



**HAL**  
open science

# Approches neurocomputationnelles de la prise de décision sociale

Rémi Philippe

► **To cite this version:**

Rémi Philippe. Approches neurocomputationnelles de la prise de décision sociale. Neurosciences. Université de Lyon, 2021. Français. NNT : 2021LYSE1121 . tel-03856253

**HAL Id: tel-03856253**

**<https://theses.hal.science/tel-03856253v1>**

Submitted on 16 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°d'ordre NNT: 2021LYSE1121



THESE DE DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'UNIVERSITÉ CLAUDE BERNARD LYON 1

ÉCOLE DOCTORALE NEUROSCIENCES ET COGNITION ED 476

---

Thèse de Doctorat de Neurosciences Cognitives

*Présentée et soutenue publiquement par*

**Rémi PHILIPPE**

*Le 1er Juillet 2021*

---

APPROCHES NEUROCOMPUTATIONNELLES DE LA PRISE  
DE DÉCISION SOCIALE

---

*Directeurs de thèse :*

**Jean-Claude DREHER** - Directeur de recherche CNRS, Université Claude Bernard Lyon 1  
**Edmund DERRINGTON** - Professeur des Universités, Université Claude Bernard Lyon 1

*Président du jury:*

**BOULINGUEZ Philippe** - Professeur des Universités, Université Claude Bernard Lyon 1

*Jury:*

**GIRARD Benoit** - Directeur de Recherche CNRS, Sorbonne Universités - Rapporteur  
**HOPFENSITZ Astrid** - Maître de Conférences Toulouse School of Economics - Rapportrice  
**AMIEZ Céline** - Chargée de Recherche CNRS, Université Claude Bernard Lyon 1  
**PALMINTERI Stefano** - Chargé de Recherche INSERM, École Normale Supérieure de Paris



*L'université Lyon 1 ne donnera aucune approbation ou désapprobation sur les réflexions exprimées dans cette thèse. Elles ne sont que celles de l'auteur et doivent être considérées comme tels.*

Je dédie ce travail à la personne qui a su  
prendre la place la plus importante dans  
ma vie. Je t'en remercie.

---

*À toi Athina MARTIN...*

# Remerciements

Je vais profiter de ces quelques lignes souvent appréciées pour la liberté qu'elles offrent, pour expliquer à quel point j'ai apprécié ces quelques années de doctorat et pour remercier toutes les personnes qui me sont chères.

Le doctorat est un projet de plusieurs années durant lequel le doctorant entreprend de tout savoir sur rien. L'autonomie, la curiosité mais aussi la patience sont de rigueur. Le plus important n'est pas le sujet sur lequel porte la thèse et auquel on pourrait réduire le docteur, comme un éminent spécialiste de l'insignifiant. Ce qui compte est le chemin parcouru pour en arriver là. La méthode de pensée, l'esprit critique et le contradictoire, l'esprit de synthèse et de vulgarisation, les capacités de communication sont les principaux atouts du jeune docteur. Toutes ces années d'étude appellent à la modestie et à rester humble face à la complexité du monde.

Je voudrais commencer par remercier mon directeur thèse, Jean-Claude DREHER avec qui les opportunités et les projets étaient toujours multipliés par cent, pour la confiance qu'il a eu en moi et pour m'avoir accompagné dans l'élaboration de ce projet. Un remerciement spécial au professeur Christian SCHEIBER qui a été le médecin responsable de toutes les expériences que j'ai effectuées à Lyon et que j'ai dérangées de nombreuses, nombreuses, nombreuses fois... Merci aussi à Edmund DERRINGTON d'avoir été mon co-directeur de thèse, d'avoir notamment joint ses capacités rédactionnelles aux nôtres pour sublimer nos résultats.

J'aimerais remercier mon binôme de travail qui est devenu mon ami, Rémi JANET, dont les conseils et les aides m'ont permis de rattraper d'un côté le temps que nous

avons perdu à profiter de la vie de l'autre. Je voudrais remercier toute les personnes de l'équipe avec qui j'ai plus ou moins collaborer, Élodie BARAT, Julien BENISTANT, Toan NONG, Valentin GUIGON, Andres POSADA, Izem MANGIONE, Alexia GERARDIN, Yang HU Jiawei LIU, Sasa ZHAO, Zixuan TANG.

J'aimerais aussi remercier l'équipe de levé de coudes "*Les Morues*" pour leur support émotionnel et leur manque de sérieux à toute épreuve. Équipe qui a mainte et mainte fois prouvé qu'on pouvait être efficace même en travaillant nu.

Une pensée spéciale pour ma chérie Athina Martin qui aura réussi à trouver plus de fautes d'orthographe qu'il n'y avait de mot dans ce manuscrit. À elle qui comprend que les chercheurs ne sont pas des gens comme tout le monde, à elle qui me soutient dans mes choix, qui me fait confiance, à elle avec qui je partage les petits "b" d'un grand bonheur au quotidien.

Je voudrais remercier ma famille avec qui j'apprends jour après jour la résilience, pour m'avoir soutenu dans le choix du doctorat qui n'était pas un choix évident. Leur fierté et leur reconnaissance est un moteur important du succès de ce projet.

Et pour finir je voudrais remercier tous les membres de mon jury : Céline AMIEZ; Philippe BOULINGUEZ; Stefano PALMINTERI et tout particulièrement mes rapporteurs Benoit GIRARD et Astrid HOPENSITZ pour le travail que cela leur a demandé. Et toute l'université de Lyon 1, parce que l'esprit universitaire c'est la recherche, la fabrication de savoir et l'interdisciplinarité mais aussi la transmission, la formation. La production de connaissances, le doute et la méthode sont des remparts contre l'obscurantisme, les préjugés ou l'immobilisme. Et même si les sciences et technologies ne sont pas toujours sources de progrès, une meilleure compréhension du monde qui nous entoure permet de faire des choix éclairés. C'est pourquoi je suis fier d'ajouter ma petite pierre à l'édifice. . .

# Table des matières

|   |          |
|---|----------|
| <b>Introduction générale</b>  | <b>1</b> |
| <b>1 Introduction</b>   | <b>3</b> |
| 1.1 La prise de décision . . . . .  | 3        |
| 1.1.1 Cadre normatif de l'étude de la prise de décision . . . . .                           | 4        |
| 1.1.1.1 Hypothèse de la rationalité . . . . .   | 5        |
| 1.1.1.2 Cadre de la théorie des jeux . . . . .  | 6        |
| 1.1.1.3 Décisions optimales, "sous-optimales" et équilibres . . . . .                       | 7        |
| 1.1.1.4 Processus d'apprentissage . . . . .   | 9        |
| 1.1.2 Structures cérébrales impliquées . . . . .  | 12       |
| 1.1.2.1 Bases neurales de l'évaluation d'un état ou d'une décision                          | 12       |
| 1.1.2.2 Bases neurales des émotions . . . . .   | 15       |
| 1.1.2.3 Bases neurales sensibles ou spécifiques à la prise de<br>décision sociale . . . . . | 19       |
| 1.1.2.4 Autre réseaux neuraux . . . . .   | 21       |
| 1.2 Spécificité du contexte social . . . . .  | 22       |
| 1.2.1 Différentes situations d'interaction . . . . .  | 23       |
| 1.2.1.1 Du point de vue d'un observateur . . . . .  | 23       |
| 1.2.1.2 Être observé . . . . .  | 34       |
| 1.2.1.3 Les interactions directes . . . . .   | 37       |
| 1.2.2 Quelles intentions derrière l'interaction . . . . .                                   | 38       |



|          |  |            |
|----------|--|------------|
| 1.2.2.1  | Les intentions compétitives . . . . .  | 39         |
| 1.2.2.2  | Les intentions coopératives . . . . .  | 42         |
| 1.2.3    | Émergence d'une hiérarchie et relation de dominance . . . . .  | 46         |
| 1.2.4    | Théorie de l'esprit . . . . .  | 57         |
| 1.3      | La modélisation . . . . .  | 62         |
| 1.3.1    | Qu'est ce qu'un modèle . . . . .   | 63         |
| 1.3.2    | Utilisation de la modélisation en neurosciences . . . . .  | 65         |
| 1.3.3    | Sélectionner parmi les modèles . . . . .   | 73         |
| 1.3.3.1  | Les mesures utilisées . . . . .  | 73         |
| 1.3.3.2  | Optimisation des paramètres libres . . . . .   | 76         |
| 1.3.3.3  | Critère de sélection des modèles . . . . .   | 78         |
| 1.3.4    | Que sont les modèles et inférences bayésiennes ? . . . . .   | 80         |
| 1.4      | Présentation des résultats obtenus . . . . .   | 83         |
| <b>2</b> | <b>Neurocomputational mechanisms engaged in detecting cooperative and competitive intentions of others</b>                             | <b>95</b>  |
| <b>3</b> | <b>Modeling other minds: Bayesian inference explains human choices in group decision-making</b>  | <b>145</b> |
| <b>4</b> | <b>Causal role of the medial prefrontal cortex in learning social hierarchies</b>  | <b>165</b> |
| <b>5</b> | <b>Regulation of social hierarchy learning by serotonin transporter availability</b>   | <b>213</b> |
|          | <b>Conclusion générale</b>   | <b>259</b> |
|          | <b>Appendix 1 - Perturbation of Right Dorsolateral Prefrontal Cortex (rDLPFC) Makes Power-Holders Less Resistant to Tempting Bribe</b> | <b>263</b> |

