2011) and homemade Matlab scripts (Matlab, The MathWork, Inc.). Trials longer than 5 sec were excluded from the analysis. Movement artifacts were removed by decomposing ECoG recordings with an independent component analysis, using the logistic infomax algorithm (Bell and Sejnowski 1995).

Data were also analyzed in the time-frequency (TF) domain by convolution with complex Gaussian Morlet's wavelets with a ratio $f/\delta f$ of 12. Single-trial TF analysis was aligned to the target onset (Targets ON signal) for analysis of the decision period, and on the onset of the feedback signal for analysis of the post-feedback period. The continuous ECoG data were epoched from -2000 to 2000 ms (by steps of 2 ms), and the power of each frequency ranging from 1 to 80 Hz using a log scale was computed. Inspection of power spectrum density representations revealed different oscillatory activities, used to define frequencies of interest. Oscillations in the beta and theta range were analyzed and averaged over the decision period (-1200 to -200ms before the ON signal. Beta: 10-17 Hz and theta: 4-6 Hz) and the post-feedback period (0–600ms after the target touch. Beta: 10-20 Hz and 10-24Hz for prefrontal and parietal electrodes respectively and theta: 6-9Hz). This procedure led to a mean beta and theta power for each trial during the 2 periods of interest. We focused on the decision period in this task in particular, because this is the period where the monkey sees the stimulus and can plan his response accordingly. Data were acquired 12 sessions (days) in monkey P.

▷ *Mixed-Effects Models*

We observed modulations of beta and theta power depending on different factors. To evaluate those modulations properly, trial-by-trial beta power measures were fitted with Linear Mixed-Effects models (Zuur et al., 2009). Such models allow us to analyze hierarchically organized data and to explicitly model variance inherent to repeated measures. In our data set, several sessions of recordings were used to analyze the within-session effect in each monkey. Characteristics such as the slope of power change over time could vary from one day to another. The random-

effects terms in these models are specifically useful to capture this sort of variation. Meanwhile, the fixed effect can separately capture the presence of the slope in of itself. We produced models with 2 sets of fixed effects, the Shift/Stay model, the Explore/Exploit model and the Trap model. The data in the 2 models are the same; they simply treat the behavioral factors differently, allowing us to capture the variance in beta and theta in different ways. Mixed models are of the form $Y_i = \beta.X_i + b.Z_i + \epsilon_i$ where X and Z correspond to fixed- and random-effects design matrices, respectively, and ϵ the random variations for each day i. All statistical procedures were performed using R (R Development Core Team 2008, R foundation for Statistical computing) and the relevant packages (nlme, MASS).

Different combinations of the following factors were included as explanatory variables to calibrate the different models: (1) FB_valence (positive/negative), corresponding to the valence of the feedback (it refers to the feedback the monkey has just received for the analysis of the feedback period, and the feedback the monkey has received at the previous trial for the analysis of the decision period), (2) Shift/Stay (2 levels: shift and stay). For the analysis of the feedback period, it corresponds to whether the monkey has repeated (stay) or changed (shift) his response compared to the previous trial with to the same stimulus. For the analysis of the decision period, it corresponds to whether the monkey will repeat and change in response in the current trial compared to the previous trial with the same stimulus. (3) Phase (exploration/ exploitation/ re-exploration), referring to the phase within the problem, as explained in the "Behavioral analysis" section. (4) Normal_or_Trap (normal/Trap) was used for distinguishing the effects of normal versus Trap feedback. (5) Time_In_Session corresponds to the time (converted in hour) of the 'Targets ON' code and was used to take time-on-task effects into account. Models were fitted to all the electrodes at once.

We tested the data using 3 different models to understand the modulations of beta and theta oscillations during the task.

'Shift/Stay' model . This model tested whether beta or theta oscillations, in the feedback and decision periods, were modulated based on trial to trial variations. This model included the factors 'Shift/Stay', 'FB_valence' and 'Time_In_Session'.

'Explore/Exploit' model . This model tested whether beta or theta oscillations, in the feedback and decision periods, were modulated based on more global variations during a problem, by including the factor 'Phase'. The factors 'FB_valence' and 'Time_In_Session' were also included in the model.

'Trap' model . This model tested whether beta or theta oscillations, in the feedback period only, were modulated depending

on whether the trial was a Trap trial or not. It was run on trials on which the monkey could possibly distinguish a Trap from a normal trial, that is, in exploitation trials only. The model was run for trials that received a negative feedback only (either Trap or normal), as there were not enough trials with a positive feedback in exploitation to run a second model.

The dependent variable Beta (or Theta) (trial-by-trial beta power measured in the time window of interest) was tested in a linear model to evaluate the need for a power transformation, that is, in particular to improve the "normality" of the data distribution. We computed and examined the profile log-likelihoods for the parameter of the Box–Cox power transformation (function box- cox in package MASS), which revealed the need to log transform the data before fitting a linear model adequately for both mon- keys. Hence, log(Beta) (or log(Theta)) was used as dependent variable in the following analyses.

▷ *Model Selection*

Models were selected using a standard procedure of constructing the model starting with all possible interactions between the included factors as described above. In a stepwise manner we evaluated the contribution of each level of fixed effect. We used the drop1 function, repeatedly testing the effect of dropping the highest-order interaction fixed-effect term on the fit (Zuur et al 2008). Models were selected using AIC, and changes in AIC between models were tested using a chi-square test ($P < 0.05$). The principle of model selection was identical for all models.

▷ *Latencies of effect detection*

We analyzed the latency of the effect of factors that were found significant in the models, for both FRP and TF data. A permutation test (with 1000 permutation) was run on the difference of the signal between the 2 levels of each factor, using unpaired samples t-test. Latencies of the effect were taken as the first significant time bin after 0.1ms after the feedback signal.

References

Bastos, André Moraes, Vezoli, Julien, Bosman, Conrado Arturo, Schoffelen, Jan-Mathijs, Oostenveld, Robert, Dowdall, Jarrod Robert, De Weerd, Peter, Kennedy, Henry, and Fries, Pascal. Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron*, 85 (2):390–401, 2015.

Benchenane, Karim, Peyrache, Adrien, Khamassi, Mehdi, Tierney, Patrick L., Gioanni, Yves, Battaglia, Francesco P., and Wiener, Sidney I. Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning. *Neuron*, 66(6):921–936, 2010.

Bosman, Conrado a., Schoffelen, Jan Mathijs, Brunet, Nicolas, Oostenveld, Robert, Bastos, Andre M., Womelsdorf, Thilo, Rubehn, Birthe, Stieglitz, Thomas, De Weerd, Peter, and Fries, Pascal. Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas. *Neuron*, 75(5):875–888, 2012.

Buschman, Timothy J and Miller, Earl K. Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science*, 315(Ci):1860–1862, 2007.

Cavanagh, James F, Cohen, Michael X, and Allen, John J B. Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(1):98–105, jan 2009.

Cavanagh, James F., Figueroa, Christina M., Cohen, Michael X., and Frank, Michael J. Frontal theta reflects

uncertainty and unexpectedness during exploration and exploitation. *Cerebral Cortex*, 22(11):2575–2586, 2012.

Cavanagh, James F, Sanguinetti, Joseph L, Allen, John J B, Sherman, Scott J, and Frank, Michael J. The Subthalamic Nucleus contributes to post-error slowing. *Journal of Cognitive Neuroscience*, pages 1–8, 2014.

Cohen, Jonathan D, McClure, Samuel M, and Yu, Angela J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1481):933–942, 2007.

Cohen, Michael X, Elger, Christian E, and Fell, Juergen. Oscillatory activity and phase-amplitude coupling in the human medial frontal cortex during decision making. *Journal of cognitive neuroscience*, 21(2):390–402, 2009.

Collins, Anne and Koechlin, Etienne. Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10(3):e1001293, 2012.

Cunillera, Toni, Fuentemilla, Lluís, Periañez, Jose, Marco-Pallarès, Josep, Krämer, Ulrike M., Càmara, Estela, Münte, Thomas F., and Rodríguez-Fornells, Antoni. Brain oscillatory activity associated with task switching and feedback processing. *Cognitive, Affective, & Behavioral Neuroscience*, 12(1):16–33, 2012.

Fries, Pascal. A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480, 2005.

Hajihosseini, Azadeh and Holroyd, Clay B. Frontal midline theta and N200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology*, 50(6):550–562, 2013.

HajiHosseini, Azadeh and Holroyd, Clay B. Sensitivity of frontal beta oscillations to reward valence but not probability. *Neuroscience Letters*, 602:99–103, 2015.

HajiHosseini, Azadeh, Rodríguez-Fornells, Antoni, and Marco-Pallarés, Josep. The role of beta-gamma oscillations in unexpected rewards processing. *NeuroImage*, 60(3):1678–1685, 2012.

Hosseini, Azadeh Haji and Holroyd, Clay B. Reward feedback stimuli elicit high-beta EEG oscillations in human dorsolateral prefrontal cortex. *Scientific Reports*, 5(April):13021, 2015.

Johnston, Kevin, Levin, Helen M., Koval, Michael J., and Everling, Stefan. Top-Down Control-Signal Dynamics in Anterior Cingulate and Prefrontal Cortex Neurons following Task Switching. *Neuron*, 53(3):453–462, 2007.

Leicht, Gregor, Troschütz, Stefan, Andreou, Christina, Karamatskos, Evangelos, Ertl, Matthias, Naber, Dieter, and Mulert, Christoph. Relationship between Oscillatory Neuronal Activity during Reward Processing and Trait Impulsivity and Sensation Seeking. *PLoS ONE*, 8(12):e83414, 2013.

Luft, Caroline Di Bernardi, Nolte, Guido, and Bhattacharya, Joydeep. High-learners present larger mid-frontal theta power and connectivity in response to incorrect performance feedback. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(5):2029–2038, 2013.

Luu, Phan, Tucker, Don M, and Makeig, Scott. Frontal midline theta and the error-related negativity: neurophysiological mechanisms of action regulation. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 115(8):1821–35, aug 2004.

Micheli, Cristiano, Kaping, Daniel, and Westendorff, Stephanie. Inferior-frontal cortex phase synchronizes with the temporal – parietal junction prior to successful change detection. (August), 2015.

Miller, Earl K and Cohen, Jonathan D. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24:167–202, 2001.

Munneke, Gert-Jan, Nap, Tanja S., Schippers, Eveline E., and Cohen, Michael X. A statistical comparison of EEG time- and time–frequency domain representations of error processing. *Brain Research*, 1618:222–230, 2015.

Pesaran, Bijan, Nelson, Matthew J., and Andersen, Richard a. Free choice activates a decision circuit between frontal and parietal cortex. *Nature*, 453(7193):406–409, 2008.

Pfurtscheller, G and Lopes da Silva, F H. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 110 (11):1842–57, nov 1999.

Phillips, Jessica M., Vinck, Martin, Everling, Stefan, and Womelsdorf, Thilo. A long-range fronto-parietal 5- to 10-Hz network predicts "top-down" controlled guidance in a task-switch paradigm. *Cerebral Cortex*, 24(8):1996–2008, 2014.

Quilodran, René, Rothé, Marie, and Procyk, Emmanuel. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2):314–325, 2008.

Rothé, Marie, Quilodran, René, Sallet, Jérôme, and Procyk, Emmanuel. Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(31):11110–11117, 2011.

Siegel, Markus, Donner, Tobias H., and Engel, Andreas K. Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, 13(February):20–25, 2012.

Stokes, Mark G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, pages 1–12, 2015.

Stoll, F. M., Wilson, C. R. E., Faraut, M. C. M., Vezoli, J., Knoblauch, K., and Procyk, E. The Effects of Cognitive Control and Time on Frontal Beta Oscillations. *Cerebral Cortex*, pages 1–18, 2015.

van de Vijver, Irene, Ridderinkhof, K. Richard, and Cohen, Michael X. Frontal Oscillatory Dynamics Predict Feedback Learning and Action Adjustment. *Journal of Cognitive Neuroscience*, 23(12):4106–4121, 2011.

van Driel, Joram, Cox, Roy, and Cohen, Michael X. Phase-clustering bias in phase–amplitude cross-frequency coupling and its removal. *Journal of Neuroscience Methods*, 254:60–72, 2015.

Voloh, Benjamin, Valiante, Taufik a., Everling, Stefan, and Womelsdorf, Thilo. Theta–gamma coordination between anterior cingulate and prefrontal cortex indexes correct attention shifts. *Proceedings of the National Academy of Sciences*, (MAY):201500438, 2015.

Voytek, Bradley, Kayser, Andrew S, Badre, David, Fegen, David, Chang, Edward F, Crone, Nathan E, Parvizi, Josef, Knight, Robert T, and D'Esposito, Mark. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nature Neuroscience*, 18(July):1–10, 2015.

Zuur, Alain F., Ieno, Elena N., Walker, Neil J., Saveliev, Anatoly a, Smith, Graham M., and Ebooks Corporation. *Mixed Effects Models and Extensions in Ecology with R*. 2009.

Chapitre 8

Discussion

8.1 Résumé

Le but de ce travail était d'obtenir des singes qu'ils maitrisent une tâche complexe de prise de décision dans un environnement incertain afin de caractériser : 1) le développement de la réponse à l'incertitude lors du processus d'apprendre à apprendre ; et 2) les corrélats neurophysiologiques de ces processus de décision via l'analyse des activités oscillatoires et évoquées en provenance des aires frontale et pariétale. Le premier objectif a été rempli, le deuxième est encore en cours.

Etude du développement de la réponse à l'incertitude au cours du processus d'apprendre à apprendre

Les façons dont s'acquièrent les stratégies pour apprendre de l'environnement incertain sont encore mal connues. Nous avons donc étudié l'acquisition, par trois singes, d'un *learning-set* pour une tâche d'inversions (*reversals*) probabilistes. Nous avons ainsi pu mettre en évidence le développement d'une stratégie exploratoire en réaction aux *feedback* inattendus induits par l'environnement stochastique et volatil. De manière intéressante, cette réactivité aux *feedback* inattendus a augmenté lors de l'acquisition d'un *learning-set* pour la tâche. L'acquisition de cette réponse a potentiellement été bénéfique puisque les singes ont pu transférer, sans coût, à une tâche de *reversals* probabilistes dans laquelle les changements ne sont plus indiqués. Nous proposons que l'augmentation de cette réponse exploratoire soit liée à l'augmenta-

tion conjointe de la volatilité environnementale, et à l'apprentissage de la structure statistique de l'environnement.

Etude des corrélats électrophysiologiques des processus de décision dans un environnement incertain

Le potentiel évoqué au *feedback*, enregistré en surface au niveau de la ligne fronto-centrale, a été suggéré comme étant le reflet des mécanismes permettant le suivi et à l'évaluation des performances. Cependant, l'implication de ce signal dans l'adaptation du comportement reste encore controversée. Deux singes ont été implantés de manière chronique afin d'enregistrer le signal électrocorticographique continu en provenance des cortex frontal et pariétal lors de la réalisation de la tâche. A l'heure actuelle, nous avons analysé le signal d'un seul des deux animaux. Néanmoins, ces analyses nous ont permis d'identifier des modulations intéressantes des potentiels évoqués par l'apparition du *feedback* (*Feedback Related Potentials* FRP, observés entre 200 et 350 ms post-*feedback*). Ces FRP sont en effet modulés par la valence du *feedback* reçu et son degré de surprise. De plus, dans le cas des *feedback* positifs, les FRP sont prédictifs de l'ajustement comportemental nécessaire en réaction à ce *feedback* (une amplitude plus négative étant associée aux *feedback* négatifs, aux *feedback* les plus inattendus, et lorsque le singe va changer de stratégie). Cela conforte le rôle de ces potentiels dans la détection des évènements pertinents pour l'adaptation. Pour autant, leur implication dans l'ajustement comportemental est moins évidente. Nous discuterons en particulier dans quelles mesures ils sont également reliés à une implémentation du contrôle cognitif.

Les oscillations pourraient notamment porter le mécanisme permettant de faire le lien entre évaluation des performances et ajustement comportemental. Nous avons donc étudié les réponses oscillatoires dans les bandes beta (10-25Hz) et thêta (4-8Hz) après le *feedback*, mais aussi au début de l'essai, dans la période de décision. Ces bandes de fréquence ont toutes deux été impliquées dans l'exercice du contrôle cognitif dans la littérature, mais leur contribution respective dans ce processus reste à clarifier. Dans notre tâche, les oscillations beta au *feedback* sont modulées par la valence du *feedback* reçu, mais reflètent également des modulations de l'encodage de ce *feedback* en fonction du degré d'apprentissage du problème, suggérant un enco-

dage du degré de surprise ou de la pertinence du *feedback*. Les oscillations beta sont également prédictives de la stratégie utilisée à l'essai suivant, mais seulement lorsque le *feedback* reçu est positif. Les oscillations thêta au *feedback* représentent également la valence de celui-ci, et leur puissance est modulée en fonction du degré d'apprentissage du problème mais uniquement pour les *feedback* positifs. Contrairement au beta, elles ne sont pas prédictives de la stratégie qui sera utilisée à l'essai suivant. Par contre, elles discriminent si un *feedback* négatif reçu en exploitation consiste en un *Trap* ou en un vrai *feedback*.

Au moment de la décision, c'est-à-dire dans la période de l'essai précédant le choix, les deux bandes sont prédictives de la stratégie qui va être utilisée en réaction à l'essai précédent, mais seulement en réaction à un *feedback* négatif. De plus, les oscillations beta, mais pas thêta, encodent le *feedback* reçu à l'essai précédent, de manière différente selon le degré d'apprentissage du problème. Des analyses complémentaires sont encore nécessaires afin de tirer au clair les rôles respectifs de ces différentes oscillations.

Un résumé de ces modulations est présenté dans la Figure 8.1

X: factor not significant *- : factor not tested* *Exr: exploration* *Exr: exploitation* *ReExr: re-exploration*		BETA		THETA		FRP
		Feedback	**Decision**	**Feedback**	**Decision**	**Feedback**
FB valence		Neg>Pos	-	Neg>Pos	-	Neg>Pos
Strategy	FB pos	Stay>Shift	x	x	x	Shift>Stay
	FB neg	x	Stay>Shift	x	Stay>Shift	x
Phase	FB pos	Exr + Ext +++ ReExr +++	Exr +++ Ext +++ ReExr +	Exr + Ext +++ ReExr +	x	Exr + Ext + ReExr +++
	FB neg	Exr + Ext + ReExr +++	Exr + Ext +++ ReExr +++	x	x	Exr + Ext +++ ReExr +
Trap (inc ext)		x	-	Normal> Trap	-	x

Figure 8.1 – Résumé des effets des modulations des potentiels évoqués au feedback et des oscillations beta et thêta en fonction des facteurs de la tâche de Switch. Données obtenue avec le singe D. X : facteur testé mais non significatif. - : facteur non testé. Exr : exploration, Ext : exploitation ; ReExr : ré-exploration.

8.2 Discussion sur la partie comportement

Pour commencer cette discussion, il semble important de souligner que nous avons réussi à apprendre à des singes une tâche particulièrement difficile consistant en l'apprentissage d'options concurrentes dans un environnement à la fois stochastique et volatil. Ceci est le résultat d'un entrainement progressif et contrôlé sur plus de deux ans (les singes étant naïfs au début de ma thèse) ; et ce pari n'était pas gagné d'avance. Tout d'abord, les tâches utilisées habituellement chez le macaque sont relativement simples et se déroulent souvent dans des environnements déterministes. Il n'existait donc pas vraiment de méthode d'entrainement à ce genre de tâche. Nous avons donc développé une stratégie d'entrainement en décidant de donner un *learning-set* à nos singes pour une version plus simple de la tâche finale (dans laquelle les *reversals* sont signalés) mais en utilisant dès le départ un environnement stochastique, en espérant que les singes acquièrent la structure de la tâche pour transférer plus facilement vers la tâche finale. Pour le choix de cette stratégie d'entrainement, nous nous sommes inspirés de la littérature sur le *learning-set*, et en particulier des travaux de Shrier chez le singe montrant que le transfert d'une tâche de discrimination à une tâche de *reversals* était possible si les animaux avaient auparavant acquis un solide *learning-set* dans la première tâche (Schrier, 1966). Cependant, il est important de noter que cette étude, ainsi que la plupart des études de la littérature du *learning-set* s'est construite dans le cadre de tâches déterministes. Les environnements stochastiques impliquent de modifier un peu la façon dont les *learning-set* ont été envisagés.

Learning-set dans un environnement stochastique

Un critère de définition pour dire qu'un *learning-set* a été acquis dans une tâche de *reversal* consiste notamment en un comportement optimal - ou tendant vers l'optimalité – telles que des performances à 100% dès le 2ème essai *post-Shift* (Harlow, 1949). De telles performances sont obtenues chez plusieurs espèces (Duncan, 1960; Slotnick et al., 2000; Wilson and Gaffan, 2008). Dans un environnement stochastique, un tel niveau de performance est impossible à atteindre, de par la nature même de l'environnement, mais il semble tout de même raisonnable de dire que nos

singes ont acquis un *learning-set*. En effet, le cœur de la définition d'un *learning-set* consiste en une amélioration progressive puis une stabilisation des performances à un niveau élevé (Harlow, 1949), ce que nous observons. Une autre propriété du *learning-set*, moins établie dans la littérature, est qu'il confère une capacité de transfert à d'autres tâches avec la même structure, se manifestant par des performances déjà hautes dès les premiers problèmes de la nouvelle tâche (Schrier, 1966), ce que nous observons également. Ce transfert devrait être positif ou négatif, selon les similarités entre les deux tâches. Cela a par exemple été montré dans une étude de Collins et Frank, dans laquelle les sujets montrent un transfert positif ou négatif à une nouvelle tâche selon la façon dont ils ont naturellement appris la structure de la tâche d'entrainement (même si, dans cette étude, le fait que les sujets aient acquis ou non un *learning-set* n'a pas été spécifiquement testé) (Collins and Frank, 2013). Des études futures pourraient continuer sur la lancée de cette étude et utiliser les expériences de transfert, comme mesure de ce qui a véritablement été appris par les sujets lors de l'acquisition du *learning-set* et de leur capacité de généralisation.

Pour poursuivre cette discussion sur les implications de l'acquisition d'un *learning-set* dans un environnement stochastique par rapport à déterministe, on peut se demander si nos résultats questionnent le fait que deux essais doivent nécessairement être contigus afin de permettre la formation de la mémoire prospective dans l'acquisition d'un *learning-set*. Murray et Gaffan ont proposé que la mémoire prospective, en tant que mémoire propagée d'un essai à l'autre entre deux essais consécutifs, jouait un rôle critique dans la mise en place des *learning-set* dans ce genre de tâche, en montrant que séparer les essais de 24h bloquait la formation du *learning-set* (Murray and Gaffan, 2006). En réalité, la contiguïté entre deux essais n'est pas une condition strictement indispensable à l'acquisition d'un *learning-set*, comme en témoigne, par exemple, le fait que des singes peuvent acquérir des *learning-set* sur des problèmes de discrimination présentés de manière concurrente (Browning et al., 2005). De manière similaire, dans notre tâche, outre le fait que nous avons également utilisé des problèmes concurrents, les *Traps* peuvent aussi être considérés comme des perturbateurs de la mémoire prospective, puisqu'ils empêchent une bonne continuité entre les essais. Si ces deux types de perturbateurs de la mémoire prospective n'ont

pas bloqué l'acquisition du *learning-set*, il est par contre fortement possible qu'ils en aient ralenti l'acquisition (Treichler, 2005). On peut se demander si les singes n'ont acquis le *learning-set* que grâce aux essais non *Trap* pour lesquels des stimuli identiques se suivaient. Des analyses calculant l'information mutuelle permettant de déterminer le poids des choix précédents sur les choix en cours pourraient peut-être répondre à cette question (Berlyne, 1957). Dans l'équipe, nous sommes également en train de tester expérimentalement l'influence des *Trap* sur la vitesse d'acquisition du *learning-set* car deux singes sont actuellement entrainés sur une version déterministe de la même tâche.

Si, comme nous le proposons, les *Traps* sont des perturbateurs de la mémoire prospective, et donc de l'acquisition des *learning-sets*, on peut se demander pourquoi nous avons choisi d'entrainer les singes dans cet environnement stochastique. La première raison est que nous voulions que nos singes apprennent à apprendre dans un contexte d'incertitude. Comme nous l'avons vu dans l'introduction de cette thèse, la façon de prendre des décisions dans un environnement incertain dépend de manière cruciale de la façon dont l'organisme a appris à le faire, et notamment du degré d'incertitude de l'environnement d'apprentissage (Biernaskie et al., 2009; Tebbich and Teschke, 2014). Cette exigence d'entrainement était un choix risqué : il aurait été certainement plus rapide d'entrainer d'abord les singes à la tâche en version déterministe, afin qu'ils intègrent sa structure de base, puis d'introduire les *feedback* probabilistes. Mais procéder de cette manière ne nous aurait pas permis d'obtenir un comportement de prise de décision en contexte d'incertitude qui se soit développé de manière naturelle en réaction à ce même contexte d'incertitude. Ce que nous aurions obtenu aurait peut-être été un comportement approprié à une tâche déterministe, qui s'adapte tant bien que mal à une tâche devenue probabiliste.

Comme nous n'avons pas, pour l'instant, les données de transfert des singes ayant appris la tâche en contexte déterministe, nous ne pouvons pas dire que notre stratégie soit la plus efficace. Il nous est par contre possible de dire que notre stratégie a été efficace, puisque cela a permis d'obtenir un excellent transfert, sans coût, à la tâche finale (stochastique, avec les changements non signalés). Les données à venir nous apporteront des informations précieuses quant à l'intérêt d'apprendre à apprendre

en contexte probabiliste. En effet, ces singes pourraient montrer différents déficits au moment du transfert. Si, lors du premier problème dans la nouvelle tâche, ils mettent plus de temps à atteindre le critère de performance que ceux ayant appris en contexte stochastique, cela signifiera que leur déficit est lié au fait d'apprendre avec la stochasticité. Par contre, peut être que le fait d'acquérir un *learning-set*, même dans un environnement déterministe, pourrait être suffisant pour qu'ils parviennent à se débrouiller dans une version stochastique de la même tâche. Cela voudra dire que le fait d'apprendre à apprendre dans un environnement stochastique n'est pas critique pour savoir prendre des décisions dans un contexte stochastique. Dans une étude réalisée chez l'homme, il a d'ailleurs été rapporté des niveaux similaires de transfert à une tâche probabiliste selon que les sujets avaient appris la tâche dans un environnement déterministe ou probabiliste (Mehta and Williams, 2002). Les auteurs montrent en effet que le facteur important pour un bon transfert était plutôt le niveau final des performances juste avant le transfert, en montrant qu'à niveau de performance final égal, le transfert était équivalent. Ainsi, si on se réfère aux conclusions de cette étude, le bon niveau de transfert chez nos singes seraient plus dû au fait qu'ils aient un bon *learning-set* qu'au fait qu'ils aient appris dans un environnement probabiliste. Mais une critique de cette étude est que, comme les sujets ne pouvaient pas avoir le même niveau de performance dans la tâche probabiliste et la tâche déterministe, les auteurs ont arrêté l'apprentissage de la tâche déterministe de telle façon à ce que les performances soient au même niveau que celles de la tâche probabiliste. Dans ces conditions, le transfert est équivalent. Cela pose problème dans le sens où les sujets ne sont donc pas au maximum de leur performance dans la tâche déterministe, ce qui est à l'origine d'un transfert moins bon que celui obtenu au maximum de l'apprentissage. Dans tous les cas, cela conforte l'idée que plus d'études sont nécessaires afin de déterminer l'impact réel de l'environnement d'apprentissage et du *learning-set* sur les mécanismes développés pour apprendre de l'environnement.

Pour finir avec les prédictions concernant les performances de transfert de nos singes entrainés sur une version déterministe de la tâche, il est possible qu'ils parviennent à réaliser le premier problème sans difficulté, mais que leurs performances

chutent dès lors que le changement de problème a eu lieu. Deux interprétations sont alors possible. La première interprétation est que ce déficit serait dû au fait qu'ils ont un problème avec le *reversal* non signalé. Cela ne devrait pas être le cas puisque des singes ayant acquis un *learning-set* sur une tâche d'associations sont capables de transférer sans coût à une tâche de *reversals*, comme l'a montré l'étude de Schrier dont nous nous sommes inspirés (Schrier, 1966). L'autre interprétation est que le déficit serait dû à un problème pour réagir aux changements parce qu'ils sont difficiles à distinguer des *Traps* (problème lié à l'interaction stochasticité x *reversals*). Cela montrerait qu'apprendre à apprendre en contexte stochastique permet d'apprendre à réagir à un environnement incertain, où l'incertitude peut être due à la stochasticité ou à la volatilité.

Si notre interprétation est bonne, notre stratégie d'entrainement pourrait permettre des transferts à des tâches nécessitant de remettre en cause des associations apprises. Une question serait par exemple de savoir si ce *learning-set* acquis sur un écran tactile serait généralisable à une tâche avec de vrais objets (comme dans les dispositifs de type Wisconsin general test apparatus, décrit par Harlow (Harlow, 1949)). Les études ayant montré que les performances de rats dans une expérience de *learning-set* étaient fortement améliorées lorsque le test était effectué via la modalité olfactive plutôt que visuelle (Slotnick and Katz, 1974) suggèrent que l'acquisition d'un *learning-set* n'est pas un processus complètement abstrait. Au contraire, on peut voir le *learning-set* comme l'amélioration de la capacité de créer des *task-sets*, c'est-à-dire d'associer des stratégies avec les commandes motrices, les processus attentionnels spécifiques à une tâche. Ainsi, il est peu probable qu'avoir un *learning-set* pour une tâche apprise sur écran tactile permette un transfert parfait à une tâche avec des objets. Néanmoins, si les règles régissant les liens entre les objets et les récompenses sont les mêmes que celles qui régissaient les stimuli à l'écran et les récompenses, il est possible qu'une partie du *learning-set* soit transférable, même si cela devrait exiger un petit temps d'adaptation pour acquérir les bonnes commandes motrices par exemple.

L'augmentation de la réactivité au Trap est-elle le reflet de l'apprentissage de la structure ?

La deuxième raison du choix de stratégie d'entrainement dans un environnement stochastique était que nous voulions suivre l'évolution de la réponse à l'incertitude au cours de l'acquisition d'un *learning-set*. L'idée étant que comprendre la façon dont ces processus se mettent en place éclaire sur leur fonctionnement une fois qu'ils sont développés. Les résultats que nous avons obtenus sont contre-intuitifs aux premiers abords, ce qui contribue à justifier l'intérêt de s'intéresser à ces processus. Nos résultats montrent que nos singes réagissent aux *feedback* inattendus en choisissant d'explorer immédiatement une autre option par rapport à la stratégie utilisée avant le *Trap*, avant de revenir à leur choix initial s'ils se rendent compte que les règles n'ont pas changé. Le résultat surprenant a été de voir qu'avec l'acquisition du *learning-set*, les singes n'ont pas appris à ignorer ces *feedback* inattendus. Ils auraient pu, par exemple, utiliser une stratégie de persévération sur un essai en réaction à tout *feedback* inattendu, pour confirmer un maintien ou un changement des règles. Ce résultat est particulièrement étonnant parce que ce comportement a l'air non optimal, en particulier parce que, dans la version d'entrainement de la tâche, tout changement de problème était signalé (par un changement conjoint de l'identité des stimuli). Ainsi, il aurait été moins surprenant d'observer une diminution de cette réponse au *Trap* avec l'apprentissage, du moins dans les périodes d'exploitation. Or, de manière intéressante, la réponse au *Trap* est toujours présente en exploitation, et augmente avec l'apprentissage.

Nous avons interprété l'augmentation de la réponse au *Trap*, comme la conséquence directe du niveau de la volatilité, qui a augmenté de manière parallèle à l'apprentissage. En effet, dans les versions d'entrainement de la tâche, les changements (*Shifts*) entre problèmes ne se réalisaient que lorsque le singe avait atteint un critère de performance sur le problème. Ainsi, au fur et à mesure que les singes devenaient de plus en plus efficaces au cours de l'acquisition du *learning-set*, les changements de problèmes se réalisaient de plus en plus rapidement. Une fois l'apprentissage arrivé à une asymptote, la volatilité a atteint une valeur stable élevée, tout comme la réponse aux *feedback* inattendus. C'est pourquoi, dans notre article,

nous proposons que l'augmentation de la réponse aux *feedback* inattendus au cours de l'acquisition du *learning-set* reflète l'augmentation du degré d'incertitude de l'environnement (rapport *Trap*/*Shift*). Cette interprétation va dans le sens de résultats chez l'homme montrant que les humains sont plus réactifs aux *feedback* dans les environnements volatils que stables (Behrens et al., 2007).

Cependant, cette interprétation est soumise à critique. En effet, le fait que les changements de problèmes soient conditionnés à un critère de performance fait que la volatilité dépend des performances. L'apprentissage et l'augmentation de la volatilité sont donc des facteurs confondants. Nous ne pouvons donc pas être certains que l'augmentation de la réponse aux *feedback* inattendus soit uniquement liée à la volatilité sans décorréler changements de volatilité et amélioration des performances. L'augmentation de la réponse aux *feedback* inattendus pourrait être due à d'autres changements ayant lieu en parallèle lors de l'apprentissage. Un premier contrôle a été de montrer que l'augmentation de la réponse n'était pas liée à une augmentation des performances. En effet, lorsque l'on considère uniquement les périodes d'exploitation de chaque problème (période avec des performances stables au cours du *learning-set*), la réponse aux *feedback* inattendus est toujours présente, et montre quand même une augmentation au cours du *learning-set*. Ainsi, l'augmentation de la réponse au *Trap* n'est pas le résultat d'un effet de seuil causé par un niveau de performances plus bas en début d'apprentissage.

Si ce choix d'entrainement au critère nous empêche de trancher complètement sur cette interprétation, il reste pertinent comme modèle d'apprentissage écologique. Dans la nature, et en particulier dans les cas de recherche de nourriture (*foraging*), l'animal n'est contraint de changer d'environnement que lorsqu'il en a épuisé toutes les ressources (Stephens, 2008). Or, un animal en train d'apprendre à apprendre va devenir de plus en plus efficace pour épuiser les ressources de son environnement, et va par conséquent devoir changer d'environnement (ou de "patch") de plus en plus souvent. Ainsi, dans ces situations naturelles, amélioration des performances et volatilité sont fortement corrélées.

Ceci étant dit, l'apprentissage au critère, s'il a traditionnellement été utilisé par les expérimentateurs pour contrôler précisément l'avancée de l'apprentissage d'un

animal, n'est pas forcément indispensable à l'acquisition d'un *learning-set*. Harlow argumente en effet que l'apprentissage au critère n'est pas une condition nécessaire pour l'acquisition d'un *learning-set* car selon lui, seul le nombre total d'essais serait le facteur important (Harlow, 1949). Ainsi, nous pouvons proposer des expériences qui permettraient de trancher entre un rôle de la volatilité et un rôle de l'apprentissage dans les changements de la réponse au *Trap* en essayant de décorréler les deux, en enlevant le critère de performance en tant que déterminant du changement de problème. Dans une nouvelle expérience, on pourrait par exemple, donner un *learning-set* pour une tâche de *reversal* probabiliste dans laquelle les changements entre les problèmes surviendraient de manière aléatoire. Cette approche permettrait de décorréler amélioration des performances et volatilité. Dans ces conditions, si la réponse aux *feedback* inattendus augmente toujours au cours de l'acquisition du *learning-set*, alors cela signifie que c'est l'apprentissage qui était responsable de son augmentation dans notre expérience. Au contraire, si sa valeur reste haute et constante tout du long de l'apprentissage, cela voudra dire que l'apprentissage n'était pas responsable de son augmentation dans notre expérience. Afin de trancher si c'est vraiment le niveau de la volatilité environnementale qui conditionne le niveau de réponse aux *feedback* inattendus, une expérience serait de comparer l'amplitude de cette réponse chez deux groupes de singes ayant acquis la tâche avec des niveaux de volatilités différents (par exemple, 5% versus 15%), et dans laquelle les changements entre problèmes surviendraient également de manière aléatoire. Si notre interprétation est juste, les singes ayant appris à apprendre dans l'environnement le plus incertain (15% de stochasticité) devraient montrer une plus grande réactivité aux *feedback* inattendus.

On peut se demander si la réactivité au *Trap* ne peut pas être considérée comme un comportement sous-optimal, puisque les changements sont signalés. Une étude par Acuna et Schrater propose une interprétation très intéressante de la sous-optimalité des sujets dans les tâches de prise de décision séquentielles (Acuna and Schrater, 2010). Les auteurs proposent que les sujets, en parallèle de la recherche de l'option la plus récompensante, apprennent en fait la structure liant les relations entre les options et leurs probabilités de récompense. Ils catégorisent les tâches de prise de

décision séquentielle utilisées actuellement en trois types de structures et proposent que les stratégies d'exploration utilisées par les sujets dépendent de cette structure, et de si le sujet est en train d'essayer de l'apprendre. La première structure est rencontrée dans les tâches où il existe une dépendance temporelle entre les probabilités passées et présentes de récompenses (comme dans le cas des tâches de Bandit à plusieurs bras. Ex : (Payzan-LeNestour and Bossaerts, 2011)). Dans ces situations non-stationnaires, l'agent doit utiliser un fort taux d'apprentissage, c'est-à-dire baser son apprentissage en majorité sur les informations récentes. La deuxième structure est rencontrée dans les tâches où les probabilités de récompenses sont affectées par les actions. Par exemple, choisir une option conduit à réduire temporairement sa probabilité de récompense. C'est le cas des environnements de *foraging*, dans lesquels les ressources sont progressivement épuisées. Dans ces conditions, un comportement rationnel consiste par exemple en un des stratégies de matching des probabilités. La dernière structure proposée par les auteurs est appelée "couplage de la récompense", et illustre les cas où recevoir de l'information sur une option est informatif sur l'autre option (par exemple, si l'une est juste, l'autre est forcément fausse). [il est intéressant de noter que cette structure est celle acquise lors d'un *learning-set* permettant d'utiliser la mémoire prospective d'un essai à l'autre]. Dans ces conditions, une exploration dite "passive" est suffisante puisqu'il n'est pas utile d'essayer l'autre option. Le cas inverse est rencontré lorsque recevoir de l'information sur une option n'est pas informatif pour connaitre l'autre option. Dans ces conditions, une exploration active est nécessaire. Ainsi, différents types de structures conduisent à différents comportements d'exploration. Les auteurs montrent que le comportement d'exploration des sujets dans ce genre de tâche est en fait mieux expliqué par un modèle prenant en compte le fait que les sujets essaient d'apprendre cette structure, en parallèle de chercher la meilleure option. Ainsi, dans notre expérience, il est possible que l'augmentation de la réponse au *Trap* soit la résultante observable du fait que les singes essaient d'apprendre la structure de l'environnement, qui se modifie au cours de l'apprentissage, en plus de chercher la meilleure option.

