L’apprentissage par renforcement implique de l’effort mental

III.1 Introduction

L’apprentissage par renforcement est un mécanisme relativement universel. Il consiste dans l’idée simple qu’un évènement immédiatement suivi d’une récompense sera associé à l’obtention de cette récompense et qu’un évènement immédiatement suivi d’une punition sera

associé au fait de subir cette punition. L'universalité de ce mode d'apprentissage dans le règne animal, des pigeons aux Primates, explique que ce mode d'apprentissage soit rattaché à des structures « primaires » qu'on retrouve chez les plupart des Vertébrés, comme les ganglions de la base et les neurones dopaminergiques (Averbeck and Costa, 2017; Grillner and Robertson, 2016; Mannella et al., 2013). De ce fait, on pourrait aussi s'attendre à ce que ce mode d'apprentissage, ayant prouvé son efficacité au fil de l'évolution, soit invariant à travers la plupart des espèces chez lesquelles on peut l'observer.

À l'inverse, la capacité d'exercer du contrôle cognitif est restreinte à quelques espèces. Cette capacité consiste plus ou moins à être capable d'établir une stratégie d'actions, notamment dans des situations où des distractions ou des automatismes doivent être inhibés en vue d'atteindre un but désiré. Le contrôle cognitif fait partie des capacités qui ont été massivement développées chez les Primates, et l'homme en particulier, grâce à l'expansion du cortex préfrontal (Carlén, 2017). L'apprentissage par renforcement, dans sa version la plus simple, sans changement dynamique des probabilités associées à chaque évènement, peut être vu comme un processus automatique d'apprentissage par essai-erreur où les valeurs associées à une action donnée sont automatiquement mises à jour en fonction des feedbacks associés à chaque action. De ce fait, l'apprentissage par renforcement ne nécessiterait pas l'implémentation de contrôle cognitif. D'un autre côté, certaines versions de l'apprentissage par renforcement partent du principe que, dans le cadre d'une situation nouvelle ou lorsque l'incertitude est élevée, l'apprentissage par renforcement va de pair avec l'allocation de ressources mentales, jusqu'à ce que, à force de répétitions, le processus devienne éventuellement automatisé. John M. Pearce et Geoffrey Hall, dans le cadre du conditionnement classique, proposent par exemple que, tant qu'un stimulus n'est pas complètement prédictif, il implique l'utilisation d'un processeur cognitif qui permette de comparer les options à disposition et de délibérer (Pearce and Hall, 1980).

Plus généralement, la détection du fait d'avoir commis une erreur entraîne une augmentation de l'activité de l'insula antérieure et du cortex préfrontal dorsomédian (dmPFC) (Bastin et al., 2016; Iannaccone et al., 2015). Cette activité provoque ensuite un phénomène classiquement dénommé comme un ralentissement post-erreur (*post-error slowing*). Ce ralentissement pourrait notamment refléter le fait que les individus utilisent plus de contrôle cognitif après avoir identifié une situation où ils ont commis une erreur en vue d'adapter leur performance future. Dans des tâches d'apprentissage par renforcement, le dmPFC pourrait être capable de moduler, de manière dynamique, le taux d'apprentissage α en fonction du degré de

volatilité de la tâche (Behrens et al., 2007; Khamassi et al., 2013). L'activité du dmPFC a par ailleurs été corrélée à l'allocation de contrôle cognitif (Botvinick et al., 2001; Shenhav et al., 2013). De plus, des temps de réaction plus longs lors d'une décision sont généralement associés à l'idée que ces décisions sont plus difficiles et nécessitent plus d'effort mental. Dans notre première étude, nous avons d'ailleurs pu mettre en évidence que le temps de délibération était corrélé à l'activité du dmPFC dans des tâches nécessitant de prendre des décisions basées sur la valeur (voir *La valeur, la confiance et le temps de délibération au moment du compromis coûts/bénéfices ont des supports neuraux distincts dans le cortex préfrontal médian*). Nous avions ainsi dissocié, dans le cadre d'une prise de décision, les corrélats cérébraux de la valeur subjective, situés en partie dans le vmPFC, et les corrélats de l'effort lié à la délibération, situé dans le dmPFC et l'insula antérieure.

Dans cette étude, nous avons utilisé une tâche d'apprentissage par renforcement requérant des participants d'apprendre, par essai-erreur, à identifier les symboles leur permettant de maximiser leurs gains et de minimiser leurs pertes. Nous voulions voir si cette tâche impliquait, ou non, l'activité de structures liées à l'allocation de contrôle cognitif. L'apprentissage par renforcement, dans un cas simple qui n'incluait pas de changement dynamique des probabilités, pourrait avoir lieu de manière automatique sans nécessiter l'activité de structures liées à l'allocation de contrôle cognitif. Nous nous attendrions alors que les structures du réseau des valeurs, comme le vmPFC identifié dans notre première étude, permettent de représenter la valeur subjective associée aux différents symboles de la tâche et de la mettre automatiquement à jour. L'activité du dmPFC, que nous avons liée à l'effort implémenté au cours de la délibération dans une tâche impliquant une prise de décision (voir **Figure 27**), pourrait ne pas être impliqué dans cette tâche si elle se fait de manière automatique. A l'inverse, si cette tâche implique un effort au moment où les participants délibèrent, la tâche pourrait impliquer l'activité du dmPFC. Comme dans notre première étude, nous avons utilisé le temps de délibération comme un reflet des ressources éventuellement engagées au cours de la délibération. Nous avons ensuite regardé quels étaient les corrélats cérébraux de ce temps de délibération pour voir si le dmPFC était ou non impliqué au moment de la délibération suggérant que la tâche impliquait un certain effort mental.

III.2 Méthodes

III.2.a Participants

Les participants qui ont participé à cette étude étaient les mêmes que ceux recrutés pour la deuxième étude (voir *Participants*). Ils ont effectué, au cours de la même journée les tâches d'effort physique et d'effort mental détaillées dans la deuxième partie et la tâche d'apprentissage détaillée ici.

Nous avons aussi exclu le participant qui avait des résultats déplorables dans la tâche d'effort physique de l'analyse de la tâche d'apprentissage. Sa performance était en effet, là aussi, déplorable suggérant qu'il n'avait soit rien compris aux tâches que nous lui avions demandé d'exécuter soit qu'il n'avait pas la volonté de les effectuer correctement.

Pour les données d'IRMf, nous avons exclu 1 sujet en plus car il bougeait trop dans le scanner (mouvement supérieur à 5 mm). Ce sujet a cependant été inclus dans les études de comportement. Les données d'imagerie et du diamètre pupille sont donc basées sur 22 participants au total (12 femmes, 10 hommes) de 26 ± 4 ans (moyenne ± écart-type).

III.2.b Tâche

La tâche a été programmée dans la toolbox Psychtoolbox (http://psychtoolbox.org/) au sein du logiciel Matlab 2012 (The MathWorks, Inc., USA). Les participants ont reçu des instructions à l'écrit et à l'oral. Le principe de l'expérience tel qu'il était indiqué aux participants était de maximiser leurs gains.

Nous avons utilisé une version similaire à celle déjà été utilisée dans notre équipe lors d'études précédentes (Palminteri et al., 2012; Pessiglione et al., 2006). Dans les instructions, nous disions aux participants que, pour une session donnée, ils allaient être confrontés à 6 symboles différents. Ils devaient chercher à identifier quels étaient les symboles qui leur permettraient de maximiser leurs gains. Nous leur disions que certains symboles étaient principalement associés à des gains de 10€, d'autres à rien (0€) et d'autres à des pertes de 10€. Ils devaient ainsi découvrir, par essai-erreur, quels étaient les symboles à sélectionner à chaque essai afin de maximiser leurs gains. Nous leur précisions que le sens associé à chaque symbole ne variait pas au cours de l'expérience, mais qu'un symbole donné ne permettait pas toujours d'obtenir le résultat escompté. Le but de ces instructions était d'essayer d'expliquer aux participants que les symboles étaient des prédicteurs probabilistes, mais que les probabilités qui y étaient associées étaient fixes pour éviter qu'ils cherchent à estimer la volatilité de la tâche.