# List of Tables

|     |   |    |
|-----|---|----|
| 1.1 | Matrice de récompenses d'un jeu dans lequel deux stratégies sont possibles : coopérer ou faire défection. . . . . | 42 |
| 1.2 | Matrice de récompense du jeu du dilemme du prisonnier dans (Rilling et al., 2002). . . . .                        | 43 |
| 1.3 | Matrice de récompense du jeu du cache cache (Devaine et al., 2014a). . .  | 67 |



# List of Figures

|      |  |    |
|------|--|----|
| 1.1  | Neurones dopaminergiques dans la substance noire (SN) et l'aire tegmentale ventrale et prédiction de récompense. <i>Extrait de (Schultz et al., 1997)</i>  | 11 |
| 1.2  | Structures des réseaux de la récompense dans le cerveau humain <i>Extrait de Arias-Carrián et al. (2010)</i>   | 13 |
| 1.3  | Structures des régions principales et étendues liées aux émotions dans le cerveau humain. En rouge, ce sont les régions que l'on retrouve le plus fréquemment dans la littérature. En orange, ce sont celles que nous retrouvons régulièrement, mais de manière moins fréquente. <i>Extrait de Pessoa (2008)</i> | 17 |
| 1.4  | Multiplés interactions entre circuit d'évaluation et d'émotion.  | 18 |
| 1.5  | Structures des réseaux des interactions sociales dans le cerveau humain.   | 20 |
| 1.6  | Composition de différents réseaux cérébraux  | 22 |
| 1.7  | Tâche expérimentale de Suzuki et al. (2012).   | 25 |
| 1.8  | Activité neurale corrélée avec l'erreur de prédiction simulé de la récompense ou de l'erreur de la prédiction de l'action de l'autre lors de la condition <i>autrui</i> (où le participant observe les choix d'un autre).  | 26 |
| 1.9  | Tâche expérimentale, performance du modèle et localisation des électrodes la dans l'article de Hill et al. (2016).   | 28 |
| 1.10 | Cinq images extraites d'une des animations scénarisées (mère et enfant)  | 29 |
| 1.11 | Représentation des quatre facteurs variant dans la conception factorielle de la tâche de Baker et al. (2017).  | 30 |

|      |   |    |
|------|---|----|
| 1.12 | Tracés et corrélations (Coefficient de corrélation de Pearson $r$ , pour $n=21$ )<br>en utilisant les paramètres maximisant l'ajustement du modèle aux données ( $\beta = 2.5$ ) dans l'article de Baker et al. (2017). | 31 |
| 1.13 | Conception de la tâche de Charpentier et al. (2020).  | 33 |
| 1.14 | Signal de mise à jour de l'émulation et de l'imitation pendant l'observation.   | 34 |
| 1.15 | Principaux résultats de l'expérience de Li et al. (2020) sur l'effet d'audience.  | 36 |
| 1.16 | Résultats de Bault et al. (2011).   | 40 |
| 1.17 | Effet principal de l'adversaire au moment de la récompense (Fareri and Delgado, 2014).  | 41 |
| 1.18 | Principal réseau engagé dans les études de neuroimagerie lors de représentation de hiérarchie sociale basé sur la perception des rangs sociaux (Qu et al., 2017).   | 48 |
| 1.19 | Tâche expérimentale de Kumaran et al. (2012, 2016).   | 50 |
| 1.20 | Résumé des résultats de Kumaran et al. (2012).  | 52 |
| 1.21 | Résumé des résultats de Kumaran et al. (2016).  | 53 |
| 1.22 | Déroulement de l'expérience IRMf d'apprentissage d'une hiérarchie par interaction dyadique (Ligneul et al., 2016).  | 55 |
| 1.23 | Résultats of Ligneul et al. (2016).   | 56 |
| 1.24 | Comportement et activité cérébrale pour les basses et hautes profondeurs de théorie de l'esprit.  | 60 |
| 1.25 | Penser et apprendre : calculs et corrélats neuronaux des différentes stratégies de penser et d'apprentissage.   | 61 |
| 1.26 | Comparaison bayésienne de modèles.  | 68 |
| 1.27 | Conception expérimental de Palminteri et al. (2015).  | 71 |
| 1.28 | Résultats comportementaux et de la simulation de modèles de Palminteri et al. (2015).   | 72 |
| 1.29 | Compromis entre complexité du modèle et erreur de prédiction.   | 74 |
| 1.30 | Illusion d'optique 1  | 82 |
| 1.31 | Illusion d'optique 2  | 82 |

# Introduction générale

L'homme est un animal politique.

---

*Aristote*

De nombreux philosophes ont travaillé sur la nature sociale de l'Homo sapiens tant cette caractéristique est présente et intrigante dans notre espèce. Dans *ses Politiques* Aristote écrit que *L'homme est un animal politique* et insiste sur le caractère naturel du trait social chez l'Homo sapiens. Dans son cadre de pensée appelé *le finalisme*, dans lequel toute chose a un but, une fin en soi, celle de l'Homme est d'atteindre le bonheur qui n'est selon lui atteignable que par la vie en société. Il en veut pour preuve que l'Homme est le seul à avoir la capacité d'exprimer des pensées complexes grâce au langage. Nous verrons dans cette thèse par ailleurs que l'Homme a bien d'autres capacités sociales que celle de la parole.

Plus tard, Thomas Hobbes ira contre Aristote et son idée selon laquelle l'Homme est social par nature, sans jamais remettre en cause l'aspect social de notre espèce. Selon lui c'est par crainte réciproque d'une mort causée par un conspécifique que l'Homme s'est vu obligé de devenir social pour sa propre conservation.

Enfin, d'autres travaux notables sur l'Homme et sa spécificité sociale nous viennent de Jean-Jacques Rousseau. Dans son deuxième discours il soutient la thèse qu'en société l'Homme est dégradé, comme corrompu, mais sans elle il n'est pas Homme, il n'est pas l'animal intelligent capable de contrôler ses instincts. Comme une synthèse, Rousseau voit donc l'Homme actuel comme l'émergence des synergies d'interactions au sein d'une

société. Ce n'est qu'à partir d'ici que nous pouvons selon lui attribuer des propriétés (bonté ou méchanceté, égalité inégalité, ...) à l'Homme.

Nous étudierons dans cette thèse, à la lumière de ces travaux philosophiques et des principes fondamentaux des neurosciences cognitives comment l'attribution d'intentions à autrui entre en compte lors de nos décisions sociales puis nous verrons comment par l'attribution de capacités, expertises ou autres se mettent en place les liens de hiérarchie qui structurent nos sociétés.

# Chapter 1

## Introduction

### 1.1 La prise de décision

La prise de décision est un processus largement étudié dans des disciplines telles que la psychologie, les neurosciences cognitives, l'économie ou les sciences informatiques. La prise de décision est un terme générique pour désigner le processus de sélection d'une option parmi d'autres options dont les attentes ne sont pas identiques. Elle peut s'appliquer tant pour un individu ou pour une société que pour un organisme unicellulaire ou une machine. Selon le dictionnaire *Larousse*, la prise de décision est l'"Action de décider après délibération ; acte par lequel une autorité prend parti après examen". En psychologie, la théorie de la prise de décision est "fondée sur la prise en considération conjointe des probabilités et des utilités <sup>1</sup> des diverses éventualités en vue de la prise de décision".

Comme l'explique Schultz (2015) deux approches complémentaires sont souvent utilisées pour étudier la prise de décision. L'approche normative, qui fixe a priori un cadre formel de la prise de décision, comme la théorie des jeux ou l'hypothèse de maximisation de l'utilité. Puis l'approche expérimentale qui consiste à s'intéresser au comportement réel qui diffère souvent des comportements normatifs. Nous aborderons

---

<sup>1</sup>J'ai parsemé dans ce travail de thèses de nombreuses annotations, pour donner un avis ou apporter une précision. Ces notes de bas de pages ont notamment pour but de rendre plus vivant cet essai. Ici, l'utilité représente l'évaluation subjective des gains, par opposition à l'espérance mathématique



dans ce chapitre une description de la prise de décision d'un point de vue normatif puis nous verrons en quoi les comportements réels s'éloignent ou se rapprochent de la théorie. Dans les chapitres suivants nous analyserons des données expérimentales sous la lunette du cadre normatif afin d'étudier l'attribution d'intentions et de capacité à autrui.