Il est plutôt difficile de rattacher notre tâche à l'une des 3 structures proposées par Acuna et Schrater. Dans cette tâche, obtenir de l'information sur une des deux

associations est informatif sur l'autre association, en particulier après un *feedback* positif. En effet, si une cible est trouvée pour un stimulus, alors elle ne sera pas correcte pour l'autre stimulus. Par contre, ce n'est pas le cas pour les *feedback* négatifs (car il y a 3 cibles), ce qui différencie les deux types de *feedback* en terme de contenu informationnel porté. De plus, la présence de la stochasticité rend cette déduction plus difficile. De plus, la présence des *reversals* à la suite desquels une seule ou les deux associations peuvent changer renverse régulièrement cette inter-dépendance des options.

Au final, dans notre protocole, la structure de la tâche évolue à la fois entre les problèmes, et au cours de l'apprentissage (avec l'augmentation de la volatilité), et un comportement adapté face à cette incertitude permanente est peut être bien d'avoir une très grande réactivité comportementale en réagissant à tout *feedback* surprenant.

Choix du critère exploration/exploitation

Un autre point mérite discussion : il s'agit de notre critère d'exploration/exploitation. Nous appuyons nos analyses de la réponse au *Trap*, ainsi que nos analyses des données d'électrophysiologie, sur un critère d'exploration/exploitation que nous calculons sur la base des performances (l'exploitation correspondant à des performances supérieures à 70% sur les essais avec le même stimulus). L'idée de l'utilisation de ce critère est de permettre de distinguer des périodes qui pourraient correspondre à différents degrés de contrôle engagé dans la tâche au cours de la résolution d'un problème. Cependant, une limite de ce critère est la valeur arbitraire de ce seuil. Une autre critique est d'arriver à dissocier l'exploration active, qui serait un processus contrôlé, du comportement proche de l'aléatoire (causé par de l'inattention par exemple), qui permet potentiellement un gain d'information sur l'environnement mais ne correspond possiblement pas au même état physiologique. Une façon basique de les dissocier est de mettre un seuil inférieur en considérant que si le singe reste à des performances au niveau du hasard pendant une certaine période, il ne s'agit probablement pas d'exploration active.

Plusieurs méthodes existent pour améliorer ce critère. La première consiste à utiliser le modèle PROBE créé pour la version originale de cette tâche par l'équipe de E.

Koechlin (Collins and Koechlin, 2012). Ce modèle permet d'estimer quand le sujet passe d'un comportement d'exploration à exploitation et vice versa. Ce travail a été commencé par Maxime Maheu, à l'époque étudiant en master, ayant travaillé dans notre équipe. Il s'agit d'un travail conséquent, encore en développement, mais qui pourrait donner des résultats précieux. Mais il existe également des méthodes plus simples qui permettent de limiter les défauts de notre critère, et qui permettraient de raffiner notre niveau d'explication du comportement du singe. C'est ce que nous avons essayé de faire dans l'étude d'électrophysiologie. Nous avons essayé d'améliorer le critère d'exploration/exploitation basé sur les performances, en ajoutant deux nouvelles catégories : la persévération (lorsque le singe continue à sélectionner les réponses qui étaient correctes dans le problème précédents pour les stimuli de l'essai en cours, après un changement de problème) et la ré-exploration (lorsque les performances redescendent en dessous du seuil des 70% après avoir eu une période d'exploitation). Ce dernier critère dépend lui aussi de paramètres libres. Une autre méthode que nous avons également essayée est celle développée par Gallistel pour la détection de "points de changement" (Gallistel et al., 2001). Cette méthode définit un critère statistique permettant d'identifier des périodes de "*Shift*" par rapport à des périodes de "stay". Pour l'instant, ce critère n'a pas donné des classifications très satisfaisantes ; nous travaillons encore à son optimisation.

Les singes utilisent-ils les task-sets ?

Un autre point intéressant concerne le fait que cette réponse aux *feedback* inattendus, c'est-à-dire le changement de réponse à l'essai suivant un *Trap* (ou un *Shift*) n'est en fait présente que lorsqu'on considère les essais qui ont le même stimulus que le *Trap*. En effet, si l'autre stimulus est présenté à l'essai suivant, le singe ne changera pas sa réponse. Cela signifie que l'animal ne lie pas les deux associations ensemble en un *task-set*, sinon, il interpréterait un *feedback* inattendu dans l'une des deux associations, comme le signe qu'un changement global des règles, donc des deux associations, et réagirait au *Trap* à l'essai suivant, quel que soit le stimulus. Cela montre que le singe considère les deux associations comme deux problèmes à résoudre de manière indépendante. D'après ce résultat, nos singes ne "forment pas de *task-set*", c'est-à-dire qu'ils ne lient pas particulièrement ces deux associations

en les associant à une même cause. Mais, avant de conclure que les singes ne sont pas capables de générer des *task-set* en général, il faut souligner que la façon dont fonctionnait la tâche ne poussait pas à lier les deux associations. En effet, si les transitions entre les problèmes pouvaient consister en un changement simultané des deux associations, dans la plupart des cas, une seule des deux associations était en fait changée (la seule condition pour les transitions était qu'au moins une des deux associations devait être différente). Ainsi, comme les associations pouvaient varier de manière indépendante, il n'est pas surprenant que les singes ne les aient pas forcément liées en un *task-set*. Dans la version de la tâche avec laquelle les enregistrements sont réalisés actuellement, nous avons changé les conditions de transition de telle sorte que les deux associations changent au moment des *Shifts*, afin de pousser vers l'utilisation des *task-sets*. Cependant, sachant que l'état de l'environnement dans lequel on apprend à apprendre pourrait conditionner de manière forte la façon de prendre des décisions par la suite, et nos singes ayant appris à apprendre dans un environnement où il n'était pas pertinent de lier les associations, il est fort probable qu'ils ne les lient pas dans la version finale de la tâche. Un autre test important de l'utilisation de *task-set* que nous réaliserons bientôt est celui utilisé dans l'article de Collins et Koechlin dont nous nous sommes inspirés pour notre étude (Collins and Koechlin, 2012). Collins et Koechlin ont utilisé deux types de sessions : des sessions, dites 'récurrentes', au cours desquelles certains "*task-sets*" (consistant dans cette étude en le groupement de 3 associations) étaient répétés plusieurs fois au cours de la session ; et des sessions dites 'ouvertes' au cours desquelles les "*task-sets*" utilisés n'étaient présentés qu'une seule fois. Les auteurs ont montré que les performances après un *Shift* étaient légèrement meilleures dans les sessions récurrentes par rapport aux sessions ouvertes. Ainsi, les sujets ont pris avantage de la répétition des *task-sets* suggérant qu'ils liaient bien les associations en un *task-set*. De manière intéressante, cet avantage n'a pas été retrouvé chez tous les sujets, ce qui a conduit les auteurs à classer les sujets en deux groupes : ceux qui exploitaient les *task-sets*, et les autres, très explorateurs. Cela montre qu'il existe de grandes différences inter-individuelles chez l'homme, pour ces comportements (qui seraient peut-être liées à la façon dont ces différents sujets ont appris à apprendre...). Il existe peut-être la

même variabilité chez le singe, ce que nous ne pourrons pas savoir, avec seulement 3 individus. Un avantage que nous avons est que les 3 singes ont été entrainés sur la tâche de la même façon, et montrent des comportements similaires. Ainsi, cela réduit peut-être le risque d'obtenir des comportements (exploiteur ou explorateur) différents lorsqu'on réalisera ce test. Il est difficile de prédire si nos singes seraient plutôt explorateurs ou exploiteurs. En effet, comme nous l'avons déjà évoqué, le type de transition pendant l'acquisition de la tâche n'a peut-être pas poussé à utiliser les *task-sets*. Mais on peut envisager qu'après quelques sessions où certains *task-set* se répètent, les singes pourraient apprendre rapidement à les exploiter.

L'importance des consignes

Enfin, on peut se demander si les résultats présentés dans cette thèse concernant l'acquisition d'un *learning-set* chez le singe sont pertinents pour comprendre ce genre de processus chez l'Homme. Il existe en effet une différence forte entre l'apprentissage chez un singe et chez un homme dans les tâches de prise de décision : les performances d'un sujet humain passent de 0 à 100% en quelques essais (changement brusque), alors que celles d'un singe évoluent petit à petit pour atteindre le même niveau. L'homme adulte n'est pas naïf : il possède déjà de solides *learning-sets* pour toutes sortes de tâches, qui lui permettent d'identifier rapidement les structures sous-jacentes et de les appliquer à de nouvelles tâches. Cependant, l'apprentissage des macaques adultes est similaire à celui d'un enfant humain (Ochoki et al., 1975). Ainsi, étudier l'apprentissage chez le singe peut donner des informations sur les origines et la mise en place de telles capacités chez l'homme. Il faut être aussi conscient que les données sur l'acquisition d'un *learning-set* chez l'homme, si elles existent (montrant de fortes similarités avec le singe (Duncan, 1960)), sont assez peu nombreuses, et notamment pour la simple raison que, dans la plupart des études, on donne des consignes aux sujets pour réaliser une tâche (par exemple, (Collins and Koechlin, 2012)). Ceci élimine d'emblée l'étape où le sujet pourrait apprendre la tâche par lui-même. L'influence des consignes données aux sujets sur leur vision de leur environnement et donc sur leur comportement ultérieur a été montrée (Hertwig et al., 2004; Payzan-LeNestour and Bossaerts, 2011). Une étude en cours réalisée par Karim N'Diaye et Eric Burguière dans l'équipe de Luc Mallet consiste justement à

mettre des sujets humains dans les mêmes conditions qu'un animal testé en laboratoire : sans qu'aucune consigne ne lui soit donnée, le sujet entre dans une pièce dans laquelle écrans tactiles sont placés sur les murs et il doit se débrouiller pour trouver ce qu'on attend de lui (sachant que, contrairement à l'animal, il sait déjà qu'on attend quelque chose de lui, et sait déjà se servir d'un écran tactile !). Les résultats de cette étude ne sont pas encore connus, mais il y a fort à parier que si on mettait des sujets humains devant la même tâche que celle utilisée avec nos singes dans cette thèse par exemple, leurs performances seraient bien plus similaires à celles des singes que ce à quoi on pourrait s'attendre. Ce serait une perspective très intéressante à tester. Pour finir, on peut se demander pourquoi les consignes contribuent de manière si importante à l'accélération de l'apprentissage. Une proposition serait qu'elles facilitent le recrutement ou l'identification d'un *task-set* pertinent pour la tâche en cours, qui aurait déjà été utilisé auparavant dans une situation similaire. Le rôle du *learning-set* pourrait être exactement celui-ci, et serait court-circuité par l'utilisation des consignes.

8.3 Discussion sur la partie électrophysiologie

Au cours de cette thèse, nous avons collecté un set très riche de données électrophysiologiques. Nos résultats préliminaires sont prometteurs car ils révèlent une variété de modulations spécifiques à notre tâche, à la fois dans les potentiels évoqués et les oscillations dans différentes bandes de fréquence. Des analyses plus poussées, que j'espère réaliser au cours des 6 mois pendant lesquels je resterai au laboratoire après ma soutenance, ainsi que le rajout des données (déjà acquises) du 2$^{\text{ème}}$ singe, permettront d'approfondir ces résultats.

Analyse des potentiels évoqués au feedback

La littérature sur les potentiels évoqués au *feedback* est très large (probablement parce qu'ils sont relativement faciles à obtenir expérimentalement) et il est très difficile d'en extraire une image claire étant donné les nombreuses études contradictoires (voir par exemple la revue (Walsh and Anderson, 2012) pour une liste d'articles avec des résultats contradictoires). Les potentiels liés à l'évaluation des performances

ont d'abord été découverts en réaction aux erreurs de mouvements (*Error Related Negativity ERN*), puis un potentiel évoqué aux *feedback* (externe) négatifs a été mis en évidence (*Feedback Related Negativity FRN*) (Falkenstein et al., 1991; Gehring et al., 1993; Miltner et al., 1997). Il a été proposé que les deux signaux reflètent le même mécanisme de détection qu'une erreur a été commise ou un mécanisme permettant l'utilisation de cette information pour empêcher la commission d'erreurs futures (Holroyd and Coles, 2002). Plusieurs travaux, utilisant notamment des techniques de localisation de source, se sont succédés pour identifier la source de ce signal, et les plus récents confirment l'implication du CCM (Hauser et al., 2014). Holroyd et Coles ont proposé une théorie de l'apprentissage par renforcement et du potentiel évoqué aux erreurs RL-ERN dans laquelle ils postulent que l'amplitude de la déflection dépend directement des arrivées dopaminergiques dans le CCM lors de la détection des erreurs de prédiction (Holroyd and Coles, 2002). Dans cette théorie, une erreur de prédiction négative, conduisant à une chute de l'activité dopaminergique, entrainerait une désinhibition des neurones du CCM et donc la déflection du FRP.

Par la suite, des études ont non seulement montré qu'un FRP peut aussi être observé après un *feedback* positif, mais que son amplitude peut être équivalente à celle d'un FRP évoqué par un *feedback* négatif (même s'il existe toujours une différence de niveau), à la condition que les deux types de *feedback* soient surprenants, suggérant que le FRP refléterait plutôt une erreur de prédiction absolue, ou non-signée (Ferdinand et al., 2012; Hauser et al., 2014; Holroyd and Krigolson, 2007; Oliveira et al., 2007; Walsh and Anderson, 2011). En lien avec cette proposition, notre protocole permet de comparer l'amplitude de *feedback* de même valence mais reflétant des degrés de surprise différent. En effet, dans notre tâche, le fait que l'environnement soit à la fois stochastique et volatil fait que l'animal alterne entre des périodes d'exploration et d'exploitation, mais aussi, de manière intéressante, des périodes de ré-exploration (correspondant aux périodes où l'animal pense probablement qu'un changement a eu lieu). Au cours de ces périodes, le singe reçoit à la fois des *feedback* positifs et négatifs mais son degré d'attente vis-à-vis de ces *feedback* diffère possiblement entre les périodes. Un *feedback* négatif reçu en exploration est probablement

moins surprenant, car très fréquent, par rapport à l'exploitation où il devient rare. Nos résultats vont dans ce sens car l'amplitude du FRP après un *feedback* négatif est plus négative en phase d'exploitation par rapport aux phases d'exploration ou de re-exploration. Inversement, un *feedback* positif est probablement plus surprenant en exploration qu'en exploitation. Mais dans nos résultats, l'amplitude du FRP après un *feedback* positif n'est pas différente entre exploration et exploitation, mais est plus grande en ré-exploration. Si on considère l'interprétation qui dit que l'amplitude du FRP reflète le degré de surprise du *feedback*, alors ces résultats suggèrent qu'un *feedback* positif est considéré par le singe comme porteur du même degré de surprise en exploration et en exploitation, et serait plus surprenant en ré-exploration. Un moyen de donner du sens à ces résultats serait de dire qu'en re-exploration, le singe estime que les contingences ont changé, alors qu'elles n'ont pas changé, et est par conséquent surpris par un *feedback* positif (reçu pour une réponse qu'il considère comme correspondant à l'ancien problème). Au contraire, en exploration, le singe devrait avoir moins d'attentes (puisqu'il ne sait rien), et donc moins de surprise, pour un *feedback* positif. Ces résultats suggèrent que l'amplitude du FRP semble être potentiellement modulée par le degré de surprise pour les *feedback* négatifs, mais cela semble moins clair pour les *feedback* positifs. Une façon de comprendre ce que signifie la ré-exploration pour le singe serait de réaliser des analyses comportementales fines de cette période, afin de déterminer les évènements qui déclenchent une ré-exploration (un *Trap* ? un certain temps passé à exploiter ?) et si des enchainements d'essais particuliers peuvent être observés (comme par exemple, quelques essais d'exploration, puis un retour en exploitation).

Si l'encodage du degré de surprise d'un évènement dépend des neurones dopaminergiques, le lien entre dopamine et FRP n'a jamais vraiment été testé directement. Une étude de l'équipe a bien montré des modulations du FRP (diminution de l'amplitude du pic de la différence incorrect-correct) par des injections d'halopéridol, un antagoniste dopaminergique, mais ces injections ont été réalisées de manière systémique, empêchant de faire le lien directement entre la dopamine dans le CCM et le FRP (Vezoli and Procyk, 2009). Le but final de ce projet est justement de tester si des perturbations pharmacologiques ciblant les récepteurs dopaminergiques dans

le CCM vont conduire à des modifications du FRP et du comportement. Walsh et Anderson proposent que si la dopamine est directement responsable de l'émission du FRP par les neurones du CCM, alors cela relie le FRP au système dit d'habitude (ou *model-free*) (Schultz et al., 1997; Nassar et al., 2012). Un moyen de tester cela serait de regarder si le FRP est atténué dans des conditions favorisant les comportements dirigés vers un but, par rapport à des conditions pouvant être réalisées en mode plus automatique. Nous ne pouvons pas vraiment tester cette hypothèse avec notre tâche à l'heure actuelle. En effet, il n'est pas vraiment possible de résoudre notre tâche de manière automatique tout en conservant de bonnes performances, puisque l'environnement changeant et stochastique pousse à un engagement permanent du contrôle cognitif. Cependant, il nous serait possible de réaliser quelques sessions sans changement des règles, de telles sortes à ce que le singe puisse réaliser la tâche en mode automatique et regarder si l'amplitude du FRP est plus élevée que dans les sessions avec des changements des règles. Mais un problème serait que nous aurions très peu de *feedback* négatifs à analyser. Dans l'introduction, nous avons également proposé que nous pourrions regarder si le développement d'un *learning-set* correspond à un basculement entre les modes *model-based* et *model-free*. Un moyen de tester ceci pourrait donc être de suivre le FRP au cours du développement d'un *learning-set*.

Des travaux de l'équipe ont notamment suggéré que le FRP serait un reflet de la détection des *feedback* pertinents, en lien avec le rôle du CCM dans le signalement du besoin adaptation comportementale (Vezoli and Procyk, 2009). Ainsi, une autre question cruciale qui reste encore sans réponse claire consiste à comprendre à quels moments le FRP coïncide avec le comportement, et à quels moments il en diffère, puisque les études publiées montrent soit l'un (Cohen et al., 2007; Frank et al., 2005) soit l'autre (Chase et al., 2010; Mars et al., 2004; Walsh and Anderson, 2011). Nos résultats montrent que l'amplitude du FRP (après un *feedback* positif) est prédictive de la stratégie que le singe adoptera à l'essai suivant (conserver ou changer sa réponse). Mais, sachant que l'amplitude du FRP est également modulée par le degré de surprise du *feedback* et qu'un *feedback* surprenant est plus probablement associé à un changement de comportement qu'un *feedback* attendu (Courville et al., 2006),

il est difficile de distinguer l'effet de la surprise de l'effet de l'ajustement comportemental à venir dans l'amplitude du FRP. Ainsi, si l'on peut affirmer que le FRP reflète le besoin d'ajustement comportemental provoqué par un *feedback* surprenant, on ne peut pas affirmer avec certitude que le FRP indique en quoi doivent consister ces ajustements. Une fois encore, les manipulations pharmacologiques prévues dans ce projet permettront de faire avancer notre compréhension car, si le FRP n'est pas (directement ou indirectement) relié aux ajustements comportementaux, alors des perturbations pharmacologiques de ce signal devraient être sans conséquence comportementale. Cela a déjà été montré de manière indirecte dans des résultats de l'équipe déjà mentionnés, puisque des injections systémiques d'halopéridol, un antagoniste dopaminergique, ont provoqué des changements d'amplitude du FRP sans provoquer de changement dans les stratégies comportementales (Vezoli and Procyk, 2009). Cependant, il est important de noter que la tâche utilisée dans cette étude est une tâche très simple et déterministe pour laquelle de forts niveaux de contrôle ne sont pas nécessaires, en particulier chez des singes surentrainés. Peut-être que notre tâche, nécessitant une implication plus soutenue du contrôle cognitif, conduira à d'autres résultats.

Il est important d'avoir à l'esprit que le suivi et l'évaluation des performances et la mise en place d'un contrôle adapté en conséquence sont deux processus qui sont pour l'instant considérés comme distincts, et même portés par des régions différentes, en particulier le CCM et le CPFl (Landmann et al., 2006; Rothé et al., 2011). Ainsi, il n'est pas forcément attendu que le FRP, signal supposé être émis par le CCM, soit prédictif de l'ajustement comportemental en réaction à un *feedback*. Si l'ajustement comportemental n'est pas pertinent par rapport au *feedback* reçu, et surtout par rapport à l'amplitude du FRP, cela ne signifie pas forcément un problème du côté du CCM, mais cela pourrait être le résultat d'un mécanisme en aval du FRP, comme le recrutement du CPFl par exemple, ou bien l'intégration de l'information par le CPFl. Si tel est le cas, il semble pertinent de chercher des implémentations du contrôle en réaction au *feedback* dans les liens unissant le CCM aux autres régions, en particulier le CFPl. Une analyse que nous réaliserons consistera à chercher des liens entre le FRP en provenance du CCM et les oscillations frontales. L'implémentation

du contrôle pourrait être en effet réalisée via des modulations des oscillations en fonction de l'amplitude du FRP. Des liens entre potentiels évoqués et oscillations ont déjà été proposés, comme par exemple l'idée que les potentiels évoqués sont générés par un reset des oscillations cérébrales (Sauseng et al., 2007). Dans notre cas, nous pourrions observer des liens entre l'amplitude du FRP et des variations de puissance ou de la phase des oscillations, ou bien différents couplages des oscillations entre les régions préfrontales latérales, préfrontales médianes, sensorimotrices, ou pariétales. De manière intéressante, une prédiction serait aussi que ce genre de couplage est modulé lors de l'acquisition d'un *learning-set* au cours duquel le comportement de l'animal devient de plus en plus efficace.

Un autre intérêt de cette tâche consiste à comparer le signal dans les essais suivant un *Trap* et un *Shift*, afin de trouver des corrélats du moment où le singe parvient à déterminer si les règles ont réellement changé ou pas. Concernant l'analyse du *Shift*, le fait qu'il n'y ait que 3 ou 4 *Shifts* par session implique qu'il faille analyser plus de données que celles utilisées actuellement (pour l'instant, nous n'avons analysé que 12 sessions chez le singe D). Une des premières choses à faire dans les analyses à venir sera de commencer par intégrer plus de sessions.

Il a été montré que plusieurs pics évoqués sont en fait identifiables après un *feedback* (ici, nous n'avons décrit que celui correspondant à la *Feedback Related Negativity* FRN, observable à partir de 200-300ms post-*feedback*). Ces différents pics ont été associés à différents mécanismes cognitifs, dont le rôle n'est pas toujours très bien différencié de celui de la FRN, comme c'est le cas en particulier de la P300 (Sutton et al., 1965). La P300 est un signal observable entre 300 et 600 ms post-*feedback*, c'est-à-dire qu'elle peut, dans certains cas, chevaucher en partie le FRP, et elle a été associée à des rôles proches de la FRN, en particulier pour ce qui concerne le signalement des évènements nécessitant une adaptation comportementale (Nieuwenhuis et al., 2005). Il est possible que la P300 montre des modulations qui ne sont pas présentes dans le FRP. Une autre méthode consiste à décomposer le signal évoqué en ses composantes induites observables dans les différentes bandes oscillatoires (Tallon-Baudry et al., 1999). Les oscillations décuplent les possibilités d'encodage de l'information dans le signal, grâce à la modulation spécifique de l'amplitude, de

la phase, et des relations de synchronie entre les différentes bandes et les différentes régions. Dans la partie suivante, nous discutons des premiers résultats obtenus grâce aux analyses réalisées sur les modulations de la puissance des oscillations en fonction des facteurs de la tâche.

Analyse des oscillations corticales

Le travail sur les oscillations présenté dans cette thèse représente une première étape de description globale des données. Cela a permis d'identifier des facteurs importants qui modulent la puissance des différentes bandes, lors des différentes périodes de la tâche. Cela a aussi permis d'identifier des différences entre les bandes thêta et beta, qui ont été toutes les deux associées au contrôle cognitif dans la littérature.

Concernant l'encodage de la valence du *feedback*, si la plupart des études s'accordent pour dire que des augmentations de thêta sont associées aux *feedback* négatifs, les études reportent des résultats contradictoires concernant le beta. En effet, certaines études ont reporté une puissance beta plus élevée après les *feedback* positifs (Cohen et al., 2007; Hosseini and Holroyd, 2015; van de Vijver et al., 2011) et d'autres après les *feedback* négatifs (Cohen et al., 2009; Leicht et al., 2013). Une analyse approfondie de cette littérature a révélé que les auteurs ne parlent en fait pas du même "beta". Il semble important de discuter de ce manque de précision des articles quant à la bande de fréquence à laquelle les auteurs font référence sous les dénominations à partir des lettres grecques. En particulier, cela a probablement conduit au fait que le beta s'est vu attribuer une grande variété de fonctions (du moteur (Pfurtscheller and Lopes da Silva, 1999), aux mécanismes top-down dans les aires visuelles (Bosman et al., 2012), au contrôle cognitif (Stoll et al., 2015)...). Ce qu'on nomme "beta" semble en fait consister en un ensemble complexe de mécanismes (Engel and Fries, 2010). Par exemple, dans une étude de l'équipe, nous avons observé des oscillations beta dans les aires sensorimotrices chez deux singes, mais à des fréquences différentes entre les singes (Stoll et al., 2015). En particulier, plusieurs bandes "beta" étaient présentes chez les deux animaux (2 bandes dans chaque animal), et que les deux bandes dans un même animal n'étaient pas forcément modulées de la même façon en fonction des facteurs de la tâche. Pour en

revenir aux études mentionnant différents beta, il semble que celles montrant un lien entre fort beta et *feedback* positif font en fait référence à du «haut beta", c'est-à-dire à une bande au-dessus de 25Hz (Cunillera et al., 2012; HajiHosseini and Holroyd, 2015; Hosseini and Holroyd, 2015; Leicht et al., 2013; van de Vijver et al., 2011) (à l'exception de (HajiHosseini et al., 2012; Skoblenick et al., 2015)), alors que les auteurs ayant trouvé plus de beta après les *feedback* négatifs font référence à du "bas beta", c'est à dire à une bande entre 10 et 20Hz (Cavanagh et al., 2012; Cohen et al., 2009; Leicht et al., 2013; Rothé et al., 2011; Stoll et al., 2015). Dans notre étude, le beta fait référence à une bande entre 10 et 20Hz, dont la puissance est plus forte après les *feedback* négatifs, ce qui correspond aux données de la littérature du "bas beta". Ce bas beta est en fait rarement analysé dans les études citées ci-dessus, bien qu'il soit visible sur les représentations temps-fréquence et semble montrer des modulations inverses par rapport au haut beta (Cavanagh et al., 2012, 2010; Cunillera et al., 2012; van de Vijver et al., 2011). Ceci étant dit, certains points restent encore à éclaircir. Le lecteur doit probablement se demander pourquoi nous n'observons qu'une bande beta chez notre singe actuel. En réalité, une observation plus précise de la bande beta dans notre étude révèle une dynamique plus complexe que celle décrite jusqu'à présent (voir figure 3.A dans le papier sur les oscillations). En effet, il semble d'abord que ce que nous appelons beta soit en fait constitué de 2 sous-bandes, dont les pics seraient aux alentours de 10Hz et 15Hz. Ensuite, on peut voir que ces bandes évoluent de manière dynamique au sein de la fenêtre temporelle que nous avons sélectionnée. Des analyses prenant en compte cette complexité seraient donc pertinentes. Par exemple, nous pourrons analyser le profil de ces oscillations en essai par essai en détectant les crêtes des oscillations (*wavelet ridges*), ou en détectant les pics (*bursts*) des oscillations. Cependant, il est vrai que, chez ce singe, nous n'observons pas de bande beta au-dessus de 25Hz. Ce profil oscillatoire du beta évoque celui du singe R dans notre précédente étude (voir figure 3.B de l'annexe (Stoll et al., 2015)). Mais il faut garder à l'esprit qu'il existe une grande variabilité entre individus concernant les fréquences des bandes oscillatoires (Kilavik et al., 2013), et il a même été montré que cela été influencé par des facteurs génétiques (De Gennaro et al., 2008; van Pelt et al., 2012).

Des différences entre les bandes thêta et beta au *feedback* ont aussi été trouvées en corrélation avec la stratégie comportementale à venir (consistant à changer ou répéter la même réponse). Cette information était présente dans les oscillations beta, après un *feedback* positif seulement (puissance plus grande quand le singe répète sa réponse à l'essai d'après) ; et absente des oscillations thêta. Dans l'étude de van de Vijver et coll., la puissance des oscillations est prédictive de l'apprentissage (puissance plus grande quand l'essai suivant est correct) après les *feedback* positifs et négatifs pour le beta (18-24Hz), et après un *feedback* négatif seulement pour le thêta (4-8Hz) (van de Vijver et al., 2011). Malgré les différences, et notamment l'absence d'effet en thêta dans nos données, nos résultats recoupent en partie ceux de cette étude dans le sens où une augmentation de la puissance beta après un *feedback* positif correspond à un bon maintien de la stratégie en cours. De manière intéressante, nous montrons aussi que les oscillations beta et thêta dans la période de décision au début d'un essai sont prédictives de la stratégie qui va être appliquée (avec toujours plus de puissance pour un maintien par rapport à un changement de réponse), mais seulement suite à un essai qui avait reçu un *feedback* négatif. On peut se demander si ces résultats montrent un décalage temporel d'encodage de la stratégie à appliquer après un *feedback* négatif ou positif. En effet, dans notre tâche, sachant qu'il existe 3 options pour chaque stimulus, un *feedback* positif indique que la bonne cible a été trouvée, mais un *feedback* négatif ne permet d'éliminer qu'une seule des 3 options. Ainsi, un *feedback* positif est immédiatement informatif concernant la stratégie à implémenter au prochain essai, et cette information peut potentiellement être encodée immédiatement après le *feedback*. Au contraire, savoir comment réagir à un *feedback* négatif est plus complexe, et nécessite probablement de se remémorer les choix précédents pour le stimulus en question. Ceci pourrait expliquer pourquoi, dans le cas des *feedback* négatifs, l'encodage de la stratégie est présent plus tard, en début de l'essai suivant. Ainsi, ces différentes modulations pourraient être le reflet d'un même mécanisme permettant la décision déclenché à des moments différents. Un moyen de tester cette proposition serait d'utiliser une tâche dans laquelle les *feedback* négatifs seraient aussi informatifs que les *feedback* positifs et de voir si les oscillations beta encodent la stratégie au même moment pour les deux types de

feedback.

Enfin, nos données permettent également la comparaison directe des oscillations thêta et des FRP, puisqu'un débat actuel cherche à comprendre s'ils représentent différents aspects du même mécanisme ou non (Hajihosseini and Holroyd, 2013; Luu et al., 2004). Munneke et collaborateurs, par exemple, ont comparé directement les deux types de signaux en essai par essai (Munneke et al., 2015) en les intégrant à des modèles pour prédire les performances de chaque essai. Ils ont montré que les deux mesures utilisées dans des modèles séparés permettaient de prédire avec la même précision les performances, mais qu'un modèle utilisant les deux mesures en même temps avait de meilleures performances. Cela suggère que les deux types de signaux reflètent à la fois des processus communs et distincts. Dans notre étude, les FRP et thêta encodent de manière significative la valence du *feedback*. Mais alors que l'effet est très fort dans les FRP, l'effet est réparti sur quelques électrodes pour le thêta. Cela suggère que la valence est plus fortement encodée dans les FRP que dans les oscillations thêta. Cela ne correspond pas aux résultats de l'étude de Munneke puisqu'ils trouvent des effets co-localisés pour les deux types de signaux. Un plus fort encodage de la valence dans les FRP par rapport au thêta a par contre été reporté en partie par Hosseini et Holroyd dans une étude dans laquelle les FRP reflétaient majoritairement la valence du *feedback* et le thêta, les probabilités (Hajihosseini and Holroyd, 2013). Les auteurs proposent que les FRP reflètent les signaux liés à l'erreur de prédiction alors que le thêta reflèterait la détection du degré de surprise par le CCM, ce que proposent aussi Cavanagh et coll. (Cavanagh et al., 2012). Dans nos données, au *feedback*, les oscillations thêta sont plus fortes en exploitation qu'en exploration après un *feedback* positif, ce qui ne soutient pas vraiment un rôle du thêta dans l'encodage de la surprise si on considère qu'un *feedback* positif est probablement plus surprenant en exploration (où il est rare) qu'en exploitation (où il est fréquent). Une autre modulation intéressante du thêta est observée pour les *feedback* négatifs en exploitation. La puissance thêta est en effet plus forte lorsqu'un de ces *feedback* négatifs est dû à un essai normal par rapport à quand il est dû à un essai *Trap*. Ici encore, ces résultats sont durs à interpréter dans le sens d'une augmentation de thêta pour signaler un plus grand degré de surprise puisque cela

voudrait dire que le *feedback* négatif normal est plus surprenant que celui reçu à cause d'un *Trap*. On pourrait plutôt penser qu'un *feedback* négatif reçu en exploitation à la suite d'un *Trap* devrait être très surprenant pour le singe, qui a fait une réponse correcte qu'il sait être correcte (puisqu'il est en exploitation), et qui reçoit pourtant un *feedback* négatif. Une explication serait alors que le *feedback* négatif normal est plus surprenant parce que le singe n'est en fait pas en exploitation mais en re-exploration, et que dans ce cas, il a des attentes opposées sur ses réponses. Mais une interprétation parcimonieuse serait que le thêta ne reflète pas la surprise.

Il est possible que ce contraste entre les phases ne donne pas de résultats très clairs parce notre tâche requiert en permanence un fort engagement du contrôle cognitif, que le singe soit en exploration, exploitation ou ré-exploration, et qu'il y a en permanence des évènements surprenants. Une autre façon d'étudier l'implication du contrôle cognitif dans ces réseaux serait de comparer ces oscillations dans les problèmes réussis et dans les problèmes où le singe ne réussit jamais à apprendre les associations. En étudiant la puissance des oscillations au moment du délai dans notre tâche (c'est-à-dire la période où le singe enclenche le levier avant la période de décision), nous pourrons peut-être mettre en évidence des modulations de puissance des oscillations beta qui représenteraient le niveau de contrôle cognitif engagé dans l'essai. Cela permettrait de conforter les résultats que nous avons obtenus avec une tâche déterministe (Stoll et al., 2015). Une autre façon de séparer les essais avec différents niveaux de contrôle serait de les classer en fonction des temps de réaction, comme dans l'étude de Phillips et coll. (Phillips et al., 2014).

Un dernier point global sur ces données concerne les différences que nous pouvons observer entre les aires frontales et pariétales. Une observation détaillée des cartographies des effets montrent que les deux structures ne sont pas tout le temps modulées de la même façon. Pour l'instant, une analyse simple a consisté à comparer les latences des effets entre les structures, mais cela n'a pas permis de révéler des différences majeures. Pourtant, le cortex pariétal joue probablement un rôle très important dans le réseau de l'adaptation, même s'il a été moins étudié que le cortex préfrontal. Plusieurs études en témoignent. Tout d'abord, un reflet de l'activité du cortex pariétal est la P300, supposée émerger de la région temporo-pariétale, dont

nous avons évoqué le rôle dans la détection du besoin d'adaptation et dans les processus attentionnels (Polich, 2007). Ensuite, le pariétal émerge quasi-systématiquement dans les expériences en IRMf cherchant à mettre en évidence les régions clés soutenant le contrôle cognitif, et fait par exemple partie des structures impliquées dans le système de *task-set* proposé par Dosenbach et coll. (Dosenbach et al., 2006). Enfin, un réseau fronto-pariétal dont la synchronie reflète le niveau de contrôle engagé dans la tâche a été mis en évidence à plusieurs reprises (Cooper et al., 2015; Phillips et al., 2014; Szczepanski and Knight, 2014). Des analyses temporelles plus fines, et surtout des analyses de la synchronie entre ces régions permettront peut-être de révéler les spécificités de cette région et sa dynamique au cœur du réseau incluant le CCM et le CPFl.

8.4 Perspectives

Manipulations comportementales

Une expérience qu'il faut mener à son terme est l'entrainement des deux singes apprenant la tâche de *reversal* indiqué, dans un environnement déterministe, afin d'étudier leurs performances lors du transfert à la tâche finale (de *reversal* non indiqué dans un environnement probabiliste). Comme nous l'avons détaillé, cette expérience pourrait servir de contrôle à notre étude comportementale pour nous permettre d'affirmer que l'excellent transfert que nous avons observé est bien dû à l'apprentissage dans l'environnement stochastique.