Nous voulions ainsi être sûrs que les participants avaient bien compris la structure de la tâche et qu'ils ne seraient pas en train d'apprendre la structure de la tâche, en plus d'apprendre les valeurs associées à chaque symbole.

Cela n'était pas précisé dans les instructions, mais chaque symbole était systématiquement présenté au sein d'une même paire. Il y avait 3 types de paires : des paires associées à des gains (+10€ ou 0€), des paires associées à un résultat neutre (0€) et des paires associées à des pertes (-10€ ou 0€). Au sein de chaque paire, les probabilités des symboles étaient associées (75/25% et 25/75%). En d'autres termes, quand, dans une paire associée à des gains, le symbole A permettait d'obtenir un gain (+10€), le symbole B de la même paire était nécessairement associé à un résultat neutre (0€) pour l'essai concerné.

Chaque essai démarrait avec la présentation d'une croix de fixation pendant une durée de 500 millisecondes. Ensuite, les deux symboles d'une même paire apparaissaient à l'écran. Pour répondre, le sujet devaient appuyer sur le bouton correspondant à la localisation spatiale du symbole qu'il voulait sélectionner (gauche/droite) et garder le doigt appuyé sur ce bouton jusqu'à la fin de la période de sélection. Pour répondre, les participants disposaient d'un boîtier de réponse (fORP 932, Current Designs Inc, Philadelphia, USA) placé sous leur main droite au niveau du torse au début de l'expérience. Leur index était posé sur le bouton permettant de sélectionner l'option de gauche et leur majeur sur le bouton permettant de sélectionner l'option de droite. Une fois cette période terminée, le symbole choisi apparaissait entouré d'un cadre rouge pendant 500 millisecondes. Enfin, le résultat obtenu suite à la sélection du symbole était affiché à l'écran (voir **Figure 36**). Si le participant n'avait appuyé sur aucun bouton pendant la période de sélection, s'il avait cessé d'appuyer avant la fin de la période de sélection ou s'il avait appuyé sur les deux boutons simultanément, nous considérions cela comme une omission de la part du sujet. Il recevait alors un message précisant l'erreur commise (« Réveillez-vous !», « Maintenez votre appui » ou « Un seul bouton ! ») et, au moment du feedback, il obtenait le pire résultat possible pour la paire considérée. Pour les paires associées à des gains ou à un résultat neutre, il n'obtenait donc rien (0€) et pour les paires associées à des pertes, ils subissaient une perte de 10€. Ceci était effectué afin de motiver les participants à répondre à tous les essais, notamment dans le cas des paires associées à des pertes.

Chaque session de la tâche d'apprentissage était composée de 60 essais. Les symboles étaient des stimuli abstraits tirés de l'alphabet Agathodaimon. Les participants ont dû effectuer 4 sessions de cette tâche. La première session était effectuée sur un ordinateur portable en dehors du scanner. Elle servait à familiariser les participants avec la tâche avant l'expérience.

Pour répondre, les participants utilisaient alors les flèches gauche et droite du clavier de l'ordinateur à la place du boîtier de réponse. Les trois sessions suivantes étaient effectuées dans le scanner. Entre chaque session, les participants effectuaient aussi une session de la tâche d'effort physique et une session de la tâche d'effort mental (voir *L'effort, tant physique que mental, est plus motivé par l'attrait de gagner que par la peur de perdre*). Les 6 symboles utilisés dans chaque session étaient différents pour chaque session. Pour effectuer ces 4 sessions, nous avons donc sélectionné 24 symboles identiques pour tous les participants. Pour l'entraînement, nous utilisions toujours les mêmes symboles pour tous les participants. Pour les sessions en IRMf cependant, le sens associé à chaque symbole était contrebalancé entre les participants. Nous voulions ainsi éviter que des propriétés visuelles liées aux symboles puissent éventuellement créer des biais dans l'apprentissage. Au sein d'une paire, la position de chaque symbole sur l'écran (gauche/droite) variait aléatoirement d'un essai à l'autre. Cela nous permettait de nous assurer que les participants apprenaient bien à associer des valeurs aux symboles et pas à leur localisation spatiale. Les symboles associés aux gains ainsi que ceux associés aux pertes apparaissaient dans 24 essais chacun. Les symboles toujours associés au résultat neutre n'apparaissaient, eux, que dans 12 essais. Nous avions réparti les essais en 12 mini-blocs de 5 essais. Chaque mini-bloc était constitué de 2 essais avec la paire associée à des gains, 2 essais avec la paire associée à des pertes et 1 essai avec la paire associée à un résultat neutre. L'ordre de présentation des essais au sein d'un mini-bloc était aléatoire. La structure en mini-blocs n'était pas connue des participants mais elle nous permettait de nous assurer que l'ordre de présentation des différents types de paires et de symboles était à peu près équivalent.