### 1.1.1 Cadre normatif de l'étude de la prise de décision

La prise de décision implique un examen des probabilités et des utilités, par l'intégration de nombreuses sources d'information parmi lesquelles des entrées multisensorielles (e.g. vue, touché, ouïe, ...), des réponses autonomiques et émotionnelles (e.g. changement du rythme cardiaque, peur, faim, ...), des associations passées et des objectifs futurs. Ces options et informations sont aussi associées à une incertitude, une notion temporelle, et une balance coût bénéfice. La prise de décision doit souvent être rapide mais rester flexible pour que l'individu puisse s'adapter à une grande variété de situations. Étant un processus complexe, elle peut être décomposée en sous-processus dont nous étudierons certains par la suite (e.g. processus d'évaluation, processus d'apprentissage, processus d'intégration d'informations sociales, ...).

Nous commencerons ici par faire des hypothèses, parfois discutables, mais indispensables pour pouvoir commencer à faire des inférences sur les données que nous observons, comme en mathématique lorsque nous posons des axiomes pour qu'en découlent des théorèmes. Ce sont les hypothèses connues sous le nom d'axiome de von Neumann et Morgenstern (von Neumann et al., 1944). C'est l'axiome selon laquelle chaque individu est un agent économique, parfois aussi appelé *Homo economicus*. L'hypothèse se décompose en trois axes. Premièrement, l'agent économique est parfaitement informé, deuxièmement, il est infiniment sensible, et enfin, troisièmement, il est rationnel.

**Agent complètement informé.** L'agent complètement informé est un agent qui non seulement connaît l'ensemble des actions possibles mais aussi leurs conséquences. Dans certains cas particuliers, nous n'appliquerons pas cette hypothèse. En effet, si l'agent

n'est pas complètement informé, nous permettons une possibilité d'apprentissage que nous étudierons plus en détail dans la section 1.1.1.4.

**Agent infiniment sensible.** L'agent infiniment sensible est un agent pour qui les alternatives dont il dispose sont continue, infiniment divisible (ce qui permet à l'agent de distinguer deux options en terme de risque ou de ce qui est enviable par exemple, même si ces deux options sont très ressemblantes). Le but de cette hypothèse est principalement de rendre les fonctions de préférence de l'individu continues et différentiables<sup>2</sup>. Concrètement, un agent infiniment sensible est capable de différencier n'importe quels stimuli.

#### 1.1.1.1 Hypothèse de la rationalité

L'hypothèse selon laquelle les individus seraient rationnels est l'hypothèse centrale. C'est pourquoi elle mérite une section à part entière. Elle implique principalement que les individus soient capables d'ordonner les états dans lesquels ils souhaitent être. Elle suppose aussi que l'*Homo economicus* essaie de maximiser quelque chose (e.g. son bien-être, ses récompenses, ...).

Ainsi, si un agent économique est face à deux choix, A ou B dont les désidérabilités respectives sont  $a$  et  $b$ , il sera capable de distinguer s'il préfère être dans la situation A ou la situation B. De plus, si un troisième choix s'offre à lui (C;c), il doit respecter la règle de transitivité. C'est-à-dire que si  $a > b$  et  $b > c$  alors  $a > c$ .

Enfin, le principe central de l'agent économique et de sa rationalité est qu'il agit dans le but de maximiser quelque chose. Dans la théorie des choix non risqués, ce quelque chose est appelé *utilité*. Si nous ajoutons du risque ou de l'incertitude dans le choix, on parle alors d'*utilité espérée*. Nous nuancerons ces propos et nous verrons pourquoi cette hypothèse aussi peu être remise en question dans la partie 1.1.1.3 et comment les travaux notamment de Daniel Kahneman lauréat du prix Nobel d'économie ont contribué à cette remise en question. Plus récemment, une nouvelle théorie appelée "inférence active" propose que l'agent par ses actions et sa perception, maximise non pas

---

<sup>2</sup>au sens mathématique du terme.

une utilité, mais la proximité de son modèle interne <sup>3</sup> avec le monde extérieur. Le but est ainsi de minimiser l'incertitude sur les futurs états (agréables ou non) du monde. Ce qui lui permet alors d'agir pour maximiser ses récompenses futures. Ainsi, les observations expliquées par les premières théories le sont aussi par celle-ci et les divergences entre la théorie et les observations sont réduites, ce qui en fait une théorie importante des neurosciences.

### 1.1.1.2 Cadre de la théorie des jeux

Le principe de la théorie des jeux est de poser un cadre formel afin d'étudier mathématiquement une gamme très générale de problème, que l'on peut appeler problème de stratégie. Ici, le terme "jeux" est utilisé pour parler d'un concept très général. Ainsi, toute situation dans laquelle de l'argent (ou une valeur équivalente) peut être gagné grâce à un choix de stratégie approprié est considérée comme un jeu. Dans la théorie des jeux, un jeu dans sa forme normale est défini selon trois critères : les joueurs, l'ensemble des stratégies possibles pour chaque joueur ainsi que leurs préférences (Edwards, 1954).

**Les joueurs** sont considérés comme indépendants (si plus qu'un unique joueur), interagissant dans leur propre intérêt. Ils sont conscients que leur gain dépend de leurs actions ainsi que de celle des autres.

**Une stratégie** est un ensemble d'action possible pour jouer au jeu. Une stratégie peut être pure, c'est-à-dire que le joueur fixe une action pour chaque état du jeu et l'exécutera ou bien, il pourra utiliser une stratégie mixte qui consistera à exécuter des actions de manière probabiliste en fonction des états du jeu.

**Les préférences** sont définies comme une fonction d'utilité ou une fonction de gain associée à chaque stratégie. En effet, au cours des étapes d'une stratégie, le joueur peut passer par plusieurs états à chacun desquels il attribue une préférence. Les préférences comprennent notamment les paiements s'il y en a.

---

<sup>3</sup>En neurosciences, on considère parfois qu'un individu construit un modèle qui selon lui génère le monde qu'il observe, on l'appelle modèle génératif. Grâce à ce modèle génératif, il peut faire des prédictions sur les états du monde futur et sur l'impact de ses actions

### 1.1.1.3 Décisions optimales, "sous-optimales" et équilibres

De nombreuses études montrent le caractère optimal de la prise de décision chez l'Homo sapiens lors de décisions simples (e.g. lors de décisions perceptives avec différentes alternatives proposées et dans un contexte non-social) (Bogacz, 2007). L'optimal est ici vu en terme de ratio  $\frac{\text{Justesse}}{\text{Temps de réaction}}$ . Cette théorie des décisions optimales s'appuie notamment sur un avantage évolutif et sur la pression sélective qui s'exerce sur les individus et leur capacité de maximiser ce ratio  $\frac{\text{Justesse}}{\text{Temps de réaction}}$ .

Cependant l'optimal peut être vue en d'autres termes. Ainsi d'autres études supportant l'hypothèse d'un "cerveau bayésien", que nous détaillerons dans la partie 1.3.4, affirment aussi l'idée que l'Homo sapiens ferait des inférences optimales ou quasi-optimales. L'optimal ici est vu en terme de minimisation de la mesure fréquentiste des pertes sur un long terme. En effet, il est aisément démontrable qu'en utilisant la règle de prise de décision de Bayes, nous minimisons la somme des erreurs sur un long terme. Ces différentes façons d'interpréter l'optimalité ne sont pas incompatibles, mais toutefois, il est important de bien définir l'optimalité que l'on considère.

Dans de nombreux cas, nous constatons des comportements qui dévient de la rationalité ou de l'optimalité en terme de minimisation fréquentiste des pertes. Il s'en éloigne notamment par son impulsivité, sa mauvaise évaluation du risque ou son appréhension approximative des probabilités. Ces biais cognitifs ont notamment été révélés grâce à l'approche expérimentale qui a montré une déviation des comportements par rapport aux prédictions des modèles normatifs.

Dans un cadre social, la situation est plus complexe. Commençons par un exemple amusant pour critiquer l'hypothèse de l'*Homo economicus*. Voilà comment se comporteraient deux *Homo economicus* qui se rencontreraient : "Un homme demande à un autre où se trouve la gare. Celui-ci lui indique une direction (qui est en fait la mauvaise direction) tout en lui disant que sur son chemin se trouve une boîte aux lettres et lui demande de poster une lettre. Le premier homme répond qu'il est d'accord pour lui poster sa lettre tout en pensant "Dès qu'il ne me voit plus j'ouvre la lettre et je vois ce qu'il y a dedans". ". Ce genre d'interaction est tout à fait invraisemblable, c'est pourtant ainsi

que se comporteraient deux agents, maximisant chacun leurs utilités indépendamment. Les limites de l'hypothèse de l'*Homo economicus* apparaissent alors de manière assez évidente dans les relations sociales<sup>4</sup>.