Un autre test très important à réaliser est celui utilisé par Collins et Koechlin pour tester l'utilisation du *task-set* chez leurs sujets (Collins and Koechlin, 2012). Un comportement d'utilisation d'un *task-set* se manifeste par de meilleurs performances au cours des sessions 'récurrentes', au cours desquelles certains groupements d'associations sont répétés par rapport aux sessions dites 'ouvertes' au cours desquelles les différents groupements d'association ne sont présentés qu'une seule fois. La difficulté est qu'actuellement, les singes ne résolvent que 5 problèmes par sessions, ce qui va poser problème pour que les singes voient plusieurs fois le même "*task-set*". Une façon d'augmenter le nombre de "*task-sets*" réalisés par jour sera de réduire le

nombre d'essais entre deux *Shifts*, ce qui va augmenter la difficulté de la tâche. Il faudra donc tester que le singe arrive toujours à réaliser la tâche avec de bonnes performances.

Ensuite, il serait intéressant de modifier certains paramètres de la tâche réalisée en ce moment avec les enregistrements, afin de tester quelques hypothèses. Par exemple, on pourrait étudier les changements dans les patrons oscillatoires dans une version de la tâche où les changements sont de nouveau signalés. Si les singes arrivent à prendre en compte ce signal de changement (ce qui devrait être le cas, puisqu'ils le faisaient dans la tâche d'entrainement), cela devrait peut-être modifier la balance exploration/exploitation/ré-exploration. En particulier, nous devrions observer moins de ré-exploration et des phases d'exploitation plus pures. En fixant un nombre assez grand d'essais entre deux *Shifts*, cela permettra d'obtenir une phase où le singe est vraiment parfaitement certain des bonnes réponses. A supposer que les singes soient assez sûrs des bonnes réponses en exploitation pour ignorer les *Trap* (ce qui n'est pas assuré, vu leur comportement obtenus jusque-là), cela nous permettra d'obtenir des *feedback* surprenants ne conduisant pas à une adaptation comportementale, et nous pourrons tester si la FRP reflète la surprise ou le besoin d'adaptation.

Analyses du signal électrophysiologique

Comme nous l'avons vu, une des raisons majeures d'étudier les oscillations est de mettre en évidence la dynamique du réseau entre les différentes régions. Cette dynamique entre régions frontales, pariétales et motrices en fonction des demandes de la tâche a commencé à être mise en évidence (Voloh et al., 2015; Voytek et al., 2015). Des analyses de connectivité seront donc au cœur des analyses à venir, en commençant par des études de cohérence de phase entre les différentes régions et les différentes bandes. Une prédiction serait, par exemple, que la connectivité fonctionnelle intra-frontale, ou entre les aires frontale et pariétale serait la plus forte au cours des problèmes réussis par rapport aux problèmes avec de mauvaises performances. Il serait également intéressant de regarder si des changements de connectivité dans le réseau seraient prédictifs du moment où le singe distingue que le *feedback* inattendu qu'il vient de recevoir était un *Trap* ou un *Shift*. Ces deux évènements devraient

provoquer des changements puisque le premier nécessite de continuer avec la stratégie en cours, alors que le deuxième nécessite de repartir à la recherche des nouvelles règles, ce que montre l'analyse du comportement des singes (voir papier comportement). Pour ce projet, nous sommes en collaboration avec D.Marinazzo, un physicien spécialisé en particulier dans le développement d'analyses permettant de mettre en évidence des relations de causalité dans les signaux temporels continus, telles que les analyses de causalité de Granger ou DCM (*Dynamic Causal Modeling*). Ces outils nous permettront d'aller plus loin que les analyses de synchronie en permettant d'établir un lien "causal" entre les activités de nos régions. Des analyses plus simples de comparaison d'apparition de la significativité des effets ou de décodage d'effet (grâce à une toolbox réalisée par Frederic Stoll) pourront également être réalisées.

Couplage électrophysiologie et pharmacologie

Une manipulation cruciale est la réalisation des injections pharmacologiques dans le CCM car la combinaison du comportement, de l'électrophysiologie et de la pharmacologie peuvent apporter des réponses causales à plusieurs questions clés de la littérature. La première question consiste en la localisation de la source émettrice du FRP. Les techniques de localisation de source les plus évoluées combinent actuellement EGG et IRM et semblent confirmer que la source du FRP est bien le CCM (Hauser et al., 2014). Seulement, la seule façon d'être certain que le CCM joue un rôle *causal* dans l'émission de ce signal passe par une perturbation locale du CCM et par l'observation des conséquences de ces perturbations sur le FRP. Nous commencerons donc d'abord nos manipulations pharmacologiques par des injections transitoires de muscimol, un inhibiteur du GABA, à plusieurs niveaux du CCM afin de l'inactiver complètement. Si le CCM est causalement impliqué dans l'émission de ce potentiel, ces injections devraient conduire à sa disparation des enregistrements. Par contre, si le FRP est toujours présent ou diminué lors de l'inactivation, il sera difficile de trancher entre la possibilité d'une inactivation incomplète de la structure et le fait que le CCM ne soit pas l'unique responsable du signal. Le deuxième point critique concerne l'implication de la dopamine dans le rôle du CCM, dont nous pouvons observer les effets à plusieurs niveaux. Un lien entre l'émission du FRP par les neurones du CCM et les arrivées dopaminergiques dans le CCM en réaction à une

erreur de prédiction a été proposé et sert de socle à toute une théorie (Holroyd and Coles, 2002), mais cela n'a jamais été testé directement. Des injections localisées dans le CCM d'antagonistes dopaminergiques pendant la réalisation de la tâche permettront de tester directement l'influence de la dopamine sur les modulations du signal FRP. Enfin, un rôle des oscillations thêta dans la synchronisation du système dopaminergique avec des régions cibles, et notamment le cortex préfrontal, a été proposé (Fujisawa and Buzsáki, 2011). Ainsi, ces manipulations pharmacologiques nous permettront de tester directement cette proposition, ainsi que les conséquences sur la dynamique du réseau. Une implantation de la chambre d'injection sur un premier singe pilote est prévue en novembre.

Couplage acquisition d'un learning-set et électrophysiologie

Enfin, une perspective intéressante pour un projet de recherche futur serait de lier les deux aspects de cette thèse : c'est-à-dire d'étudier les corrélats électrophysiologiques de l'amélioration des processus de décision dans un contexte d'incertitude lors de l'établissement d'un *learning-set*. Pour réaliser ce projet, il faudrait implanter des singes dès le début de l'entrainement et suivre l'évolution des marqueurs évoqués et oscillatoires identifiés dans cette thèse au cours de l'acquisition d'un *learning-set*. En combinant les résultats des travaux de la littérature sur le *learning-set* et les données comportementales et électrophysiologiques acquises au cours de cette thèse, des prédictions testables peuvent être proposées. Tout d'abord, nous avons vu que l'acquisition d'un *learning-set* dépend de manière cruciale de l'intégrité du CPF dans son ensemble. En effet, des lésions séparées de ses différentes sous-régions n'induisent pas de déficits contrairement à une lésion entière du CPF (Wilson et al., 2010). Ainsi, des corrélats neurophysiologiques des effets de l'acquisition d'un *learning-set* seraient plus à chercher dans des modifications de la connectivité entre les régions du CPF que dans des modifications localisées d'activités (Browning et al., 2007). Dans notre expérience, cela pourrait consister en une augmentation progressive de la synchronie entre les aires préfrontale, fronto-médiane et pariétale, qui pourrait refléter une amélioration de l'efficacité du réseau. Pour faire des prédictions un peu plus spécifiques, nous avons également vu que l'acquisition d'un *learning-set* mène à une amélioration des performances qui ne correspond pas à une amélioration du simple processus

d'apprentissage mais à la mise en place d'une stratégie implémentée *en avance* d'un choix afin de rendre ce choix plus efficace (Murray and Gaffan, 2006; Wilson et al., 2010). Ainsi, si des changements électrophysiologiques sont à prédire au cours de l'acquisition du *learning-set*, ceux-ci ne devraient pas consister en des changements des processus simples d'apprentissage par renforcement (qui sont réactifs plutôt que prospectifs) comme la détection des *feedback*, mais plutôt des processus permettant d'agir de manière prospective. Les résultats de notre étude comportementale montrent que l'acquisition du *learning-set* dans un environnement incertain s'accompagne du développement d'une réponse aux *feedback* incertains, potentiellement liée à l'anticipation d'un *Shift*. Ainsi, une prédiction serait que nous observerions des changements liés au choix de la stratégie à appliquer après un *feedback* incertain, *Trap* ou *Shift*. Un premier candidat semble être la puissance des oscillations thêta au *feedback*, puisque nos résultats suggèrent qu'elle discrimine entre un *feedback* négatif normal et un *Trap* en exploitation. Ainsi, une prédiction pourrait être que cette réponse différentielle entre essais normaux et *Trap*, si elle est le reflet d'un processus prospectif, apparaisse progressivement au cours de l'acquisition du *learning-set*, en lien avec l'augmentation de cette réponse identifiée dans le comportement. En contraste, les modulations (des FRP, du beta ou du thêta) en fonction de la valence du *feedback*, si elles sont le reflet d'un processus réactif, devraient être immédiatement présentes et stables tout au long de l'acquisition du *learning-set*. Une étude récente a utilisé un paradigme de *task-switching* pour contraster le contrôle proactif (en préparation d'un changement de règles) par rapport au contrôle réactif (pour mettre à jour le *task-set* et contrer l'interférence entre deux cibles) et a montré que les deux types de contrôles impliquaient des modulations de l'activité oscillatoire en thêta mais au sein de différents réseaux fronto-pariétaux différents (Cooper et al., 2015). C'est précisément ces deux types de réseaux qui devraient être modulés différemment lors de l'acquisition d'un *learning-set*. C'est ce que permettrait de révéler une étude longitudinale étudiant l'évolution de la connectivité fronto-pariétale en thêta lors de l'acquisition d'un *learning-set*.

8.5 Conclusion

En conclusion, ce travail de thèse propose des bases pour lier deux domaines qui ne se sont pas encore vraiment rencontrés, alors qu'ils ont beaucoup à apporter l'un à l'autre : celui de la prise de décision dans les environnements complexes et celui du *learning-set*. En effet, nos résultats suggèrent que les mécanismes que nous utilisons pour apprendre de l'environnement semblent fortement dépendants de la façon dont ils ont été acquis lors du processus d'apprendre à apprendre, même si des recherches supplémentaires sont encore nécessaires afin de déterminer précisément dans quelles mesures cela semble être le cas.

De plus, ce travail contribue à avancer l'énorme travail de décryptage de l'implication des oscillations cérébrales en tant que mécanisme permettant la réalisation des processus cognitifs. Répondre à cette question constitue probablement le nouveau défi des neurosciences : comprendre vraiment le cerveau exige avant tout de comprendre son fonctionnement en tant que réseau.

Bibliographie

Acuna, Daniel E. and Schrater, Paul. Structure Learning in Human Sequential Decision-Making. *PLoS Computational Biology*, 6(12), 2010.

Adams, Christopher D. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, 34B :77–98, 1982.

Alexander, William H and Brown, Joshua W. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(September) :1338–1344, sep 2011.

Allport, Alan, Styles, Elizabeth a, and Hsieh, Shulan. Shifting Intentional Set : Exploring the Dynamic Control of Tasks. *Attention and Performance XV*, (August) :421–452, 1994.

Amiez, Céline and Petrides, Michael. Neuroimaging evidence of the anatomo-functional organization of the human cingulate motor areas. *Cerebral Cortex*, 24(3) :563–578, 2014.

Amiez, Céline, Joseph, Jean-paul P., and Procyk, Emmanuel. Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex*, 16(7) :1040–1055, 2006.

Apicella, P, Scarnati, E, Ljungberg, T, and Schultz, W. Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *Journal of neurophysiology*, 68 (3) :945–960, 1992.

Apicella, Paul. Leading tonically active neurons of the striatum from reward detection to context recognition. *Trends in Neurosciences*, 30(6) :299–306, 2007.

Apicella, Paul, Deffains, Marc, Ravel, Sabrina, and Legallet, Eric. Tonically active neurons in the striatum differentiate between delivery and omission of expected reward in a probabilistic task context. *European Journal of Neuroscience*, 30(3) :515–526, 2009.

Apicella, Paul, Ravel, Sabrina, Deffains, Marc, and Legallet, Eric. The role of striatal tonically active neurons in reward prediction error signaling during instrumental task performance. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(4) :1507–1515, 2011.

Asaad, W F, Rainer, G, and Miller, E K. Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6) :1399–1407, 1998.

Asaad, W F, Rainer, G, and Miller, E K. Task-specific neural activity in the primate prefrontal cortex. *Journal of neurophysiology*, 84(1) :451–459, 2000.

Aston-Jones, Gary and Cohen, Jonathan D. An integrative theory of locus coeruleus-norepinephrine function : adaptive gain and optimal performance. *Annual review of neuroscience*, 28 :403–450, 2005.

Azuar, C., Reyes, P., Slachevsky, a., Volle, E., Kinkingnehun, S., Kouneiher, F., Bravo, E., Dubois, B., Koechlin, E., and Levy, R. Testing the model of caudo-rostral organization of cognitive control in the human with frontal lesions. *NeuroImage*, 84 :1053–1060, 2014.

Badre, David and D'Esposito, Mark. Is the rostro-caudal axis of the frontal lobe hierarchical ? *Nature reviews. Neuroscience*, 10(9) :659–669, 2009.

Bailey, a M and Thomas, R K. The effects of nucleus basalis magnocellularis lesions in Long-Evans hooded rats on two learning set formation tasks, delayed matching-to-sample learning, and open-field activity. *Behavioral neuroscience*, 115(2) :328–340, 2001.

Balleine, Bernard W. and Dickinson, Anthony. Goal-directed instrumental action : Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5) :407–419, 1998.

Barone, P. and Joseph, J. P. Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, 78(3) :447–464, 1989.

Barron, Greg and Erev, Ido. Small Feedback-based Decisions and Their Limited Correspondence to Description-based Decisions. *Journal of Behavioral Decision Making*, 16(3) :215–233, 2003.

Bastos, Andre M, Vezoli, Julien, and Fries, Pascal. Communication through coherence with inter-areal delays. *Current Opinion in Neurobiology*, 31 :173–180, 2015.

Bateson, Melissa and Kacelnik, Alex. Risk-sensitive foraging : decision making in variable environments. In *Cognitive Ecology*, pages 297–340. 1998.

Baxter, Mark G, Gaffan, David, Kyriazis, Diana a, and Mitchell, Anna S. Orbital prefrontal cortex is required for object-in-place scene memory but not performance of a strategy implementation task. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(42) : 11327–11333, 2007.

Baxter, Mark G, Browning, Philip G F, and Mitchell, Anna S. Perseverative interference with object-in-place scene learning in rhesus monkeys with bilateral ablation of ventrolateral prefrontal cortex. *Learning & memory (Cold Spring Harbor, N.Y.)*, 15(3) :126–132, 2008.

Bechara, A, Damasio, a R, Damasio, H, and Anderson, S W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3) :7–15, 1994.

Behrens, Timothy E J, Woolrich, Mark W, Walton, Mark E, and Rushworth, Matthew F S. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9) :1214–1221, 2007.

Benchenane, Karim, Peyrache, Adrien, Khamassi, Mehdi, Tierney, Patrick L., Gioanni, Yves, Battaglia, Francesco P., and Wiener, Sidney I. Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning. *Neuron*, 66(6) :921–936, 2010.

Berlyne, D E. Uncertainty and conflict : a point of contact between information-theory and behavior-theory concepts. *Psychological review*, 64, Part 1(6) :329–339, 1957.

Biernaskie, Jay M., Walker, Steven C., and Gegear, Robert J. Bumblebees learn to forage like Bayesians. *The American Naturalist*, 174(3) :413–423, 2009.

Blaisdell, a P, Bristol, A S, Gunther, L M, and Miller, R R. Overshadowing and latent inhibition counteract each other : support for the comparator hypothesis. *Journal of experimental psychology. Animal behavior processes*, 24(3) :335–351, 1998.

Blundell, Pam, Hall, Geoffrey, and Killcross, Simon. Preserved Sensitivity to Outcome Value after Lesions of the Basolateral Amygdala. *The Journal of Neuroscience*, 23(20) :7702–7709, 2003.

Bogacz, Rafal, Brown, Eric, Moehlis, Jeff, Holmes, Philip, and Cohen, Jonathan D. The physics of optimal decision making : A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4) :700–765, 2006.

Bonini, Francesca, Burle, Boris, Liégeois-Chauvel, Catherine, Régis, Jean, Chauvel, Patrick, and Vidal, Franck. Action monitoring and medial frontal cortex : leading role of supplementary motor area. *Science (New York, N.Y.)*, 343(6173) :888–91, 2014.

Bosman, Conrado a., Schoffelen, Jan Mathijs, Brunet, Nicolas, Oostenveld, Robert, Bastos, Andre M., Womelsdorf, Thilo, Rubehn, Birthe, Stieglitz, Thomas, De Weerd, Peter, and Fries, Pascal. Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas. *Neuron*, 75(5) :875–888, 2012.

Bouret, Sebastien and Sara, Susan J. Network reset : A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, 28(11) :574–582, 2005.

Bouton, Mark E. Context and behavioral processes in extinction. *Learning & memory (Cold Spring Harbor, N.Y.)*, 11(5) :485–494, 2004.

Braun, Daniel a., Mehring, Carsten, and Wolpert, Daniel M. Structure learning in action. *Behavioural Brain Research*, 206(2) :157–165, 2010.

Braver, Todd S, Cohen, Jonathan D, and Barch, Deanna M. The role of prefrontal cortex in normal and disordered cognitive control : A cognitive neuroscience perspective. *Principles of Frontal Lobe Function*, pages 428–47, 2002.

Braver, Todd S, Reynolds, Jeremy R, and Donaldson, David I. Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron*, 39(4) :713–26, aug 2003.

Browning, P. G F, Easton, Alexander, Buckley, Mark J., and Gaffan, David. The role of prefrontal cortex in object-in-place learning in monkeys. *European Journal of Neuroscience*, 22(12) :3281–3291, 2005.

Browning, Philip G F and Gaffan, David. Prefrontal cortex function in the representation of temporally complex events. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(15) :3934–3940, 2008.

Browning, Philip G F, Easton, Alexander, and Gaffan, David. Frontal-temporal disconnection abolishes object discrimination learning set in macaque monkeys. *Cerebral Cortex*, 17(4) :859–864, apr 2007.

Brozoski, T J, Brown, R M, Rosvold, H E, and Goldman, P S. Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey. *Science (New York, N.Y.)*, 205 (4409) :929–932, 1979.

Buschman, Timothy J, Denovellis, Eric L, Diogo, Cinira, Bullock, Daniel, and Miller, Earl K. Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron*, 76(4) : 838–46, nov 2012.

Caracheo, Barak F., Emberly, Eldon, Hadizadeh, Shirin, Hyman, James M., and Seamans, Jeremy K. Abrupt changes in the patterns and complexity of anterior cingulate cortex activity when food is introduced into an environment. *Frontiers in Neuroscience*, 7(May) :1–14, 2013.

Caraco, Thomas. Energy budgets, risk and foraging preferences in dark-eyed juncos (Junco hyemalis). *Behavioral Ecology and Sociobiology*, 8(3) :213–217, 1981.

Cavanagh, James F, Frank, Michael J, Klein, Theresa J, and Allen, John J B. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*, 49(4) :3198–209, feb 2010.

Cavanagh, James F., Bismark, Andrew J., Frank, Michael J., and Allen, John J B. Larger error signals in major depression are associated with better avoidance learning. *Frontiers in Psychology*, 2(NOV) :1–6, 2011.

Cavanagh, James F., Figueroa, Christina M., Cohen, Michael X., and Frank, Michael J. Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral Cortex*, 22(11) :2575–2586, 2012.

Cavanagh, James F, Sanguinetti, Joseph L, Allen, John J B, Sherman, Scott J, and Frank, Michael J. The Subthalamic Nucleus contributes to post-error slowing. *Journal of Cognitive Neuroscience*, pages 1–8, 2014.

Chase, Henry W, Swainson, Rachel, Durham, Lucy, Benham, Laura, and Cools, Roshan. Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of cognitive neuroscience*, 23(4) :936–946, 2010.

Chen, M. Keith, Lakshminarayanan, Venkat, and Santos, Laurie R. How Basic Are Behavioral Biases ? Evidence from Capuchin Monkey Trading Behavior. *Journal of Political Economy*, 114 (3) :517–537, 2006.

Cisek, Paul. Integrated neural processes for defining potential actions and deciding between them : a computational model. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(38) :9761–9770, 2006.

Cisek, Paul. Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, 22(6) :927–936, 2012.

Cisek, Paul and Kalaska, John F. Neural correlates of reaching decisions in dorsal premotor cortex : Specification of multiple direction choices and final selection of action. *Neuron*, 45(5) :801–814, 2005.

Cisek, Paul and Kalaska, John F. Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33(March) :269–298, 2010.

Cohen, J D, Dunbar, K, and McClelland, J L. On the control of automatic processes : a parallel distributed processing account of the Stroop effect. *Psychological review*, 97(3) :332–61, 1990.

Cohen, Jonathan D, McClure, Samuel M, and Yu, Angela J. Should I stay or should I go ? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1481) :933–942, 2007.

Cohen, Michael X, Elger, Christian E, and Fell, Juergen. Oscillatory activity and phase-amplitude coupling in the human medial frontal cortex during decision making. *Journal of cognitive neuroscience*, 21(2) :390–402, 2009.

Collins, a. G. E., Brown, J. K., Gold, J. M., Waltz, J. a., and Frank, M. J. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *Journal of Neuroscience*, 34 (41) :13747–13756, 2014.

Collins, Anne and Koechlin, Etienne. Reasoning, learning, and creativity : Frontal lobe function and human decision-making. *PLoS Biology*, 10(3) :e1001293, 2012.

Collins, Anne G E and Frank, Michael J. Cognitive control over learning : creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1) :190–229, 2013.

Cooper, Patrick S., Wong, Aaron S.W., Fulham, W.Ross, Thienel, Renate, Mansfield, Elise, Michie, Patricia T., and Karayanidis, Frini. Theta frontoparietal connectivity associated with proactive and reactive cognitive control processes. *NeuroImage*, 108 :354–363, 2015.

Cos, Ignasi, Bélanger, Nicolas, and Cisek, Paul. The influence of predicted arm biomechanics on decision making. *Journal of neurophysiology*, 105(6) :3022–3033, 2011.

Courville, Aaron C., Daw, Nathaniel D., and Touretzky, David S. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7) :294–300, 2006.

Coutureau, Etienne and Killcross, Simon. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, 146(1-2) :167–174, 2003.

Coutureau, Etienne, Esclassan, Frederic, Di Scala, Georges, and Marchand, Alain R. The role of the rat medial prefrontal cortex in adapting to changes in instrumental contingency. *PLoS ONE*, 7(4), 2012.

Crone, Eveline a, Wendelken, Carter, Donohue, Sarah E, and Bunge, Silvia a. Neural evidence for dissociable components of task-switching. *Cerebral cortex (New York, N.Y. : 1991)*, 16(4) : 475–86, apr 2006.

Cunillera, Toni, Fuentemilla, Lluís, Periañez, Jose, Marco-Pallarès, Josep, Krämer, Ulrike M., Càmara, Estela, Münte, Thomas F., and Rodríguez-Fornells, Antoni. Brain oscillatory activity associated with task switching and feedback processing. *Cognitive, Affective, & Behavioral Neuroscience*, 12(1) :16–33, 2012.

Dalley, J W, McGaughy, J, O'Connell, M T, Cardinal, R N, Levita, L, and Robbins, T W. Distinct changes in cortical acetylcholine and noradrenaline efflux during contingent and noncontingent performance of a visual attentional task. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(13) :4908–4914, 2001.

Dalton, Gemma L, Phillips, Anthony G, and Floresco, Stan B. Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(13) :4618–26, 2014.

Daw, Nathaniel D, Niv, Yael, and Dayan, Peter. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12) :1704–1711, 2005.

De Gennaro, Luigi, Marzano, Cristina, Fratello, Fabiana, Moroni, Fabio, Pellicciari, Maria Concetta, Ferlazzo, Fabio, Costa, Stefania, Couyoumdjian, Alessandro, Curcio, Giuseppe, Sforza, Emilia, Malafosse, Alain, Finelli, Luca a, Pasqualetti, Patrizio, Ferrara, Michele, Bertini, Mario, and Rossini, Paolo Maria. The electroencephalographic fingerprint of sleep is genetically determined : a twin study. *Annals of neurology*, 64(4) :455–60, 2008.

Dehaene, Stanislas, Posner, Michael I., and Tucker, Don M. Localization of a Neural System for Error Detection and Compensation, sep 1994.

Dehaene, Stanislas, Artiges, Eric, Naccache, Lionel, Martelli, Catherine, Viard, Armelle, Schürhoff, Franck, Recasens, Christophe, Martinot, Marie Laure Paillère, Leboyer, Marion, and Martinot, Jean-Luc. Conscious and subliminal conflicts in normal subjects and patients with schizophrenia : the role of the anterior cingulate. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23) :13722–13727, 2003.

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. J., and Stephan, K. E. Striatal Prediction Error Modulates Cortical Coupling. *Journal of Neuroscience*, 30(9) :3210–3219, 2010.

di Pellegrino, G and Wise, S P. A neurophysiological comparison of three distinct regions of the primate frontal lobe. *Brain : a journal of neurology*, 114 (Pt 2 :951–978, 1991.

Dickinson, A. Actions and Habits : The Development of Behavioural Autonomy, 1985.

Donoso, Maël, Collins, Anne G E, and Koechlin, Etienne. Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science (New York, N.Y.)*, 344(6191) :1481–6, 2014.

Dosenbach, Nico U F, Visscher, Kristina M, Palmer, Erica D, Miezin, Francis M, Wenger, Kristin K, Kang, Hyunseon C, Burgund, E Darcy, Grimes, Ansley L, Schlaggar, Bradley L, and Petersen, Steven E. A core system for the implementation of task sets. *Neuron*, 50(5) :799–812, jun 2006.

Duncan, Carl P. Description of learning to learn in human subjects. *Am J Psychol*, 73 :108–114, 1960.

Durstewitz, Daniel, Vittoz, Nicole M., Floresco, Stan B., and Seamans, Jeremy K. Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. *Neuron*, 66(3) :438–448, 2010.

E, Bell D and Bell, D. E. Regret in Decision Making under Uncertainty. *Operations Research*, 30 (5) :961–981, 1982.

Ellsberg, Daniel. Risk, Ambiguity, and the Savage Axioms, 1961.

Engel, Andreas K and Fries, Pascal. Beta-band oscillations–signalling the status quo ? *Current opinion in neurobiology*, 20(2) :156–65, apr 2010.

Espinosa-Parrilla, Juan-Francisco, Baunez, Christelle, and Apicella, Paul. Modulation of neuronal activity by reward identity in the monkey subthalamic nucleus. *European Journal of Neuroscience*, pages n/a–n/a, 2015.

Falkenstein, M, Hohnsbein, J, Hoormann, J, and Blanke, L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, 78(6) :447–455, 1991.

Ferdinand, N. K., Mecklinger, a., Kray, J., and Gehring, W. J. The Processing of Unexpected Positive Response Outcomes in the Mediofrontal Cortex. *Journal of Neuroscience*, 32(35) : 12087–12092, 2012.

Fiorillo, Christopher D, Tobler, Philippe N, and Schultz, Wolfram. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, 299(5614) :1898–1902, 2003.

Fischer, AdrianG and Ullsperger, Markus. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron*, 79(6) :1243–1255, 2013.

Flavell, Steven W., Pokala, Navin, Macosko, Evan Z., Albrecht, Dirk R., Larsch, Johannes, and Bargmann, Cornelia I. Serotonin and the neuropeptide PDF initiate and extend opposing behavioral states in C. Elegans. *Cell*, 154(5) :1023–1035, 2013.

Forstmann, Birte U, Brass, Marcel, Koch, Iring, and von Cramon, D Yves. Voluntary selection of task sets revealed by functional magnetic resonance imaging. *Journal of cognitive neuroscience*, 18(3) :388–398, 2006.

Frank, Michael J., Woroch, Brion S., and Curran, Tim. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, 47(4) :495–501, 2005.

Fujisawa, Shigeyoshi and Buzsáki, György. A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. *Neuron*, 72(1) :153–165, 2011.

Funahashi, S, Bruce, Charles J., and Goldman-rakic, Patricia S. Neuronal activity related to saccadic eye movements in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, 65 (6) :1464–1483, 1991.

Fuster, J M and Jervey, J P. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science (New York, N.Y.)*, 212(4497) :952–955, 1981.

Fuster, Joaquín M. The prefrontal cortex - An update : Time is of the essence, 2001.

Gallagher, M, McMahan, R W, and Schoenbaum, G. Orbitofrontal cortex and representation of incentive value in associative learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(15) :6610–6614, 1999.

Gallistel, C R and Gibbon, J. Time, rate, and conditioning. *Psychological review*, 107(2) :289–344, 2000.

Gallistel, C R, Mark, T a, King, a P, and Latham, P E. The rat approximates an ideal detector of changes in rates of reward : implications for the law of effect. *Journal of experimental psychology. Animal behavior processes*, 27(4) :354–372, 2001.

Gallistel, C R, Krishan, Monika, Liu, Ye, Miller, Reilly, and Latham, Peter E. The perception of probability. *Psychological review*, 121(1) :96–123, 2014.

Garcin, Béatrice, Volle, Emmanuelle, Dubois, Bruno, and Levy, Richard. Similar or different ? the role of the ventrolateral prefrontal cortex in similarity detection. *PLoS ONE*, 7(3), 2012.

Gehring, William J, Goss, Brian, and Coles, Michael G H. A neural system for error detection and compensation. *Psychological Science*, 4 :385–390, 1993.

Genovesio, Aldo, Brasted, Peter J., Mitz, Andrew R., and Wise, Steven P. Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47(2) :307–320, 2005.

Gershman, Samuel J, Blei, David M, and Niv, Yael. Context, learning, and extinction. *Psychological review*, 117(1) :197–209, 2010.

Gilbert, Sam J and Shallice, Tim. Task switching : a PDP model. *Cognitive psychology*, 44(3) : 297–337, 2002.

Gittins, Jc. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2) :148–177, 1979.

Haber, Suzanne N and Behrens, Timothy E J. Review The Neural Network Underlying Incentive-Based Learning : Implications for Interpreting Circuit Disruptions in Psychiatric Disorders. *Neuron*, 83(5) :1019–1039, 2014.

Hajihosseini, Azadeh and Holroyd, Clay B. Frontal midline theta and N200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology*, 50 (6) :550–562, 2013.

HajiHosseini, Azadeh and Holroyd, Clay B. Sensitivity of frontal beta oscillations to reward valence but not probability. *Neuroscience Letters*, 602 :99–103, 2015.

HajiHosseini, Azadeh, Rodríguez-Fornells, Antoni, and Marco-Pallarés, Josep. The role of beta-gamma oscillations in unexpected rewards processing. *NeuroImage*, 60(3) :1678–1685, 2012.

Hall, G and Pearce, J M. Latent inhibition of a CS during CS-US pairings. *Journal of experimental psychology. Animal behavior processes*, 5(1) :31–42, 1979.

Hampton, Alan N, Bossaerts, Peter, and O'Doherty, John P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(32) :8360–8367, 2006.

Harlow, H F. The formation of learning sets. *Psychological review*, 56(1) :51–65, jan 1949.

Harlow, H F. Learning set and error factor theory. *Psychology : A study of a science*, 2 :492–537, 1959.

Hauser, Tobias U., Iannaccone, Reto, Stämpfli, Philipp, Drechsler, Renate, Brandeis, Daniel, Walitza, Susanne, and Brem, Silvia. The feedback-related negativity (FRN) revisited : New insights into the localization, meaning and network organization. *NeuroImage*, 84 :159–168, 2014.

Hayden, Benjamin Y, Heilbronner, Sarah R, Pearson, John M, and Platt, Michael L. Surprise signals in anterior cingulate cortex : neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(11) :4178–87, mar 2011.

Heilbronner, Sarah R. and Hayden, Benjamin Y. The description-experience gap in risky choice. *Psychon Bull Rev*, 2015.

Heilbronner, Sarah R, Rosati, Alexandra G, Stevens, Jeffrey R, Hare, Brian, and Hauser, Marc D. A fruit in the hand or two in the bush ? Divergent risk preferences in chimpanzees and bonobos. *Biology letters*, 4(3) :246–249, 2008.

Herrnstein, R J. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, 4 :267–272, 1961.

Hertwig, Ralph, Barron, Greg, Weber, Elke U., and Erev, Ido. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8) :534–539, 2004.

Hikosaka, K and Watanabe, M. Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral cortex (New York, N.Y. : 1991)*, 10(3) :263–271, 2000.

Holroyd, Clay B. and Coles, Michael G.H. The neural basis of human error processing : Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4) : 679–709, 2002.

Holroyd, Clay B. and Krigolson, Olave E. Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, 44(6) :913–917, 2007.

Hoshi, E, Shima, K, and Tanji, J. Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of neurophysiology*, 83(4) :2355–2373, 2000.

Hosseini, Azadeh Haji and Holroyd, Clay B. Reward feedback stimuli elicit high-beta EEG oscillations in human dorsolateral prefrontal cortex. *Scientific Reports*, 5(April) :13021, 2015.

Hsu, Ming, Bhatt, Meghana, Adolphs, Ralph, Tranel, Daniel, and Camerer, Colin F. Neural systems responding to degrees of uncertainty in human decision-making. *Science (New York, N.Y.)*, 310 (5754) :1680–1683, 2005.

Huettel, Scott a, Song, Allen W, and McCarthy, Gregory. Decisions under uncertainty : probabilistic context influences activation of prefrontal and parietal cortices. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(13) :3304–3311, 2005.

Huettel, Scott a., Stowe, C. Jill, Gordon, Evan M., Warner, Brent T., and Platt, Michael L. Neural signatures of economic preferences for risk and ambiguity. *Neuron*, 49(5) :765–775, 2006.

Isoda, Masaki and Hikosaka, Okihide. Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience*, 10(2) :240–248, jan 2007.

Izquierdo, a and Murray, E a. Opposing effects of amygdala and orbital prefrontal cortex lesions on the extinction of instrumental responding in macaque monkeys. *Eur J Neurosci*, 22(9) : 2341–2346, 2005.

Izquierdo, Alicia, Suda, Robin K, and Murray, Elisabeth a. 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) :7540–7548, 2004.

Jang, a. I., Costa, V. D., Rudebeck, P. H., Chudasama, Y., Murray, E. a., and Averbeck, B. B. The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals. *Journal of Neuroscience*, 35(33) :11751–11760, 2015.

Jensen, Robert. Behaviorism, Latent Learning, and Cognitive Maps : Needed Revisions in Introductory Psychology Textbooks. *The Behavior Analyst*, 29(2) :187–209, 2006.

Ji, Jinzhao and Maren, Stephen. Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. *Learning & Memory*, 12(3) :270–276, 2005.

Johnston, Kevin, Levin, Helen M., Koval, Michael J., and Everling, Stefan. Top-Down Control-Signal Dynamics in Anterior Cingulate and Prefrontal Cortex Neurons following Task Switching. *Neuron*, 53(3) :453–462, 2007.

Kahneman, Daniel. Maps of Bounded Rationality : Psychology for Behavioral Economics. 93 (December 2003) :1449–1475, 2003.

Kahneman, Daniel and Tversky, Amos. Prospect Theory : An Analysis of Decision under Risk. *Econometrica*, 47(2) :263–292, 1979.

Kakade, Sham and Dayan, Peter. Dopamine : Generalization and bonuses. *Neural Networks*, 15 (4-6) :549–559, 2002.

Kamigaki, Tsukasa, Fukushima, Tetsuya, Tamura, Keita, and Miyashita, Yasushi. Neurodynamics of Cognitive Set Shifting in Monkey Frontal Cortex and Its Causal Impact on Behavioral Flexibility. *Journal of Cognitive Neuroscience*, 24(11) :2171–2185, 2012.

Karlsson, M. P., Tervo, D. G. R., and Karpova, A. Y. Network Resets in Medial Prefrontal Cortex Mark the Onset of Behavioral Uncertainty. *Science*, 338(6103) :135–139, 2012.

Kemp, Charles, Perfors, Amy, and Tenenbaum, Joshua B. Learning domain structures. *Proceedings of the 26th annual conference of the Cognitive Science Society*, pages 672–677, 2004.

Kemp, Charles, Goodman, Noah D., and Tenenbaum, Joshua B. Learning to Learn Causal Models. *Cognitive Science*, 34(7) :1185–1243, 2010.

Kennerley, Steven W, Dahmubed, Aspandiar F, Lara, Antonio H, and Wallis, Jonathan D. Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of cognitive neuroscience*, 21(6) :1162–1178, 2009.

Kennerley, Steven W S.W., Walton, Mark E M.E., Behrens, Timothy E J T.E.J., Buckley, M.J. Mark J, and Rushworth, M.F.S. Matthew F S. Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9(7) :940–947, jul 2006.

Kepecs, Adam, Uchida, Naoshige, Zariwala, Hatim a, and Mainen, Zachary F. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210) :227–231, 2008.

Khamassi, Mehdi, Enel, Pierre, Dominey, Peter Ford, and Procyk, Emmanuel. *Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters*, volume 202. Elsevier B.V., 1 edition, 2013.

Kiani, Roozbeh, Hanks, Timothy D, and Shadlen, Michael N. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(12) :3017–3029, 2008.

Kilavik, Bjørg Elisabeth, Zaepffel, Manuel, Brovelli, Andrea, MacKay, William a, and Riehle, Alexa. The ups and downs of β oscillations in sensorimotor cortex. *Experimental neurology*, 245 :15–26, 2013.

Killcross, Simon and Coutureau, Etienne. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13(4) :400–408, 2003.

Kimberg, Daniel Y., Aguirre, Geoffrey K., and D'Esposito, Mark. Modulation of task-related neural activity in task-switching : An fMRI study. *Cognitive Brain Research*, 10(1-2) :189–196, 2000.