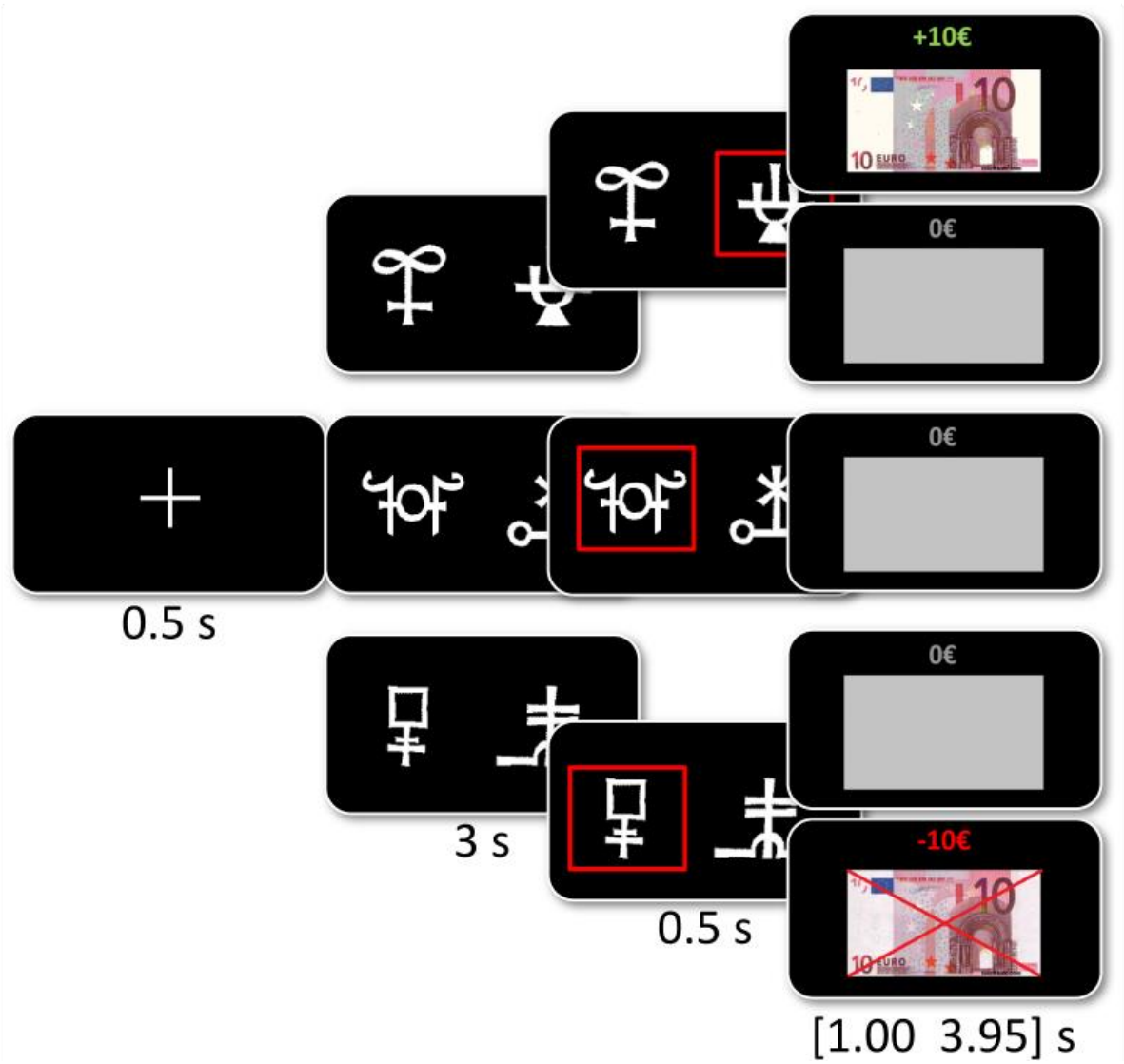

Figure 36 : Tâche d'apprentissage. Un essai démarrait toujours avec la croix de fixation affichée pendant 500 millisecondes. Ensuite, les deux symboles d'une paire apparaissaient à l'écran. Le participant devait sélectionner un des deux symboles à l'aide d'un boîtier de réponse où deux boutons correspondaient à la localisation spatiale (gauche/droite) de chaque option. Cette période de sélection durait 3 secondes. Le participant devait garder le doigt appuyé sur le bouton de l'option choisie, sans appuyer sur l'autre bouton, jusqu'à la fin de la période de sélection pour qu'une réponse soit considérée comme valide. Lorsque cette période était terminée, l'option sélectionnée, le cas échéant, était entourée d'un cadre rouge. Cette période durait 500 millisecondes. Si la réponse n'était pas valide, alors un message apparaissait à l'écran indiquant aux participants pourquoi la réponse n'était pas considérée comme valide et ils obtenaient le pire résultat possible pour la paire en question. Dans le cas de la paire de symboles associées à des gains (haut), l'un des deux symboles permettait d'obtenir une récompense de +10€ dans 75% des essais et rien (0€) dans les 25% d'essais restants, alors que l'autre symbole avait les probabilités inverses. Pour la paire de symboles associés à un résultat neutre (milieu), quelle que soit l'option choisie, le résultat était systématiquement 0€. Enfin, dans le cas de la paire associée à des pertes (bas), l'un des deux symboles faisait perdre 10€ dans 75% des essais et ne provoquait rien (0€) dans les 25% restants, alors que l'autre faisait de même mais avec les probabilités inverses. Le résultat final était affiché pour une durée variant aléatoirement entre

1.00 et 3.95 secondes. Une fois cette période terminée, un nouvel essai démarrait avec à nouveau l'affichage de la croix de fixation.

La session d'entraînement avant de rentrer dans le scanner contenait 60 essais. A la fin de cette session, nous vérifions que le participant sélectionnait bien le meilleur symbole possible dans plus de 50% des essais pour les paires associées à des gains et les paires associées à des pertes. Si ce n'était pas le cas, nous demandions aux participants de refaire une session complète, avec les mêmes symboles associés aux mêmes résultats, jusqu'à ce que nous soyons assurés qu'ils avaient bien compris la structure de la tâche.

À la fin de chaque session d'apprentissage, les participants pouvaient voir le total des gains accumulés sur l'ensemble des sessions d'apprentissage effectuées dans le scanner. Ce total était indépendant des deux autres tâches effectuées.

III.2.c Analyse comportementale

- **Modèle computationnel**

Nous avons modélisé le comportement de chaque individu séparément en utilisant un modèle de « Q-learning » standard (Watkins and Dayan, 1992) à l'aide de la VBA_toolbox (Daunizeau et al., 2014). Le modèle a été inversé en utilisant l'approche variationnelle bayésienne avec une approximation de Laplace. Ce modèle dit de « Q-learning » est utilisé de longue date pour modéliser l'apprentissage par renforcement et rend bien compte des choix opérés par les participants dans des tâches de ce type (FitzGerald et al., 2012; Palminteri et al., 2009; Pessiglione et al., 2006). Le principe du modèle est que chaque option de chaque paire est associée à un état caché, une valeur Q, et que cet état évolue au fur et à mesure de la tâche en fonction de la séquence de choix et de feedbacks. Ces états cachés représentent la récompense (ou la punition) attendue du fait de sélectionner l'option en question. Toutes les valeurs Q initiales ont été définies avec une valeur de zéro. Nous partions ainsi du principe que comme les participants ne savaient pas quelle est la valeur associée à chaque option au début de chaque session, ils démarraient avec une valeur attendue nulle pour chaque option. Par la suite, la valeur associée à chaque option évolue, au sein d'une fonction d'évolution, selon la formule :

$$Q_{choisie}(t+1) = Q_{choisie}(t) + \alpha * PE(t)$$

(Équation 10)

$Q_{choisie}(t+1)$ est la valeur de l'option choisie à l'essai t après l'avoir mise à jour au cours de l'essai, $Q_{choisie}(t)$ est la valeur attendue de l'option choisie à l'essai t, α est le taux

d'apprentissage et *PE(t)* est l'erreur de prédiction. *PE(t)* se décompose selon la formule $PE(t) = feedback(t) - Q_{choisie}(t)$ où *feedback(t)* est une variable égale à 1 lorsque le feedback obtenu est un gain (+10€), à 0 lorsque le feedback obtenu est neutre (0€) et à -1 lorsque le feedback obtenu est une perte (-10€). Nous avons aussi défini le modèle selon le raisonnement contrefactuel, de sorte que les valeurs des deux options affichées soient mises à jour à la suite d'un essai, partant du principe que les participants ont compris, au cours de la session d'entraînement que les deux symboles d'une même paire étaient associés. La valeur de l'option non-choisie évoluait ainsi selon la formule :

$$Q_{non-choisie}(t+1) = Q_{non-choisie}(t) + \alpha * PE\_bis(t)$$

(Équation 11)

$Q_{non\text{-}choisie}(t+1)$ est la valeur de l'option non-choisie à l'essai *t* après l'avoir mise à jour au cours de l'essai, $Q_{non\text{-}choisie}(t)$ est la valeur attendue de l'option non-choisie à l'essai *t*, α est le taux d'apprentissage estimé par le modèle et *PE_bis(t)* est l'erreur de prédiction pour la valeur de l'option non-choisie. *PE_bis(t)* est basée sur l'idée que les participants mettaient à jour la valeur de l'option non-choisie comme s'ils avaient aussi pu observer le feedback associé à l'option non-choisie de sorte que $PE_{bis(t)} = feedback_{bis}(t) - Q_{non-choisie}(t)$ où $feedback_{bis}(t)$ vaut 1 lorsque l'option non-choisie était associée à un gain (+10€), à 0 lorsque elle était associée à un feedback neutre (0€) et à -1 lorsqu'elle était associée à une perte (-10€). Bien évidemment, dans le cas des paires neutres, aucun apprentissage ne pouvait être modélisé étant donné que les feedbacks obtenus étaient toujours de 0€.

Ce modèle nous a aussi permis de modéliser les choix effectués par les participants au sein d'une fonction d'observation qui permet de dériver la probabilité, au sein d'une paire, de choisir la meilleure des deux options selon la formule :

$$p(choix = meilleure\ option) = \frac{1}{1 + \exp\left(-\frac{DV}{\beta}\right)}$$

(Équation 12)

Où β est la température estimée par le modèle qui capte le fait que les participants répondent plus ou moins aléatoirement, *DV* est la différence entre l'état caché associé à la meilleure des deux options et l'état caché associé à la moins bonne des deux options. β était contraint à être positif dans le modèle.

Nous avons utilisé les mêmes paramètres α et β pour les symboles associés aux gains et les symboles associés aux pertes. Nous avons aussi utilisé les mêmes paramètres à travers les 3 sessions d'un individu donné. Le taux d'apprentissage α permet de modéliser le fait que chaque individu pourrait accorder un poids différent aux feedbacks obtenus. La VBA_toolbox permet de trouver, pour un participant donné, le set de paramètres α et β qui maximise l'énergie libre du modèle selon l'approche variationnelle bayésienne.

- **Analyse des temps de réaction**

Nous avons voulu voir si les variables dérivées du modèle permettaient d'expliquer les temps de réaction. Pour la mesure du temps de réaction, nous avons pris en compte le temps mis entre l'apparition des stimuli à l'écran et le moment du premier appui sur une des touches de réponse. Pour l'analyse, nous avons d'abord regroupé les données à travers les trois sessions. Ensuite, nous avons regroupé les essais avec une paire associée à des gains avec les essais contenant une paire associée à des pertes, en laissant les essais avec une paire associée à un résultat toujours nul (paires neutres) à part. Nous avons ensuite effectué un GLM pour voir quelles variables expliquent le temps de réaction en incluant les régresseurs suivants :

1) une constante pour chaque session, au cas où il existe des différences d'une session à l'autre ;
2) Val : la sommes des valeurs des deux options présentées à chaque essai ;
3) Conf : une mesure indirecte de la confiance dans la décision estimée selon la formule $[p(gauche) - 0.5]^2$ où la probabilité *p(gauche)* était la probabilité de choisir l'option située à gauche de l'écran qui a été dérivée du modèle d'apprentissage détaillé précédemment.