L'un des exemples d'expérience de laboratoire le plus courant pour mettre en exergue l'éloignement du comportement de l'humain par rapport aux hypothèses de l'*Homo economicus* est celui du jeu de l'ultimatum. Le principe est simple et met en jeu une et une seule interaction entre deux individus. Le premier reçoit une somme d'argent qu'il doit partager entre lui et l'autre individu. L'autre individu peut alors à son tour décider d'accepter ou de refuser le partage. Dans le premier cas, les deux participants gagnent le montant attribué lors du partage, dans le second cas personne ne gagne. Il n'est pas rare de voir dans cette expérience le deuxième joueur refuser une offre qu'il juge inéquitable. Comportement qui s'éloigne de la rationalité normative de l'*Homo economicus* qui devrait accepter puisqu'il maximise ses gains. En effet, il gagnera plus en acceptant une offre injuste qu'en la refusant. Nous verrons dans le chapitre 1.2.2.2 comment, en changeant de point de vue dans ce cas précis, nous pouvons retrouver de l'optimalité et de la rationalité dans ce comportement.

Malgré les divergences entre théorie et expérimentation, le cadre formel de la théorie des jeux permet d'étudier les stratégies en les classant les unes par rapport aux autres. Dans un jeu social, les agents économiques tentent de maximiser leur fonction d'utilité en choisissant rationnellement parmi les stratégies possibles en fonction des actions des autres. Comme chaque joueur en fait autant, la stratégie optimale pour chacun dépend de la stratégie des autres joueurs. Il existe donc des stratégies meilleures que d'autres que l'on regroupe dans un sous-ensemble de stratégie appelé *solution*.

**Optimalité de Pareto.** Du point de vue d'un observateur extérieur au jeu, il est possible de classer les stratégies avec une relation appelée la domination de Pareto. Avec cette relation, au sens mathématique du terme, une stratégie est dominée si un des joueurs peut augmenter son utilité sans que les autres ne diminuent les leurs. Ainsi, si une stratégie n'est dominée par aucune autre au sens de Pareto alors elle est optimale

---

<sup>4</sup>c'est pourtant une croyance sous-jacente et persistante de nombreux économistes actuels.

au sens de Pareto. Cette relation tente ainsi de maximiser l'utilité globale (sommées des utilités) lors du jeu et reste vraie si l'on applique des transformations affines aux fonctions d'utilité des joueurs <sup>5</sup>.

**Équilibre de Nash.** Si l'on se place du point de vue de chacun des joueurs, l'équilibre de Nash est atteint lorsque tous les joueurs ont davantage de gain dans la présente stratégie qu'en changeant seul de stratégie. Si l'équilibre de Nash est maintenu par des stratégies pures, alors nous avons un équilibre de Nash pur. S'il est maintenu avec des stratégies mixtes, alors c'est un équilibre de Nash mixte.

#### 1.1.1.4 Processus d'apprentissage

Comme vu précédemment, si un agent économique ne possède pas une connaissance parfaite de son environnement, alors nous induisons théoriquement une possibilité d'apprentissage. Nous ne rentrerons pas dans le détail de l'apprentissage tant le sujet est vaste et complexe. L'apprentissage peut notamment se décliner en fonction des points de vue (l'apprentissage peut être celui des individus eux-mêmes, mais l'évolution d'une espèce peut aussi être vue comme un apprentissage) ou des espèces étudiées (Des espèces allant de l'organisme uni-cellulaire comme le *Physarum polycephalum* connu sous le nom de "blob" jusqu'à des organismes plus complexe comme l'*Homo sapiens*). En considérant uniquement l'apprentissage chez l'*Homo sapiens* il existe déjà différentes formes d'apprentissage possibles (connaissance ou mémoire, par renforcement ou par construction d'un modèle du monde, méta-apprentissage, ...). Toutefois, selon le dictionnaire *Larousse*, l'apprentissage est l' "Ensemble des processus de mémorisation mis en œuvre par l'animal ou l'Homme pour élaborer ou modifier les schèmes <sup>6</sup> comportementaux spécifiques sous l'influence de son environnement et de son expérience." Il existe

---

<sup>5</sup>Cet optimal est intéressant pour comprendre la différence entre certaine forme de communisme et certaine forme de libéralisme. En effet, il semblerait que l'objectif de l'idéologie communiste soit de maximiser la somme des utilités, tandis que dans une idéologie libérale, il semblerait que l'objectif soit de maximiser les utilités indépendamment. Par exemple, alors qu'*Uber* fixe les rémunérations des courses pour ses chauffeurs en fonction de l'offre et de la demande, pour que chacun maximise ses profits, *Didi chuxing* attribut directement une courses au chauffeur pour maximiser la répartition des chauffeurs dans une ville.

<sup>6</sup>Selon J. Piaget (Piaget, 1977), régularité construite par tâtonnement dans l'action du sujet et qui peut être généralisée à d'autres situations.

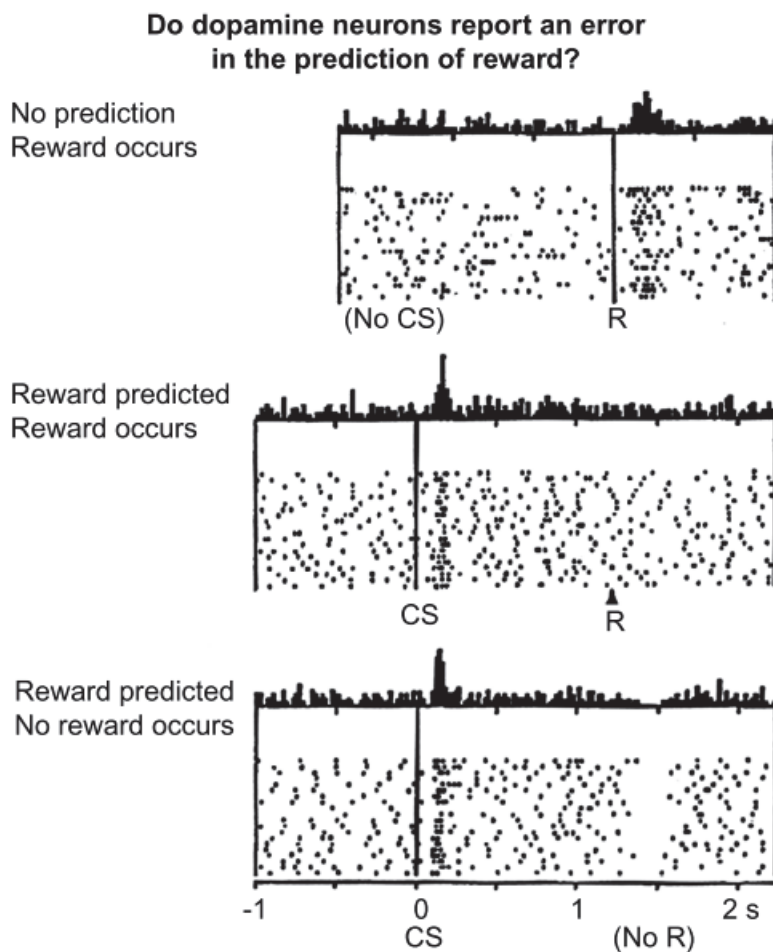
de nombreux exemple d'apprentissage, mais l'un des plus connu est l'apprentissage Pavlovien, car il a permis de révéler les mécanismes d'acquisition et de perte des réflexes conditionnés (Pavlov, 1927). Dans cette expérience Pavlov commence par tester deux types de stimuli. L'un neutre, une cloche (stimulus conditionné), l'autre significatif, un os (stimulus inconditionné). Une fois les réponses aux stimuli enregistrées, Pavlov associe de manière répétée la présentation du stimulus inconditionné avec la cloche, qui devient alors le stimulus conditionné. Une fois le processus terminé, le stimulus conditionné (la cloche) provoque la même réaction que le stimulus inconditionné (l'os) c'est-à-dire la salivation du chien. Cela révèle une mise en mémoire de l'association des deux stimuli appelé apprentissage Pavlovien. Pavlov a reçu le prix Nobel de médecine pour ses études<sup>7</sup>. Ce n'est qu'à partir de 1972 que Robert A. Rescorla et Allan R. Wagner (Rescorla and Wagner, 1972) proposent une explication théorique formalisée mathématiquement avec un modèle qui porte aujourd'hui leurs noms<sup>8</sup>. Il a ensuite fallu encore attendre 1997 que Schultz et al. (1997) découvre les bases neurales de cet apprentissage : les neurones dopaminergiques. Ces neurones produisent un neurotransmetteur associé à la prédiction de récompense (stimulus conditionné) et la récompense elle-même (stimulus inconditionné), la dopamine. La théorie prédit l'augmentation d'une certaine valeur face à une récompense non conditionnée, le décalage de cette valeur vers le stimulus conditionné après conditionnement (salivation à la cloche et non plus à l'os) et aucune variation si la prédiction est correcte. Par contre, s'il y a présence d'un stimulus conditionné qui n'est pas suivi d'une récompense, alors il y a diminution de la valeur du stimulus conditionné (déconditionnement). Or, c'est ce type de comportement des neurones dopaminergiques qui a été observé comme on le voit figure 1.1. Plus tard, d'autres recherches viendront remettre en cause la spécificité de cette association entre dopamine et erreur de prédiction. En effet, Matsumoto and Hikosaka (2009), notamment, confirment l'hypothèse précédente pour certain neurones dopaminergique, mais montrent aussi que certains sont actif à la fois face aux récompenses et face aux stimuli aversifs.