Klaes, Christian, Westendorff, Stephanie, Chakrabarti, Shubhodeep, and Gail, Alexander. Choosing Goals, Not Rules : Deciding among Rule-Based Action Plans. *Neuron*, 70(3) :536–548, 2011.

Knight, Frank. Risk, Uncertainty, and Profit. *Hart Schaffner Marx prize essays*, XXXI :1–173, 1921.

Knight, R T. Decreased response to novel stimuli after prefrontal lesions in man., 1984.

Kobayashi, Shunsuke, Pinto de Carvalho, Ofelia, and Schultz, Wolfram. Adaptation of reward sensitivity in orbitofrontal neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(2) :534–544, 2010.

Koechlin, Etienne, Ody, Chrystèle, and Kouneiher, Frédérique. The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, 302(5648) :1181–5, nov 2003.

Koo, Minjung and Fishbach, Ayelet. Dynamics of Self-Regulation : How (Un)accomplished Goal Actions Affect Motivation. *Journal of Personality and Social Psychology*, 94(1) :1–5, 2008.

Koronakos, Chris and Arnold, William J. The Formation Of Learning Sets In Rats. 1957.

Landman, Janet. Regret : the persistence of the possible. *Philosophical Quarterly*, 47(188) :0, 1997.

Landmann, C., Dehaene, S., Pappata, S., Jobert, a., Bottlaender, M., Roumenov, D., and Le Bihan, D. Dynamics of Prefrontal and Cingulate Activity during a Reward-Based Logical Deduction Task. *Cerebral Cortex*, 17(4) :749–759, 2006.

Leicht, Gregor, Troschütz, Stefan, Andreou, Christina, Karamatskos, Evangelos, Ertl, Matthias, Naber, Dieter, and Mulert, Christoph. Relationship between Oscillatory Neuronal Activity during Reward Processing and Trait Impulsivity and Sensation Seeking. *PLoS ONE*, 8(12) :e83414, 2013.

Lejuez, C W, Read, Jennifer P, Kahler, Christopher W, Richards, Jerry B, Ramsey, Susan E, Stuart, Gregory L, Strong, David R, and Brown, Richard a. Evaluation of a behavioral measure of risk taking : the Balloon Analogue Risk Task (BART). *Journal of experimental psychology. Applied*, 8(2) :75–84, 2002.

Levy, R and Goldman-Rakic, P S. Association of storage and processing functions in the dorsolateral prefrontal cortex of the nonhuman primate. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(12) :5149–5158, 1999.

Levy, Richard and Dubois, Bruno. Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits. *Cerebral Cortex*, 16(7) :916–928, 2006.

Lhermitte, F. 'Utilization behaviour' and its relation to lesions of the frontal lobes. *Brain : a journal of neurology*, 106 (Pt 2) :237–255, 1983.

Lisman, John E. and Jensen, Ole. The Theta-Gamma Neural Code. *Neuron*, 77(6) :1002–1016, 2013.

Liston, Conor, Matalon, Shanna, Hare, Todd a., Davidson, Matthew C., and Casey, B. J. Anterior Cingulate and Posterior Parietal Cortices Are Sensitive to Dissociable Forms of Conflict in a Task-Switching Paradigm. *Neuron*, 50(4) :643–653, 2006.

Loomes, Graham and Sugden, Robert. Regret Theory : An Alternative Theory of Rational Choice Under Uncertainty. *Economic journal*, 92(368) :805–824, 1982.

Louie, Kenway and Glimcher, Paul W. Separating value from choice : delay discounting activity in the lateral intraparietal area. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(16) :5498–5507, 2010.

Luft, Caroline Di Bernardi, Nolte, Guido, and Bhattacharya, Joydeep. High-learners present larger mid-frontal theta power and connectivity in response to incorrect performance feedback. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(5) :2029–2038, 2013.

Luksys, Gediminas, Gerstner, Wulfram, and Sandi, Carmen. Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. *Nature neuroscience*, 12(9) : 1180–1186, 2009.

Luu, Phan, Tucker, Don M, and Makeig, Scott. Frontal midline theta and the error-related negativity : neurophysiological mechanisms of action regulation. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 115(8) :1821–35, aug 2004.

MacLeod, C M. Half a century of research on the Stroop effect : an integrative review. *Psychological bulletin*, 109(2) :163–203, 1991.

Markov, Nikola T, Ercsey-Ravasz, Mária, Van Essen, David C, Knoblauch, Kenneth, Toroczkai, Zoltán, and Kennedy, Henry. Cortical high-density counterstream architectures. *Science (New York, N.Y.)*, 342(6158) :1238406, 2013.

Mars, Rb, De Bruijn, Era, Hulstijn, W, Miltner, Whr, and Coles, Mgh. What if I told you : "You were wrong" ? Brain potentials and behavioral adjustments elicited by feedback in a time-estimation task. pages 129–134, 2004.

Martino, Benedetto De, Kumaran, Dharshan, Seymour, Ben, and Dolan, Raymond J. UKPMC Funders Group Frames , Biases , and Rational Decision-Making in the Human Brain. 313(5787) : 684–687, 2009.

Matsumoto, Kenji, Suzuki, Wataru, and Tanaka, Keiji. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science (New York, N.Y.)*, 301(5630) :229–232, 2003.

Matsumoto, Madoka, Matsumoto, Kenji, Abe, Hiroshi, and Tanaka, Keiji. Medial prefrontal cell activity signaling prediction errors of action values. *Nature neuroscience*, 10(5) :647–656, 2007.

McCoy, Allison N and Platt, Michael L. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature neuroscience*, 8(9) :1220–1227, 2005.

McLinn, Colleen M. and Stephens, David W. What makes information valuable : signal reliability and environmental uncertainty. *Animal Behaviour*, 71(5) :1119–1129, 2006.

McPeek, Robert M and Keller, Edward L. Deficits in saccade target selection after inactivation of superior colliculus. *Nature neuroscience*, 7(7) :757–763, 2004.

Mehta, Rick and Williams, Douglas a. Elemental and configural processing of novel cues in deterministic and probabilistic tasks. *Learning and Motivation*, 33(4) :456–484, 2002.

Michelet, Thomas, Duncan, Gary H, and Cisek, Paul. Response competition in the primary motor cortex : corticospinal excitability reflects response replacement during simple decisions. *Journal of neurophysiology*, 104(1) :119–127, 2010.

Micheli, Cristiano, Kaping, Daniel, and Westendorff, Stephanie. Inferior-frontal cortex phase synchronizes with the temporal – parietal junction prior to successful change detection. (August), 2015.

Miller, E K, Erickson, C a, and Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16) :5154–5167, 1996.

Miller, Earl K and Cohen, Jonathan D. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24 :167–202, 2001.

Miltner, Wolfgang H. R., Braun, Christoph H., and Coles, Michael G. H. Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task : Evidence for a "Generic" Neural System for Error Detection, 1997.

Monsell, Stephen. Task switching. *Trends in Cognitive Sciences*, 7(3) :134–140, mar 2003.

Munneke, Gert-Jan, Nap, Tanja S., Schippers, Eveline E., and Cohen, Michael X. A statistical comparison of EEG time- and time–frequency domain representations of error processing. *Brain Research*, 1618 :222–230, 2015.

Murray, Elisabeth a and Gaffan, David. Prospective memory in the formation of learning sets by rhesus monkeys (Macaca mulatta). *Journal of experimental psychology. Animal behavior processes*, 32(1) :87–90, 2006.

Narayanan, Nandakumar S, Cavanagh, James F, Frank, Michael J, and Laubach, Mark. Common medial frontal mechanisms of adaptive control in humans and rodents. *Nature Neuroscience*, 16 (12) :1–10, 2013.

Nassar, Matthew R, Wilson, Robert C, Heasly, Benjamin, and Gold, Joshua I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(37) :12366–12378, 2010.

Nassar, Matthew R, Rumsey, Katherine M, Wilson, Robert C, Parikh, Kinjan, Heasly, Benjamin, and Gold, Joshua I. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, 15(7) :1040–1046, 2012.

Nieuwenhuis, Sander, Aston-Jones, Gary, and Cohen, Jonathan D. Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, 131(4) :510–532, 2005.

Ochoki, Miller, and Cotter. Discrimination learning in children, 1975.

Olds, James and Milner, Peter. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, 47 : 419–427, 1954.

Oliveira, Flavio T P, McDonald, John J, and Goodman, David. Performance monitoring in the anterior cingulate is not all error related : expectancy deviation and the representation of action-outcome associations. *Journal of cognitive neuroscience*, 19(12) :1994–2004, 2007.

Ott, Torben, Jacob, Simon Nikolas, and Nieder, Andreas. Dopamine Receptors Differentially Enhance Rule Coding in Primate Prefrontal Cortex Neurons. *Neuron*, 84(6) :1317–28, 2014.

Otto, A. Ross, Skatova, Anya, Madlon-Kay, Seth, and Daw, Nathaniel D. Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27(2) :319–333, 2014.

Padoa-Schioppa, Camillo and Assad, John a. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090) :223–226, 2006.

Parker, Amanda and Gaffan, David. Memory after frontal/temporal disconnection in monkeys : Conditional and non-conditional tasks, unilateral and bilateral frontal lesions. *Neuropsychologia*, 36(3) :259–271, 1998.

Paton, Joseph J, Belova, Marina a, Morrison, Sara E, and Salzman, C Daniel. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, 439(7078) : 865–870, 2006.

Payzan-LeNestour, Elise and Bossaerts, Peter. Risk, unexpected uncertainty, and estimation uncertainty : Bayesian learning in unstable settings. *PLoS Computational Biology*, 7(1) :e1001048, 2011.

Payzan-LeNestour, Élise and Bossaerts, Peter. Do not bet on the unknown versus try to find out more : Estimation uncertainty and "unexpected uncertainty" both modulate exploration. *Frontiers in Neuroscience*, 6(OCT) :1–6, 2012.

Pearce, J M and Hall, G. A model for Pavlovian learning : variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6) :532–552, 1980.

Pearson, John M., Watson, Karli K., and Platt, Michael L. Decision making : The neuroethological turn. *Neuron*, 82(5) :950–965, 2014.

Pesaran, Bijan, Nelson, Matthew J., and Andersen, Richard a. Free choice activates a decision circuit between frontal and parietal cortex. *Nature*, 453(7193) :406–409, 2008.

Petrides, Michael, Tomaiuolo, Francesco, Yeterian, Edward H., and Pandya, Deepak N. The prefrontal cortex : Comparative architectonic organization in the human and the macaque monkey brains. *Cortex*, 48(1) :46–57, 2012.

Pfurtscheller, G and Lopes da Silva, F H. Event-related EEG/MEG synchronization and desynchronization : basic principles. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 110(11) :1842–57, nov 1999.

Phillips, Jessica M., Vinck, Martin, Everling, Stefan, and Womelsdorf, Thilo. A long-range fronto-parietal 5- to 10-Hz network predicts "top-down" controlled guidance in a task-switch paradigm. *Cerebral Cortex*, 24(8) :1996–2008, 2014.

Platt, M L and Glimcher, P W. Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741) :233–238, 1999.

Polich, John. Updating P300 : an integrative theory of P3a and P3b. *Clin Neurophysiol.*, 118(10) : 2128–2148, 2007.

Prescott, Tony J, Bryson, Joanna J, and Seth, Anil K. Modelling natural action selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1485) : 1521–1529, 2007.

Preuschoff, Kerstin, Bossaerts, Peter, and Quartz, Steven R. Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*, 51(3) :381–390, 2006.

Preuschoff, Kerstin, Bernard Marius, and Einhäuser, Wolfgang. Pupil dilation signals surprise : Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, 5(SEP) :1–12, 2011.

Procyk, E, Tanaka, Y L, and Joseph, J P. Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nature neuroscience*, 3(5) :502–508, may 2000.

Procyk, E., Wilson, C. R. E., Stoll, F. M., Faraut, M. C. M., Petrides, M., and Amiez, C. Midcingulate Motor Map and Feedback Detection : Converging Data from Humans and Monkeys. *Cerebral Cortex*, pages 1–10, 2014.

Quilodran, René, Rothé, Marie, and Procyk, Emmanuel. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2) :314–325, 2008.

Ratcliff, R. and Rouder, J. N. Modeling Response Times for Two-Choice Decisions. *Psychological Science*, 9(5) :347–356, 1998.

Redish, David a, Jensen, Steve, Johnson, Adam, and Kurth-Nelson, Zeb. "Reconciling reinforcement learning models with behavioral extinction and renewal : Implications for addiction, relapse, and problem gambling" : Correction. *Psychological review*, 114(3) :784–805, 2007.

Rescorla, R a and Wagner, a R. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, 21(6) :64–99, 1972.

Rogers, Robert d and Monsell, Stephen. Costs of a predictable switch between simple cognitve tasks. *Journal of experimental psychology. General*, 1995.

Roitman, Jamie D and Shadlen, Michael N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, 22 (21) :9475–89, 2002.

Rothé, Marie, Quilodran, René, Sallet, Jérôme, and Procyk, Emmanuel. Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(31) :11110–11117, 2011.

Rudebeck, Peter H, Behrens, Timothy E, Kennerley, Steven W, Baxter, Mark G, Buckley, Mark J, Walton, Mark E, and Rushworth, Matthew F S. Frontal cortex subregions play distinct roles in choices between actions and stimuli. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(51) :13775–13785, 2008.

Rushworth, M F S, Walton, M E, Kennerley, S W, and Bannerman, D M. Action sets and decisions in the medial frontal cortex. *Trends in cognitive sciences*, 8(9) :410–7, sep 2004.

Sakagami, M and Tsutsui, K. The hierarchical organization of decision making in the primate prefrontal cortex. *Neuroscience research*, 34 :79–89, 1999.

Sakai, Katsuyuki. Task set and prefrontal cortex. *Annual review of neuroscience*, 31 :219–45, jan 2008.

Sakai, Katsuyuki and Passingham, Richard E. Prefrontal interactions reflect future task operations. *Nature neuroscience*, 6(1) :75–81, 2003.

Sallet, Jérôme, Quilodran, René, Rothé, Marie, Vezoli, Julien, Joseph, Jean-Paul, and Procyk, Emmanuel. Expectations, gains, and losses in the anterior cingulate cortex. *Cognitive, affective & behavioral neuroscience*, 7(4) :327–336, 2007.

Samejima, Kazuyuki and Doya, Kenji. Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, 1104 :213–228, 2007.

Samejima, Kazuyuki, Ueda, Yasumasa, Doya, Kenji, and Kimura, Minoru. Representation of action-specific reward values in the striatum. *Science (New York, N.Y.)*, 310(5752) :1337–1340, 2005.

Sauseng, P, Klimesch, W, Gruber, W R, Hanslmayr, S, Freunberger, R, and Doppelmayr, M. Are event-related potential components generated by phase resetting of brain oscillations ? A critical discussion. *Neuroscience*, 146(4) :1435–44, jun 2007.

Sawaguchi, T and Goldman-Rakic, P S. D1 dopamine receptors in prefrontal cortex : involvement in working memory. *Science*, 251(4996) :947–950, 1991.

Schrier, Allan M. Transfer by macaque monkeys between learnin-set and repeated -reversal tasks. *Perceptual and Motor Skills*, (23) :787–792, 1966.

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

Schultz, W, Dayan, P, and Montague, P R. A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306) :1593–9, mar 1997.

Schultz, Wolfram, Preuschoff, Kerstin, Camerer, Colin, Hsu, Ming, Fiorillo, Christopher D, Tobler, Philippe N, and Bossaerts, Peter. Explicit neural signals reflecting reward uncertainty. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1511) : 3801–3811, 2008.

Seo, Hyojung and Lee, Daeyeol. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(31) :8366–8377, 2007.

Shima, K and Tanji, J. Role for cingulate motor area cells in voluntary movement selection based on reward. *Science (New York, N.Y.)*, 282(5392) :1335–1338, 1998.

Sigman, Mariano and Dehaene, Stanislas. Parsing a cognitive task : A characterization of the mind's bottleneck. *PLoS Biology*, 3(2) :0334–0349, 2005.

Simon, H, Scatton, B, and Moal, M L. Dopaminergic A10 neurones are involved in cognitive functions., 1980.

Skoblenick, K. J., Womelsdorf, T., and Everling, S. Ketamine Alters Outcome-Related Local Field Potentials in Monkey Prefrontal Cortex. *Cerebral Cortex*, pages 1–10, 2015.

Skvortsova, Vasilisa, Palminteri, Stefano, and Pessiglione, Mathias. Learning To Minimize Efforts versus Maximizing Rewards : Computational Principles and Neural Correlates. *The Journal of Neuroscience*, 34(47) :15621–15630, 2014.

Slotnick, B, Hanford, L, and Hodos, W. Can rats acquire an olfactory learning set ? *Journal of experimental psychology. Animal behavior processes*, 26(4) :399–415, 2000.

Slotnick, B M and Katz, H M. Olfactory learning-set formation in rats. *Science (New York, N.Y.)*, 185(153) :796–798, 1974.

Sohn, M H, Ursu, S, Anderson, J R, Stenger, V a, and Carter, C S. The role of prefrontal cortex and posterior parietal cortex in task switching. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24) :13448–53, nov 2000.

Steinhauser, Marco and Yeung, Nick. Error awareness as evidence accumulation : effects of speed-accuracy trade-off on error signaling. *Frontiers in Human Neuroscience*, 6(August) :1–12, 2012.

Stephens, David W. Decision ecology : foraging and the ecology of animal decision making. *Cognitive, affective & behavioral neuroscience*, 8(4) :475–484, 2008.

Stevens, Jeffrey R. Rational decision making in primates : the bounded and the ecological. *Primate neuroethology*, pages 96–116, 2010.

Stoet, Gijsbert and Snyder, Lawrence H. Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron*, 42(6) :1003–1012, 2004.

Stoll, F. M., Wilson, C. R. E., Faraut, M. C. M., Vezoli, J., Knoblauch, K., and Procyk, E. The Effects of Cognitive Control and Time on Frontal Beta Oscillations. *Cerebral Cortex*, pages 1–18, 2015.

Sugrue, Leo P, Corrado, Greg S, and Newsome, William T. Matching behavior and the representation of value in the parietal cortex. *Science (New York, N.Y.)*, 304(5678) :1782–1787, 2004.

Sun, Yanlong, O'Reilly, Randall C., Bhattacharyya, Rajan, Smith, Jack W., Liu, Xun, and Wang, Hongbin. Latent structure in random sequences drives neural learning toward a rational bias. *Proceedings of the National Academy of Sciences*, 112(12) :201422036, 2015.

Sutton, Richard S and Barto, Andrew G. Chapter 12 : Introductions. *Acta Physiologica Scandinavica*, 48(Mowrer 1960) :57–63, 1960.

Sutton, S, Braren, M, Zubin, J, and John, E R. Evoked-potential correlates of stimulus uncertainty. *Science (New York, N.Y.)*, 150(700) :1187–1188, 1965.

Szczepanski, Sara M and Knight, Robert T. Review Insights into Human Behavior from Lesions to the Prefrontal Cortex. *Neuron*, 83(5) :1002–1018, 2014.

Tallon-Baudry, C, Kreiter, a, and Bertrand, O. Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans. *Visual neuroscience*, 16 (3) :449–459, 1999.

Tanaka, Saori C, Doya, Kenji, Okada, Go, Ueda, Kazutaka, Okamoto, Yasumasa, and Yamawaki, Shigeto. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature neuroscience*, 7(8) :887–893, 2004.

Tebbich, Sabine and Teschke, Irmgard. Coping with Uncertainty : Woodpecker Finches (Cactospiza pallida) from an Unpredictable Habitat Are More Flexible than Birds from a Stable Habitat. *PLoS ONE*, 9(3) :e91718, 2014.

Tinklepaugh, O. L. An experimental study of representative factors in monkeys., 1928.

Tobler, Philippe N, Fiorillo, Christopher D, and Schultz, Wolfram. Adaptive coding of reward value by dopamine neurons. *Science (New York, N.Y.)*, 307(5715) :1642–1645, 2005.

Tobler, Philippe N, Doherty, John P O, Dolan, Raymond J, and Schultz, Wolfram. Reward Value Coding Distinct From Risk Attitude-Related Uncertainty Coding in Human Reward Systems. pages 1621–1632, 2007.

Tobler, Philippe N, Christopoulos, George I, O'Doherty, John P, Dolan, Raymond J, and Schultz, Wolfram. Risk-dependent reward value signal in human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17) :7185–7190, 2009.

Tolman, Edward C. and Honzik, C. H. Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, 4 :257–275, 1930.

Tom, Sabrina M, Fox, Craig R, Trepel, Christopher, and Poldrack, Russell a. The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, 315(5811) :515–518, 2007.

Tom Schonberg, Fox, Craig R., and Poldrack, Russell A. Mind the gap : bridging economic and naturalistic risk-taking with cognitive neuroscience. *Russell The Journal Of The Bertrand Russell Archives*, 15(1) :11–19, 2011.

Tomita, H, Ohbayashi, M, Nakahara, K, Hasegawa, I, and Miyashita, Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, 401(6754) :699–703, 1999.

Treichler, F Robert. Successive reversal of concurrent discriminations by macaques (Macaca mulatta) : proactive interference effects. *Animal cognition*, 8(2) :75–83, apr 2005.

Tsujimoto, Satoshi, Genovesio, Aldo, and Wise, Steven P. Monkey orbitofrontal cortex encodes response choices near feedback time. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(8) :2569–2574, 2009.

Tversky, a and Kahneman, D. Advances in Prospect-Theory - Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4) :297–323, 1992.

Ullsperger, Markus, Danielmeier, Claudia, and Jocham, Gerhard. Neurophysiology of performance monitoring and adaptive behavior. *Physiological reviews*, 94(1) :35–79, 2014.

van de Vijver, Irene, Ridderinkhof, K. Richard, and Cohen, Michael X. Frontal Oscillatory Dynamics Predict Feedback Learning and Action Adjustment. *Journal of Cognitive Neuroscience*, 23(12) : 4106–4121, 2011.

van Pelt, S., Boomsma, D. I., and Fries, P. Magnetoencephalography in Twins Reveals a Strong Genetic Determination of the Peak Frequency of Visually Induced Gamma-Band Synchronization. *Journal of Neuroscience*, 32(10) :3388–3392, 2012.

Varazzani, C., San-Galli, a., Gilardeau, S., and Bouret, S. Noradrenaline and Dopamine Neurons in the Reward/Effort Trade-Off : A Direct Electrophysiological Comparison in Behaving Monkeys. *Journal of Neuroscience*, 35(20) :7866–7877, 2015.

Vezoli, Julien and Procyk, Emmanuel. Frontal feedback-related potentials in nonhuman primates : modulation during learning and under haloperidol. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(50) :15675–83, dec 2009.

Voloh, Benjamin, Valiante, Taufik a., Everling, Stefan, and Womelsdorf, Thilo. Theta–gamma coordination between anterior cingulate and prefrontal cortex indexes correct attention shifts. *Proceedings of the National Academy of Sciences*, (MAY) :201500438, 2015.

Voytek, Bradley, Kayser, Andrew S, Badre, David, Fegen, David, Chang, Edward F, Crone, Nathan E, Parvizi, Josef, Knight, Robert T, and D'Esposito, Mark. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nature Neuroscience*, 18(July) : 1–10, 2015.

Wallis, Jonathan D and Miller, Earl K. From Rule to Response : Neuronal Processes in the Premotor and Prefrontal Cortex. pages 1790–1806, 2003.

Wallis, Jonathan D, Anderson, Kathleen C, and Miller, Earl K. Single neurons in prefrontal cortex encode abstract rules. 411(June) :953–956, 2001.

Walsh, M. M. and Anderson, J. R. Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, 108(47) :19048–19053, 2011.

Walsh, Matthew M. and Anderson, John R. Learning from experience : Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, 36(8) :1870–1884, 2012.

Walton, Mark E., Behrens, Timothy E J, Buckley, Mark J., Rudebeck, Peter H., and Rushworth, Matthew F S. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, 65(6) :927–939, 2010.

Wan Lee, Sang, Shimojo, Shinsuke, and O'Doherty, John P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, 81(3) :687–699, 2014.

Warren, J M. The Comparative Psychology of Learning. *Annual review of psychology*, 16 :95–118, 1965.

Warren, J M and Baron, Alan. The Formation Of Learning Sets By Cats. *J. comp. physiol. Psycho*, 45(119-128), 1952.

Watanabe, M. Prefrontal unit activity during associative learning in the monkey. *Experimental Brain Research*, 80(2) :296–309, 1990.

Watanabe, M. Reward expectancy in primate prefrontal neurons., 1996.

Watanabe, Masataka. Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research*, 89(2) :233–247, 1992.

Weber, Elke U and Johnson, Eric J. *Decisions Under Uncertainty : Psychological, Economic, and Neuroeconomic Explanations of Risk Preference*. 2009.

Weissman, D. H., Gopalakrishnan, a., Hazlett, C. J., and Woldorff, M. G. Dorsal anterior cingulate cortex resolves conflict from distracting stimuli by boosting attention toward relevant events. *Cerebral Cortex*, 15(2) :229–237, 2005.

Werchan, D. M., Collins, a. G. E., Frank, M. J., and Amso, D. 8-Month-Old Infants Spontaneously Learn and Generalize Hierarchical Rules. *Psychological Science*, 2015.

White, Ilsun M and Wise, Steven P. Rule-dependent neuronal activity in the prefrontal cortex. pages 315–335, 1999.

White, N M. A functional hypothesis concerning the striatal matrix and patches : mediation of S-R memory and reward. *Life sciences*, 45(21) :1943–1957, 1989.

Wilkinson, L S, Humby, T, Killcross, a S, Torres, E M, Everitt, B J, and Robbins, T W. Dissociations in dopamine release in medial prefrontal cortex and ventral striatum during the acquisition and extinction of classical aversive conditioning in the rat. *The European journal of neuroscience*, 10 (3) :1019–1026, 1998.

Wilson, Charles R E and Gaffan, David. Prefrontal-inferotemporal interaction is not always necessary for reversal learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(21) :5529–38, may 2008.

Wilson, Charles R E, Gaffan, David, Browning, Philip G F, and Baxter, Mark G. Functional localization within the prefrontal cortex : Missing the forest for the trees ? *Trends in neurosciences*, 33(12) :533–40, dec 2010.

Witte, E. a., Davidson, M. C., and Marrocco, R. T. Effects of altering brain cholinergic activity on covert orienting of attention : Comparison of monkey and human performance. *Psychopharmacology*, 132(4) :324–334, 1997.

Womelsdorf, Thilo, Johnston, Kevin, Vinck, Martin, and Everling, Stefan. Theta-activity in anterior cingulate cortex predicts task rules and their adjustments following errors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(11) :5248–5253, 2010.

Yang, Chung-Hui, Belawat, Priyanka, Hafen, Ernst, Jan, Lily Y, and Jan, Yuh-Nung. Drosophila egg-laying site selection as a system to study simple decision-making processes. *Science (New York, N.Y.)*, 319(5870) :1679–1683, 2008.

Yeung, Nick, Nystrom, Leigh E, Aronson, Jessica a, and Cohen, Jonathan D. Between-task competition and cognitive control in task switching. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(5) :1429–38, feb 2006.

Yin, Henry H., Ostlund, Sean B., Knowlton, Barbara J., and Balleine, Bernard W. The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22(2) : 513–523, 2005.

Yu, Angela J. Adaptive behavior : humans act as bayesian learners. *Current Biology*, 17(22) : 977–980, 2007.

Yu, Angela J. and Dayan, Peter. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4) : 681–692, 2005.

Chapitre 9

Annexes

Article 1 :

Midcingulate Motor Map and Feedback Detection : Converging Data from Humans and Monkeys.

Procyk E., Wilson C, Stoll F., **Faraut M.**, Petrides M. and Amiez C. Cerebral Cortex, 2014

Article 2 :

The Effects of Cognitive Control and Time on Frontal Beta Oscillations.

Stoll F., Wilson C, **Faraut M.**, Vezoli J., Knoblauch K., Procyk E. Cerebral Cortex, 2015

Cerebral Cortex Advance Access published September 12, 2014

Cerebral Cortex
doi:10.1093/cercor/bhu213

Midcingulate Motor Map and Feedback Detection: Converging Data from Humans and Monkeys

Emmanuel Procyk[1,2], Charles R. E. Wilson[1,2], Frederic M. Stoll[1,2], Maïlys C. M. Faraut[1,2], Michael Petrides[3] and Céline Amiez[1,2]

[1]Institut National de la Santé et de la Recherche Médicale U846, Stem Cell and Brain Research Institute, 69500 Bron, France, [2]Université de Lyon, Lyon 1, Unité Mixte de Recherche S-846, 69003 Lyon, France and [3]Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada H3A2B4

Address correspondence to Dr Procyk Emmanuel. Email: emmanuel.procyk@inserm.fr; or to Dr Céline Amiez. Email: celine.amiez@inserm.fr

The functional and anatomical organization of the cingulate cortex across primate species is the subject of considerable and often confusing debate. The functions attributed to the midcingulate cortex (MCC) embrace, among others, feedback processing, pain, salience, action-reward association, premotor functions, and conflict monitoring. This multiplicity of functional concepts suggests either unresolved separation of functional contributions or integration and convergence. We here provide evidence from recent experiments in humans and from a meta-analysis of monkey data that MCC feedback-related activity is generated in the rostral cingulate premotor area by specific body maps directly related to the modality of feedback. As such, we argue for an embodied mechanism for adaptation and exploration in MCC. We propose arguments and precise tools to resolve the origins of performance monitoring signals in the medial frontal cortex, and to progress on issues regarding homology between human and nonhuman primate cingulate cortex.

Keywords: decision, learning, prefrontal, primate, reward

Introduction

Primates show a remarkable ability to adapt in the face of rapidly changing environments. Evaluation of decisions and of their outcomes, so-called performance monitoring, lies at the heart of such abilities. The search for computational and neurobiological principles of performance monitoring has been fruitful in the last 30 years, largely due to parallel research in rodents, monkeys, and humans (reviewed in Holroyd and Coles 2002; Montague et al. 2004; Rushworth et al. 2004; Shenhav et al. 2013).

Several studies have highlighted one subdivision of the cingulate cortex, the midcingulate cortex (MCC), as a central element of the performance monitoring network (Rushworth et al. 2007; Bush 2009; Shackman et al. 2011). Understanding the specific contribution of the MCC is an important challenge because of its putative key role in several aspects of human cognition, its association to a wide range of pathological conditions (Vogt 2009b) and, also, because physiological activity in parts of the cingulate cortical region might serve as markers of developmental and individual behavioral traits (Segalowitz and Dywan 2009).

In the search for MCC functions, discrepancies between human and monkey studies, and between functional and lesion data (Fellows and Farah 2005; di Pellegrino et al. 2007; Nachev 2011), have fueled debates on the exact contribution of this subdivision and, to some extent, on the validity of the nonhuman primate as a model of human cingulate functions (Cole et al. 2009; Schall and Emeric 2010). The debates have confronted multiple anatomical definitions of cingulate areas, as well as different functional interpretations of data obtained with multiple techniques. Important theoretical attempts have been made to integrate various pools of data (Botvinick 2007; Shenhav et al. 2013). However, we think that it is essential to clarify the fundamental issues in comparing empirical data obtained in humans and monkeys. These are the precision of anatomical descriptions and the experimental equivalence. In particular, the provision of juice reward and reward omission are central to the study of decision making in animals. The computational basis of adaptation relies on teaching signals that have been mostly studied using juice with animals. Juice reward and feedback must thus be taken into account as such when comparing human and monkey brain functions.

In the present contribution, we specifically address the issue of functional homology between human and monkey MCC, and its functional organization. To achieve this, we first deal with some difficulties in the anatomical and functional subdivisions of the cingulate region in the 2 species. Second, we show that the functional organization of the human MCC for juice feedback follows a systematic rule. The studies had 2 crucial constraints: behavioural protocols in human functional studies that are similar to those used in monkeys; and parsing the results on the basis of human interindividual morphological variability. Finally, we perform a meta-analysis of cingulate feedback-related unit activity in monkey to show a functional homology with human anterior MCC. This approach then allows us to discuss a possible functional organization principle in MCC, and to provide testable hypotheses.

Overview of Cingulate Cortical Organization in Primates

Part of the confusion in the functional definition of MCC arose from the multiplicity of labels naming subdivisions of the cingulate cortex (Laird et al. 2005; Vogt 2009b). It has become virtually impossible to understand what part of the medial frontal cortex is referred to when one uses the label anterior cingulate cortex (ACC). The label dorsal ACC (dACC) emerged in an attempt to reduce confusion, but it is based only on a rough estimate from brain imaging studies. Use of a common and consistent terminology is mandatory for further progress in this field. The regional model proposed by Vogt et al. is to date the clearest and most rigorous. It is based on multidimensional mappings in human and nonhuman species, including nonhuman primates (Vogt et al. 1995, 2005; Palomero-Gallagher et al. 2009; Vogt 2009c). The model describes 4 cingulate regions among which the most anterior is labeled ACC (See Fig. 1). The region just posterior, dorsal to the corpus callosum, is the MCC (mostly equivalent to dACC) with its most anterior part (aMCC) being the subject of the present study.

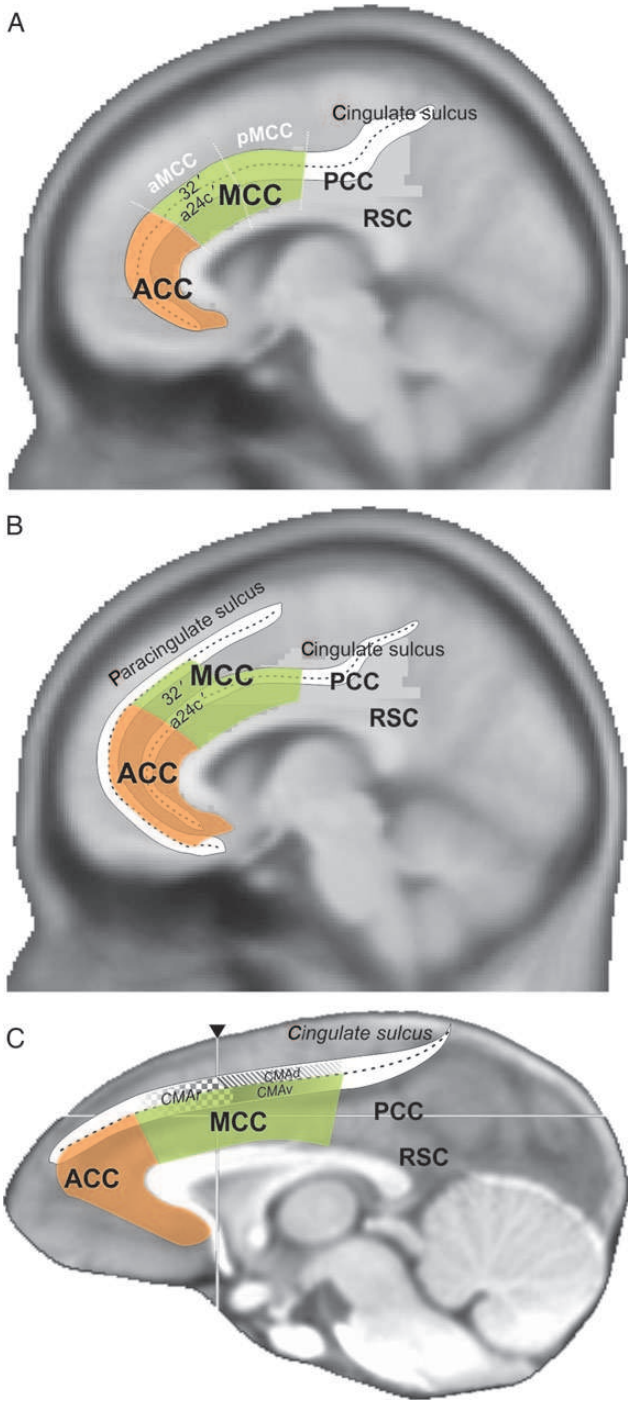

Figure 1. Schematic representations of the ACC and MCC region in the human (*A*,*B*) and macaque (*C*) brains according to Vogt et al. Overlap on brain anatomical scans average in MNI standard spaces for both species. The regions ACC, MCC, PCC, and RSC are based on the 4 regions subdivision by Vogt et al. (Vogt et al. 2005; Palomero-Gallagher et al. 2009; Vogt 2009b). The human representations schematize the organization of cingulate subdivisions in the case of absence (*A*) or presence (*B*) of the paracingulate sulcus. Area 32′ and a24c′ were defined by the same authors. In *A*, the schematic limits of anterior and posterior MCC (aMCC and pMCC) are shown. In *C*, the schematic position of cingulate motor areas (CMAr, CMAd, CMAv) are presented as in He et al. (1995).

It is important to note that the cytoarchitectonic limits of the MCC appear to relate to the morphology of sulci in primates and such relationships have been in fact another important source of confusion regarding the functional organization of MCC. Based on classical cytoarchitectonic studies and their own research, Vogt et al. propose that the human MCC comprises cytoarchitectonic areas 24a′, 24b′, 24c′, 24d, 33′, and 32′ (Vogt et al. 1995, see schema Fig. 1*B*; Palomero-Gallagher et al. 2008). According to these studies, area 32′ is always dorsal to area 24c′, but the relationship between these areas and the sulci on the medial wall is not trivial. Specifically, this is because of individual variations in morphology. Whereas all humans possess a cingulate sulcus in each hemisphere, a double cingulate sulcus known as the paracingulate sulcus is variably present (Petrides 2012). The paracingulate sulcus is observed in ~70% of subjects at least in one hemisphere, and runs dorsal and parallel to the cingulate sulcus through the MCC (Vogt et al. 1995; Paus et al. 1996; Fornito et al. 2008). A paracingulate sulcus can be observed in both hemispheres in some brains, in only one hemisphere in most cases (see Supplementary Material), or in neither hemisphere. Morphological variability appears clearly in surface-based standardized analyses (Hill et al. 2010).