Ensuite, nous avons moyenné les betas issus de ce GLM à travers les participants. Puis, nous avons testé leur significativité à l'aide d'un t.test en les comparant à zéro.

III.2.d IRM

- **Acquisition et pré-traitement**

Les sessions de cette tâche ayant été acquises en même temps que la tâche d'effort physique et la tâche d'effort mental, le lecteur est invité à se reporter à cette partie pour voir les paramètres d'acquisition des données (voir *Acquisition*). De plus, nous avons prétraité les données conjointement pour les 3 tâches et avons ainsi aussi utilisé les mêmes paramètres pour le pré-traitement que pour ces tâches (voir *Pré-traitement*).

- **Analyse des données d'IRMf : les modèles utilisés**

Après le prétraitement, nous avons analysé les données à l'aide de modèles linéaires généralisés (*general linear model*, GLM) dans SPM12 au premier niveau pour chaque sujet. Pour chaque sujet et chaque session, 6 paramètres de mouvement étaient directement estimés par SPM12 et rajoutés au sein du GLM. Ensuite, nous avons effectué une analyse au second niveau en regardant la significativité des paramètres issus du premier niveau au niveau du groupe.

Dans le GLM principal (**GLM1**) que nous avons effectué, nous avons modélisé le moment de l'affichage des symboles associés aux paires de gains et de pertes ensemble avec une fonction boxcar enveloppant la durée de l'affichage (3s). Cette fonction était modulée avec les régresseurs suivants, rentrés dans cet ordre : 1) Val, 2) Conf, 3) DT. Nous avons aussi modélisé séparément les paires neutres avec une fonction boxcar enveloppant la durée de l'affichage des symboles à l'écran (3s). Cet évènement était modulé uniquement par le temps de délibération à chaque essai. Le modèle incluait aussi un évènement pour le moment où le stimulus choisi apparaissait en rouge à l'écran. Cet événement était modélisé avec une fonction boxcar enveloppant la durée d'affichage de la réponse (0.5s). Enfin, nous avons modélisé le moment d'affichage du feedback pour les paires gains et les paires pertes ensemble avec une fonction delta de durée nulle. Cette fonction était modulée par 2 régresseurs rentrés dans cet ordre :1) l'erreur de prédiction non-signée pour l'option choisie, 2) l'erreur de prédiction signée pour l'option choisie. Enfin nous avons aussi modélisé le moment où les participants voyaient le feedback des paires neutres avec une fonction delta de durée nulle mais sans modulateur paramétrique.

Les régresseurs étaient orthogonalisés par SPM de manière sérielle au sein de chaque bloc, c'est-à-dire que le temps de délibération DT était orthogonalisé par rapport aux régresseurs Val et Conf et l'erreur de prédiction signée était orthogonalisée par rapport à l'erreur de prédiction non-signée.

Nous avons effectué un deuxième GLM, le **GLM2**, dont le but était de voir dans quelle mesure les corrélations observées dans le GLM1 étaient, ou non, impactées par l'ordre des régresseurs dans le GLM1. Ce GLM était ainsi strictement identique au GLM1 à part pour l'ordre des modulateurs paramétriques modulant l'apparition des symboles à l'écran dont l'ordre était inversé de la manière suivante : 1) DT, 2) Conf et 3) Val. Dans ce GLM aussi, toutes les variables étaient orthogonalisées de manière sérielle.

Nous avons aussi effectué un autre GLM, le **GLM3**, dont le but était d'extraire l'activité de nos régions d'intérêt pour chaque essai. Dans cette optique, ce modèle n'incluait aucun modulateur paramétrique. Nous avons modélisé indépendamment chaque apparition de stimuli à l'écran, chaque apparition de l'option choisie avec un cadre rouge et chaque apparition du feedback à l'écran. Tous ces événements étaient modélisés avec une fonction delta, c'est-à-dire avec une durée nulle. Nous pouvions ainsi obtenir l'activité moyenne d'une région d'intérêt donnée pour chaque période de chaque essai.

- **Régions d'intérêt**

Nous avons établi nos régions d'intérêt sur la base des corrélats cérébraux de Val et DT dans notre première étude (voir **Figure 27**). Le vmPFC a ainsi été défini comme une sphère avec un rayon de 8 millimètres autour des coordonnées MNI (-10 ; 48 ; -12) et le dmPFC comme une sphère avec un rayon de 8 millimètres autour des coordonnées MNI (10 ; 12 ; 48).

III.3 Résultats

III.3.a Le comportement

Les participants ont effectué une tâche d'apprentissage par renforcement mêlant à la fois des symboles associés à des gains, des symboles neutres et des symboles associés à des pertes. Ils devaient essayer, à chaque essai, de choisir l'option qui leur permettrait de maximiser leurs gains et de minimiser leurs pertes. Comme on peut le voir sur la figure (voir **Figure 37**), les participants ont appris à sélectionner plus souvent le symbole qui leur permettait de maximiser leurs gains dans le cas des paires associées à des gains. Ils ont aussi appris à sélectionner plus souvent le symbole qui leur permettait de minimiser leurs pertes dans le cas des paires associées à des pertes (voir **Figure 37**). Le taux d'apprentissage α estimé par le modèle était de 0.12 ± 0.07 (moyenne ± écart-type). La température moyenne β estimée par le modèle était de 0.17 ± 0.09. Le R^2 moyen du modèle était de 0.29 ± 0.15 (moyenne ± écart-type) et son AIC de -73.49 ± 20.66 (moyenne ± écart-type). En résumé, les participants, à l'exception de celui que nous avons exclu, ont bien compris la tâche. Ils ont réussi à apprendre comment maximiser leurs gains et minimiser leurs pertes en sélectionnant, dans chaque paire, le meilleur symbole possible. De plus, le modèle que nous avons utilisé semble relativement bien rendre compte du comportement des participants.