---

<sup>7</sup>Pavlov, et non son chien qui l'aurait aussi mérité peut-être.

<sup>8</sup>Nous étudierons plus en détails les apprentissages par renforcement dans la partie concernant la modélisation 1.3.1

Ils en concluent donc que la dopamine doit avoir un rôle motivationnel. Des théories plus récentes suggèrent que la dopamine a un rôle de modulation et de hiérarchisation des processus cognitifs de niveaux différents (Cools, 2011).



**Figure 1.1** – Neurones dopaminergiques dans la substance noire (SN) et l’aire tegmentale ventrale et prédiction de récompense. *Extrait de (Schultz et al., 1997)*

La substance noire et l’aire tegmentale ventrale contenant les corps cellulaires des neurones dopaminergiques sont les deux principaux centres de production de la dopamine qui est ensuite diffusée dans le cerveau par trois grandes voies constituées d’axones : le système mésocortical et mésolimbique ainsi que le système nigrostriatale.



## 1.1.2 Structures cérébrales impliquées

Comme vu dans l'introduction générale, la prise de décision est un mécanisme qui implique de nombreux processus cognitifs, et plus encore lorsqu'elle a lieu en contexte social. Depuis le début du XIX<sup>e</sup> siècle et la phrénologie<sup>9</sup>, les scientifiques essaient d'attribuer ces différents processus cognitifs à des régions cérébrales distinctes ou à des réseaux de régions. Il existe aujourd'hui de nombreuses manières de segmenter le cerveau en différentes parties, par exemple de manière anatomique ou de manière fonctionnelle. C'est principalement cette deuxième méthode que nous emploierons dans cette thèse. Lorsque plusieurs régions sont impliquées dans un même processus cognitif, nous parlons alors de réseau. Dans les sections qui suivent, nous détaillerons notamment les régions impliquées dans l'évaluation des stimuli, celles impliquées dans les émotions, et enfin celles engagées par la décision en contexte sociale. Il est important de préciser que ces réseaux peuvent interagir.

### 1.1.2.1 Bases neurales de l'évaluation d'un état ou d'une décision

Les aires d'évaluation sont les aires impliquées lorsque le cerveau attribue une valeur à un stimulus (à l'os ou à la cloche du chien de Pavlov par exemple) ou lorsqu'il compare les valeurs de deux stimuli. L'évaluation d'un stimulus ainsi que l'appréhension d'une récompense sont intimement liées. En effet, l'évaluation des stimuli consiste en l'examen du stimulus dans le but d'estimer la valeur de la récompense associée, c'est une anticipation de la récompense. L'évolution de l'évaluation du stimulus et donc le transfert de la valeur depuis la récompense vers le processus d'évaluation, s'appelle l'apprentissage. L'évaluation des stimuli étant donc au centre du processus d'apprentissage par récompense, son réseau est donc aussi intimement lié au réseau de la récompense et donc aux projections des neurones dopaminergiques. Comme illustré dans la figure 1.2 :

---

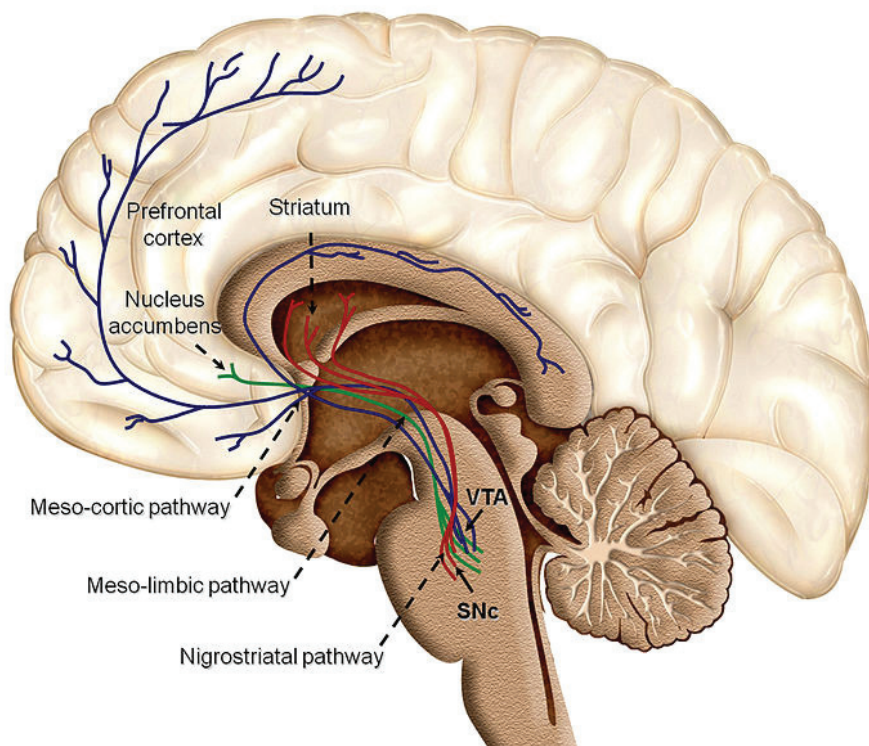
<sup>9</sup>La phrénologie est selon le *Larousse* [une théorie qui] relie chaque fonction mentale à une zone du cerveau et soutient que la forme même du crâne indique l'état des différentes facultés.

**La voie mésocorticale** représente les axones des neurones dopaminergiques situées dans l'aire tegmentale ventrale (VTA en anglais) qui se projettent vers le cortex préfrontal, cingulaire et perirhinal (Arias-Carrián et al., 2010).

**La voie mésolimbique** représente les axones des neurones dopaminergiques situées dans l'aire tegmentale ventrale qui se projettent principalement vers les noyaux accumbens, mais aussi vers le tubercule olfactif qui innerve à son tour notamment l'amygdale, l'hippocampe et le septum (Arias-Carrián et al., 2010).

Cependant, ces deux systèmes se superposent partiellement, c'est pourquoi on parle régulièrement de voie mesocorticolimbique.

**La voie nigrostriatale** représente les fibres partant de la substance noire et projetant dans les noyaux caudés et ceux du putamen (Arias-Carrián et al., 2010).



**Figure 1.2** – Structures des réseaux de la récompense dans le cerveau humain *Extrait de Arias-Carrián et al. (2010)*

Nous avons vu ce qu'il en est d'un point de vue anatomique et donc des régions auxquelles nous aurons à faire, mais qu'en est-il d'un point de vue fonctionnel ? Le système dopaminergique joue plusieurs rôles qui commencent à être bien connus notamment grâce aux études faites sur la maladie de Parkinson qui est une maladie dégénérative du système dopaminergique. La dégénérescence se fait souvent dans un ordre précis, et les apparitions synchrones des symptômes permettent de différencier les multiples rôles de la dopamine.

Une partie de mon travail de thèse à d'ailleurs consisté à l'utilisation d'une base de données de patients récoltée par le professeur Christian Scheiber à l'hôpital Neurologique Pierre Wertheimer à Lyon. Ses patients ont effectué 3 examens combinés afin de diagnostiquer la présence ou non de la maladie de Parkinson notamment dans ses cas atypiques. Une scintigraphie et un scanner sont deux examens de contrôle que les patients passent pour vérifier si les signes cliniques peuvent être dus à d'autres anomalies telles que des troubles de l'irrigation sanguine cérébrale ou à des anomalies des tissus. L'examen d'intérêt pour étudier le système dopaminergique est appelé le DaTSCAN et consiste à l'injection d'un traceur radio actif [123I]-FP-CIT qui se fixe sur les transporteurs pré-synaptiques de la dopamine<sup>10</sup> et permet donc d'extraire un potentiel de liaison du système dopaminergique. L'objectif de notre travail était d'obtenir une base dite "normale" de sujet sain afin de pouvoir comparer leurs potentiels de liaison de différentes structures cérébrales (notamment celle du striatum ventral) à celui de patients ayant des signes cliniques ou pré-cliniques. Je ne m'attarderai pas plus sur ce projet dans cette thèse, le travail est encore en cours d'écriture.