The question of interest here is the relationship between these sulci and the cytoarchitecture, and although this requires further study, current understanding is depicted in Figure 1. Areas 32′ and 24c′ cover the dorsal and ventral banks of the cingulate sulcus in the absence of paracingulate sulcus (Fig. 1*A*). In contrast, area 32′ was observed in the paracingulate gyrus above the cingulate sulcus and in the paracingulate sulcus when the latter is present, with area 24c′ covering both banks of the cingulate sulcus (Fig. 1*B*, Vogt et al. 1995). In the standard stereotaxic space (MNI), as used in brain imaging experiments, the cortex lying in the paracingulate and cingulate sulci have different coordinates. This suggests that the location of area 32′ in standard space is different for the 2 types of morphology. Crucially, this means that population averaging procedures should significantly decrease the reliability of activation measures in that region, unless individual morphology is rigorously taken into account (Shackman et al. 2011; Amiez et al. 2013).

In monkeys, the cingulate cortex presents important similarities in anatomical organization with the human cingulate region. Figure 1*C* represents the subdivision of the rostral cingulate region as proposed by Vogt et al. who studied a comparative anatomy in humans and monkeys. The different subdivisions of the cingulate cortex are organized around the single cingulate sulcus as there is no paracingulate sulcus in macaque monkeys. This cingulate sulcus contains several cytoarchitectonnic areas that have been mostly shown to be comparable with human cytoarchitectonic subdivisions. The exceptions are areas 32′ in MCC and 33′. Earlier cytoarchitectonic studies of the macaque monkey midcingulate region had not identified area 32′, and it was attributed to the human species only (Vogt 2009a). In macaque, the 4-region model uses the fundus of the cingulate sulcus as the dorsal limit of the MCC, excluding the dorsal bank of the sulcus (Vogt et al. 2005). However, neuroanatomical studies from several groups observed that the cortex in the dorsal bank of the cingulate sulcus includes cingulate or transition areas (Matelli et al. 1991; Petrides and Pandya 1994; Zilles et al. 1995; Geyer et al. 1998; Paxinos et al. 2009; for review Sallet et al. 2011). In the human brain, the cortex above the cingulate sulcus when there is a paracingulate sulcus, that is, on the paracingulate gyrus, is a transitional dysgranular area that separates agranular cingulate cortex (classical area 24) from medial dorsal frontal areas (see

Petrides and Pandya 1994, 1999). The corresponding region in the macaque brain lies in the dorsal bank of the cingulate sulcus, above the anterior part of the corpus callosum (Petrides and Pandya 1994). Interestingly, the dorsal bank is where the most dorsal MCC lies in the human brain when there is only a single cingulate sulcus (Vogt et al. 1995; Palomero-Gallagher et al. 2008 and see Fig. 1*B*).

In conclusion, some architectonic studies suggest important primate species difference in the cingulate cortex, with a dorsal limit in monkey cingulate sulcus, supporting the theoretical argument on primate interspecies difference (Cole et al. 2009). However, based on cytoarchitectonic studies (e.g., Petrides and Pandya 1994), this dorsal limit can be challenged. Also, as we shall see, most of the physiological recordings in monkey cingulate cortex that are compared with human functional neuroimaging data have been performed in the dorsal bank and fundus of the cingulate sulcus, that is, outside of Vogt's definition of the MCC. In addition, the layout of cingulate motor areas (CMAs) also favors the extension of MCC onto the dorsal bank.

Cingulate Motor Areas

Crucially, the MCC region overlaps with or includes CMAs. The CMAs have been defined in monkeys using intracortical microstimulation, as well as by anatomical demonstration of connection to the premotor cortex, the primary motor cortex, and the spinal cord (Woolsey et al. 1952; Hutchins et al. 1988; Mitz and Godschalk 1989; Dum and Strick 1991, 1996; Godschalk et al. 1995; He et al. 1995; Hatanaka et al. 2001). Cortical labeling following tracer injections in the cervical or lumbar segment of the spinal cord showed that several representations of the arm and of the lower limb are present in the cortex of the dorsal and ventral banks of the cingulate sulcus, which contrasts with a cytoarchitectonic limit in the fundus of the sulcus. Three major subdivisions were defined by Strick et al.: CMAr, CMAd, and CMAv (for rostral, dorsal, and ventral CMAs) each containing somatomotor representations (Hutchins et al. 1988; Dum and Strick 1991). It is unclear how Vogt's borders relate to motor areas on the medial wall in nonhuman primates (Fig. 1*C*). For instance, because the posterior representation of the limbs (in CMAd) are found in the dorsal bank of the sulcus (He et al. 1995; Hatanaka et al. 2001), a rigorous application of the dorsal border in the cingulate sulcus results in double arm and leg representations in the primary motor cortex and supplementary area, respectively.

Although CMAr (the main subject of this paper) is often discussed in relation to its well-known arm and hand representations, a representation of the face/eye field has also been described using experimental anatomical tract tracing and microstimulation, just anterior to the arm representation (Mitz and Godschalk 1989; Morecraft et al. 1996, 2007; Tokuno et al. 1997). Further information on the rostral cingulate face representation is provided below.

Until recently, the definition of human equivalents of the monkey CMAs relied mainly on the comprehensive meta-analysis by Picard and Strick (1996), who provisionally identified based on positron emission tomography studies 3 subdivisions labeled anterior and posterior rostral cingulate zones (RCZa and RCZp) and a caudal cingulate zone (CCZ) that might be equivalent to the respective CMAr, CMAv, and CMAd identified in the macaque monkey. These investigators predicted the presence of face (related to studies on eye movements and speech) and arm representations in the 2 RCZ and an arm representation in the CCZ. However, no single study had ever tested such somatomotor mappings in individual human subjects paying particular attention to the sulcal morphological variability. Amiez and Petrides (2014) have recently performed this experiment by mapping, with functional magnetic resonance imaging (fMRI) activations in the cingulate regions for eye, tongue, arm, and foot movements (Amiez and Petrides 2014). As in the monkey, they uncovered 3 cingulate motor regions, each including a focus for arm and foot movements and with the 2 anterior regions including in addition foci for movements of the eye and of the tongue. (Note that it was not possible to find a significant spatial dissociation between eye and tongue-related activations, unpublished observation.) This suggests that limb representations are present in all cingulate zones but that only the most anterior ones (RCZa and RCZp) appear to have representations of the face, including eye-related fields, although a separation between face and eye remains to be investigated.

Importantly, using a single-subject approach, it was shown that the face activations are located in the paracingulate sulcus when it is present, but in the cingulate sulcus in the absence of the paracingulate sulcus. In either case, the eye/face focus is always located at the junction of the cingulate or paracingulate sulci with a small perpendicular sulcus (Fig. 2*A*) (Amiez and Petrides 2014). Arm and foot activations were always in the cingulate sulcus. Together with the observations of displacement in the presence of a paracingulate sulcus (Fig. 1), these functional data suggest that the rostral eye/face representation in aMCC is displaced in a similar way to the displacement of area 32′ described by Vogt et al. (1995). A clear morphological landmark might thus be used to track the location of RCZ face area in humans and allow precise functional mappings.

Feedback Evaluation and the MCC

Functional data have accumulated on the role of MCC in outcome-based decisions and adaptation, in both humans and monkeys. Single-unit and local field potential recordings in the anterior section of the dorsal bank of the cingulate sulcus in monkeys have revealed particularly prominent activity related to outcome or feedback detection and evaluation (e.g., Amiez et al. 2005; Matsumoto et al. 2007; Seo and Lee 2007; Kennerley and Wallis 2009; Luk and Wallis 2009). During an explore/repeat task (searching for a rewarded response and then repeat), single-unit activity has been shown to be related to the coding and discrimination of various forms of feedback relevant for adaptation (negative feedback: no reward, positive feedback: juice delivery, etc.), in particular during exploration (Quilodran et al. 2008). How does this relate to human MCC? Learning and decision behavioral protocols in monkeys use juice reward as feedback and incentive. Delivery or omission of reward provides the relevant information to guide behavior. Thus, a proper comparison of human and macaque studies requires the use of reward in similar ways in both species. For a direct comparison with monkey studies, we recently adapted the task used in monkey experiments, including exploration (trial and error) and repetition periods and using fruit juice as outcome or feedback, in a human fMRI protocol (Amiez et al. 2013). As predicted from monkey electrophysiological results, very reliable aMCC activation was observed at feedback during

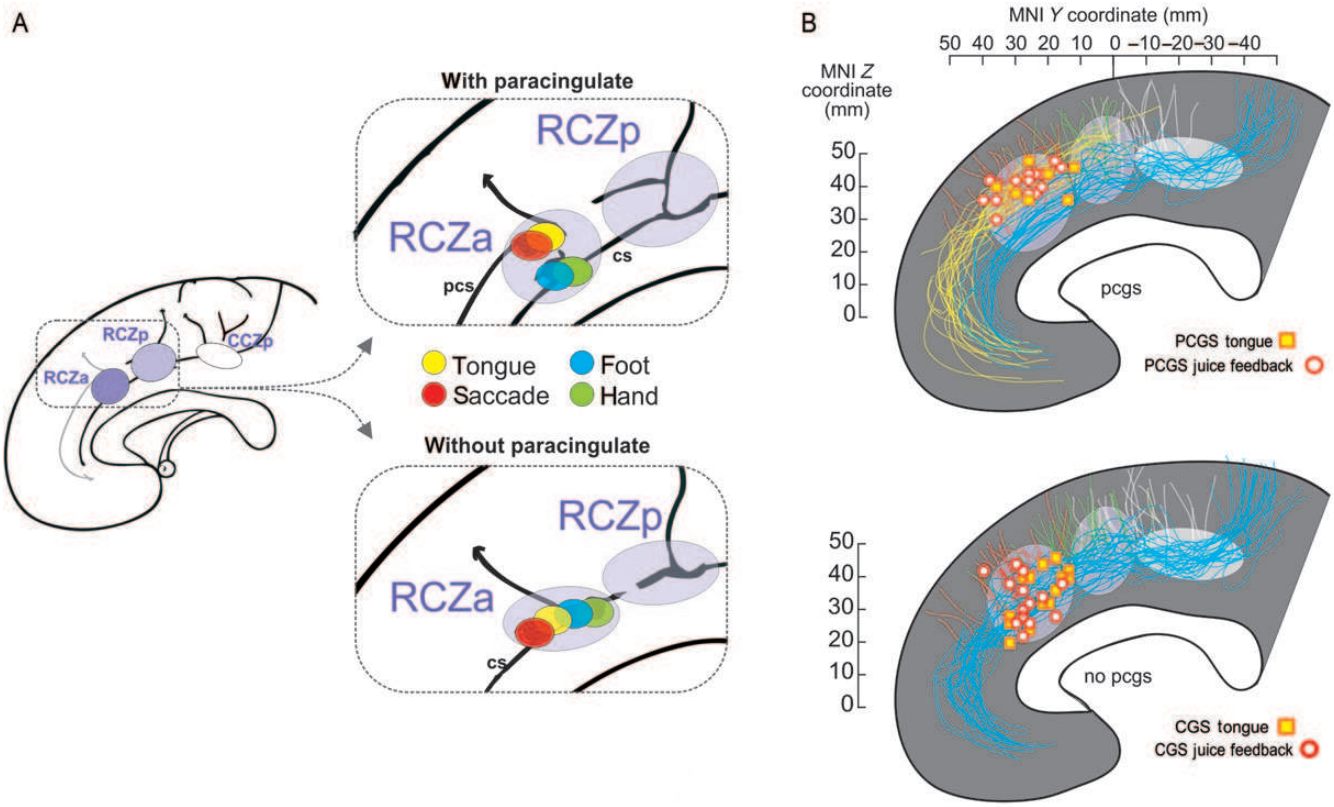

Figure 2. Human cingulate motor areas, feedback activity, and sulcal morphology. (*A*) Schematic illustration of the 3 human cingulate motor areas (RCZa, RCZp, and CCZ) as described by Amiez and Petrides (2014). Colored disks represent the average location of activation peaks in response to simple voluntary movements for hemispheres with (top) and without (bottom) a paracingulate sulcus. cs: cingulate sulcus, pcs: paracingulate sulcus. (*B*) Overlap of tongue movement-related activation peaks (individual peak locations are represented by squares) and peaks for feedback-related activation (circles) during exploration for hemispheres with and without a paracingulate sulcus. Each individual sulcus path has been retraced, and all paracingulate (blue) and cingulate (yellow) sulci, as well as vertical branches (red, green, and white), have been overlapped for the populations of subjects. Data taken from Amiez et al. (2013) and Amiez and Petrides (2014). Note, the activation data come from 2 separate experiments. The approximate location of RCZa, RCZp, and CCZ is indicated by ellipses (rostrocaudal extent estimated from Amiez and Petrides 2014).

exploration, but not at feedback during repetition. Most importantly, the location of feedback-related activation could be related to the local sulcal morphology in the cingulate region as was the case for the location of the rostral CMA. The feedback-related activation was located in the paracingulate sulcus when present, or in the cingulate sulcus in the absence of the secondary paracingulate sulcus.

Taken together, our recent functional neuroimaging experiments in human subjects reveal organizational principles in MCC that stimulate a reconsideration of data in monkeys (Amiez et al. 2013; Amiez and Petrides 2014). The data suggest that the juice feedback activations and the CMAs are related, and that this relationship could be similar in human and nonhuman primates. To address this homology issue, we proceeded in 2 ways: 1) we combined single human subject fMRI data to test the relationships between feedback-related activations and cingulate motor areas and 2) we performed a meta-analysis of monkey outcome-related and CMA-related data. The aim was to provide a comparative assessment of the relationship between juice feedback-related activity and CMAs in the 2 species.

Materials and Methods

Brain Imaging

Individual peaks of statistically identified clusters reported in Amiez and Petrides (2014) and Amiez et al. (2013) were plotted in the MNI standard stereotaxic space (Fig. 2*B*). The course of the cingulate sulcus, paracingulate sulcus, and 3 major branching vertical sulci were drawn from single-subject T_1 sagittal views. The most posterior of the 3 vertical sulci is the paracentral sulcus (pacs), followed by the preparacentral sulcus (prpacs), and then the vertical paracingulate sulcus (vpcgs), which is the most anterior (see Amiez et al. 2013).

Principles of Monkey Meta-analysis

We evaluated, from the literature, the location of reported outcome or feedback-related single-unit activity in relation to CMAr representations. To do this, we performed a meta-analysis of published neurophysiological and neuroanatomical data obtained in monkeys.

Our aim was to co-register, using the same anatomical reference framework, data from unit recordings, microstimulation mappings, and neuroanatomical tract tracing, and to investigate whether outcome-related activity was likely to come from recordings in CMAr/face region. Data from 26 articles were used (see Supplementary Material). Reconstructions of recording sites were based on the data available in those published articles.

For unit recordings, the selection of articles was based on whether the investigators reported outcome-, feedback-, or more generally juice-related changes in single-unit activity and also on whether there was sufficient information to reconstruct the recording zone. The recording zone retained for analysis corresponds, for each article, to the entire extent of recordings that included outcome- or feedback-related activity.

Published articles reporting anatomical data were selected for this review when they presented cortical map reconstructions or sufficient comprehensive data to reconstruct the rostrocaudal extent of the face/eye or arm representation identified by retrograde tracing or microstimulation mapping.

Co-registration

There is unfortunately no accepted standard method to report the location of data in monkeys, although an effort is made to provide an MNI standard monkey stereotaxic space (Frey et al. 2011). The investigators report either the extent of recordings relative to morphological landmarks (genu of the arcuate sulcus, genu of the corpus callosum, anterior commissure), relative to stereotaxic binaural zero, or both, or none of the above. The most comprehensive approach is to report all that information on a reconstructed cortical surface map.

In order to co-register the rostrocaudal coordinates reported in all articles considered, we have taken the level of the genu of the arcuate sulcus (ArcGen) as a reference. This landmark is indeed the most reported landmark. When the position of recordings relative to ArcGen was available, we aligned data to the ArcGen position. When stereotaxic coordinates were provided but the location of ArcGen was absent we realigned data on the average ArcGen location obtained from a database of 11 monkey MRIs. This average was AP + 24 (SD 2.6). The average location for the genu of corpus callosum (32.69 mm) was also used in some cases (on average 8.67 mm between ArcGen and Ccgen).

Results

Human fMRI

In Amiez et al. (2013), a single-subject analysis revealed that the feedback-related activation was systematically (15/15 subjects) located in the paracingulate sulcus when present, or in the cingulate sulcus in the absence of the paracingulate sulcus. The activation was also always observed at the junction with a specific short perpendicular sulcus, the vertical paracingulate sulcus. A less consistent (6 of 15 subjects) posterior peak was systematically located at the intersection between the cingulate sulcus (if there was no paracingulate sulcus) or the paracingulate sulcus (if present) and the preparacentral sulcus. It important to note here that we observed 2 distinct peaks and not a single peak that spread out.

Based on the description of the cingulate motor zones described above, such properties suggest that the juice feedback-related activation in the aMCC overlaps with an orofacial representation of RCZa. To evaluate this overlap, we compared the activation coordinates obtained for tongue movements in Amiez and Petrides (2014) and for juice feedback (Amiez et al. 2013) provided in Figure 2*B*. We chose to represent only the anterior activation peak because of its consistency in 100% of subjects bilaterally. In the explore/exploit task, activation of RCZp was obtained only in 50% of subjects and is not considered further. The single-subject data reported on individual morphology for hemispheres with and without a paracingulate sulcus, and in relation to the extent of the 3 cingulate motor zones as described by Amiez and Petrides (2014) reveal that both activations for juice feedback in exploration and for tongue movements are located in RCZa and are associated with the paracingulate sulcus when this sulcus is present. Experiments are currently being performed to further test this overlap.

Monkey Meta-analysis

If the orofacial representation in human aMCC processes feedback provided by juice, then can we find the same correspondence in monkeys? If so this would converge towards a clear anatomical and functional homology between human and monkey performance monitoring systems, in particular regarding the aMCC/RCZa subregion. Most unit recording experiments in monkeys reported data acquired close to or just anterior to CMAr in the dorsal bank and fundus of the cingulate sulcus, a region often referred to as the dACC. Because an eye/face representation exists anterior to the hand representation of CMAr, we re-evaluated from the literature the location of outcome or feedback-related activity relative to CMAr representations. We performed a meta-analysis of published data acquired in macaques (see Materials and Methods). This approach is quite rare in the monkey literature, and is in fact quite difficult, mostly because of a lack of a convention in the reporting of the location of recordings or of anatomical data. Nevertheless, this approach allowed us to synthesize and map available functional data in the cingulate sulcus.

Our aim was to co-register, using the same anatomical reference framework, data from unit recordings, microstimulation mappings, and neuroanatomical tract tracing, and to investigate whether outcome-related activity was likely to come from recordings in the CMAr/face region. As pointed out above, 26 articles formed the basis of this meta-analysis (see Supplementary Material regarding selection criteria and methods).

Figures 3 and 4 present the major findings. The rostrocaudal extent of regions of interest collected from the 26 articles are grouped according to whether they reported data on outcome-/feedback-related unit activity, data on the location of a face or eye-related area (Face: tracing studies or microstimulations) and, data on the hand region of CMAr (Forelimb) (Fig. 3). Note that the figure reports several specific points regarding each study, including the effectors used to respond in single-unit recording studies (see also Supplementary Material). This information is provided because the effector might be a key factor in determining the functional organization of CMAr. The raw data show that the eye/face representation clearly overlaps with the recordings reporting outcome-related activity. These 2 regions are somewhat anterior to the forelimb representation in CMAr. Most recordings were performed in the dorsal bank of the cingulate sulcus, and most neuroanatomical data regarding CMAr were in the dorsal bank with some extensions in the ventral bank (see Supplementary Material).

We calculated the overall rostrocaudal distributions of reported regions of interest, and display them in Figure 4*A* on a flat map reporting the main anatomical landmarks on a macaque brain. Statistical comparisons of the antero-posterior distributions (Fig. 4*B*) show a small difference between feedback/outcome recordings and face regions but a highly significant difference between those locations and the distribution for reports related to the forelimb CMAr representation (Distributions were compared by Wilcoxon rank-sum tests. Recordings versus Eye/Face: $P = 0.011$, ns; Recordings versus Forelimb: $P < 10^{-9}$, zval: −7.65 ; Eye/Face versus Forelimb: $P < 10^{-8}$, zval: 6.14. Two-sample Kolmogorov–Smirnov tests led to the exact same conclusions with $P = 0.019$ for Recordings versus Eye/Face and all $P < 10^{-8}$ for tests against Forelimb). Indeed, some authors have specifically noted the drop of prevalence of outcome encoding when recording in posterior parts of the cingulate sulcus (see the Conclusion in Luk and Wallis 2009). Selected articles for which clear (anatomical) maps were provided reveal an eye/face-related area mostly in the dorsal bank and fundus of the cingulate sulcus and for some study in the ventral bank, overlapping with the licking activity obtained with 2-deoxyglucose by Picard and Strick (1997) (Fig. 4*C*). Note that discriminating between putative eye and face fields remains difficult with the analyzed data.

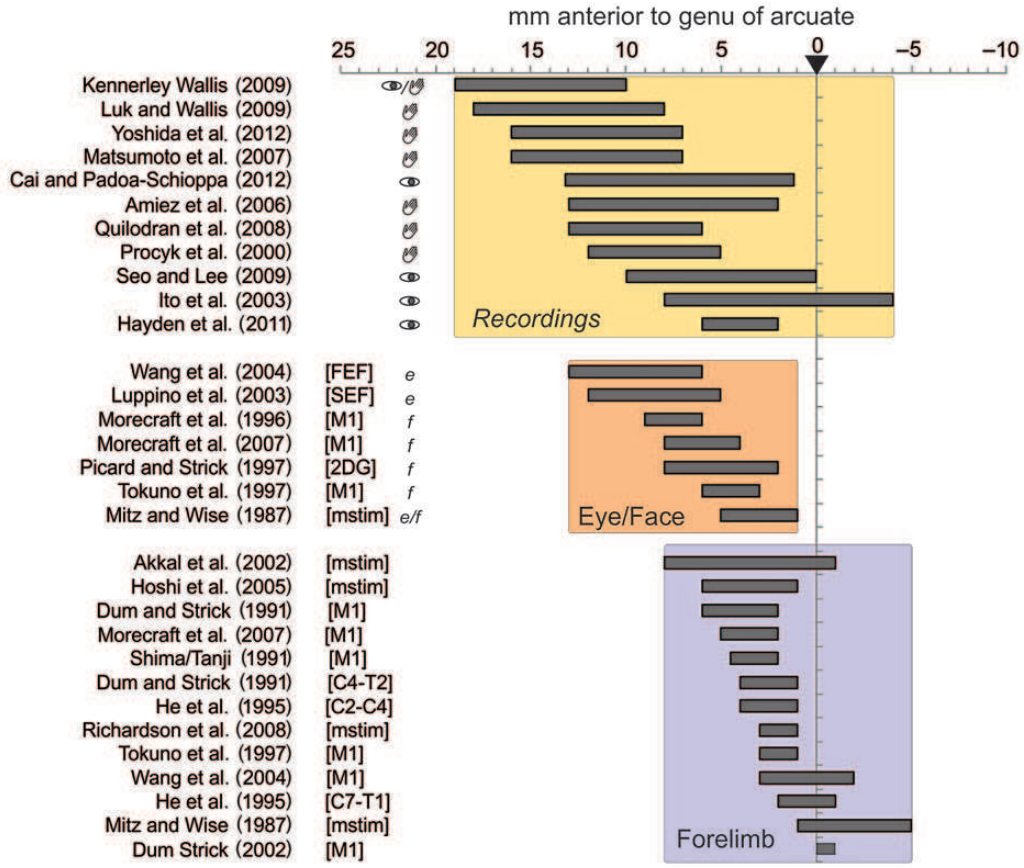

Figure 3. Database for meta-analysis in monkeys. Rostrocaudal extent of (top) recording sites in studies reporting feedback/outcome-related activity, (middle) regions with face-related effects of microstimulations and regions with anatomical connections with face-related areas and nuclei, and (bottom) regions with arm-related effects of microstimulations and regions with connections with arm-related areas and spinal levels. On the left of recording sites extent, symbols of an eye and of a hand indicate the effector used by animals to respond. On the left of Eye/Face studies, "e" and "f" relate to studies focusing on eye-related data (e.g., connections to FEE) or to face-related data (e.g., connection to M1 face), respectively. The specificity of anatomical studies is indicated in brackets (FEF, SEF, M1, C4-T2, C2-C4, C7-T1: injections of tracer in the respective cortical or spinal regions; mstim: microstimulation study; 2DG: study using 2-deoxyglucose). All data are aligned to the level of the genu of the arcuate sulcus (anterior 0, ArcGen). See Supplementary Information for details.

In conclusion, the meta-analysis strongly suggests that most recordings of feedback-related activity (juice in all cases but in Seo and Lee 2009) were most likely overlapping with the eye/face representation of CMAr, a functional overlap comparable with the one found in humans.

Discussion

We have provided evidence for the functional organization of the midcingulate cortical region in humans (Amiez et al. 2013; Amiez and Petrides 2014) and for a functional homology between the human and the monkey MCC by using comparable behavioral protocols in both humans and monkeys, and by taking into account the interindividual sulcal variability in humans. Based on this research, we propose that, in both species, the anterior MCC processes feedback provided by juice in a specialized somatomotor orofacial field. We further argue that feedback processing in general is embodied in the rostral cingulate motor area (CMAr) which is a specialized area of the MCC that may have evolved for higher control of motor action and decision making in both species.

Monkey Cingulate Maps

The data suggest that juice feedback is processed by homologous areas in both human and nonhuman primates, namely in the rostral cingulate premotor field. In contrast to previous suggestions (Cole et al. 2009), the present data support a functional homology between the aMCC in humans and a part of the dorsal bank of the cingulate sulcus in macaque monkeys and suggest an extension of MCC in the dorsal bank of the cingulate sulcus in the monkey brain. This finding is consistent with cytoarchitectonic studies showing that the upper bank of the cingulate sulcus in macaque monkeys is distinct from the dorsomedial frontal gyrus above and the cingulate gyrus below (Petrides and Pandya 1994) as is area 32′ defined by Vogt et al. (1995) in the MCC of the human brain. It also agrees with the coherent scheme of premotor field organizations observed in both monkeys and humans within the cingulate sulcus (He et al. 1995; Amiez and Petrides 2014). To clarify this point further, future studies will have to combine neural recordings during reward-based decision tasks and control sensorimotor tasks, in both the arm and face representations of CMAs. Ideally such experiments would include microstimulation and neuroanatomical tracing as performed in the study by Shima and Tanji (1998) on voluntary arm movement selection. Interestingly, these authors mentioned anatomical and microstimulation data supporting a location of recordings in the forelimb representation of CMAr. Their report of cingulate activity selectively modulated by changing arm movement after a reward decrease suggests that the arm representation would be involved

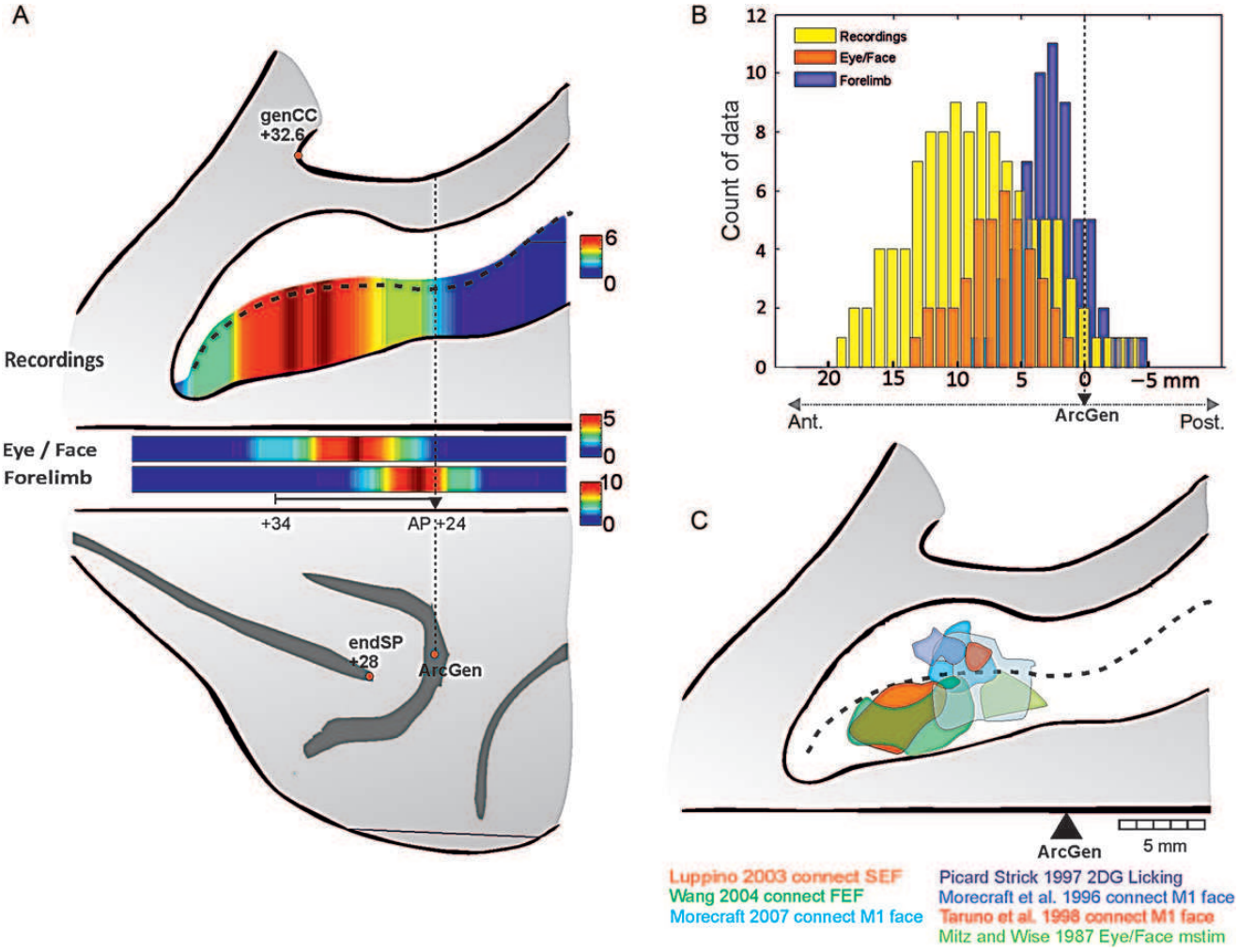

Figure 4. Meta-analysis of functional and anatomical data in monkeys. (*A*) Number of studies covering the rostrocaudal regions of the dorsal bank of the cingulate sulcus (data from Fig. 3). Single-unit recording studies are represented in the opened sulcus. Eye/face data and forelimb-related data are shown just below. Red in the color scale indicates a greater number of studies. AP coordinates for genu of the arcuate (genArc), caudal end of principalis (endSP), and genu of the Corpus Callosum (genCC) are averages taken from a population of 11 rhesus monkeys (from MRI images). (*B*) Histogram of data reported along the cingulate sulcus for recordings related to outcome/feedback (yellow), and for anatomical maps for Eye/Face representation (orange) and Forelimb (purple). Comparing distributions reveals that data for Eye/Face and outcome/feedback are different in terms of antero-posterior coverage at $P < 0.01$, but that both differ from the distribution related to Forelimb at $P < 10^{-8}$. The rostrocaudal extent is aligned on ArcGen. (*C*) Schematic overlap of eye/face-related data reconstructed from 7 studies. Maps are aligned on the rostrocaudal level of the genu of the arcuate sulcus (ArcGen).

when specific movement selection is required. In any case, the data reviewed above weakens previous arguments on species differences regarding MCC functions.

MCC Integrative Function and Embodied Feedback Processing

The value of a single individual approach and of using comparable protocols between human and monkeys have already been emphasized (Bush et al. 2002; Amiez et al. 2006; Bush 2009; Shackman et al. 2011). The present synthesis highlights their relevance to a better understanding of anatomo-functional relationships. The MCC is considered as an integration zone between cognition, motivation, and action (Paus 2001; Shackman et al. 2011). Orofacial fields in CMAs might contribute to the control of facial expressions (Morecraft et al. 2004). We propose that those fields process face-related information in the context of information-seeking during exploration or learning. In monkey experiments, juice reward provides important feedback information to resolve behavioral tasks. The orofacial representation in the most CMAr would participate in harvesting the information relevant for adaptive behavior. Although there is no evidence regarding a specific role of an orofacial versus eye representation, we propose that juice feedback engaged the former.

Further, a general principle can be proposed, namely that behaviorally relevant information is attended to and processed by MCC somatomotor maps, as an embodied mechanism that serves the search for information relevant for modifying behavior. This principle might be extended to other types or modalities of feedback. For instance, tactile feedback on the hand or feedback related to arm movement itself might be expected to involve the forelimb representation of CMAr/RCZa. Because of the sulcal morphology to functional relationships in humans, hand feedback-related activation should appear near the cingulate sulcus even if a paracingulate sulcus is present. Current experiments in our laboratories are evaluating these hypotheses.

Relationship to 'Other' MCC Functions

The embodied mechanism clarifies the often disregarded presence of premotor fields (CMAs) in a region often associated with higher cognitive functions. Yet, meta-analyses have shown notably that verbal and manual Stroop tasks often recruit the anterior and posterior parts of RCZa, respectively, which fits with the scheme, presented in Figure 2, of a dissociated face and arm representation (Laird et al. 2005).

Key questions can be tested using the proposed framework. The first concerns the role of eye-related fields in CMAr. There is currently no clear information regarding whether an eye field is segregated from a face field in the CMAr, and the data collected in Figure 4 are indeed unclear in this regard. Because our approach focused specifically on juice reward, we parsimoniously hypothesized that our data reflect activations of the orofacial subdivision. Yet more experiments are required to directly test this segregation. Using the structure-to-function relationship in humans, one can test for instance whether eye movement control in the context of active information-seeking activates a specific representation in aMCC. Indeed, eye movements are major tools for information-seeking in primates (Gottlieb et al. 2013). A further extension to be tested is that the rostral cingulate face representation processes others' facial expressions as feedback for specific adaptation. Recent experiments suggest that CMAs or more anterior parts of the cingulate cortex might be involved in face processing (Mies et al. 2011; Morita et al. 2014). Also, the mapping of body-specific behaviorally relevant information, such as pain, could be processed by specific subdivisions of CMA maps. This is suggested by recent experiments (Misra and Coombes 2014), and by an anatomical overlap between CMAs fields and the spinothalamic pain-related inputs in monkeys (Dum et al. 2009). Similarly, motor error-related activity observed in MCC in most decision tasks might be processed in the cingulate representation of the corresponding effector.

However, the proposed mapping does not resolve certain aspects of MCC functions. If primary, physical feedback, is processed in cingulate somatomotor areas, then what about visual or abstract secondary feedback? Money, power or other types of feedback often used in human studies could be processed by generalized CMA processes or in regions specific to processing more abstract information. Only precise single-subject analyses using voluntary movement tasks to produce specific localizers can answer these questions. Moreover, most theoretical approaches of MCC have emphasized its role in producing teaching signals but also in updating value functions to drive positive and negative feedback-based adaptations (Botvinick 2007; Alexander and Brown 2011; Shackman et al. 2011; Khamassi et al. 2014). For instance, MCC is proposed to monitor control-relevant information to estimate values necessary for optimal control selection (Shenhav et al. 2013). Other investigators suggest that MCC promotes searching or exploring the environment based on value signals estimated from the environment (Rushworth et al. 2012). Searching for the relationship between these valuation functions, the cingulate motor maps, and the varying sulcal patterns and their relationships to cytoarchitectonic areas in the human brain will certainly contribute to major improvement of our comprehension of MCC function. This will also be one key route for a clear resolution of the homology between human and nonhuman primate cingulate cortex. Toward that goal, and as mentioned above, methodological issues and differences between the 2 species will have to be taken seriously. In addition to proper protocol designs, using fMRI in monkeys combined with traditional neurophysiological approach will provide major information.

Finally, just as interindividual variability is important for precise investigations of the anatomo-functional organization in the human brain, it should be useful also to clinical approaches such as deep brain stimulation or relatively localized lesions as currently performed in patients with behavioral or mood disorders (Richter et al. 2004). The precise functional mapping of aMCC will be crucial to the planning of targeted and efficient interventions and to the understanding of their differential clinical effects.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

This work was supported by Agence National de la Recherche, project ANR-11-BSV4-0006 LU2, and by the labex CORTEX ANR-11-LABX-0042, Canadian Institutes of Health Research (CIHR) grant FRN 37753 (M.P.), and Fondation Neurodis (C.A.). CREW is funded by a Marie Curie Intra-European Fellowship (PIEF-GA-2010-273790). F.M.S. is funded by Fondation pour la Recherche Médicale and ANR DECCA project ANR-10-SVSE4-1441. M.C.M.F. is funded by Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche. E.P. and C.A. are employed by the Centre National de la Recherche Scientifique.

Notes

The authors thank the reviewers for their important comments and suggestions.

References

Alexander WH, Brown JW. 2011. Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci. 14:1338–1344.

Amiez C, Joseph JP, Procyk E. 2005. Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci. 21:3447–3452.

Amiez C, Kostopoulos P, Champod AS, Petrides M. 2006. Local morphology predicts functional organization of the dorsal premotor region in the human brain. J Neurosci. 26:2724–2731.

Amiez C, Neveu R, Warrot D, Petrides M, Knoblauch K, Procyk E. 2013. The location of feedback-related activity in the midcingulate cortex is predicted by local morphology. J Neurosci. 33:2217–2228.

Amiez C, Petrides M. 2014. Neuroimaging evidence of the anatomo-functional organization of the human cingulate motor areas. Cereb Cortex. 24:563–578.

Botvinick MM. 2007. Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci. 7:356–366.