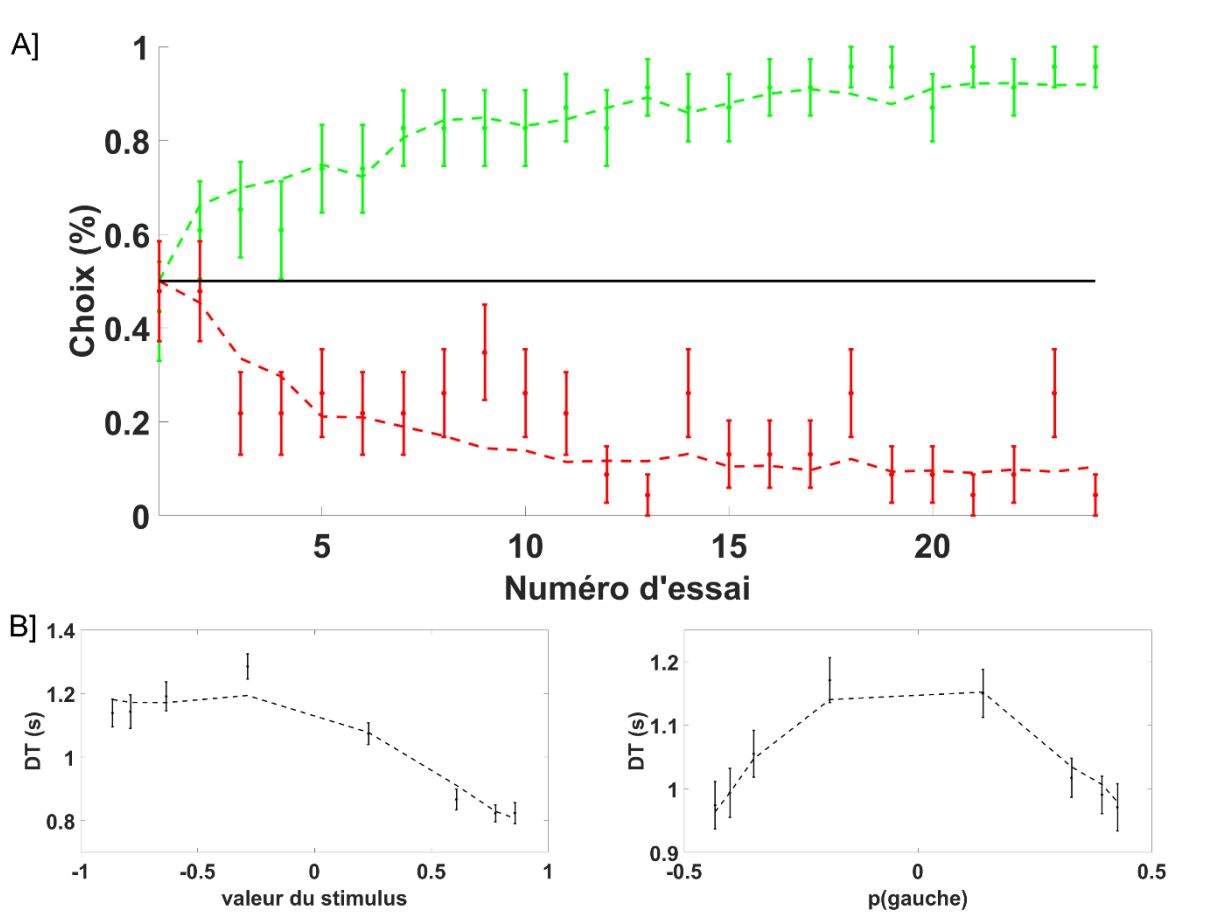

Figure 37 : Le comportement dans la tâche d'apprentissage.

A] **L'apprentissage au cours du temps.** Au sein de chaque session, chaque paire apparaît 24 fois. Les points de données en vert montrent le taux moyen, à travers les sessions et les participants, où l'option associée à des gains dans 75% des occurrences a été choisie pour chaque occurrence de la paire gain à l'écran. Les traits en pointillés vert montrent le résultat moyen à travers les sessions et les participants de l'estimation du modèle pour la probabilité de choisir la meilleure des deux options de la paire associée à des gains pour chaque occurrence. Les points de données en rouge montrent le taux moyen, à travers les sessions et les participants, où l'option associée à des pertes dans 75% des essais a été choisie pour chaque occurrence de la paire perte à l'écran. Les traits en pointillés rouge montrent le résultat moyen à travers les sessions et les participants de l'estimation du modèle pour la probabilité de choisir la pire des deux options de la paire associée à des pertes pour chaque occurrence. Les barres d'erreurs représentent l'erreur type de la moyenne à travers les participants dans les deux conditions. Cette figure a été établie sur la base des 23 participants inclus dans l'étude.

B] **Temps de délibération (DT)** en fonction de la somme des valeurs dérivées du modèle computationnel, à gauche, et en fonction de la probabilité de choisir l'option de gauche centrée, à droite. Les deux graphiques regroupent les données issues des 23 participants de l'étude pour les essais avec des paires associées à des gains ou à des pertes au cours des 3 sessions d'apprentissage. Les points avec les marges d'erreur représentent la moyenne ainsi que l'erreur type de la moyenne des temps de délibération divisés en 8 bins chez chaque participant en fonction de la somme des valeurs des stimuli à gauche et de la probabilité de choisir l'option de gauche centrée à droite. Les traits en pointillés représentent la moyenne à travers les participants, pour chaque bin, de la prédiction estimée par le modèle.

Nous avons ensuite effectué une régression linéaire sur le temps de délibération (DT) pour voir s'il variait avec la valeur subjective (Val) et la confiance dans la réponse (Conf). Nous avons constaté que plus la valeur des options inférées par le modèle était élevée, plus le temps de délibération était rapide (β = -0,221 ± 0.019 ; p.value = $1 \cdot 10^{-10}$). Le DT était aussi d'autant plus rapide que Conf était élevée (β = -1.195 ± 0.284 ; p.value = $4 \cdot 10^{-4}$). Le DT variait donc avec la valeur du choix et avec la confiance (voir **Figure 37**).

III.3.b Résultats neuraux

Dans un premier temps, nous avons voulu vérifier que nous arrivions à répliquer les résultats de notre première étude en reproduisant la triple dissociation, dans le cortex préfrontal médian, entre Val, Conf et le DT (voir **Figure 27**). Nous avons constaté que, dans cette tâche, Val et Conf partageaient les mêmes corrélats cérébraux. Nous avons en effet identifié un cluster dans le vmPFC avec un cluster de 747 voxels dont le pic était localisé aux coordonnées (-8 ; 56 ; 2) pour Val et un cluster de 299 voxels dont le pic était localisé aux coordonnées (0 ; 56 ; -2) pour Conf. Nous avons d'ailleurs confirmé ce résultat avec une conjonction entre ces deux contrastes (voir **Figure 38A**) qui ne révélait qu'un cluster unique de 289 voxels dont le pic était situé aux coordonnées (0 ; 52 ; -4). Au seuil utilisé, le seul autre cluster significatif qui corrélait avec Val était un cluster de 476 voxels localisé dans le cortex cingulaire postérieur avec un pic aux coordonnées MNI (-4 ; -54 ; 16). Cette aire est régulièrement associée au vmPFC en lien avec l'encodage de la valeur subjective. Pour Conf, au seuil utilisé, seuls deux autres clusters apparaissaient, dont un situé à la frontière entre les ventricules et la matière blanche, qui constituait probablement un artéfact, et un autre de 440 voxels dans le précuneus, avec un pic aux coordonnées (2 ; -46 ; 54). Le DT était principalement corrélé à un cluster plus dorsal localisé dans le cortex préfrontal dorsomédian (dmPFC) (voir **Figure 38B**), malgré le fait que le DT était orthogonalisé par rapport à Val et Conf. La carte d'activation en lien avec le DT à $p < 0.05$ corrigé au niveau des voxels pour les comparaisons multiples révélait aussi plusieurs autres clusters dont un cluster bilatéral dans l'insula antérieure avec un cluster de 37 voxels dans l'insula droite aux coordonnées (32 ; 22 ; 4) et un cluster de 3 voxels dans l'insula gauche aux coordonnées (-34 ; 14 ; 6). Contrairement aux clusters associés à Val et Conf, ce cluster survivait même pour un seuil avec p.value <0.05 corrigée au niveau des voxels (voir **Figure 38B** bas).