La maladie de Parkinson se traduit en une dégénérescence de la substance grise et affecte donc principalement la voie nigrostriatale, ce qui permet notamment d'en évaluer son rôle fonctionnel. Ainsi, nous savons que cette voie influe principalement sur le contrôle moteur volontaire direct (facilitation des mouvements voulus) et indirect (suppression des mouvements involontaires) (Gerfen and Surmeier, 2011). Cette voie est

---

<sup>10</sup>Les transporteurs permettent la recapture de la dopamine qui a été libérée dans la synapse afin de réguler la transmission du signal et de pouvoir la réutiliser.

aussi associée à l'apprentissage suite aux renforcements positifs et négatifs (Mhatre V. Ho and Kelsey C. Martin, 2012), à la créativité et au circuit de la récompense (Gerfen and Surmeier, 2011).

Quant au système mésocorticolimbique, il a été suggéré qu'il joue un rôle dans l'initiation volontaire de la locomotion, dans la cognition liée à la récompense (notion de plaisir, renforcement positif) et à l'aversion. Il jouerait aussi un rôle descendant, c'est-à-dire de motivation en promouvant les comportements dirigés vers un but et les fonctions exécutives de haut niveau notamment la modulation des comportements liés aux émotions. Un dysfonctionnement de ses voies peut impliquer un trouble du déficit de l'attention, une dépendance ou des troubles schizophrénique (Arias-Carrián et al., 2010; Phillips et al., 2008; Chi Yiu Yim and Mogenson, 1980; D'Ardenne et al., 2008). Ainsi, les deux voies n'ont pas un rôle distinct, mais complémentaire comme par exemple l'apprentissage et la créativité<sup>11</sup> (D'Ardenne et al., 2008; Boot et al., 2017).

### 1.1.2.2 Bases neurales des émotions

La théorie cartésienne du philosophe René Descartes qui a dominé la pensée européenne pendant plusieurs siècles et selon laquelle seule la raison, qu'il semble opposer à l'émotion, importe pour l'Homme est remise en question par les neurosciences. Nous connaissons maintenant et notamment grâce à Antonio Damasio l'importance du rôle des émotions dans la prise de décision et plus encore leur caractère indispensable dans la validité de nos raisonnements<sup>12</sup>. Il est important de définir le concept d'émotion pour se mettre d'accord sur le sujet dont nous parlons. Nous nous baserons sur la définition d'Antonio Damasio pour qui l'émotion est un changement homéostatique du corps. Voici comment il décrit une émotion dans Damasio (1995) *"Le cœur d'une émotion, selon moi, est un ensemble de changements dans l'état du corps et du cerveau induits dans une myriade*

---

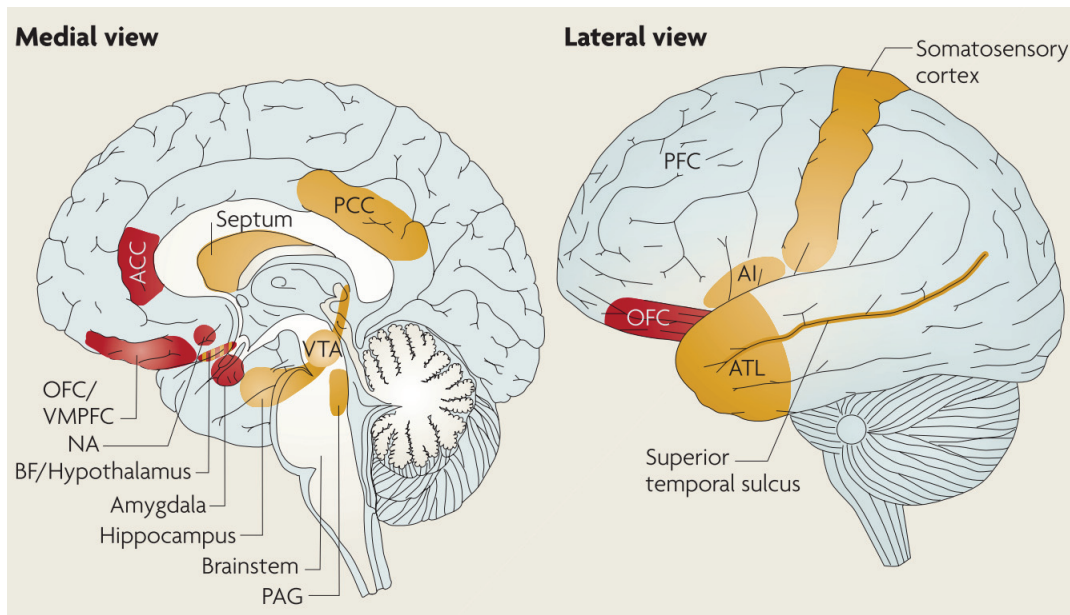
<sup>11</sup> Apprentissage et créativité ne sont en effet pas des compétences indépendantes. En voici une illustration faite par *Google Deep Dream*. La société a entraîné un réseau de neurones à classifier des images dans des catégories représentées par des mots. En inversant son réseau de neurones et en lui donnant en entrée des mots, l'algorithme a généré de nouvelles images qui ne lui avaient jamais été présentées, ce qui peut être vu comme une forme de créativité.

<sup>12</sup> Antonio Damasio est notamment auteur de *L'erreur de Descartes : la raison des émotions* (Damasio, 2010) et de *Spinoza avait raison : joie et tristesse, le cerveau des émotions* (Damasio, 2003).

*d'organes et dans certains circuits cérébraux sous le contrôle d'un système cérébral dédié, qui réagit au contenu des pensées d'une personne par rapport à une entité ou un événement particulier. Les réactions à l'égard du corps lui-même entraînent un état corporel spécifique, et celles à l'égard du cerveau lui-même entraînent un mode de fonctionnement en réseau spécifique qui implique un changement de style cognitif. Le premier produit des modifications physiologiques, dont beaucoup sont perceptibles par un observateur extérieur, par exemple des changements de couleur de la peau, de posture et d'expression du visage" . Selon Damasio, comme pour les récompenses, il y aurait deux types d'émotions, les émotions primaires (e.g. joie, peur, tristesse, colère ou dégoût) et les émotions secondaires (e.g. honte, culpabilité, fierté, ...)<sup>13</sup>. Comme pour les récompenses, les émotions primaires sont celles innées, pré-organisées dans le cerveau. Elles reposent notamment sur le système limbique, l'amygdale, le cortex cingulaire antérieur, l'hypothalamus, le tronc cérébral et le prosencéphale basal situé en avant du striatum. Encore comme pour les récompenses, les émotions secondaires sont apprises par association systématique entre des situations ou des objets et des émotions primaires. Les émotions secondaires impliquent notamment des régions des cortex pré-frontaux (OFC, ACC), les cortex insulaires et des régions somatosensorielles.*