Bush G. 2009. Dorsal anterior midcingulate cortex: roles in normal cognition and disruption in attention-deficit/hyperactivity disorder. In: Vogt BA, editor. Cingulate neurobiology and disease. New York: Oxford University Press. p. 245–274.

Bush G, Vogt BA, Holmes J, Dale AM, Greve D, Jenike MA, Rosen BR. 2002. Dorsal anterior cingulate cortex: a role in reward-based decision making. Proc Natl Acad Sci USA. 99:523–528.

Cole MW, Yeung N, Freiwald WA, Botvinick M. 2009. Cingulate cortex: diverging data from humans and monkeys. Trends Neurosci. 32:566–574.

di Pellegrino G, Ciaramelli E, Ladavas E. 2007. The regulation of cognitive control following rostral anterior cingulate cortex lesion in humans. J Cogn Neurosci. 19:275–286.
Dum RP, Levinthal DJ, Strick PL. 2009. The spinothalamic system targets motor and sensory areas in the cerebral cortex of monkeys. J Neurosci. 29:14223–14235.
Dum RP, Strick PL. 1991. The origin of corticospinal projections from the premotor areas in the frontal lobe. J Neurosci. 11:667–689.
Dum RP, Strick PL. 1996. Spinal cord terminations of the medial wall motor areas in macaque monkeys. J Neurosci. 16:6513–6525.
Fellows LK, Farah MJ. 2005. Is anterior cingulate cortex necessary for cognitive control? Brain. 128:788–796.
Fornito A, Wood SJ, Whittle S, Fuller J, Adamson C, Saling MM, Velakoulis D, Pantelis C, Yucel M. 2008. Variability of the paracingulate sulcus and morphometry of the medial frontal cortex: associations with cortical thickness, surface area, volume, and sulcal depth. Hum Brain Mapp. 29:222–236.
Frey S, Pandya DN, Chakravarty MM, Bailey L, Petrides M, Collins DL. 2011. An MRI based average macaque monkey stereotaxic atlas and space (MNI monkey space). Neuroimage. 55:1435–1442.
Geyer S, Matelli M, Luppino G, Schleicher A, Jansen Y, Palomero-Gallagher N, Zilles K. 1998. Receptor autoradiographic mapping of the mesial motor and premotor cortex of the macaque monkey. J Comp Neurol. 397:231–250.
Godschalk M, Mitz AR, van Duin B, van der Burg H. 1995. Somatotopy of monkey premotor cortex examined with microstimulation. Neurosci Res. 23:269–279.
Gottlieb J, Oudeyer PY, Lopes M, Baranes A. 2013. Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends Cogn Sci. 17:585–593.
Hatanaka N, Nambu A, Yamashita A, Takada M, Tokuno H. 2001. Somatotopic arrangement and corticocortical inputs of the hindlimb region of the primary motor cortex in the macaque monkey. Neurosci Res. 40:9–22.
He SQ, Dum RP, Strick PL. 1995. Topographic organization of corticospinal projections from the frontal lobe: motor areas on the medial surface of the hemisphere. J Neurosci. 15:3284–3306.
Hill J, Dierker D, Neil J, Inder T, Knutsen A, Harwell J, Coalson T, Van Essen D. 2010. A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. J Neurosci. 30:2268–2276.
Holroyd CB, Coles MG. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev. 109:679–709.
Hutchins KD, Martino AM, Strick PL. 1988. Corticospinal projections from the medial wall of the hemisphere. Exp Brain Res. 71:667–672.
Kennerley SW, Wallis JD. 2009. Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables. Eur J Neurosci. 29:2061–2073.
Khamassi M, Quilodran R, Enel P, Dominey PF, Procyk E. 2014. Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. Cereb Cortex. doi:10.1093/cercor/bhu114.
Laird AR, McMillan KM, Lancaster JL, Kochunov P, Turkeltaub PE, Pardo JV, Fox PT. 2005. A comparison of label-based review and ALE meta-analysis in the Stroop task. Hum Brain Mapp. 25:6–21.
Luk CH, Wallis JD. 2009. Dynamic encoding of responses and outcomes by neurons in medial prefrontal cortex. J Neurosci. 29:7526–7539.
Matelli M, Luppino G, Rizzolatti G. 1991. Architecture of superior and mesial area 6 and the adjacent cingulate cortex in the macaque monkey. J Comp Neurol. 311:445–462.
Matsumoto M, Matsumoto K, Abe H, Tanaka K. 2007. Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci. 10:647–656.
Mies GW, van der Molen MW, Smits M, Hengeveld MW, van der Veen FM. 2011. The anterior cingulate cortex responds differently to the validity and valence of feedback in a time-estimation task. Neuroimage. 56:2321–2328.
Misra G, Coombes SA. 2014. Neuroimaging evidence of motor control and pain processing in the human midcingulate cortex. Cereb Cortex. doi:10.1093/cercor/bhu001.
Mitz AR, Godschalk M. 1989. Eye-movement representation in the frontal lobe of rhesus monkeys. Neurosci Lett. 106:157–162.
Montague PR, Hyman SE, Cohen JD. 2004. Computational roles for dopamine in behavioural control. Nature. 431:760–767.
Morecraft RJ, McNeal DW, Stilwell-Morecraft KS, Gedney M, Ge J, Schroeder CM, van Hoesen GW. 2007. Amygdala interconnections with the cingulate motor cortex in the rhesus monkey. J Comp Neurol. 500:134–165.
Morecraft RJ, Schroeder CM, Keifer J. 1996. Organization of face representation in the cingulate cortex of the rhesus monkey. Neuroreport. 7:1343–1348.
Morecraft RJ, Stilwell-Morecraft KS, Rossing WR. 2004. The motor cortex and facial expression: new insights from neuroscience. Neurologist. 10:235–249.
Morita T, Tanabe HC, Sasaki AT, Shimada K, Kakigi R, Sadato N. 2014. The anterior insular and anterior cingulate cortices in emotional processing for self-face recognition. Soc Cogn Affect Neurosci.
Nachev P. 2011. The blind executive. Neuroimage. 57:312–313.
Palomero-Gallagher N, Mohlberg H, Zilles K, Vogt B. 2008. Cytology and receptor architecture of human anterior cingulate cortex. J Comp Neurol. 508:906–926.
Palomero-Gallagher N, Vogt BA, Schleicher A, Mayberg HS, Zilles K. 2009. Receptor architecture of human cingulate cortex: evaluation of the four-region neurobiological model. Hum Brain Mapp. 30:2336–2355.
Paus T. 2001. Primate anterior cingulate cortex: where motor control, drive and cognition interface. Nat Rev Neurosci. 2:417–424.
Paus T, Tomaiuolo F, Otaky N, MacDonald D, Petrides M, Atlas J, Morris R, Evans AC. 1996. Human cingulate and paracingulate sulci: pattern, variability, asymmetry, and probabilistic map. Cereb Cortex. 6:207–214.
Paxinos G, Huang XF, Petrides M, Toga A. 2009. The Rhesus monkey brain in stereotaxic coordinates. 2nd ed. New York: Academic Press.
Petrides M. 2012. The human cerebral cortex. An MRI atlas of the sulci and gyri in MNI stereotaxic space. London: Academic Press. p. 168.
Petrides M, Pandya DN. 1994. Comparative architectonic analysis of the human and the macaque frontal cortex. In: Boller F, Grafman J, editors. Handbook of neuropsychology. Amsterdam: Elsevier. p. 17–58.
Petrides M, Pandya DN. 1999. Dosrsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. Eur J Neurosci. 11:1011–1036.
Picard N, Strick PL. 1997. Activation on the medial wall during remembered sequences of reaching movements in monkeys. J Neurophysiol. 77:2197–2201.
Picard N, Strick PL. 1996. Motor areas of the medial wall: a review of their location and functional activation. Cereb Cortex. 6:342–353.
Quilodran R, Rothé M, Procyk E. 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron. 57(2):314–325.
Richter EO, Davis KD, Hamani C, Hutchison WD, Dostrovsky JO, Lozano AM. 2004. Cingulotomy for psychiatric disease: microelectrode guidance, a callosal reference system for documenting lesion location, and clinical results. Neurosurgery. 54:622–628; discussion 628–630.
Rushworth MF, Behrens TE, Rudebeck PH, Walton ME. 2007. Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. Trends Cogn Sci. 11:168–176.
Rushworth MF, Kolling N, Sallet J, Mars RB. 2012. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Curr Opin Neurobiol. 22:946–955.
Rushworth MF, Walton ME, Kennerley SW, Bannerman DM. 2004. Action sets and decisions in the medial frontal cortex. Trends Cogn Sci. 8:410–417.
Sallet J, Mars RB, Quilodran R, Procyk E, Petrides M, Rushworth M. 2011. Neuroanatomical bases of motivational and cognitive control: a focus on the medial and lateral prefrontal cortex. In: Mars RB, Sallet J, Rushworth MFS, Yeung N, editors. Neural basis of motivational and cognitive control. The MIT Press. p. 5–20.
Schall JD, Emeric EE. 2010. Conflict in cingulate cortex function between humans and macaque monkeys: more apparent than real. Brain Behav Evol. 75:237–238.

Segalowitz SJ, Dywan J. 2009. Individual differences and developmental change in the ERN response: implications for models of ACC function. Psychol Res. 73:857–870.
Seo H, Lee D. 2009. Behavioral and neural changes after gains and losses of conditioned reinforcers. J Neurosci. 29:3627–3641.
Seo H, Lee D. 2007. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci. 27:8366–8377.
Shackman AJ, Salomons TV, Slagter HA, Fox AS, Winter JJ, Davidson RJ. 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nat Rev Neurosci. 12:154–167.
Shenhav A, Botvinick MM, Cohen JD. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron. 79:217–240.
Shima K, Tanji J. 1998. Role for cingulate motor area cells in voluntary movement selection based on reward. Science. 282:1335–1338.
Tokuno H, Takada M, Nambu A, Inase M. 1997. Reevaluation of ipsilateral corticocortical inputs to the orofacial region of the primary motor cortex in the macaque monkey. J Comp Neurol. 389:34–48.
Vogt BA, editor. 2009a. Architecture, neurocytology and comparative organization of monkey and human cingulate cortices. In: Vogt BA, editor. Cingulate neurobiology and disease. New York: Oxford University Press. p. 65–93.
Vogt BA, editor. 2009b. Cingulate neurobiology and Disease. New York: Oxford University Press.
Vogt BA, editor. 2009c. Regions and subregions of the cingulate cortex. In: Vogt BA, editor. Cingulate neurobiology and disease. New York: Oxford University Press. p. 3–30.
Vogt BA, Nimchinsky EA, Vogt LJ, Hof PR. 1995. Human cingulate cortex: surface features, flat maps, and cytoarchitecture. J Comp Neurol. 359:490–506.
Vogt BA, Vogt L, Farber NB, Bush G. 2005. Architecture and neurocytology of monkey cingulate gyrus. J Comp Neurol. 485: 218–239.
Woolsey CN, Settlage PH, Meyer DR, Sencer W, Pinto Hamuy T, Travis AM. 1952. Patterns of localization in precentral and "supplementary" motor areas and their relation to the concept of a premotor area. Res Publ Assoc Res Nerv Ment Dis. 30:238–264.
Zilles K, Schlaug G, Matelli M, Luppino G, Schleicher A, Qü M, Dabringhaus A, Seitz R, Roland PE. 1995. Mapping of human and macaque sensorimotor areas by integrating architectonic, transmitter receptor, MRI and PET data. J Anat. 187:515–537.

Cerebral Cortex Advance Access published January 30, 2015

OXFORD

Cerebral Cortex, 2015, 1–18

doi: 10.1093/cercor/bhv006
Original Article

ORIGINAL ARTICLE

The Effects of Cognitive Control and Time on Frontal Beta Oscillations

Frederic M. Stoll[1,2], Charles R.E. Wilson[1,2], Maïlys C.M. Faraut[1,2], Julien Vezoli[1,2,3], Kenneth Knoblauch[1,2], and Emmanuel Procyk[1,2]

[1]INSERM U846, Stem Cell and Brain Research Institute, Bron 69500, France, [2]Université de Lyon, Lyon 1, UMR S-846, Lyon 69003, France, and [3]Current address: Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt D-60528, Germany

Address correspondence to Frederic M. Stoll, INSERM U846, Stem Cell and Brain Research Institute, 18 avenue du doyen Lépine, Bron 69500, France. Email: frederic.stoll@inserm.fr

F.M.S. and C.R.E.W. contributed equally to this work (co-first authors).

Abstract

Frontal beta oscillations are associated with top-down control mechanisms but also change over time during a task. It is unclear whether change over time represents another control function or a neural instantiation of vigilance decrements over time, the time-on-task effect. We investigated how frontal beta oscillations are modulated by cognitive control and time. We used frontal chronic electrocorticography in monkeys performing a trial-and-error task, comprising search and repetition phases. Specific beta oscillations in the delay period of each trial were modulated by task phase and adaptation to feedback. Beta oscillations in this same period showed a significant within-session change. These separate modulations of beta oscillations did not interact. Crucially, and in contrast to previous investigations, we examined modulations of beta around spontaneous pauses in work. After pauses, the beta power modulation was reset and the cognitive control effect was maintained. Cognitive performance was also maintained whereas behavioral signs of fatigue continued to increase. We propose that these beta oscillations reflect multiple factors contributing to the regulation of cognitive control. Due to the effect of pauses, the time-sensitive factor cannot be a neural correlate of time-on-task but may reflect attentional effort.

Key words: cognitive control, effort, monkey, reward, time-on-task

Introduction

Flexibility and efficiency of behavior involves monitoring performance and changing levels of cognitive control in order to properly adapt to the current context (Miller and Cohen 2001). Cognitive control refers to a set of computational mechanisms in prefrontal cortex (PFC), proposed from and supported by computational modeling findings, implementing "active maintenance of task-relevant context and top-down biasing of local competitive interactions that occur during processing" (Braver et al. 2002). Proponents of cognitive control hold that these computational-level mechanisms provide variable mediation of more familiar cognitive-level functions of working memory, inhibition, and attention, such that the latter are applied appropriately to a given task context. Tasks that require flexible and active maintenance of goals and the means to achieve them will demand cognitive control, and neurophysiological processes that support this cognitive control can be inferred from task phases requiring varying levels of control (Miller and Cohen 2001; Procyk and Goldman-Rakic 2006).

Much work on cognitive control has been in the domain of stimulus-response tasks, where cognitive control is used, for example, to overcome response conflict, due to ambiguous stimulus features, that degrades performance (Botvinick et al. 2001; Kerns et al. 2004). But, the mechanism can be applied to a broader

class of task, in which strategy rather than stimulus features guide performance, and therefore in which feedback will continually drive the level of cognitive control required (Quilodran et al. 2008). The current study seeks to understand neural mechanisms of cognitive control in such a context.

Neurophysiological markers of cognitive control have been recorded in frontal cortex (Siegel et al. 2012; Phillips et al. 2014). Amongst these, oscillations in local field potentials (LFPs) in the beta band (20–30 Hz) have been linked to top-down control of behavior (Buschman and Miller 2007; Siegel et al. 2012; Bastos et al. 2014), and to the formation of neural ensembles representing rules (Buschman et al. 2012). A potential role of these oscillations is in the interaction of frontal cortex with other related regions, as shown by enhanced frontoparietal beta-band coherence during free decisions compared with instructed decisions (Pesaran et al. 2008).

Control processes in PFC are necessary to provide flexibility, application of strategy, and maintenance of performance despite fatigue. A proposed mechanism for adjusting control in the face of fatigue is the process of attentional effort (Sarter et al. 2006). Attentional effort is a cognitive incentive that integrates explicit and implicit motivational forces (Sarter et al. 2006). Thus, a subject will increase attentional effort in order to maintain good performance, especially when highly motivated. This is a control function, and it might well be considered to result from the computational mechanisms of cognitive control. The way in which such ideas of effort and motivation integrate into cognitive control models is the subject of current experimental and theoretical work (Shenhav et al. 2013; Kool and Botvinick 2014). The current study seeks to understand this relationship in the context of the modulation of beta oscillations by cognitive control.

Attentional effort might contribute to the compensation for the well-established effect of vigilance decrement over time, known as the time-on-task effect. This effect is variously described as leading to impairments in simple performance measures, such as increases in execution errors (Boksem et al. 2006) or response time, but also as causing reduced behavioral flexibility (Lorist et al. 2009) or reduction in cognitive control. Again, the oscillatory dynamic of large-scale networks at lower frequencies is implicated. Frontal and parietal oscillatory signals across a range of low frequencies (theta and alpha) increase in power with time-on-task (Boksem et al. 2005; Borghini et al. 2014). Notably, increasing power in beta frequencies has also been reported (Boksem et al. 2005).

The time-on-task effect is widespread and pervasive in cognitive tasks (Kahneman 1973). Attentional effort can be seen when subjects are motivated to work for a goal (Sarter et al. 2006). Time-on-task effects will lead to continual degradation of performance measures over time whereas attentional effort should maintain goal-directed performance measures and cognitive control when it is applied (Sarter et al. 2006; Boksem and Tops 2008). Hence, these 2 effects may be thought to act in opposition, and in this case, they will both change over time. Changes in neurophysiological markers over time, such as beta oscillations, might therefore seem to correlate with both effects, and separating them out to understand the role of the marker requires careful analysis.

In this study, we examine the relationship between beta oscillations and varying levels of cognitive control, attentional effort, and behavioral changes occurring during time-on-task decrements. We seek to resolve whether each of these factors can be related to changes in beta oscillations. We use chronic ECoG surface recordings in macaque monkeys carrying out a test of cognitive control in conditions that allow us to simultaneously measure time-on-task decrements. The behavioral task requires regular changes in levels of cognitive control (Procyk and Goldman-Rakic 2006), whilst within-session changes in recordings allow us to observe time-on-task decrement or attentional effort effects. Using this approach and mixed-effects statistical modeling, we show the first clear observations of the neurophysiological impact of prolonged within-session work in monkeys with a trial-by-trial resolution.

By contrasting execution with cognitive errors and, uniquely to our knowledge, by looking at the effect of spontaneous pauses in work, we demonstrate that high beta oscillations in the preparatory period at the start of trials are separately modulated by multiple factors acting within-session and in relation to behavioral control. Crucially, pauses in work temporarily reset within-session effects, while cognitive performance is maintained and cognitive control continually represented in the beta oscillations. We propose that such pauses may provide a break in the attentional effort required or applied, yet time-on-task continues to increase despite the pause. These data hence suggest how modulations of the same beta oscillations can reflect more than one single process.

Materials and Methods

Subjects and Materials

Two rhesus monkeys (*Macaca mulatta*), 1 female and 1 male, weighing 7 and 8.5 kg (monkeys R and S, respectively) were used in this study. Ethical permission was provided by the local ethical committee "Comité d'Éthique Lyonnais pour les Neurosciences Expérimentales," CELYNE, C2EA #42, under reference C2EA42-11-11-0402-004. Animal care was in accordance with European Community Council Directive (2010) (Ministère de l'Agriculture et de la Forêt), and all procedures were designed with reference to the recommendations of the Weatherall report, "The use of non-human primates in research." Laboratory authorization was provided by the "Préfet de la Région Rhône-Alpes" and the "Directeur départemental de la protection des populations" under Permit Number: #A690290402.

Monkeys were trained to perform a problem-solving task (PST). Animals were seated in a primate chair (Crist Instrument Co.) in front of a tangent touch-screen monitor (Microtouch System). An open-window in front of the chair allowed them to use their preferred hand to interact with the screen (monkey R, left-handed; monkey S, right-handed). The position and accuracy of each touch was recorded on a computer, which also controlled the presentation of visual stimuli via the monitor (CORTEX software, NIMH Laboratory of Neuropsychology). During the behavioral task, eye movements were monitored using an Iscan infrared system (Iscan, Inc.). Electrophysiological data were recorded using an Alpha-Omega multichannel system (AlphaOmega engineering).

Behavioral Tasks

Principles of Cognitive Control Tasks

The task employed here, in common with much of the cognitive control literature, contains 2 phases, which vary in their cognitive control demands. The aim of the experiment is to make comparable recordings from comparable trials in which the only variation is the level of cognitive control currently being employed. One phase ("Search") demands higher cognitive control, in that a greater level of mediation of goals, memory, attention, and rules is required. The other phase ("Repetition") makes simpler demands. The contrast of these 2 phases provides the means to detect cognitive control processes in the neurophysiological data.

Problem-Solving Task with 4 Targets (PST4)

Monkeys sought by trial and error the correct target from a choice of four, the search phase. Having found the rewarded target, they were allowed to repeat the discovered rewarded choice 3 times the repetition phase, before being instructed to search again.

Each trial, whether in the search or repetition phase, followed an identical format. Monkeys initiated the trial by touching and holding a lever, represented by a central gray triangle. A fixation point appeared and animals had to fixate it with their gaze. After a delay period of 1400 ms, 4 gray target circles were displayed at different locations, on the upper side of a circular axis (Fig. 1*A*). At the onset of targets (ON signal), monkeys had to make a saccade toward a selected target and fixate it during a random delay between 400 and 800 ms (steps of 200 ms). At this point, all targets turned from gray to white, providing the GO signal following which monkeys were permitted to touch the target already chosen by fixation. After a random delay interval of 600–1200 ms (steps of 200 ms), a visual feedback stimulus was shown to the monkey for 800 ms. Feedback consisted of horizontal (correct) or vertical (incorrect) rectangles, in the same location and of the same color and luminosity as the circular targets. If the choice was incorrect (negative visual feedback and no reward), the monkey could select another target in the following trial and so on until the solution was discovered (search phase). After discovering the correct target, the animal was allowed to repeat the correct choice 3 times (repetition phase) (Fig. 1*A* and *B*). Correct responses were rewarded after the feedback with a 1- or 1.8-mL pulse of fruit juice. Successful discovery of a rewarded stimulus and repetition of that response 3 times is hereafter termed a problem. After the completion of a given problem, a signal-to-change was displayed on screen, consisting of the negative feedback stimulus flashed 3 times. A new correct target was pseudo-randomly selected and the monkey reentered the search phase. The trial after the signal-to-change is referred to as a switch trial.

Problem-Solving Task with 2 Targets (PST2)

PST2 task was identical to the PST4 task, with the sole exception that there were only 2 stimuli presented throughout, randomly selected from the 4 possible stimuli used in PST4. The feedback and signal to change stimuli likewise had only 2 stimuli, those corresponding to the 2 targets used.

The simple effect was that the search phase was easier for the monkeys, as the rewarded target could be found among fewer possibilities. The repetition phase was unchanged, requiring 3 correct responses. All other delays and procedures were identical. PST2 and PST4 problems were presented in the same sessions, in a pseudorandom manner.

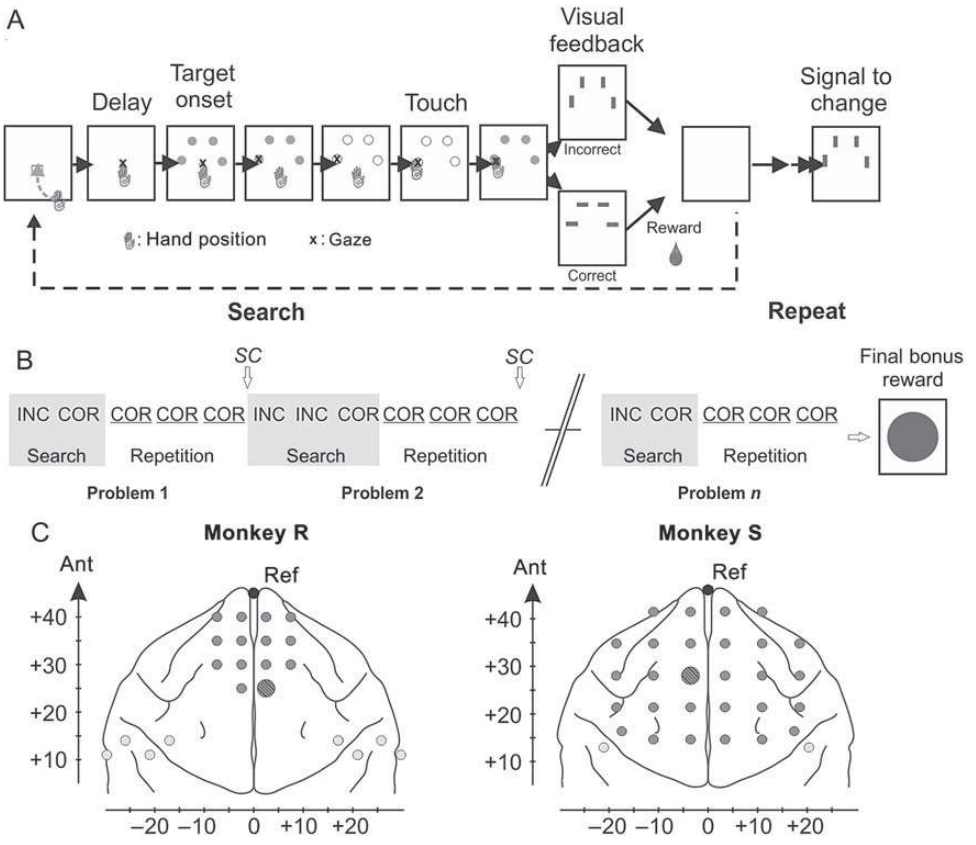

Figure 1. Problem-solving task and electrode location. (*A*) The events of an individual trial of the PST4. The task consists in seeking by trial and error which target was rewarded from a choice of 4 and then in repeating this correct choice 3 times. Frames represent the successive events for each trial during a problem. (*B*) A sample of successive problems during the PST4, extracted from a testing session. In the first problem, after making an incorrect selection (INC), the monkey discovered the correct target (COR), ending the search phase. After repeating this choice 3 times (repetition phase), a signal (SC) was presented to the animal, which indicated that a new target was going to be rewarded. After the monkey completed a fixed number of problems, a large salient green circle was presented on the screen and a final large bonus reward was delivered. (*C*) Positions of transcranial electrodes for each monkey. Electrodes (gray/white dots) are reported on stereotaxic coordinates in millimeters and displayed over a standard reconstructed brain surface. Black dots represent the location of reference electrodes in both monkeys. Dark gray dots represent the grid used in the present study. Larger dashed gray dots correspond to the position of the 2 electrodes used for mixed-model selection.

Final Reward

In order to motivate and maintain performance at a stable level throughout each daily session, animals were asked to complete a fixed number of problems each day ($n = 120$ and 60 problems for monkey R and S, respectively). Upon successfully completing this number of problems (average session duration ± sd, 132.5 ± 18.1 and 89.46 ± 29.2 min for monkey R and S, respectively), a salient green signal was displayed on the screen and monkeys received a large reward bonus (20–30 mL of fruit juice, calculated on the effectiveness in motivating the monkey). In addition, monkeys who had completed their session to this bonus received a reward of fruit immediately upon returning to the home-cage. Data in this study are derived only from sessions when the monkeys successfully completed the requested number of trials and received this juice bonus (86.4% and 59.8% of sessions for monkeys R and S, respectively).

Surgical Procedures

Surgical procedures were performed under aseptic conditions. The monkey was sedated on the morning of surgery with both ketamine (10 mg/kg) and xylazine (0.5 mg/kg) following preanesthetic treatment with glycopyrrolate (0.006 mg/kg). Once sedated, the monkey was given antibiotic (amoxicillin, 8.75 mg/kg) for prophylaxis of infection, and a nonsteroidal anti-inflammatory (ketoprofen, 2 mg/kg) agent for analgesia. The head was shaved and an intravenous cannula put in place for intraoperative delivery of fluids (sterile saline drip, 5 mL/h/kg). The monkey was intubated, placed on isoflurane anesthesia (0.5–2.75%, to effect, in an O_2 and NO_2 mix), and then mechanically ventilated. Heating blankets allowed maintenance of normal body temperature during surgery. Heart rate, oxygen saturation of hemoglobin, expired CO_2, body temperature, and respiration rate were monitored continuously throughout surgery.

Each animal was implanted with a head-holder (Crist Instrument Co.) and intracranial electrodes. Both the placement of the electrodes and their depth were determined from separately acquired structural MRI images of each monkey. Using stereotaxic guidance, holes were drilled through the skull and then stainless steel surgical screws (Synthes) were fixed into the holes, with the aim at each site of advancing the screw through the thickness of the bone to rest on the dura mater. Each electrode was connected to a standard connector constructed in-house. The ensemble was then anchored with dental acrylic to the head-holder. For monkey R, a grid of 14 electrodes spaced by 5 mm was implanted throughout the frontal cortex, and 8 electrodes were fixed over sensorimotor cortex near the central sulcus in a later surgery (Fig. 1C, left panel). Monkey S was implanted in a single procedure, and a larger grid of 31 electrodes, spaced by 7 mm, covered the frontal and sensorimotor cortex (Fig. 1C, right panel). For both monkeys, the last electrode serving as reference was screwed into the bone of the thick brow of the monkey on the midline anterior to the frontal grid. The present investigation concerns the grid of electrodes placed over the PFC for both monkeys (electrodes in gray in Fig. 1C).

Behavioral Analysis

When discussing performance on the tasks, we distinguish between "execution performance" and "cognitive performance." Cognitive performance measures the optimality of the choices made on the task itself. This therefore assesses the success of the monkey in completing problems and advancing toward the important final reward and fruit reward. Execution performance measures the successful completion of the trial itself, regardless of outcome, and so is measured by testing accurate touching of the screen, correct fixation of the stimuli, and reaction times (RT).

Cognitive performance was measured by 2 independent measures of nonoptimal choice. It is important to note that nonoptimal choice has a different definition in the 2 phases. When the monkey is searching, it is optimal to make some errors up to once per stimulus (this is a trial-and-error task), but nonoptimal to repeat those incorrect responses. Hence, in the search phase, the nonoptimal choice measure was the proportion of trials on which the monkey repeated an incorrect choice already made during that search (i.e., this is a measure of perseverative errors). In contrast, during the repetition phase, the monkey has found the correct response, and optimal performance means no errors. Hence, the nonoptimal measure is simply the proportion of incorrect choices during the repetition. A further measure of behavior tested whether monkeys re-initialized their search at the start of a new problem by using the signal-to-change. A monkey that successfully uses the signal-to-change should immediately change chosen target, because each new problem has a new pseudo-randomly assigned correct target. The measure is simply the percentage of cases where the choice is different before and after the signal-to-change.

Execution errors, which are preemptive touch responses or breaks of fixation, were also computed throughout the session. Motivational parameters, such as the number of trials not initiated by the animal over the total number of trials presented (i.e., "no start" trials), and the number and length of pauses in work during each session, were recorded.

RT (i.e., time between the appearance of white targets and the release of the lever) and movement times (MT, i.e., time of arm movements from lever to target) were computed on each trial.

We qualitatively investigated changes in behavioral parameters as a function of session progress by dividing each recording session into 3 groups of an equal number of trials. These 3 bins were made without taking into account any missed trials in which the animals did not initiate within a given start time, nor did it consider trials following pauses in work (see below for details on pause definition and analysis). PST2 and PST4 trials were pooled together for execution but not cognitive measures.

Electrophysiological Data Processing

All electrodes were referenced to the most frontal reference electrode (Fig. 1C). The signal from each electrode was amplified and filtered (1–250 Hz) and digitized at 781.25 Hz. Data analysis was performed off-line with FieldTrip toolbox (Oostenveld et al. 2011) and homemade Matlab scripts (Matlab, The MathWork, Inc.). Movement artifacts were removed by decomposing ECoG recordings with an independent component analysis, using the logistic infomax algorithm (Bell and Sejnowski 1995). Data were analyzed in the time-frequency (TF) domain by convolution with complex Gaussian Morlet's wavelets with a ratio $f/\delta f$ of 12. Single-trial TF analysis was aligned to the target onset (ON signal) for analysis of the delay period, and on the touch of the selected target for analysis of the post-touch period. The continuous ECoG data were epoched from −2500 to 2000 ms (by steps of 10 ms), and the power of each frequency ranging from 5 to 40 in 0.5-Hz steps was computed. Inspection of power spectrum density representations revealed different oscillatory activities, used to define frequencies of interest. Oscillations in the beta range (beta 1: 15–18 Hz and beta 2: 20–24 Hz for monkey R; beta 1: 10–18 Hz and beta 2: 24–32 Hz for monkey S) were analyzed and averaged over the delay period (−1200 to −200 ms before the ON

signal) and the post-touch period (0–600 ms after the target touch). This procedure led to a mean beta power for each trial during the 2 periods of interest. We focused on the delay period in this task in particular, because this is the period where the monkey is integrating feedback information from the previous trial with preparation for the upcoming trial, and therefore likely to be employing cognitive control (Procyk and Goldman-Rakic 2006). It should be noted that we consider the delay period for all trials in these analyses. This includes the first trial in each problem, even though the monkey is yet to receive feedback within that problem. This is because the signal-to-change acts as a feedback to initiate the search period, and so this initial delay is part of the search. Previous work from our laboratory on the same task has confirmed that monkeys show behavioral and neural evidence of entering into search immediately after the signal-to-change (Quilodran et al. 2008; Khamassi et al. 2014). Indeed, the response to the signal-to-change along with other forms of feedback is explicitly considered in the "response to feedback" analysis.

Data were acquired in 36 and 33 sessions (days) in monkey S and R, respectively. We rejected from our analysis 2 sessions (one for each monkey) for which data observation revealed abnormal distributions of single-trial beta powers. Moreover, the 2 most posterior electrodes in monkey S were rejected due to high signal variance. Trials with execution errors (38.60 ± 6.02% and 22.50 ± 3.48% for monkeys R and S, respectively) were included in our analysis if they occurred after the target onset, except for statistics including RT as covariate (see mixed-effects models below).

Measuring Cognitive Control Processes

As laid out in Introduction, cognitive control is a computational process rather than a psychological function, and so there is no direct behavioral measure of cognitive control. Rather, we induce variable use of cognitive control and then show how neurophysiological processes, in our case beta oscillations, are modulated with cognitive control demands. The current task is well established in this regard (Procyk and Goldman-Rakic 2006; Quilodran et al. 2008; Rothé et al. 2011; Khamassi et al. 2014; Procyk et al. 2014). The search period, during which the monkey is maintaining the goal, information on prior trials, and its progress, requires high levels of cognitive control. The repetition period, during which a single response must be maintained, requires minimal cognitive control. Frontal neurophysiological markers that vary significantly between search and repetition are therefore assumed to represent changes in cognitive control. Following our previous work, the initial analyses below employ this search–repetition contrast as the index of cognitive control. Specifically, we contrast the beta in search and repetition trials in the "phase" model.

Changes in cognitive control can also be observed following certain types of feedback that engender a change in behavior or goal (Botvinick et al. 2001; Kerns et al. 2004). For example, an incorrect choice or signal-to-change should induce greater cognitive control than a correct choice. As such we also analyze the same data in a "response-to-feedback" model. Here, we contrast the beta power on the trial "after" the monkey has received each type of feedback to show how reactive cognitive control is represented in the oscillations.

In some tasks, cognitive control is expected to lead directly to a subsequent behavioral measure. For example, in a Stroop task, correct application of cognitive control should lead to a slowing in reaction time and subsequent correct response. Here stimulus, response, and control are tightly coupled. Because the PST task is a trial-and-error task in which response is not directly determined by stimulus, this coupling should be less strong or even absent, whereas cognitive control is still employed in response to informative feedback. We nevertheless test for this form of "predictive" cognitive control using the predictive model.

Therefore, the following modeling analyses use several statistical models to capture the variance created in beta oscillations during the task, and test for different behavioral explanations.

Mixed-Effects Models

We observed modulations of beta power with task phase, within-session, and around pauses in work. To evaluate those modulations properly, trial-by-trial beta power measures were fitted with Linear Mixed-Effects models (Pinheiro and Bates 2000; Zuur et al. 2009). Such models allow us to analyze hierarchically organized data and to explicitly model variance inherent to repeated measures. In our data set, several sessions of recordings were used to analyze the within-session effect in each monkey. Characteristics such as the slope of power change over time could vary from one day to another. The random-effects terms in these models are specifically useful to capture this sort of variation. Meanwhile, the fixed effect can separately capture the presence of the slope in of itself. We produced models with 2 sets of fixed effects, the Phase model and the Response-to-feedback model. The data in the 2 models are the same; they simply treat the behavioral factors differently, allowing us to capture the variance in beta in different ways.

Mixed models are of the form $Y_i = \beta.X_i + b.Z_i + e_i$ where X and Z correspond to fixed- and random-effects design matrices, respectively, and e the random variations for each day i.

All statistical procedures were performed using R (R Development Core Team 2008, R foundation for Statistical computing) and the relevant packages (nlme, MASS).

Data Transformations

The "Phase" model included the following factors (and levels): Session, Phase (search/repetition), Task (PST2/PST4), Trial-type (Break, Incorrect, Correct), and the covariates "time" (time of target onset from the start of session, this is therefore within-session time) and RT. The "response-to-feedback" model is identical with the exception of phase and trial-type factors. These 2 factors are replaced by a previous trial feedback factor (PFB) referred to as PFB (Break, Incorrect, Correct, Switch), indicating the feedback that the monkey has just received. Here, the switch case refers to trials after a signal-to-change. Note that the first trial of the session and the first trial after the pause (where there was one) had to be removed from this analysis, as these trials do not follow any meaningful feedback.

The dependent variable Beta (trial-by-trial beta power measured in the time window of interest) was tested in a linear model to evaluate the need for a power transformation, that is, in particular to improve the "normality" of the data distribution. We computed and examined the profile log-likelihoods for the parameter of the Box–Cox power transformation (function boxcox in package MASS), which revealed the need to log transform the data before fitting a linear model adequately for both monkeys. Hence, log(Beta) was used as dependent variable in the following analyses.

In order to bring model coefficients within relatively similar ranges of values, "time" values were transformed into hours. Time values were then centered on the average time over all sessions, by simply subtracting each time value from the average

time. RT were also centered to the mean RT value. This was used in order to reduce correlations between the intercept and slope estimates.