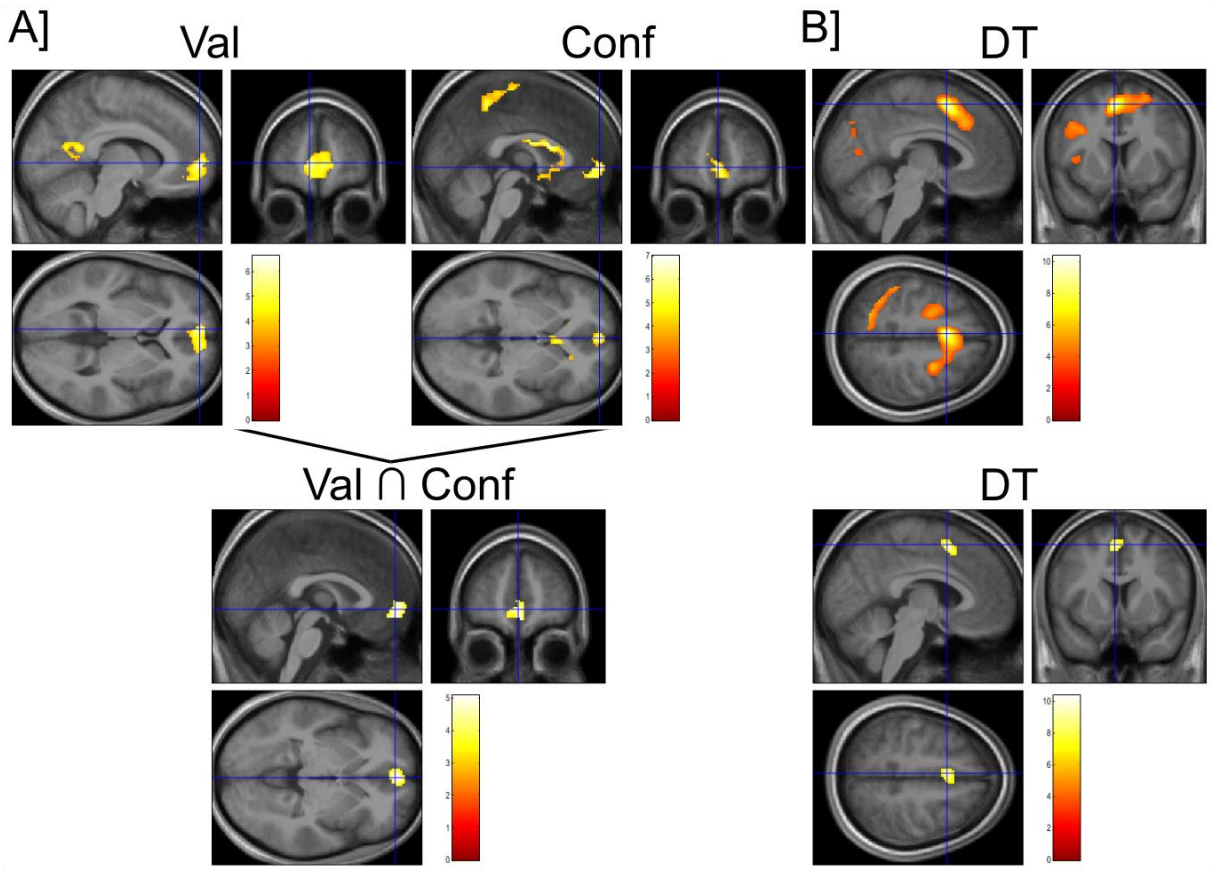

Figure 38 : Les corrélats cérébraux de Val, Conf et DT dans la tâche d'apprentissage. Les cartes d'activation sont superposées à la moyenne des scans anatomiques des 22 cerveaux inclus dans l'étude après le pré-traitement, en particulier après normalisation dans l'espace MNI. Les barres colorées montrent la correspondance entre les couleurs des cartes et les valeurs T pour chaque voxel.

A] **Val et Conf corrèlent avec l'activité du vmPFC.** En haut, ces cartes représentent la carte d'activation, pour le GLM1, correspondant à Val à gauche et à Conf à droite à travers les paires associées à des gains et celles associés à des pertes. Toutes les cartes sont seuillées à p.value < 0.05 corrigée au niveau des clusters pour les comparaisons multiples au niveau de la famille (*family wise error rate*). En bas, la figure montre le résultat d'une conjonction entre le contraste Val et le contraste Conf. La carte est aussi seuillée à p.value < 0.05, corrigée au niveau des clusters pour les comparaisons multiples au niveau de la famille. Le vmPFC est le seul cluster qui survit à ce seuil.

B] **DT corrèle avec le dmPFC.** Corrélats cérébraux du temps de réaction pour les paires associées à des gains et les paires associées à des pertes dans le GLM1. Les deux images montrent la même carte, cependant la carte du haut est seuillée à p.value < 0.05 corrigée au niveau des clusters pour les comparaisons multiples au niveau de la famille (family wise error rate), alors que la carte située en-dessous est seuillée à p.value < 0.05 corrigée au niveau des voxels pour les comparaisons multiples au niveau de la famille.

Ensuite, nous avons défini deux régions d'intérêt sous la forme d'une sphère de 8 millimètres de rayon, centrée sur les coordonnées issues de la première étude pour Val et DT (voir **Table 1**). Nous avons ainsi confirmé que l'activité du vmPFC était corrélé à Val (β =

0.523 ± 0.112, p = 0.0001) et Conf (β = 6.931 ± 2.232, p = 0.005), mais qu'il n'était pas significativement lié à DT (β = -0.148 ± 0.088, p = 0.108) (voir **Figure 39**, haut). À l'inverse, l'activité du dmPFC corrélait négativement avec Val (β = -0.288 ± 0.093, p = 0.005). L'activité du dmPFC avait aussi une tendance non-significative à décroître avec la confiance subjective (β = -1.767 ± 1.044, p = 0.105) et à augmenter avec le DT (β = 0.319 ± 0.066, p = 0.0001). De plus, l'activité du vmPFC était mieux corrélée à Val ($p = 5 \cdot 10^{-7}$) et à Conf ($p = 7 \cdot 10^{-8}$) que l'activité du dmPFC. A l'inverse, l'activité du dmPFC était mieux corrélée à DT que l'activité du vmPFC ($2 \cdot 10^{-4}$). Pour voir si ces résultats étaient liés à la manière dont nos régresseurs avaient été orthogonalisés, nous avons effectué un deuxième GLM où l'ordre des régresseurs était inversé, de sorte à ce que Val soit orthogonalisé à Conf et DT. En particulier, nous voulions voir si la corrélation entre le dmPFC et Val était liée au fait que la délibération était plus rapide à mesure que Val était plus élevée (voir **Figure 37**) ou si elle reflétait réellement un lien entre le dmPFC et la valeur subjective. Nous avons alors observé que les résultats étaient relativement inchangés dans le vmPFC (voir **Figure 39**, bas) dont l'activité était toujours liée à Val (β = 0.392 ± 0.135, p = 0.008) et à Conf (β = 6.526 ± 2.593, p = 0.020), mais pas à DT (β = -0.148 ± 0.088, p = 0.1075), malgré la corrélation qui existait entre le DT et Val et Conf (voir **Figure 37**). Par contre, l'activité du dmPFC était toujours significativement corrélée à DT (β = 0.319 ± 0.066, p = 0.0001) mais elle n'était plus significativement corrélée à Val (β = -0.114 ± 0.104, p = 0.284). Elle n'était toujours pas corrélée à Conf non plus (β = -0.105 ± 1.191, p = 0.930). Nous avons ainsi vu que le vmPFC était positivement corrélé à la fois à Val et à Conf, alors que le dmPFC était positivement corrélé avec le temps de délibération.