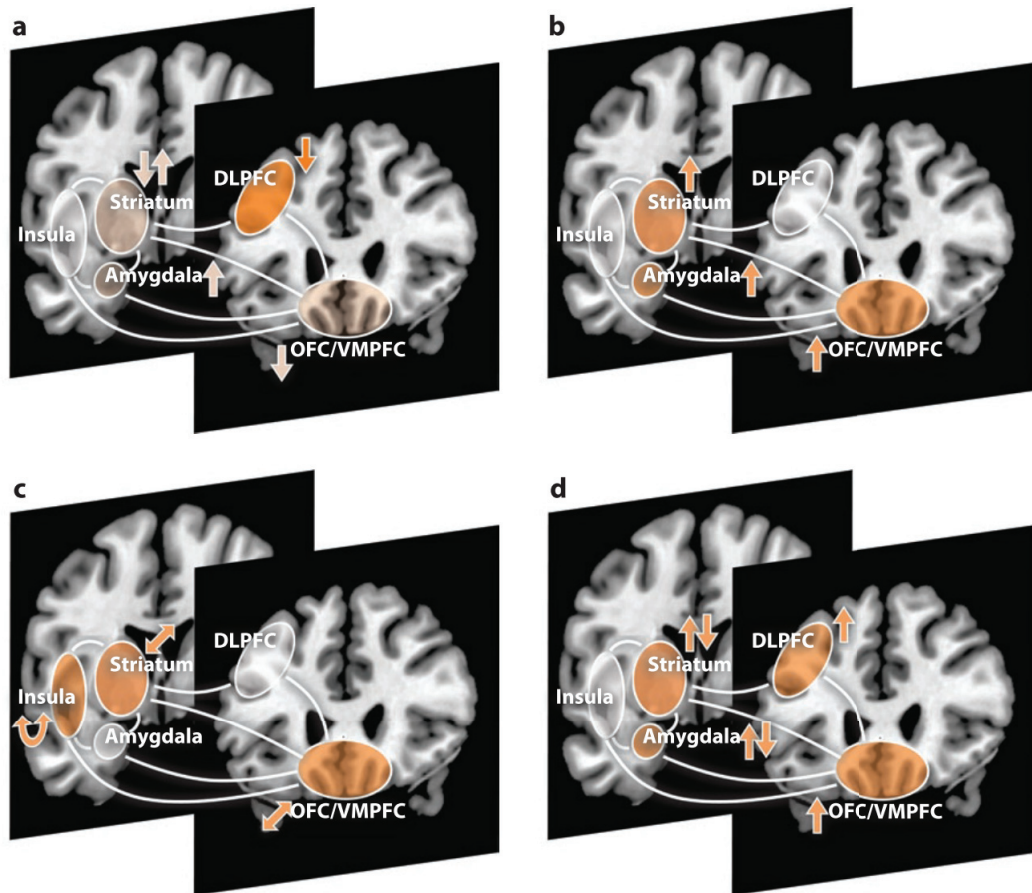
---

<sup>13</sup>La même distinction peut être faite pour les récompenses. Les récompenses primaires sont les récompenses qui remplissent un besoin vital (e.g. nourriture, abris, reproduction, ...). Celles secondaires sont celles qui permettront dans un second temps d'atteindre une récompense primaire (e.g. statut social, argent, connaissance, ...)



**Figure 1.3** – Structures des régions principales et étendues liées aux émotions dans le cerveau humain. En rouge, ce sont les régions que l'on retrouve le plus fréquemment dans la littérature. En orange, ce sont celles que nous retrouvons régulièrement, mais de manière moins fréquente. *Extrait de Pessoa (2008)*

Nous pouvons ainsi voir dans la figure 1.3 que de nombreuses régions sont en commun entre le réseau des émotions et celui de l'évaluation des stimuli et des états, ce qui suggère qu'il n'y a pas comme Descartes aimait à penser deux systèmes distincts, mais bien deux systèmes interagissant. En effet, les études tendent à montrer qu'il y a une relation réciproque entre affect et décision. Ainsi, les états affectifs peuvent se reporter sur l'évaluation de la valeur subjective et la décision. Et à l'inverse, l'évaluation des stimuli et états du monde peut moduler les émotions comme illustré dans la figure 1.4 (Pessoa, 2008).



**Figure 1.4** – Multiples interactions entre circuit d'évaluation et d'émotion. (a) Le stress perturbe le cortex préfrontal dorsolatéral (dlPFC), ce qui diminue les comportements orientés vers un but et augmente les comportements d'habitudes. Mais le stress diminue aussi l'activité du cortex orbitofrontal (OFC) et de celui préfrontal ventromédian (vmPFC) et augmente celui de l'amygdale. (b) Les émotions contribuent à l'évaluation des stimuli et états. En effet, l'amygdale influence l'évaluation des valeurs subjectives des choses dans le striatum et l'OFC/vmPFC. Les émotions modulent aussi l'apprentissage par renforcement. (c) Alors que la valeur subjective semble encoder linéairement dans l'OFC/vmPFC et dans le striatum, alors que l'insula l'encode en forme de U. Ce qui signifie que l'insula encode la valeur dans les situations d'état de vigilance élevé ou de sillance des stimuli. (d) L'influence des émotions sur les choix peut être modulée en utilisant la régulation cognitive des émotions médiée par le dlPFC et le vmPFC. La régulation peut être positive ou négative menant au changement correspondant dans l'amygdale et le striatum. *Extrait de Phelps et al. (2014)*

Non seulement le réseau des émotions interagit avec celui de l'évaluation et de la récompense, mais aussi avec les réseaux engagés dans les interactions sociales. No-

tamment, parce que les émotions constituent une information sociale de premier plan (Phelps et al., 2014; Keltne and Haidt, 2001).

### 1.1.2.3 Bases neurales sensibles ou spécifiques à la prise de décision sociale

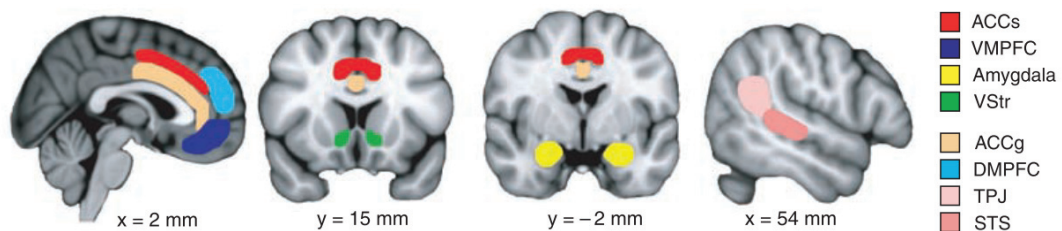
Les interactions sociales sont des processus complexes et très variés. L’Homo sapiens, comme beaucoup d’autres êtres vivants, est capable d’une grande diversité de comportements sociaux, comme l’affiliation ou l’agression, ou comme l’établissement de hiérarchie. Des bactéries, des fourmis ou certains mammifères par exemple sont aussi capables de comportements sociaux, et comme pour nous, ils sont indispensables à leur survie. Cependant, l’Homo sapiens excelle en terme de comportements sociaux allant jusqu’à créer des normes sociales et morales, partagées sur de grands territoires par de nombreuses personnes. Pour cette spécialisation de notre espèce, l’évolution a notamment doté l’humain d’une capacité de mentalisation, qui par empathie, par compassion (si les émotions entrent en jeu), ou par raisonnement (si les émotions n’interagissent pas ou peu), permettent à un individu de prendre le point de vue d’un autre et de lui prêter des intentions, des croyances, des préférences et donc de complexifier les relations sociales. On parle parfois de théorie de l’esprit pour parler de cette capacité d’intellectualiser le point de vue d’un autre. Ainsi, lors d’une interaction sociale, chaque individu calcule grâce à son réseau cérébral de l’évaluation ce qui est le mieux pour lui, mais aussi ce qui est le mieux pour l’autre en fonction des croyances qu’on lui attribut. Nous constatons la sensibilité sélective de certaines régions à ces processus sociaux. En effet, certaines zones cérébrales sont sensibles à l’évaluation relative à soi tandis que d’autres sont sensible à l’évaluation relative à autrui. Ainsi Behrens et al. (2009) reprennent cette partition du cerveau et répertorient les régions le plus souvent impliquées lorsque l’on considère les intentions et actions des autres. Ils incluent dans ce réseau le cortex cingulaire antérieur, plus précisément son gyrus<sup>14</sup>, le sulcus serait

---

<sup>14</sup>Le cerveaux étant constitué de nombreux replis, les creux sont appelés les sulci (sulcus au singulier), les bosses sont appelées les gyri (gyrus au singulier). Comme nous l’avons vu précédemment, les régions peuvent avoir une spécialisation fonctionnelle, mais souvent, leur rôle est intégré dans un réseau. Une idée intéressante, qui a entre autres poussé à la création du “*connectome project*”, est que ces plis cérébraux sont dus à la connectivité entre les différentes régions. La théorie du pliage corticale basée sur la tension propose



quant à lui plus spécifique au *soi* (voir figure 1.5). De même, la jonction temporopariétale (TPJ), le sulcus temporal supérieur (STS) et le cortex préfrontal dorsomédian (dmPFC) sont souvent engagés dans les processus de considération des intentions, croyances et préférences des autres (voir figure 1.5). Si l'humain est capable de prendre la perspective d'un conspécifique ou d'un individu d'une autre espèce, il est nécessaire qu'il puisse tout de même faire la différence entre ses propres croyances, objectifs et préférences de celles des autres. Cela peut passer par des régions cérébrales activées différemment (comme celles dont nous venons de discuter) ou bien par des populations de neurones différenciées selon l'objet (soi ou autrui) comme nous le verrons par la suite.



**Figure 1.5** – Structures des réseaux des interactions sociales dans le cerveau humain *Extrait de Behrens et al. (2009)*. Les couleurs primaires indiquent les régions impliquées fréquemment dans les interactions sociales ayant pour cadre de référence les propres actions du participant. En pastel figurent les régions impliquées considérant les intentions d'une autre personne.