Model Selection

Models were first selected using data acquired on 1 test electrode in each monkey and then applied to all electrodes (see details later). The test electrodes were selected based on observation of a clear peak of Beta power on the power spectrum density on the contralateral side, using similar sites in both monkeys (dashed gray disks in Fig. 1C).

In a first series of analyses, we evaluated the effect of the above-mentioned covariates and factors on Beta power measured in completed trials (trials in which monkeys touched a target). In this case, we focused on all data acquired at the beginning of each session, before any pause made by the monkey (see below for a definition of pauses). Because our main objective was to test the effect of time, "time" was used in the initial full Phase model in interaction with each independent variable:

$$\begin{aligned}\log(\text{Beta}) = {} & (\beta_0 + b_{0,S}) + \beta_{\phi}\text{Phase} + \beta_{Tr}\text{Trial-type} + \beta_{Tk}\text{Task} \\ & + \beta_{\log(RT)}\log(\text{RT}) + (\beta_t + b_{1,S})t + (\beta_{t:\phi} + b_{1,S})(t : \text{Phase}) \\ & + (\beta_{t:Tk} + b_{1,S})(t : \text{Task}) + (\beta_{t:Tr} + b_{1,S})(t : \text{Trial-type}) \\ & + (\beta_{t:\log(RT)} + b_{1,S})(t : \log(\text{RT})) + \varepsilon,\end{aligned} \tag{1}$$

where the β_i are fixed-effects and the b_j random-effect coefficients, each of the latter assumed to be distributed as $N(0; s_j^2)$. The subscripts are coded as S—session, ϕ—Phase, Tr—Trial-type, RT—reaction time, Tk—task, and t—time. The terms Phase, Trial-type, and Task are factors and RT and time are covariates. A key feature of the model is that the random-effects capture variation across the multiple sessions. Specifically, they show the influence of random effects on the intercept and the slope of the log(Beta) change as a function of time. A second key feature is that the interaction of the factors with time will correspond to differences in slope across the factor levels.

The procedure to select the most appropriate model consisted in evaluating each component, starting from the most complete model.

The first step evaluated random effects by comparing models with and without specific random effect terms. Models were selected using AIC and Log Likelihood ratio tests ($P < 0.05$). We evaluated the random effects in a model in which all orders of interactions of the fixed effects were retained. We found that random slopes and intercepts associated with the Sessions needed to be retained (L.ratio test, $P < 0.0001$ in both monkeys). As such, the random effects were retained in all subsequent models.

The second step evaluated the contribution of fixed effects. We used the drop 1 function, repeatedly testing the effect of dropping the highest-order interaction fixed-effect term on the fit (Zuur et al 2008). Again, models were selected using AIC, and changes in AIC between models were tested using a chi-square test ($P < 0.05$). The principle of model selection was identical for Phase and Response-to-feedback models.

The selected model was fitted on all electrodes by incorporating the factor electrode as an overall interaction term. Finally, to test the validity of applying on all electrodes the model selected on only one, we proceeded in 2 further steps. We first tested whether adding the higher-order interactions to the selected model would reveal significant interactions of any terms with any of the electrodes. A forward selection procedure (using the add1 function in R) was used that adds terms in a fashion that respects the marginality of lower-order terms and evaluates the significance of added terms using a likelihood ratio test. No significant interaction with electrodes means that no model more complicated than the selected model is needed to fit the data arising from all the electrodes. In a second step, we again took the selected model with the factor electrode incorporated into it and tested whether dropping terms using the drop 1 function (Venables and Ripley 2002) would reveal significant interactions. This is a backward elimination procedure that respects the marginality of lower-order terms, and in this case, its application might reveal differential effects of fixed parameters across electrodes. In these data, such interactions are to be expected, revealing a functional organization captured by the electrode map. If the first step revealed significant interactions, we examined the statistical table resulting from the more complicated model with factor electrode.

Effect of Pauses

In a second series of analyses, within-session effects were also studied around short pauses that monkeys made spontaneously in some sessions. In order to observe the evolution of oscillatory activity and behavioral parameters around these events, trial-by-trial measures were aligned on pauses. For statistical analyses, we included all trials completed up to the onset of the 4 targets. This therefore included trials with breaks in fixation or in lever holding before the monkey could touch a target. Hence, for these analyses, we excluded the covariate RT, as not all trials had a reaction time. The way in which the 2 monkeys worked and paused differed slightly. In principle, we chose sessions with a pause of several minutes after a sustained period of work at the start of the session (after 40 min of work in monkey R and 25 min in monkey S). Sessions could contain none, one, or more pauses (monkey R: none 18/33, one 13/33, more 2/33; monkey S: none 10/36, one 16/36, more 10/36). We focused our analyses on the data acquired before and after the first pause in a session. This therefore concerned 15 sessions in monkey R and 26 sessions in monkey S. Pauses included in the analyses were of 8.1 min on average for monkey R (minimum: 2.9 min, sd = 5.1) and of 13.5 min on average for monkey S (minimum: 6.2 min, sd = 5.7 min). We first considered all trials including execution errors.

The initial model for this analysis around pauses was chosen on the basis of the model selection carried out on the first model mentioned in Equation (1), the outcome of which is described in "Results." This allowed us to remove a number of interaction terms that had no significant influence. The Phase model selection was therefore initiated with:

$$\begin{aligned}\log(\text{Beta}) = {} & (\beta_0 + b_{0,S} + b_{0,S:Bk}) + \beta_{Tk}\text{Task} + \beta_{Tr}\text{Trial-type} \\ & + \beta_{\phi}\text{Phase} + \beta_{Bk}\text{Block} \\ & + \beta_{Bk:\phi}\text{Block:Phase} + (\beta_t + b_{t,S} + b_{t,S:Bk0})\text{time} \\ & + (\beta_{t,S:Bk} + b_{t,S} + b_{t,S:Bk1})\text{time} : \text{Block} + \varepsilon,\end{aligned} \tag{2}$$

where the β_i are fixed-effect and the b_j random-effect coefficients, each of the latter assumed to be distributed as $N(0; s_j^2)$. The subscripts are coded as above and with Bk for Block. Block is a factor with 2 levels (0, 1) representing the block of trials before (0) and after (1) the pause in work. Again it is important to note a couple of key features of the model. As in the model described in Equation (1), the random-effects capture variation across the multiple sessions, but also the variation in the relationship between the pre- and post- pause responses (time:Block interaction). An interaction between Block and Phase corresponds to a difference in

the intercept for different combinations of Block and Phase. An interaction of Block with time corresponds to a difference in slope between Block levels.

To test for the contribution of Reaction time to signal variance, we ran the same analyses on the subgroup of completed trials, i.e., including a touch on target.

Model Validation

Model validation was performed by checking that normalized residuals plotted against fitted values and factors did not show inhomogeneity or violations of independence.

The final best fitting model (see Results) was used to extract intercepts and slopes for each session and to provide global statistical evaluations. *P*-values obtained from Mixed models applied to each electrode were adjusted using Bonferroni correction.

Results

Behavior

It is important to recall the distinction between execution performance (successful completion of the trial regardless of outcome) and cognitive performance (optimal completion of the task itself).

Cognitive performance in the PST was stable across the different recording sessions (see Supplementary Fig. 1 for example). The monkeys had significant previous experience on this task prior to data collection for this study and should not be considered as being in a phase of learning about the task. Data presented in Figure 2 summarize behavior of the monkeys in the task. The low levels of nonoptimal choices (Fig. 2C) indicate that the monkeys were approaching optimality in the task, successfully completing search and repetition phases. Monkeys

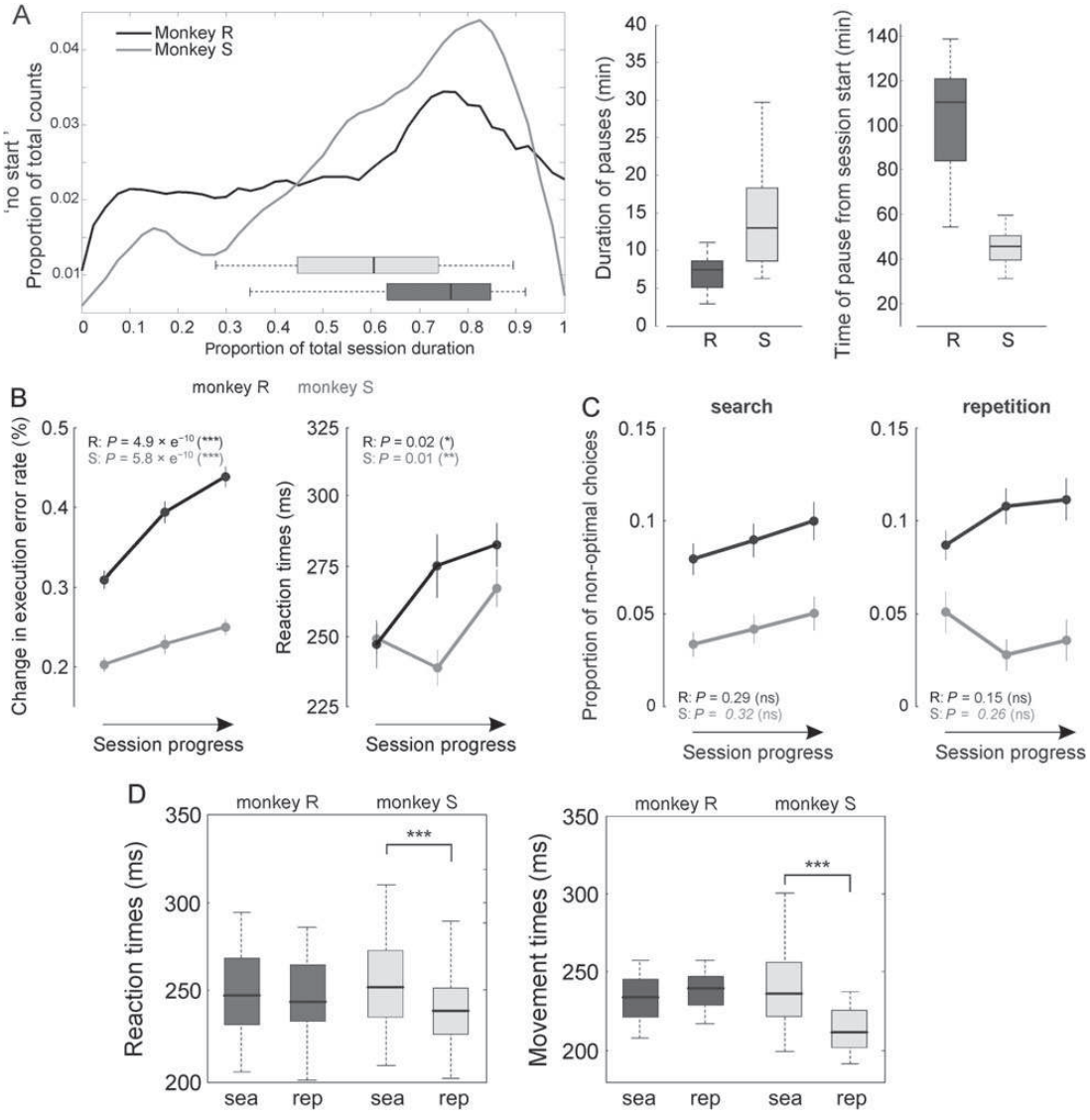

Figure 2. Behavioral effects within session and of cognitive control. (*A*) left panel, average evolution of the number of trials not initiated ("no start") during sessions for both monkeys. Data are normalized in time to the start and end of sessions. The boxplot below the curves shows the median and deviation of the normalized time of the first pause of work in the session, this being the pause considered for analysis, again for each monkey. Right panel, median and dispersion of pause duration and time of pause relative to session onset for monkeys R and S. (*B*) Evolution of execution error rates and RT (left and right plots, respectively) as the session progresses. Data are grouped into 3 bins of trials (see Materials and Methods for details). (*C*) Evolution of cognitive performance within-session, represented by the proportion of nonoptimal choices in the search (left panel) and repetition (right panel) phases. Black and gray lines show the average across days (+/−SEM) for monkeys R and S, respectively. *P*-values represent one-way ANOVA main effects (factor bin of trials). As the session progresses, cognitive performance remains stable, whereas execution performance worsens. (*D*) RT and MT for both monkeys in both phases of the PST4 task only. (ns) $P > 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

transitioned well from repetition back to search by taking into account the signal-to-change and changing their choice on the following trial (percentage of optimal transitions ± sd: monkey R: 94.19 ± 4.37%; monkey S: 75.72 ± 13.83%). Monkeys did show some execution errors as well as cognitive errors, and these were the starting point of our study of within-session changes.

Behavioral Modulations within Session

Figure 2A–C present the evolution of execution and cognitive performance within-session. First, monkeys tended to miss more trial initiations (i.e., not touching the lever when available) as time passed in the session, with a peak toward the end of the session, the end corresponding to the final bonus reward delivery (Fig. 2A left). In several sessions, the animals even stopped working for several minutes, making a voluntary pause in work (see Materials and Methods). During pauses, monkeys stayed at rest without initiating trials, neither fixating nor touching the screen. The median duration and time of occurrence of the first pause in a session varied in the 2 monkeys but, for both animals, it occurred most often during the second half of the session and largely contributed to the peak in no start trials (Fig. 2A right). Reluctance to initiate trials can be associated with motivational changes and fatigue, which contribute to time-on-task effects.

To precisely evaluate the effect of time, we then restricted behavioral analyses to the first block of trials (up to the first pause, or to the end of session when no pause occurred). For both monkeys, progress in the session was also characterized by a significant increase of execution error rates (one-way ANOVA, factor bin of trials, monkey R: $F_{2,93} = 27.21$, $P = 4.9\ e^{-10}$; monkey S: $F_{2,102} = 5.43$, $P = 5.8\ e^{-3}$) and RT (monkey R: $F_{2,93} = 4.01$, $P = 0.02$; monkey S: $F_{2,102} = 4.59$, $P = 0.01$) (Fig. 2B). These within-session effects are landmark effects of time-on-task.

Despite the drop in execution performance during the session, both monkeys succeeded in keeping stable cognitive performance as indicated by a constant proportion of nonoptimal choices during the search phase (monkey R: $F_{2,93} = 1.26$, $P = 0.29$; monkey S: $F_{2,102} = 1.15$, $P = 0.32$) and the repetition phase (monkey R: $F_{2,93} = 1.92$, $P = 0.15$; monkey S: $F_{2,102} = 1.34$, $P = 0.26$) in PST4 during the course of the session (Fig. 2C). Stable cognitive performance was also observed for PST2 trials ("search": monkey R: $F_{2,93} = 1.33$, $P = 0.27$ and monkey S: $F_{2,102} = 0.28$, $P = 0.75$; "repetition": monkey R: $F_{2,93} = 3.24$, $P = 0.04$ and monkey S: $F_{2,102} = 1.71$, $P = 0.18$).

RT were slightly different over all sessions when comparing task versions in monkey S (two-way ANOVA, factor PST2/PST4, monkey R: $F_{1,125} = 0.85$, $P = 0.36$; monkey S: $F_{1,137} = 4.5$, $P = 0.03$) and significantly higher in search than repetition phases for monkey S too (Fig. 2D left; two-way ANOVA, factor search/repetition, monkey R: $F_{1,125} = 0.13$, $P = 0.72$; monkey S: $F_{1,137} = 23.01$, $P = 4.1\ e^{-6}$). MT were similar between PST2 and PST4 for both monkeys (two-way ANOVA, factor PST2/PST4, monkey R: $F_{1,125} = 0.58$, $P = 0.45$; monkey S: $F_{1,137} = 1.81$, $P = 0.18$). MT were significantly higher in search than repetition for monkey S (Fig. 2D right; two-way ANOVA, factor search/repetition, monkey R: $F_{1,125} = 0.65$, $P = 0.42$; monkey S: $F_{1,137} = 47.52$, $P = 1.8\ e^{-10}$).

In summary, we observed that execution performance declined in both monkeys within-session but cognitive performance showed no within-session effect.

General Characteristics of Beta Oscillations

Recordings were made every day thanks to the chronically implanted array, and as we have already demonstrated, such recordings are stable across months (Vezoli and Procyk 2009). Supplementary Figure 1 underlines the stability of the recordings in this study over the individual sessions, showing that there is no practice-related modulation over time. TF analysis revealed beta oscillations during the execution of the PST (Fig. 3). Sustained beta oscillations were observed during the delay period (−1200 to −200 ms before the ON signal, Fig. 3A) when the monkey is fixating and holding touch on the screen. The power spectra revealed 2 identifiable peaks within the beta range for both monkeys (Fig. 3B). Although the average Beta power varied across sessions, the 2 peaks were clearly observed in most sessions (Supplementary Fig. 2). These 2 peaks are referred to here as Beta 1 and Beta 2 oscillations. Beta 1 and Beta 2 peaks were at 16 and 22 Hz for monkey R and at 14 and 28 Hz for monkey S (Fig. 3B). Both were located in a distributed frontal network, with a bias over the ipsilateral frontal cortex (Fig. 3C). This inter-individual difference in the peaks of power is worth noting. It is now clear that such individual differences in power spectra are to be expected (Buzsáki et al. 2013), and despite the different frequencies of these bands, we go on to show that equivalent peaks maintain equivalent properties (Fig. 3D and E). This property is further considered in the discussion.

Beta oscillations were also observed and modulated during the other periods of the trial. They were suppressed after target onset, when monkeys were allowed to make a saccade toward a selected target (Fig. 3A). This suppression [event-related beta desynchronization, ERD (Pfurtscheller and Lopes da Silva 1999)] remained until after the touch of the selected target. Beta oscillations reappeared after hand movement.

Inspection of data first revealed that delay-related beta oscillations were modulated between the search and repetition phases of the task suggesting an effect of cognitive control on beta power (Fig. 3D). This follows previous observations of prefrontal neuronal and evoked potential modulations using this task (Procyk et al. 2000; Procyk and Goldman-Rakic 2006; Quilodran et al. 2008; Vezoli and Procyk 2009).

As the within-session analysis revealed a time-on-task effect in execution performance (but notably not in cognitive performance), we wondered whether delay-related beta activity would show within-session variations correlated with time during the first block of trials. Indeed, as the session progressed, delay-related beta oscillations increased in power for both monkeys in particular for the Beta 2 oscillations (Fig. 3E). Therefore, a within-session effect on beta is present during the period of the trial in which beta also varies with cognitive control (see examples in Supplementary Fig. 3). This therefore potentially links the 2 phenomena, and we investigated this possibility using statistical modeling.

Within-Session and Cognitive Control Modulations of Beta Power

Phase Model

To characterize the modulation of beta oscillations between sessions, within sessions, and by different key elements of the task, we performed linear mixed-effects modeling (see Materials and methods). We first conducted an exploratory analysis on one selected electrode for each monkey (Fig. 1). As fixed-effects, the initial full model included Time within-session and RT as covariates and Phase, Task, and Trial-type as two- and three-level factors, respectively. Sessions were treated as a grouping factor for the random effects (see Materials and Methods). Analyses were performed separately on measures for Beta 2 and Beta 1. We selected a linear mixed-effects model to describe these data by starting from a full model and successively testing the effect of dropping the highest-order interaction fixed effect term on the fits. The

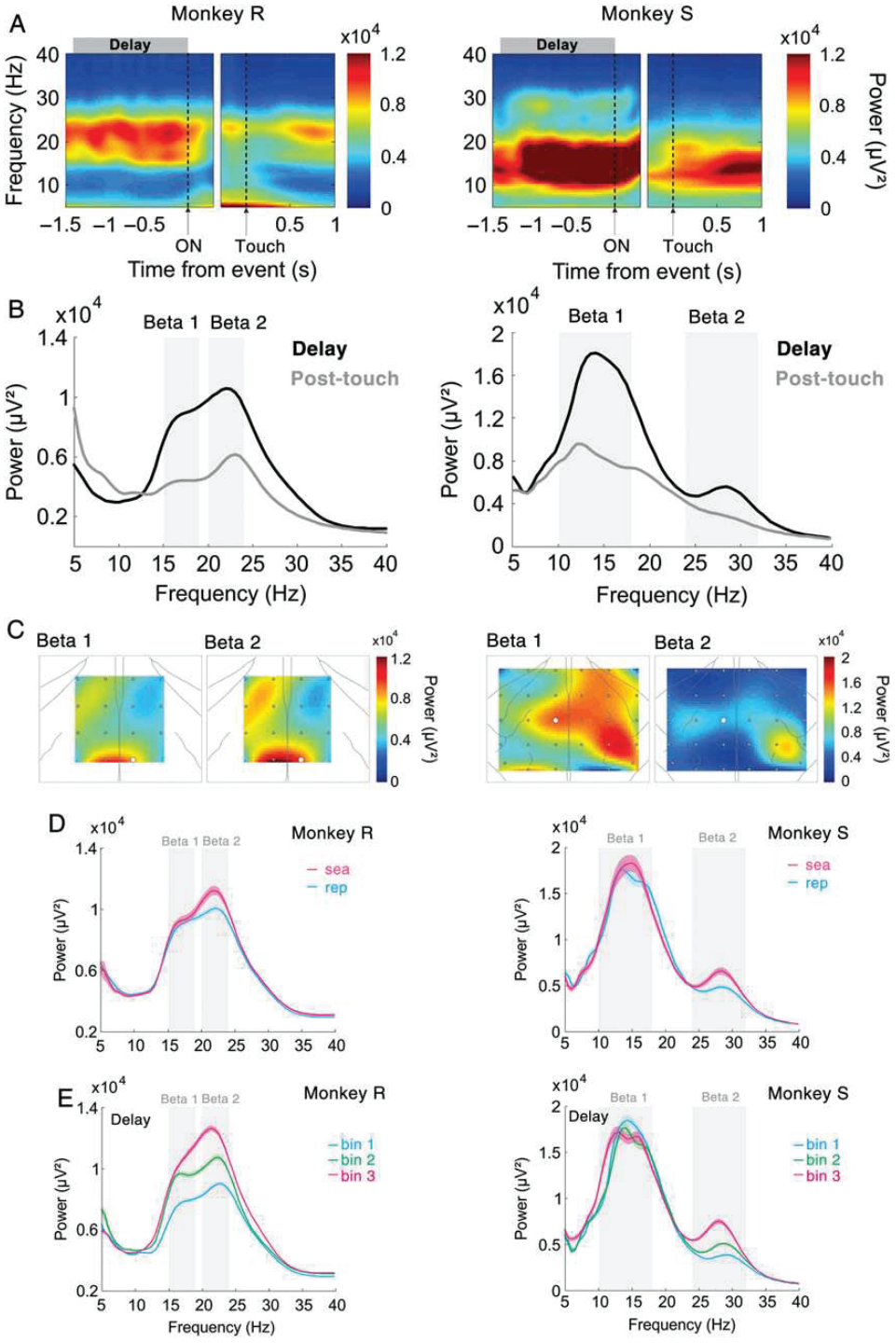

Figure 3. Description of beta oscillations. (*A*) Average TF representation of 2 contralateral electrodes (white dots in Panel C) aligned on target onset and target touch for both monkeys. The color scale shows the average power of oscillations. (*B*) Mean power spectra during the delay period (−1.2 to −0.2 s from ON) and post-touch period (0–0.6 s from Touch). We identified 2 beta frequency bands (beta 1 and beta 2) of interest from these plots (monkey R: 15–18 and 20–24 Hz; monkey S: 10–17 and 24–32 Hz). (*C*) Spatial representations of average beta power during the delay period. Note that monkey R is left-handed and monkey S right-handed. (*D*) Effect of task Phase (Search and Repeat) on the beta power. A power spectrum has been extracted from signal recorded in the delay for all trials in search (sea, magenta) and repetition (rep, blue). Note that the effect is most marked in the shaded region of beta 2. (*E*) Within-session effect on the overall power spectrum. A power spectrum has been extracted for signal recorded in the delay for all trials of the first block (before the first pause) separated in 3 terciles (bins 1, 2, and 3, see Materials and Methods). Again the effect is particularly clear in beta 2. All figures were derived from 1 representative session in both monkeys.

final model was therefore that model from which we were unable to remove any additional interactions because the remaining interactions had a significant effect on the fit.

Importantly, statistical modeling of Beta 2 showed that Time × Phase interaction was not significant (likelihood ratio test, monkey R: $P = 0.37$; monkey S: $P = 0.93$). There was no effect of RT or of

the Trial-type with time (nonsignificant interactions in both monkeys). A Task × time interaction was significant in monkey S ($P < 0.001$). There was no main effect of Task in both monkeys (monkey R: $P = 0.34$, monkey S: $P = 0.18$). RT had a significant contribution in 1 monkey (monkey S: $P = 0.11$, monkey R: $P < 0.001$). Since removing RT had no significant effect on the other conclusions and had to be removed for the second analysis on pause, we excluded this covariate for both monkeys. These tests therefore permitted us to reject these factors and interactions as having a significant contribution to the measure of Beta 2, and therefore to construct a final model.

Although no interaction was found between them, Time and Phase (search vs. Repetition) had each made a very significant contribution in both monkeys to the modulation of Beta 2 oscillations, and so these terms were retained in the final model. We therefore tested the influence of these terms within the selected model. The effect of Phase revealed a significant reduction of Beta 2 power in Repetition compared with Search (Wald statistics, Repetition vs. Search, monkey R: t-value =– 10.04, monkey S: t-value =– 4.5, $P < 0.0001$ in both monkeys). The slope coefficient for time was significant (Wald statistics, monkey R: t-value = 11.9, $P < 0.0001$; monkey S: t-value = 14.8, $P < 0.0001$) with Beta 2 power increasing within-session.

The same analyses for power in the Beta 1 range provided identical results for monkey R (Wald statistics, monkey R, Phase: t-value =– 14.2, $P < 0.0001$, Time: t-value = 16.2, $P < 0.0001$) but no effect of Time or Phase for monkey S (Wald statistics, monkey S, Phase: t-value =– 1.1 $P = 0.27$, Time: t-value = 1.1, $P = 0.27$). The Beta 1 range was very close to the Beta 2 range in monkey R (Fig. 3*B*), and the dissociation might have been impossible. Because of this confound and of the absence of within-session and phase effects in monkey S, we only focused on the Beta 2 range in the rest of the study (referred to as Beta from here on).

Thus, the model selection approach on chosen electrodes showed that Beta power increases within-session over time and is modulated by task Phase and Trial-type but that there is no interaction between the influences of these factors. Hence, the modulation of Beta oscillations with cognitive control and the modulation of Beta within-session are both significant, but these influences are statistically independent.

To evaluate the effects over all electrodes, we tested a model including the factors found significant in each monkey and fitted to Beta power acquired on each electrode. P-values were corrected for multiple testing (Bonferroni correction). The validity of applying the selected model to all electrodes was tested with a double step procedure, in particular regarding the potential interaction between time and Phase (see Materials and Methods). The validation revealed that no more complicated model was needed for monkey R and revealed in monkey S a weak interaction effect ($P = 0.02$) for time and phase produced by 3 electrodes. Overall, the procedure revealed that the selected model with no interaction between time, Phase, and Trial-type explained the great part of the variance in data over all electrodes. The second step of validation showed moreover some dissociation between electrodes as revealed in the analyses presented later.

Slope of fit, intercept, and P-values were then extracted from the model fit and used to create maps presented in Figures 4 and 5. Figure 4 shows the results for the modulation by Phase. The mean power of Beta oscillations was almost always higher during search trials than during repetition trials, as shown on the contrasts between average measures for each session (Fig. 4*A*). Across electrodes, and for both monkeys, the major effect was a reduction of Beta power in repetition compared with search, as revealed by a systematically lower intercept for the fit for repetition data compared with the fit for search data (the difference REP-SEA was negative in the majority of cases). Mapping the differences in intercept revealed a global unidirectional change in Beta between search and repetition with a slight ipsilateral hemispheric bias (Fig. 4*B*).

Changes in Beta power during the delay were systematic across sessions and for both monkeys. The topographic maps in Figure 5*A* show the within-session effect (average slope of fit) for each monkey for the Beta measured in the delay and after the touch. These maps and the data extracted for each electrode (Fig. 5*B*–*C*) revealed that the within-session effect during delay was strong and always positive, that is, Beta power increased with time. Beta changes across time (Fig. 5*A*) and between search and repetition phases (Fig. 4*B*) appeared somewhat dissociated in their topographic localization. Time-on-task variations appeared slightly more anterior and bilateral compared with search–repetition differences, and also more contralateral in monkey R.

Importantly, we tested whether the within-session variation of beta activity was specific to certain beta-related processes or reflected a general effect found in all recordings across the time of the session. We found that the effect was indeed specific to the delay period, as within-session modulations of beta power recorded after the touch were not consistent and much weaker for both monkeys. Figure 5 shows the results of mixed-effect models for each electrode for the Beta in Delay and after the touch (Post-Touch). The panels show that the effect of time (slope) was much stronger in delay, and not related to the initial Beta power (intercept) (Fig. 5*B*) and that the effect of time was always significant in delay and rarely in post-touch (Fig 5*C*). This rules out, in particular, a global effect of time on the entire spectrum (shift of the power law).

Predictive Model

To explicitly address the question of whether increased beta power in the delay leads to a subsequently more optimal outcome on that trial, we replaced the "Trial-type" factor with a behavioral outcome factor: "Optimal." This factor codes whether performance on the upcoming trial made optimal use of the feedback received on the previous trial. The factor therefore had 2 levels, optimal and nonoptimal, and the definition of optimality followed exactly that used in the behavioral analysis described in the Materials and Methods section. We applied the same model selection strategy to this model. In a model that retained terms for time and phase as before, subsequent optimality was also a significant predictor of delay beta power in monkey R ($P = 0.02$), but not in monkey S ($P = 0.6$). As such we are unable to claim that higher beta power is predictive of optimal responding, despite the fact that beta power strongly tracks task phase. To better understand the role of beta in this task, therefore, we tested whether there is a responsive rather than predictive link between beta power and cognitive control.

Response-to-Feedback Model

In order to capture the way in which beta oscillations responded to the changing cognitive control demands following different types of feedback, we also applied the Response-to-feedback model, using the same ECoG data but different behavior factors. We replaced the Phase and Trial-type factors with a single PFB, which described the feedback the monkey had received on the previous trial. Hence, the model tested how reactive cognitive control was expressed in beta oscillations following each feedback. This allows us in particular to test whether the incorrect feedback and signal-to-change do indeed induce increased cognitive control, as previously shown (Khamassi et al. 2014). The

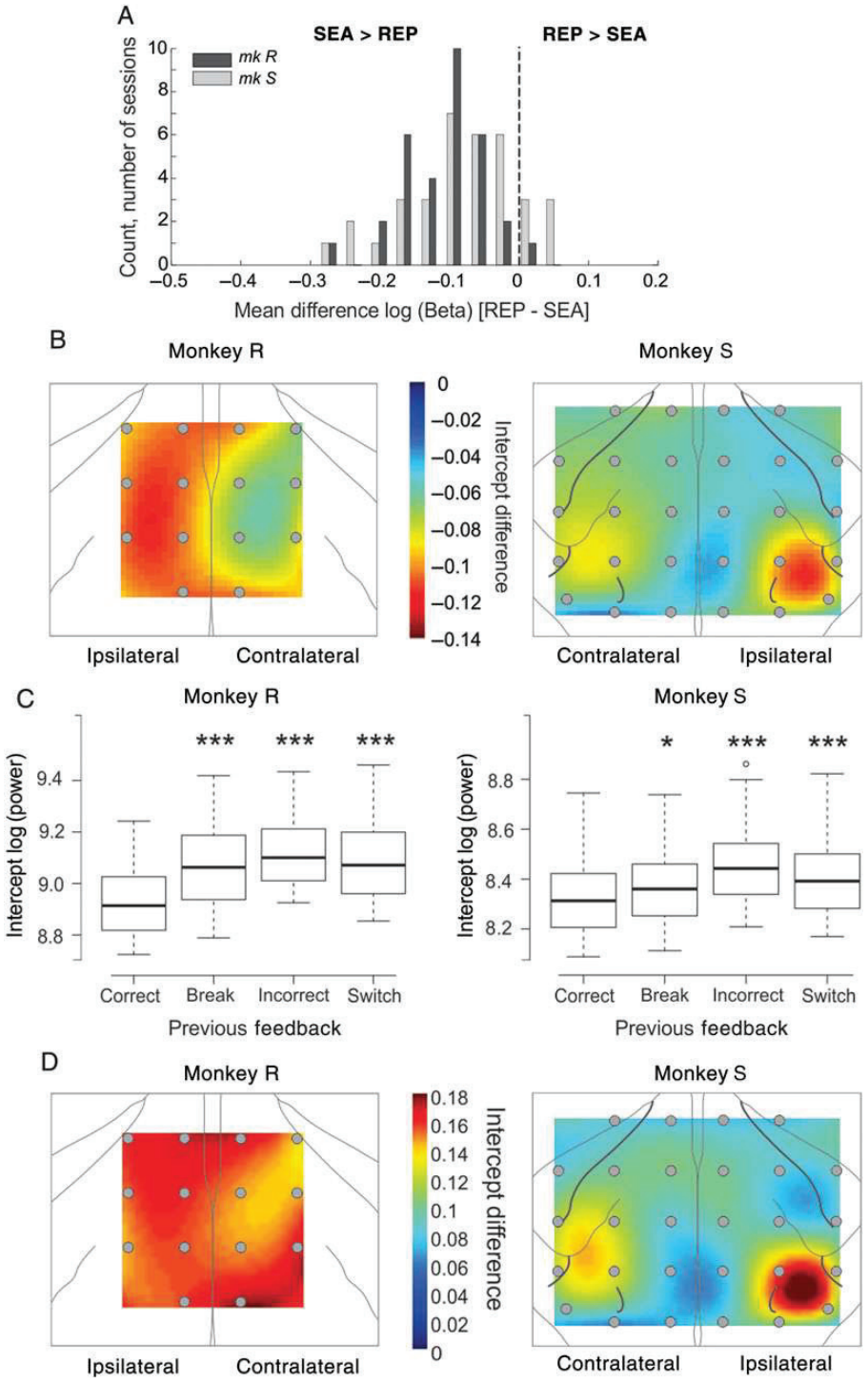

Figure 4. Modulation of delay beta oscillations by task Phase and Previous Feedback. (*A*) Distribution of the difference in mean log(Beta) calculated for each session and each monkey before the first pause. A negative difference indicates higher power in search than in repetition, so power decreases from search to repetition. (*B*) Spatial representations of the main effect of Phase for each monkey. The fixed effect of Phase is reported for each electrode and mapped onto the grid. The color scale represents the intercept difference between search and repetition (red is more negative, i.e., Beta in search is higher). Note that all electrodes show a significant effect between search and repetition. (*C*) Effect of previous feedback on intercept of log(Beta) as extracted from the Response-to-feedback model. Stars indicate level of significance compared with Correct feedback—in all cases, the Beta is increased compared with trials that follow correct feedback and so demand limited cognitive control. (*D*) Spatial representation of the difference between trials following Correct feedback and a Signal-to-change. The color grid represents the intercept difference Switch-Correct (positive values, i.e., Beta after Switch is higher). The effect is significant on all electrodes. Note that monkey R is left-handed and monkey S right-handed. (ns) $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

model was selected and tested in an identical fashion to the Phase model, on the Beta data in the delay period, and starting with the same selected electrodes.

A significant main effect of PFB was retained in both monkeys. Correct feedback should induce minimal cognitive control in the following trial, as the correct response need merely be repeated.

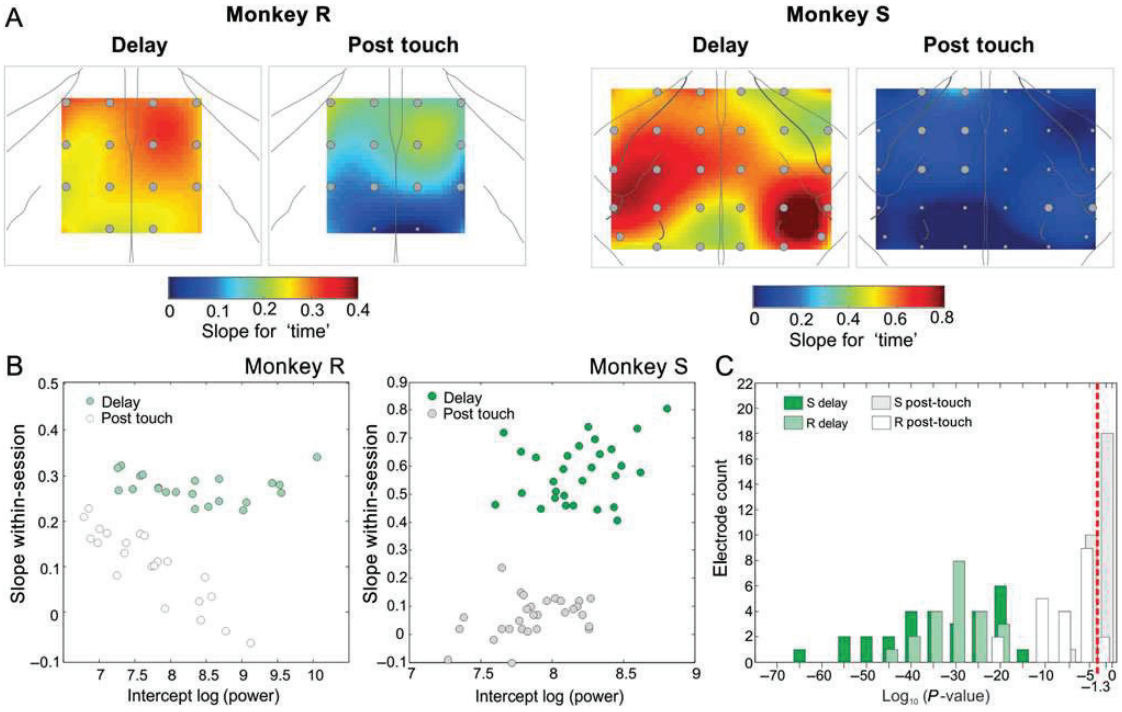

Figure 5. Changes in beta oscillations within-session. (*A*) Surface maps reconstructed with the slope of the main effect of time obtained from the linear mixed model, for monkeys R and S (left and right, respectively) and for both trial epochs (Delay and Post-Touch). Larger gray dots indicate electrodes with a significant slope. (*B*) Relationships between the level of Beta power and the within-session effect in the Delay (greens) and Post-Touch (white and gray) epochs for each monkey. The level of Beta power is estimated by the intercept, and the within-session effect is given by the slope of the regression. Note that the slope is always greater in the Delay period than the Post-Touch period. Also, there is no clear or stable relationship between beta power and slope, meaning that the size of the within-session effect does not depend on overall Beta power. (*C*) Histogram of the *P*-values from within-session slope effect for all electrodes measured for the beta power in delay (light and dark green) and post-touch (white and gray) epochs. Log of *P*-values are presented, with those above the statistical threshold ($P > 0.05$) represented on the left of the −1.3 limit (vertical dashed red line). Data are presented separately for the 2 monkeys (S and R).