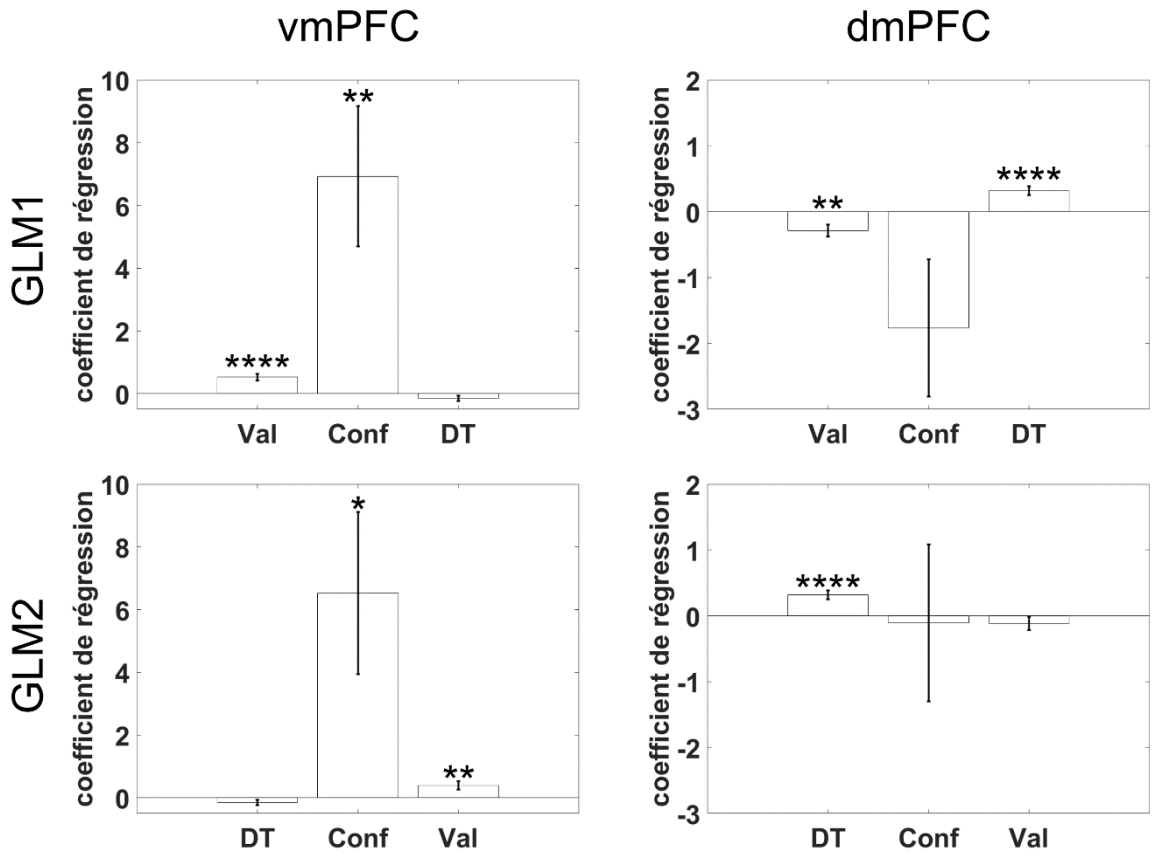

Figure 39 : Lien entre Val, Conf, DT et l'activité du vmPFC et du dmPFC. Les figures représentent la corrélation d'une sphère d'un rayon de 8 millimètres, centrée, à gauche, sur les coordonnées du vmPFC (-10 ; 48 ; -12) et, à droite, sur les coordonnées du dmPFC (10 ; 12 ; 48). Les coordonnées sont issues de notre première étude pour définir le vmPFC en lien avec Val et le dmPFC en lien avec le DT (voir **Table 1**). Les figures du haut sont issues du GLM1 où les régresseurs étaient orthogonalisés dans l'ordre sériel Val/Conf/DT. Les figures du bas sont issues du GLM2 où les régresseurs étaient orthogonalisés dans l'ordre sériel inverse DT/Conf/Val. Les barres et les barres d'erreurs montrent, respectivement, la moyenne et l'erreur type de la moyenne à travers les participants. Les étoiles indiquent la significativité à la suite d'un t-test des données contre zéro : *p.value < 0.05 ; **p.value < 0.01 ; ***p.value < 0.005 ; ****p.value < 0.001.

Nous avons aussi voulu observer les corrélats de l'erreur de prédiction signée et non-signée au moment de l'apparition du feedback à l'écran. De manière cohérente avec son rôle dans l'encodage de la valeur subjective, l'activité du vmPFC augmentait linéairement avec l'erreur de prédiction signée (voir **Figure 40**), c'est-à-dire que son activité était d'autant plus grande que le gain obtenu était plus important que ce qui était attendu et qu'elle était au contraire d'autant plus faible que la perte obtenue était plus conséquente que ce qui était attendu. On peut d'ailleurs aussi voir sur la carte deux autres composantes souvent associées au vmPFC dans ce qui a été appelé le réseau des valeurs, dans la présence du cortex postérieur cingulaire et du striatum ventral. L'activité du vmPFC était aussi négativement corrélée avec l'erreur de

prédiction non-signée (voir **Figure 40**), c'est-à-dire que son activité était d'autant moins importante que la surprise liée au résultat donné dans le feedback était grande. A l'inverse, nous n'avons pas observé de corrélation significative entre le dmPFC et l'erreur de prédiction signée (β_{PE} = 0.190 ± 0.261, p = 0.474) ou l'erreur de prédiction non-signée ($\beta_{|PE|}$ = -0.152 ± 0.219, p = 0.494).

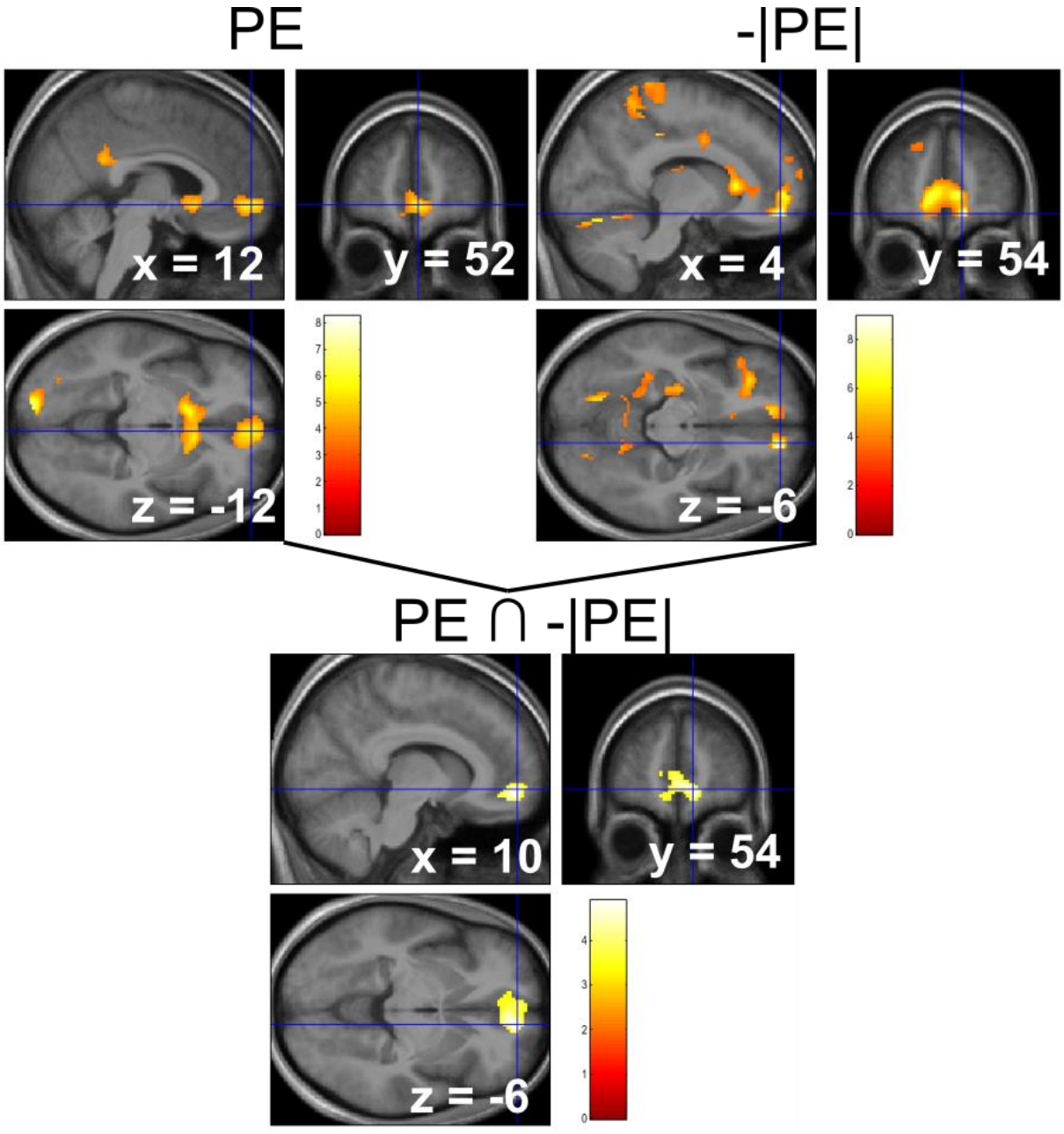

Figure 40 : Erreur de Prédiction au moment du feedback et activité du vmPFC. En haut à gauche, zoom sur le pic d'un cluster dans le vmPFC (12 ; 52 ; -12) corrélé à l'erreur de prédiction signée. En haut à droite, zoom sur le pic d'un cluster dans le vmPFC (4 ; 54 ; -6) négativement corrélé à l'erreur de prédiction non-signée. En bas, la figure représente la conjonction entre le contraste correspondant à la corrélation positive avec l'erreur de prédiction signée et celui correspondant à la corrélation négative avec l'erreur de prédiction non-signée.

Les cartes d'activation sont superposées à la moyenne des scans anatomiques des 22 cerveaux inclus dans l'étude après le pré-traitement, en particulier après normalisation dans l'espace MNI. Les barres colorées montrent la correspondance entre les couleurs des cartes et les valeurs T pour chaque voxel. Toutes les cartes sont montrées pour un seuil de p.value < 0.05 corrigée au niveau des clusters pour les comparaisons multiples au niveau de la famille (*Family Wise Error Rate*).

III.4 Discussion

III.4.a L'effort dans la délibération

L'apprentissage par renforcement est un mécanisme relativement universel. Il consiste dans l'idée simple qu'un évènement suivi d'une récompense sera associé à l'obtention de cette récompense et qu'un évènement suivi d'une punition sera associé au fait de subir cette punition, par association. Au contraire, il a été proposé que les Primates, et l'homme plus particulièrement, soit capable d'effectuer une forme plus perfectionnée de ce mode d'apprentissage, notamment en allouant plus d'effort mental (Pearce and Hall, 1980). De plus, il est connu que, dans des tâches impliquant d'explorer l'environnement par essai-erreur, le cortex préfrontal dorsomédian (dmPFC), ainsi que l'insula antérieure, qui est souvent co-activée avec lui, ont tendance à s'activer plus intensément pendant la période d'exploration que pendant la période d'exploitation de la réponse correcte (Amiez et al., 2012, 2013, 2016). Plus généralement, l'activité du dmPFC et de l'insula antérieure suite à des erreurs semblent entraîner une adaptation comportementale qui peut se résumer par un ralentissement de la vitesse de réponse et une baisse du nombre d'erreurs en vue d'améliorer la performance (Bastin et al., 2016; Iannaccone et al., 2015). Or, le dmPFC et l'insula antérieure constituent deux structures associées à l'exécution de contrôle cognitif (Menon and Uddin, 2010; Shenhav et al., 2013, 2016a), ou du moins au recrutement du dlPFC censé être la structure qui exerce le contrôle cognitif. Dans notre tâche d'apprentissage par renforcement, nous avons constaté une augmentation de l'activité du dmPFC en lien avec des temps de délibération plus longs (voir **Figure 38**). Cette corrélation était observée alors même que les temps de délibération étaient orthogonalisés par rapport à la valeur subjective et la confiance dans la décision, deux variables auxquelles ils étaient liés. Ce résultat suggère que même dans cette tâche d'apprentissage par renforcement relativement simple, les participants utilisaient du contrôle cognitif au moment de délibérer sur le symbole qu'ils allaient choisir en vue de maximiser leurs gains et de minimiser leurs pertes. Ce résultat est d'ailleurs cohérent avec les résultats de notre première étude où nous avons aussi observé une corrélation significative entre le temps de délibération dans chacune de nos trois tâches et l'activité du dmPFC et de l'insula antérieure bilatéralement

(voir **Figure 27** et **Table 1**). Enfin, nous avons vu que l'activité du dmPFC était aussi corrélée négativement avec Val dans le GLM1, mais que cette corrélation disparaissait dans notre GLM2, lorsque Val était orthogonalisé par rapport au DT. Cela suggère que, dans notre tâche du moins, si le dmPFC s'activait plus en réponse à l'anticipation de pertes plus importantes, c'était principalement lié au fait que cette condition induisait des temps de délibération plus longs. Dans notre tâche, il se pourrait ainsi que les symboles associés à des pertes impliquent une délibération plus importante. Une possibilité est que les participants présentent une aversion à la perte qui les amène à donner plus d'importance à des pertes potentielles par rapport à des gains potentiels (Kahneman and Tversky, 1984) d'où des temps de délibération plus importants. Une hypothèse alternative serait liée au fait que, dans le cas des pertes, les participants sont moins souvent exposés au feedback associé à la perte en comparaison à la situation associée à des gains. De ce fait, les participants pourraient être légèrement moins certains de la réponse à donner puisque le symbole associé à la perte n'a pas été négativement renforcée au même degré que le symbole associé au gain a été positivement renforcé. Enfin, il est possible qu'il existe un effet pavlovien qui pousse à prendre plus de temps à répondre dans l'anticipation de pertes par rapport à l'anticipation de gains (Dayan, 2008; Shadmehr et al., 2019) et que ce biais incite en quelque sorte le dmPFC à prendre plus de temps pour délibérer dans cette situation. Dans tous les cas, la corrélation entre le dmPFC et le DT au moment de la délibération suggère que notre tâche simple d'apprentissage par renforcement implique de l'effort mental. Les participants ne choisiraient pas un symbole de manière totalement automatisée, lorsqu'ils doivent choisir un symbole en vue d'optimiser leurs gains dans notre tâche, mais ils prendraient le temps de délibérer en exécutant un effort mental pour s'assurer de faire le meilleur choix possible.

III.4.b La confiance subjective pendant la délibération

Le fait de prendre plus de temps avant de prendre une décision pourrait refléter la recherche, au niveau méta-décisionnel, d'augmenter la confiance subjective dans la décision prise (Lee and Daunizeau, 2019). Nous avons d'ailleurs pu voir que les corrélats neuraux de la confiance étaient localisés dans le cortex préfrontal ventromédian (vmPFC) au sein de la même zone que celle qui corrélait avec la valeur subjective (voir **Figure 38**). Notre analyse consiste d'ailleurs, à notre connaissance, de la première analyse montrant des corrélats cérébraux de la confiance dans une tâche d'apprentissage par renforcement. La corrélation entre Val et le vmPFC est cohérente avec des résultats précédents dans une tâche semblable (Palminteri et al., 2009) ainsi qu'avec le rôle plus général du vmPFC dans l'encodage de la valeur subjective de manière générique (Bartra et al., 2013; Levy and Glimcher, 2012). Par ailleurs, dans notre

première étude incluant des tâches basées sur les préférences subjectives, le vmPFC corrélait aussi avec la valeur et avec la confiance subjective (voir **Figure 27**). Cependant, l'épicentre de la corrélation était dorsal au vmPFC. D'un autre côté, le lien entre le vmPFC et la confiance subjective est en accord avec une vaste littérature montrant un lien entre le vmPFC et la confiance subjective dans divers contextes (Chua et al., 2006; De Martino et al., 2013; Fleming et al., 2018; Gherman and Philiastides, 2018; Kuchinke et al., 2013; Lebreton et al., 2015; Moritz et al., 2006; Rolls et al., 2010; Shapiro and Grafton, 2020). Une limite de cette étude, dont ce n'était pas la question principale, est que les choix des participants avaient un rôle instrumental sur leur paiement final. Une confiance accrue correspondait ainsi à une espérance de gains plus élevée. De ce fait, dans cette tâche, contrairement à notre première étude où les décisions n'étaient pas instrumentales, le plaisir attendu à obtenir des gains, ou à éviter des pertes, et le plaisir lié au fait que la probabilité d'avoir choisi la meilleure des deux options était plus élevée, étaient confondus. Cela pourrait expliquer pourquoi l'épicentre de la corrélation avec la confiance, dans cette tâche, était centré sur le vmPFC. Il serait intéressant de voir si l'épicentre de la corrélation avec la confiance subjective varie selon

III.4.c L'activité du vmPFC au moment du feedback

D'un autre côté, nous avons observé une corrélation significative dans le sens positif entre l'activité du vmPFC et l'erreur de prédiction signée et dans le sens négatif entre l'activité du vmPFC et l'erreur de prédiction non-signée (voir **Figure 40**). Ce résultat est cohérent avec l'idée que l'activité du vmPFC reflète la valeur subjective du feedback. En effet, premièrement, une erreur de prédiction signée plus élevée indique l'obtention de gains plus élevés que ce que le modèle interne prévoyait. À l'inverse, une erreur de prédiction signée plus faible signale que des pertes plus importantes que prévu ont été subies. Ce résultat est d'ailleurs en accord avec une vaste littérature liant les deux composantes principales du réseau des valeurs, qui inclue le vmPFC et le striatum ventral, avec l'erreur de prédiction signée dans des tâches d'apprentissage par renforcement (Fouragnan et al., 2018; Garrison et al., 2013). Deuxièmement, la corrélation négative entre l'activité du vmPFC et l'erreur de prédiction non-signée pourrait représenter la satisfaction subjective liée à la confirmation que le choix effectué était le bon, puisque le décalage entre les attentes et le résultat obtenu est minime.

III.5 Conclusion

Premièrement, nous avons pu répliquer une partie des résultats de notre première étude et de la littérature dans cette tâche mélangeant des choix binaires liés à la valeur avec de