In both monkeys, each of Incorrect, Switch, and Break feedback induced significantly increased Beta compared with Correct feedback (Fig. 4C, Wald statistics, in every case t-value > 6, $P < 0.0001$; excepting Correct vs. Break in monkey S: t-value = 2.1, $P = 0.04$). Hence, in all cases where feedback instructed a change of strategy or response, requiring the use of cognitive control, the Beta was increased significantly compared with the case of repeating the same correct response.

The main effect of time remained present in this model (Wald statistics, monkey R: t-value = 10.0, $P < 0.0001$; monkey S: t-value = 14.95, $P < 0.0001$). A time × PFB interaction also survived but only in monkey R ($P = 0.017$), and a main effect of Task survived in monkey S ($P = 0.0006$). RT made no contribution to the model following selection, and all other interactions were removed in the selection process.

We then applied the selected models to all electrodes. Intercepts and *P*-values were then extracted from the model fit and used to create Figure 4D. This figure shows the difference in log (Beta) between trials following a correct response and trials following a signal-to-change. There is clear evidence for reactive cognitive control being represented by Beta oscillations, with a significant increase in Beta on switch trials. The topography of this effect is comparable with the general cognitive control effect derived from the Phase model (Fig. 4B).

In summary, Beta power is high when cognitive control is required, both during cognitive control demanding phases of the task and immediately after feedback that required cognitive control. Beta power also increases within session, but the cognitive control and within-session effects do not interact. In addition, the behavior provided within-session effects in execution measures, but not in cognitive measures. This suggests these 2 effects, cognitive control and within-session, were independently expressed in the beta oscillations.

Delay-Related Beta Oscillations around Pauses

As noted earlier, the motivational state of monkeys decreased while execution errors and RTs increased as the session progressed. This is a time-on-task effect. In some sessions, the monkeys made voluntary pauses for a few minutes before re-initiation of the task, quite possibly due to the drop in motivation. During these pauses, monkeys stayed at rest without initiating trials, neither fixating nor touching the screen. The interpretation of these pauses in the context of the time-on-task literature is unclear. We studied the effect of pauses on all measures in order to understand their impact on control and time-on-task.

Sessions with pauses in work were isolated (see Materials and Methods for details and Fig. 2A Right panels) and used to further investigate within-session effects at the level of trial series. We refer to within-session changes over time but restricted to a single block of work (before or after a pause) as "within-block" effects. Only data recorded before and after the first pause until the next pause or end of session were considered (see Materials and Methods). For descriptive purposes, single-trial measures of power in delay were aligned to the start of the first pause in each selected session before averaging (Fig. 6), thus isolating 2 blocks of work in each session—one before and one after the pause. The pattern of changes in beta oscillations around pauses

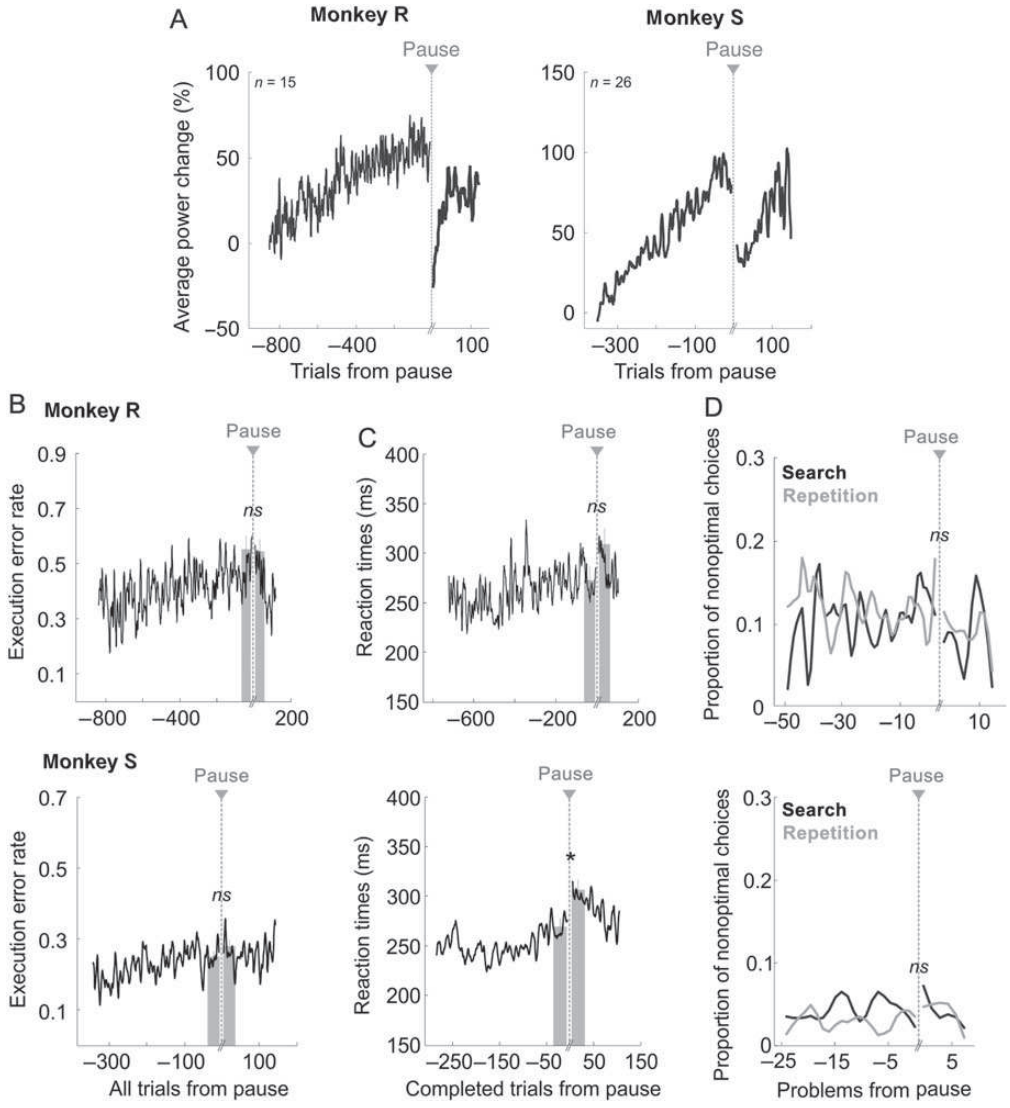

Figure 6. Electrophysiological and behavioral changes after pauses. (*A*) Average power change aligned to pauses in work during the delay period. Power of beta oscillations drops significantly then increases again after pauses (monkey R, $n = 15$; monkey S, $n = 26$ sessions; see Materials and Methods for pause detection). (*B–D*) Evolution of execution error rate (*B*), reaction time (*C*), and cognitive performance (*D*) throughout sessions and aligned to voluntary pauses for monkeys R and S (top and bottom, respectively). Gray bars represent the average measure (+sem) for each parameter that was used for statistical testing (see Materials and Methods for details). Asterisks indicate the significance level (Kruskall–Wallis test, factor "pre/post-pause," using Bonferroni post-hoc correction): (ns) $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Execution and cognitive measures do not follow the same pattern as the beta oscillatory power.

is clearly illustrated in Figure 6*A*. As shown earlier, delay-related beta power increased significantly within-block prior to the pause. The power reached a maximum just before the onset of the pause for both monkeys. When the monkey started to work again, the beta power was lower, as if it had been reset to a level comparable with session start. On subsequent trials, during the second block, beta power increased again until the end of the block (Fig. 6A, but see also sessions with 2 pauses in Supplementary Fig. 4).

We further investigated how execution and cognitive measures exhibit a similar change around pauses. Neither execution error rate nor RT were modulated in the same way as beta power around pauses, specifically we did not observe a drop in errors or RT after the pause (Fig. 6*B–C*). In fact, a linear mixed-effect model applied to RT showed that RT at the start of the second block were similar to those at the end of the first block and then decreased during the second block (see Fig. 6*C* and Supplementary Fig. 5). Also, the proportion of nonoptimal choices did not follow beta power changes (Fig. 6*D*). We evaluated the changes between the end of the first block and the beginning of the second one using a Kruskall–Wallis test (with Bonferroni post-hoc correction) and report significant comparisons on Figure 6. Statistical comparisons were made by taking the average of 30 trials for each session included (or the average of 2 problems for the cognitive measure), before and after pauses. Only a significant increase in RT was observed in monkey S ($H = 4.68$, 1 d.f., $P = 0.03$), but not in monkey R ($H = 2.41$, 1 d.f., $P = 0.12$). Other comparisons were not significant ("execution error rate," monkey R: $H = 0.17$, 1 d.f., $P = 0.68$; monkey S: $H = 0.96$, 1 d.f., $P = 0.33$; "Pre- vs. Post-cognitive performance during search," monkey R: $H = 0.5$, 1 d.f., $P = 0.48$; monkey S: $H = 2.23$, 1 d.f., $P = 0.13$; "Pre- vs. Post-cognitive performance during repetition," monkey R: $H = 0.29$, 1 d.f., $P = 0.59$; monkey S: statistical testing could not be assessed due to zero inflated data, 106 of 116 problems with a value of zero, i.e., no error during repetition).

None of the changes were comparable with the changes in beta, suggesting there was no direct relationship between execution or cognitive performance and beta power changes for either monkey.

Phase Model

In order to quantitatively assess this effect, and its variance across sessions, we varied the Phase model by including Block (before and after pause) as fixed and random factor. The Block effect was relevant in both cases to fit the data as shown by model selection (see Materials and Methods).

Only the Phase and the Time × Block fixed effects survived model selection, showing that the difference in Beta power between search and repetition remained stable even after the pause, but that the within-block effect was modified after the pause. We investigated these phenomena further by extracting fitted data and model coefficients for each session and each monkey. Figure 7 shows the key results. First, the drop in power after the pause was very reliable across sessions and monkeys (Fig. 7*A* left). In addition, the slope of fits reflecting the strength of the within-session effect was almost always stronger in the second block (after the pause) than in the first block (Fig. 7*A* right). In other words, after a drop during pauses, the increase of the Beta power with time was greater than at the beginning of the session.

This last element of results suggests that cumulated fatigue or effort, or the continuous change in motivation during work, induces slow evolving neural changes that impact on the production of beta oscillations and possibly on the effect of pauses. Thus, the time spent in work in the first block and the duration of pauses might contribute to the variance of the drop of power observed at pauses. We hence investigated the relationships between pause length or time of pause in sessions and the difference in power between the end of the first block and the beginning of the second. Figure 7*B* and *C* presents the data fitted by the above-mentioned mixed model. Fitted powers at the end the first block for each session are aligned to observe the variation in the size of power drop (symbolized with blue lines). Graphs show the effect of pause duration (left) and time of pause in sessions (right) on the distribution of drop in power. The data for both monkeys clearly show 2 important phenomena: 1) the longer the pause, the weaker the drop in power (Fig. 7*B*) and 2) the later the pause, the weaker the drop in power (Fig. 7*C*). These negative effects were tested in a multiple linear regression showing strong significance of both duration and time of pause on the power drop ($P < 0.0001$). There was no effect of power at the beginning and end of the first block. The slope of within-session change in power made a small contribution to the drop size only for one monkey (monkey R: $P < 0.002$, monkey S: $P = 0.55$). There was no correlation between timing of pauses and duration of pauses (monkey R: $P = 0.59$, monkey S: $P = 0.97$).

Response-to-Feedback Model

We also added the fixed and random factor Block to the Response-to-feedback model and selected it with the same procedure. Again Time × Block effects survived model selection, as did PFB. In monkey R, a time × PFB interaction survived selection, but not in monkey S. As in the original model, significantly lower Beta was measured in trials after Correct feedback than after other types of feedback (Wald statistics, in every case t-value > 4, $P < 0.0001$), and so the cognitive control response represented by changes in Beta was maintained after the pause.

In summary, pauses in work caused significant interruption of the within-session effect in delay-related beta oscillations, in a sense resetting it. By contrast, other measures that were modulated within-session were not modulated by the pauses in work.

Discussion

In monkeys performing a test of cognitive control, we recorded beta oscillations that reflected the different cognitive control demands of the task: When cognitive control was required (during search, when shifting between problems, and after negative feedback), beta power was increased, notably during the delay period at the start of each trial. In this same delay period, beta oscillations also showed a significant increase in power within-session over time. Behaviorally, monkeys also showed within-session increases in response times and execution errors, and we regard these changes in execution measures to reflect a time-on-task effect. Despite this, cognitive performance on the task suffered no significant within-session decrement and therefore was resistant to the time-on-task effect. When the monkey made a voluntary pause in work during the session, we observed a significant reset of the beta power, and so an apparent re-initialization of the within-session effect. This effect was absent in the behavioral data. The magnitude of the beta reset was an inverse function of the timing of the pause in the session and the length of the pause. Finally, we observed no interaction between the cognitive control and within-session effects in beta oscillations during the delay period, in that the search–repetition difference in power was conserved despite the power increase during the session.

We propose that beta oscillations during delay therefore reflect multiple factors influencing cognitive control, one task-sensitive component and one time-sensitive component that, due to the effect of pauses, cannot be interpreted as a simple time-on-task effect.

Beta and Cognitive Control

One approach to explain the relationship between cognitive functions and beta oscillations has been to relate them to top-down control of behavior (Buschman and Miller 2007; Siegel et al. 2012). Pesaran and colleagues demonstrated enhanced frontoparietal beta-band coherence in LFP during free decisions compared with instructed decisions (Pesaran et al. 2008). Free and instructed decisions require different cognitive control demands in the same way as the search and repetition periods of the current task. This proposed explanation accords with a generalized theory of beta oscillations in maintaining the "status-quo" proposed by Engel and Fries (2010). In the cognitive domain, these authors propose that such activity should be associated with the active maintenance of a cognitive set when the task involves a "strong endogenous top-down component." Our finding of higher beta power in search seems to verify this proposal (Engel and Fries 2010). But, in contrast to these authors, we do not feel able to conclude that there is a unifying hypothesis of beta function. Our findings are too specific to permit this conclusion—the effects we report are limited to Beta 2 and to the delay period, and therefore, our interpretations do not apply to other instances of beta power recorded during the trial, for example after the touch.

LFP beta power in the pre-cue period of motor tasks, equivalent to the delay in the current task, has been reported in a number of studies (Kilavik et al. 2013). In human motor cortex, pre-cue beta power peaks before an informative but not a noninformative cue, supporting the idea that this is a top-down control signal that can be linked to cognitive elements of the task (Saleh et al. 2010). In the current task, we clearly demonstrate that beta power can

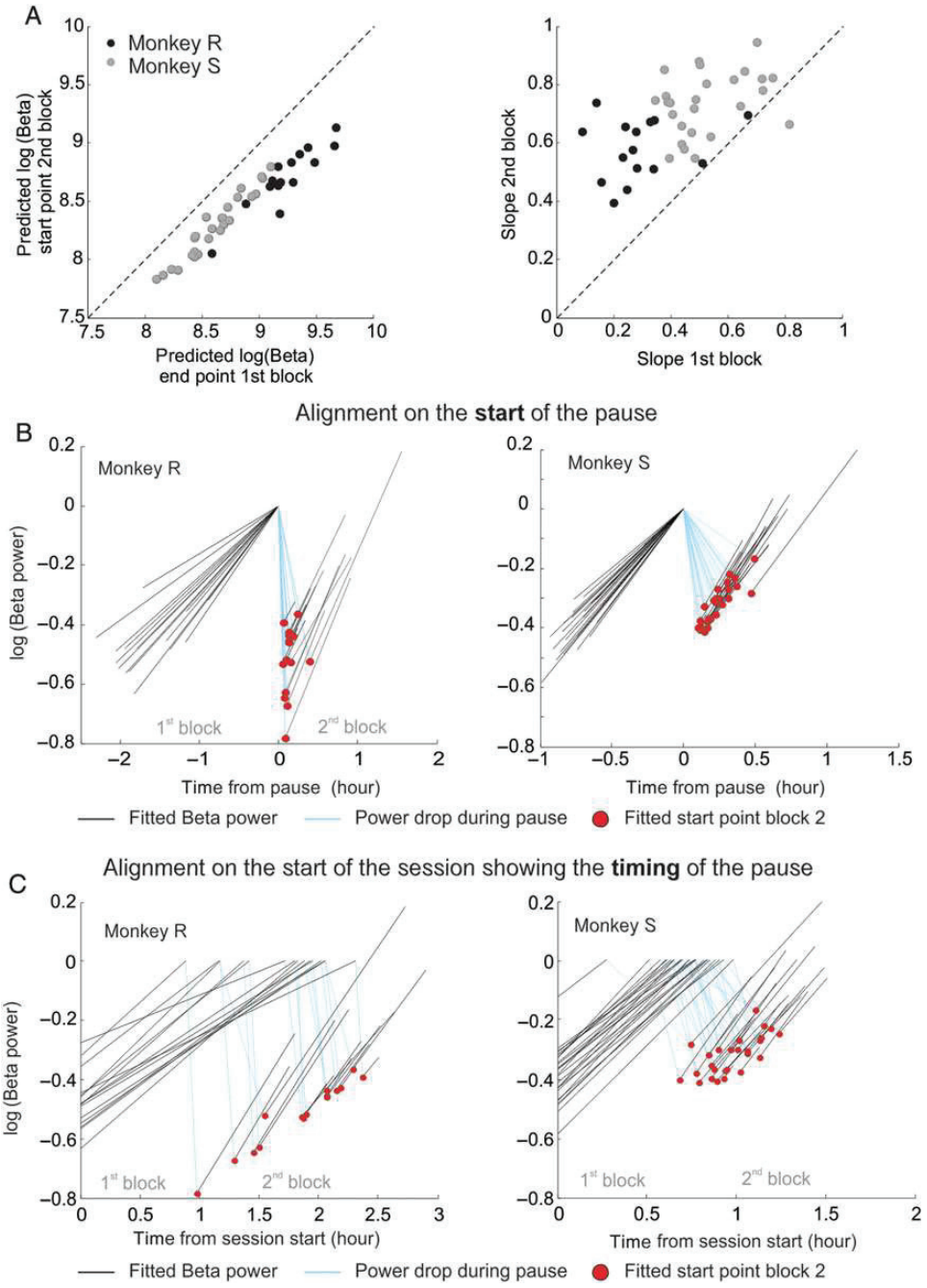

Figure 7. Effect of pauses in work on beta power drop. (A) The plot on the left presents the fitted values of log beta power at the start of the second block (after the pause), as a function of the values at the end of the first block (i.e., showing the drop in Beta power). The plot on the right compares the slope of the within-session effects in the second and first blocks of trials. (B and C). Dynamic of beta modulation around the pause for each session, as described by the model fit. The figures show the relationships between the drop in beta (indicated by a blue line) and the parameters of the pause, specifically the duration and the timing within the session of the pause. Black lines represent the linear fits for beta power both before and after the pause. This shows the increase of beta up to and after the pause. Log(Beta) power values are aligned at zero to equalize values at the start of each pause. Figures B and C show the same data with different alignments in time. (B) The longer the pause, the less the drop in Beta power—data are aligned on the time of pause to stress this effect, represented by the distribution of the red starting points. (C) The later the pause, the weaker the drop in Beta power—data are aligned on the start of the session to stress the timing of the pause within the session.

be modulated by cognitive and motivational information during the delay period, but separately by motor action later in the trial.

Beta oscillations fluctuate depending on the cognitive control requirements, specifically those determined by the phase of the task (search/repetition) and by the outcome of the previous trial (INC/COR/SWI/BK). This is indeed predicted by the regulation models of cognitive control that account for variations in control in different situations, including not only tasks involving overriding of competing responses but also tasks with underdetermined responses like in the PST (Botvinick et al. 2001). One important instance of regulation occurs after errors. This is exactly what we observed. In the PST, shifting after error is the key to solve problems and optimal performance requires using errors appropriately. The animals did that perfectly, and their levels of beta

oscillations reflect this. In this sense, we suggest that beta oscillations index cognitive control. Although beta power in the delay did not predict more optimal outcomes, nor longer RT, as might be expected of a cognitive control mechanism in a stimulus-response task driven by speed-accuracy trade off (Kerns et al. 2004), this was not necessarily expected in our case. In the PST, responses are underdetermined and driven by strategy not stimulus, and so a direct coupling between application of cognitive control (as indexed by beta power) and subsequent performance is not necessarily to be expected. Finally, in discussing the link between beta and cognitive control, we must remember that our evidence is correlational in nature, in common with most neurophysiological studies of cognitive control. Only direct experimental alterations of brain function, specifically focusing on beta oscillations, would be able to define a causal relationship between beta power, cognitive control, and performance.

Beta Within-Session

At the same time, beta power also changed within the session, showing a striking increase over time. A simple and tempting interpretation is that this is a reflection of the behavioral time-on-task effect, which we have recorded through within-session increases in response times and execution errors. As such, one might posit that beta power is an index of general fatigue, linking to previous work in human EEG (Lafrance and Dumont 2000). But, this interpretation cannot be maintained, because of the crucial effect on beta power of pauses in work.

The evaluation of physiological and behavioral changes around pauses is a particularly important test that is rarely considered in the literature. Should the pause, for example, reset time-on-task decrements in performance in the case where the subject restarts work after the pause? Or should time-on-task effects maintain regardless of pauses? Variations in task are surely critical here. In our protocol, where the objective is fixed but work is self-paced, pauses and their related physiological changes are informative about the general state and cognition of the monkey within the context of a session that the monkey is motivated to finish. Note that although pauses could be experimentally imposed, such a manipulation would remove any possibility of understanding physiological reasons for pausing.

Our data show that beta power is profoundly influenced by pauses and that there is a de-correlation between execution parameters and the beta power either side of pauses (Fig. 6). If execution error rates or response times are an index of fatigue, the beta power is clearly not reflecting fatigue, and so is not a marker of the time-on-task effect as defined in the literature. Likewise, the pause effect on beta power permits rejection of several explanations of general change over the session, such as influences of reward satiation and the expectation of the bonus. Nevertheless, beta is modulated as a function of time, yet the task remains the same throughout the session. So while beta power changes during the session might not reflect time-on-task per se, they could reflect changes induced to counteract time-on-task perturbations, especially as our monkeys maintain cognitive performance throughout the session.

In some human research, cognitive control and execution error rates are modulated together across the session (Boksem et al. 2005). In our study, they are dissociated. A hypothesis for this dissociation is that monkeys have a strong overriding motivation to perform well on the cognitive task, to gain the final bonus reward more quickly. When motivation is manipulated in human subjects, time-on-task effects can be reversed (Lorist et al. 2009). Some subjects respond to increased motivation by increasing their accuracy, whereas others increase their response speed, but subjects do not do both (Boksem et al. 2006).

This influence of motivation returns us to the concept of attentional effort: a cognitive incentive that integrates explicit and implicit motivational forces (Sarter et al. 2006). A subject will increase attentional effort in order to maintain good performance, especially when highly motivated, and we posit that this is the case with the monkeys in the current experiment. In particular, attentional effort will be goal directed—monkeys will apply attentional effort in order to attain the final bonus reward, but attentional effort may not impact measures of general fatigue, such as the increasing response times. Hence, we hypothesize that the increasing beta power in the delay within-session is an index of the monkey's increasing attentional effort. The impact of attentional effort should be limited in some way, although note that we did not find a maximum value of beta power that induced pauses. Hence, the pause allows beta power and attentional effort reset: cognitive performance can be maintained after the pause with a lowered beta and presumably lowered attentional effort. The pause allows recovery of attentional effort, but not recovery from general fatigue; hence, the behavioral time-on-task effect remains (i.e., response times are not modulated around pauses). Such an interpretation might explain the dissociations between cognitive performance, execution performance, and beta oscillations.

Other interpretations of the within-session result should be considered, but they would seem to require that beta relates to a completely independent but time-sensitive function that we have not measured. We are, for the moment, unable to conceive of what such a function might be. Nevertheless, our interpretation demands experimental confirmation—in particular consideration of explicit manipulations of both fatigue and motivation in order to change the attentional effort applied to the task and record the resulting beta oscillations. We further note that studying pause effects in experiments targeting the neural correlates of fatigue and time-on-task would seem now to be essential.

Multiple Roles of Beta Oscillations

So a single band of beta oscillations in the delay shows changes when the task demands more or less cognitive control, and when attentional effort is applied. It is established that different beta bands (as observed in our data and the data of others) have differential sensitivity in different moments of the trial, and this suggests that a large variety of mechanisms is expressed in the oscillations ranging from 12 to 30 Hz (Pfurtscheller et al. 1997). In addition to the roles cited earlier, a role in the motor system is well established (Pfurtscheller and Lopes da Silva 1999), and there is a clear link between beta oscillations and the pathophysiology of movement disorders (Brown 2007). Studies in visual cortical areas have sought to link low-frequency oscillations between 12 and 30 Hz, with a specific top-down role within the hierarchy of visual areas (Bosman et al. 2012; Bastos et al. 2014), and a specific localization of beta oscillations in the infragranular layers of cortex (Buffalo et al. 2011). These different roles are often differently localized or reflect different bands of "beta." For example, several authors have already emphasized 2 sub-bands in the beta range (Beta 1 and Beta 2) that are associated with different neural mechanisms (Roopun et al. 2008; Cannon et al. 2013). Where our results contrast is that one single beta band (Beta 2) at one point in the task (the delay) in frontal cortex can be independently modified by different behavioral factors. We propose that this occurs because the 2 factors impinge on a single control mechanism.

Our data show interesting inter-individual variability for the beta power spectrum (Fig. 3). In both animals, 2 beta-bands were observed, close together in one monkey and well separated in the other. Yet, the effects of Phase and Time were observed for both animals on the highest frequency band, Beta 2. The literature provides interesting information regarding the relevance of inter-individual variability in power spectra (Buzsáki et al. 2013), a variability that should be taken into account when studying precise task-related variations in specific frequency bands (Pfurtscheller et al. 1997) and that can also be observed in LFPs (Kilavik et al. 2012). Brain rhythms are specific to individual brains and are at some level under genetic control, representing a robust heritable phenotype (De Gennaro et al. 2008; van Pelt et al. 2012). Importantly, our data show that although the exact Beta 2 peak differed between monkeys, the properties and modulations by cognitive challenge are identical, suggesting that functions are associated with a hierarchy of oscillatory bands but are independent of the exact frequency of operation.

Although the experiment was not designed to precisely locate functional dissociations within the frontal cortex, the reconstructed maps of effects show inhomogeneity hence functional organization. Although the electrode grids had different sizes in the 2 animals, they overlapped over the dorsomedial frontal cortex. The maps of the Phase effect (Fig. 4*B*) reveal for both monkeys a slight bias in the effects on the side ipsilateral to the hand used in the task, while relative to this, the Time effect appears to be slightly more anterior and bilateral (Fig. 5*A*). The largest grid showed effects over the dorsal arcuate sulcus but the description in one single monkey precludes strong conclusion. Finally, our analyses of ECoG signals do not avoid potential volume conduction effects that would alter anatomical resolution.

Motivation and Cognitive Control

We propose that our results show a mechanism for the integration of motivational parameters into cognitive control. Beta oscillations index the implementation of cognitive control. Attentional effort acts to provide a general increase in beta over time. This is presumably necessary to combat a fatigue process, though our data do not show exactly what that might be. Crucially, this change in gain of beta allows for maintenance of the search–repetition difference necessary for the task. Motivation parameters such as attentional effort therefore act as modulators of cognitive control.

The alternative interpretation is 2 independent mechanisms each involving beta oscillations at identical frequencies and times, which we recorded as a single phenomenon on the surface of the brain. We posit that such a coincidence is unlikely. The dual influence on control interpretation is more plausible and more coherent with current theoretical positions.

The exertion of cognitive control is considered to be intrinsically costly and selected only in the presence of sufficient or adequate incentives (Westbrook et al. 2013; Kool and Botvinick 2014). In our case, decision to work or to pause might thus be based on the evaluation of the cost of exerting control in the PST in the face of fatigue, motivation, obtained and predicted outcomes, and distance to the final bonus reward. A recent proposition (Shenhav et al. 2013) integrates this cost-benefit analysis (including motivational factors) into control signal specification. Expected values of potential control signals are compared in order to select the identity and intensity of the control signal to be applied in a particular context. The specified control is then implemented, and we hypothesize that beta oscillations provide an index of this control implementation. Hence, beta modulation by 2 different factors reflects separated mechanisms that contribute to control specification. Independent inputs to control specification also echoes the dual mechanisms of control framework (Braver et al. 2007), albeit in a context where proactive and reactive mechanisms are employed concurrently (Braver 2012).

This interpretation leaves many open questions about the integration of motivation into cognitive control (Braver et al. 2014). Nevertheless, our findings here reiterate the suitability of beta oscillations as a contributing neural mechanism to these computations.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

This work was supported by Fondation de France, Agence National de la Recherche, Fondation pour la Recherche Médicale (J.V., F.M.S., M.C.M.F.), Fondation Caisse d'Epargne Rhône Alpes Lyon (J.V.), Fondation Neurodis (C.R.E.W.), and by the labex CORTEX ANR-11-LABX-0042. C.R.E.W. is funded by a Marie Curie Intra-European Fellowship (PIEF-GA-2010-273790). M.C.M.F. is funded by Ministère de l'enseignement et de la recherche. J.V. is currently funded by the LOEWE – NeFF project (Neuronale Koordination Forschungsschwerpunkt).

Notes

We thank M. Valdebenito, M. Seon, and B. Beneyton for animal care and C. Nay for administrative support. *Conflict of Interest*: None declared.

References

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P. 2014. Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron. doi:http://dx.doi.org/10.1016/j.neuron.2014.12.018

Bell AJ, Sejnowski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7:1129–1159.

Boksem MAS, Meijman TF, Lorist MM. 2005. Effects of mental fatigue on attention: an ERP study. Cogn Brain Res. 25:107–116.

Boksem MAS, Meijman TF, Lorist MM. 2006. Mental fatigue, motivation and action monitoring. Biol Psychol. 72:123–132.

Boksem MAS, Tops M. 2008. Mental fatigue: costs and benefits. Brain Res Rev. 59:125–139.

Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci Biobehav Rev. 44C:58–75.

Bosman CA, Schoffelen J-M, Brunet N, Oostenveld R, Bastos AM, Womelsdorf T, Rubehn B, Stieglitz T, De Weerd P, Fries P. 2012. Attentional stimulus selection through selective synchronization between monkey visual areas. Neuron. 75:875–888.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. Psychol Rev. 108:624–652.

Braver TS. 2012. The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn Sci (Regul Ed). 16:106–113.

Braver TS, Cohen JD, Barch DM. 2002. The role of prefrontal cortex in normal and disordered cognitive control: a cognitive

neuroscience perspective. In: Stuss DT, Knight RT, editors. Principles of Frontal Lobe Function. Oxford: Oxford University Press. p. 428–447.
Braver TS, Gray JR, Burgess GC. 2007. Explaining the many varieties of working memory variation: dual mechanisms of cognitive control. In: Conway A, Jarrold C, Kane M, Miyake A, Towse J, editors Variation in Working Memory. Oxford: Oxford University Press. p. 76–106.
Braver TS, Krug MK, Chiew KS, Kool W, Westbrook JA, Clement NJ, Adcock RA, Barch DM, Botvinick MM, Carver CS, et al. MOMCAI group. 2014. Mechanisms of motivation-cognition interaction: challenges and opportunities. Cogn Affect Behav Neurosci. 14:443–472.
Brown P. 2007. Bad oscillations in Parkinson's Disease. J Neural Trans Suppl. 70:27–30.
Buffalo EA, Fries P, Landman R, Buschman TJ, Desimone R. 2011. Laminar differences in gamma and alpha coherence in the ventral stream. Proc Natl Acad Sci. 108:11262–11267.
Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK. 2012. Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron. 76:838–846.
Buschman TJ, Miller EK. 2007. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science. 315:1860–1862.
Buzsáki G, Logothetis N, Singer W. 2013. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. Neuron. 80:751–764.
Cannon J, McCarthy MM, Lee S, Lee J, Börgers C, Whittington MA, Kopell N. 2013. Neurosystems: brain rhythms and cognitive processing. Eur J Neurosci. 39:705–719.
De Gennaro L, Marzano C, Fratello F, Moroni F, Pellicciari MC, Ferlazzo F, Costa S, Couyoumdjian A, Curcio G, Sforza E, et al. 2008. The electroencephalographic fingerprint of sleep is genetically determined: a twin study. Ann Neurol. 64:455–460.
Engel AK, Fries P. 2010. Beta-band oscillations—signalling the status quo? Curr Opin Neurobiol. 20:156–165.
Kahneman D. 1973. Attention and effort. Upper Saddle River: Prentice Hall.
Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. Science. 303:1023–1026.
Khamassi M, Quilodran R, Enel P, Dominey PF, Procyk E. 2014. Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. Cereb Cortex. doi:10.1093/cercor/bhu114.
Kilavik BE, Ponce-Alvarez A, Trachel R, Confais J, Takerkart S, Riehle A. 2012. Context-related frequency modulations of macaque motor cortical LFP beta oscillations. Cereb Cortex. 22:2148–2159.
Kilavik BE, Zaepffel M, Brovelli A, MacKay WA, Riehle A. 2013. The ups and downs of β oscillations in sensorimotor cortex. Exp Neurol. 245:15–26.
Kool W, Botvinick M. 2014. A labor/leisure tradeoff in cognitive control. JExp Psychol General. 143:131–141.
Lafrance C, Dumont M. 2000. Diurnal variations in the waking EEG: comparisons with sleep latencies and subjective alertness. J Sleep Res. 9:243–248.
Lorist MM, Bezdan E, Caat ten M, Span MM, Roerdink JBTM, Maurits NM. 2009. The influence of mental fatigue and motivation on neural network dynamics; an EEG coherence study. Brain Res. 1270:95–106.
Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 24:167–202.
Oostenveld R, Fries P, Maris E, Schoffelen J-M. 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011:1–9.
Pesaran B, Nelson MJ, Andersen RA. 2008. Free choice activates a decision circuit between frontal and parietal cortex. Nature. 453:406–409.
Pfurtscheller G, Lopes da Silva FH. 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol. 110:1842–1857.
Pfurtscheller G, Stancák A, Edlinger G. 1997. On the existence of different types of central beta rhythms below 30 Hz. Electroencephalogr Clin Neurophysiol. 102:316–325.
Phillips JM, Vinck M, Everling S, Womelsdorf T. 2014. A long-range fronto-parietal 5- to 10-hz network predicts "top-down" controlled guidance in a task-switch paradigm. Cereb Cortex. 24:1996–2008.
Pinheiro JC, Bates DM. 2000. Mixed-effects models in S and S-PLUS. New York: Springer.
Procyk E, Goldman-Rakic PS. 2006. Modulation of dorsolateral prefrontal delay activity during self-organized behavior. J Neurosci. 26:11313–11323.
Procyk E, Tanaka YL, Joseph J-P. 2000. Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat Neurosci. 3:502–508.
Procyk E, Wilson CRE, Stoll FM, Faraut MCM, Petrides M, Amiez C. 2014. Midcingulate motor map and feedback detection: converging data from humans and monkeys. Cereb Cortex. doi:10.1093/cercor/bhu213
Quilodran R, Rothé M, Procyk E. 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron. 57:314–325.
Roopun AK, Kramer MA, Carracedo LM, Kaiser M, Davies CH, Traub RD, Kopell NJ, Whittington MA. 2008. Period concatenation underlies interactions between gamma and beta rhythms in neocortex. Front Cell Neurosci. 2(1):8.
Rothé M, Quilodran R, Sallet J, Procyk E. 2011. Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation. JNeurosci. 31:11110–11117.
Saleh M, Reimer J, Penn R, Ojakangas CL, Hatsopoulos NG. 2010. Fast and slow oscillations in human primary motor cortex predict oncoming behaviorally relevant cues. Neuron. 65: 461–471.
Sarter M, Gehring WJ, Kozak R. 2006. More attention must be paid: the neurobiology of attentional effort. Brain Res Rev. 51: 145–160.
Shenhav A, Botvinick MM, Cohen JD. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron. 79:217–240.
Siegel M, Donner TH, Engel AK. 2012. Spectral fingerprints of large-scale neuronal interactions. Nat Rev Neurosci. 13:121–134.
van Pelt S, Boomsma DI, Fries P. 2012. Magnetoencephalography in twins reveals a strong genetic determination of the peak frequency of visually induced gamma-band synchronization. J Neurosci. 32:3388–3392.
Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. 4th ed. New York: Springer Science & Business Media.
Vezoli J, Procyk E. 2009. Frontal feedback-related potentials in nonhuman primates: modulation during learning and under haloperidol. J Neurosci. 29:15675–15683.
Westbrook A, Kester D, Braver TS. 2013. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. PLoS ONE. 8:e68210.
Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM. 2009. Mixed Effects Models and Extensions in Ecology with R. New York: Springer.