Dans chacun des réseaux que je vous ai présenté précédemment, il faut faire attention à plusieurs points : la sensibilité de ces régions ainsi que leur spécificité. La **sensibilité** correspond à la fréquence d'apparition de cette région d'intérêt dans une tâche impliquant le processus étudié (émotion, apprentissage, ...), tandis que la **spécificité** d'une région est sa capacité à ne pas être activée dans une tâche n'impliquant pas le processus étudié. Nous avons donc ici surtout discuté des régions sensibles à l'évaluation, aux émotions et aux interactions sociales, mais comme certaines régions se trouvent dans plusieurs réseaux, elles sont peu spécifiques, tout du moins à ces processus particuliers, puisqu'il n'est pas impossible qu'une meilleure définition des processus améliore la sensibilité et la spécificité des régions à ces processus. De plus, il ne faut pas oublier que

---

que les grandes régions connectées par peu de voies créent des plis marqués tandis que des régions plus petites et/ou connectées par des voies plus nombreuses créent des plis plus variables (Van Essen, 1997)

le cerveau est un réseau et qu'intégrée dans une configuration ou dans une autre, une même région peut jouer des rôles relativement différents, d'où l'intérêt d'étudier aussi le cerveau en réseaux (Fox et al., 2005).

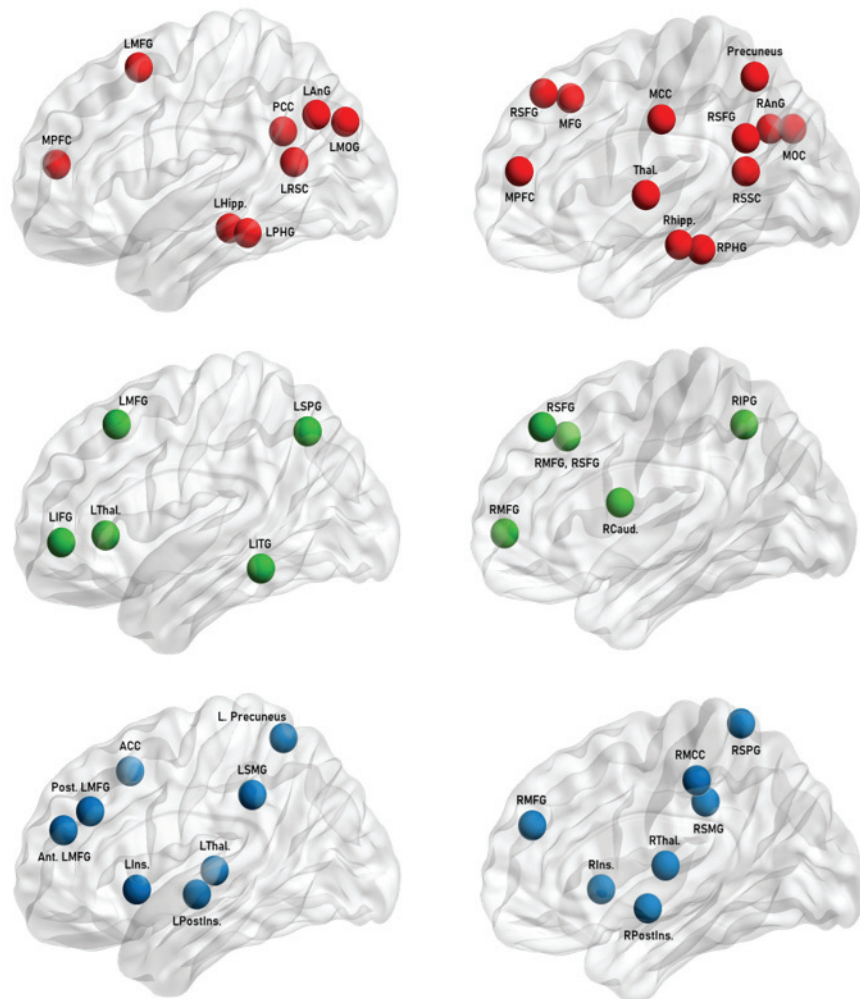
#### 1.1.2.4 Autres réseaux neuraux

Récemment, grâce à une autre méthode d'analyse<sup>15</sup> des IRMf<sup>16</sup> dans un "état de repos" par exemple, il a été proposé que l'implication des régions cérébrales ne soit pas spécifique à un processus comportemental, mais que ce soit son engagement au sein d'un réseau qui soit spécifique à un comportement. Nous pouvons par exemple distinguer un réseau de mode par défaut, qui est le réseau des régions impliquées lorsque le cerveau n'est pas focalisé sur le monde extérieur, que le cerveau n'effectue aucune tâche, mais ne dort pas. Ce réseau joue un rôle notamment dans la mémoire, la perception des émotions et l'introspection. Ce réseau contient notamment le lobe temporal médial, le cortex préfrontal médial, le cortex cingulaire postérieur (voir figure 1.6). Lors de la réalisation d'une tâche, le réseau de mode par défaut se désactive, ou plus précisément se dé-corrèle, c'est-à-dire que les régions de ce réseau fonctionnent de manière moins synchrones, et un autre réseau s'active, ou plus précisément se synchronise, par exemple le réseau exécutif. Le réseau exécutif lui, permet l'adaptation à des situations nouvelles ou complexes. En dernier exemple, je peux vous présenter le réseau de la saillance qui est un réseau qui permet de distinguer les stimuli dignes d'attention ou non. Ce troisième réseau est intéressant, car il pourrait être celui qui fait la transition depuis le réseau de fonctionnement en mode par défaut vers le réseau exécutif. Une transition depuis un mode orienté vers soi (introspectif) vers un mode orienté vers l'extérieur (Goulden et al., 2014).

---

<sup>15</sup>Habituellement les analyses des images IRM se font en comparant les activités dans différentes conditions d'une tâche ou en recherchant quelles régions covarient avec une variable d'un modèle du comportement. Ici, il s'agit de rechercher les régions qui covarient ensemble, on parle alors de connectivité fonctionnelle

<sup>16</sup>L'IRMf (Imagerie par Résonance Magnétique fonctionnelle) est une modalité d'observation cérébrale. Elle ne possède ni une grande définition temporelle (une image toutes les secondes environ), ni une grande définition spatiale (cube de 2 ou 3 millimètres de côté environ). Elle a l'avantage d'être non invasive. Elle permet plus précisément d'observer les fluctuations du taux d'oxygénation sanguin liées à la consommation de ressources par les neurones.



**Figure 1.6** – Composition de différents réseaux cérébraux *Adapté de (Shirer et al., 2012)*. Le réseau rouge représente les principales aires impliquées dans le réseau de mode par défaut. Le réseau vert représente les principales aires impliquées dans le réseau exécutif et enfin en bleu sont représentées les aires impliquées dans le réseau de la saillance.

## 1.2 Spécificité du contexte social

Comme nous l'avons vu, l'aspect social de l'Homo sapiens est une de ses caractéristiques fondamentales. Ces interactions peuvent se différencier sur plusieurs aspects, notamment sur le rôle de la personne étudiée dans l'interaction et son engagement. Par

exemple, l'individu observe-t-il uniquement le comportement d'une autre personne ou est-il lui-même observé ou bien encore sont-ils en train d'interagir directement ? Le mode d'interaction peut aussi varier. Intervient alors ici une notion d'intention dans l'interaction : quel est le but de l'interaction ? Nous étudierons plus en détail comment par ces interactions sociales, chez l'espèce *Homo sapiens* comme chez de nombreuses autres espèces, une hiérarchie peut se mettre en place, notamment par l'attribution de compétences ou de capacités. Enfin, nous étudierons les intentions et l'attribution d'intentions à d'autres au travers de la théorie de l'esprit dont nous avons parlé précédemment.

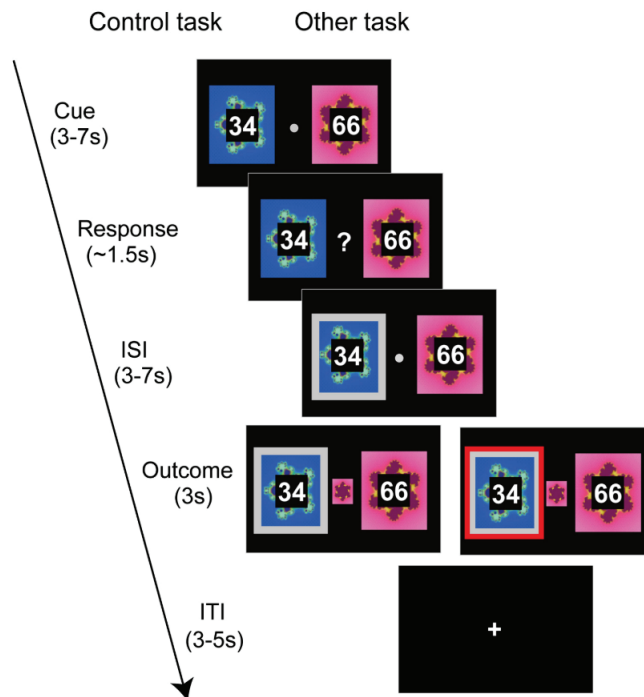
### **1.2.1 Différentes situations d'interaction**

Il existe donc différentes configurations possibles d'interaction. D'abord, dès lors que deux individus peuvent se voir sans interaction directe, voire même quand ils pensent être vu, on parle déjà d'interaction, et les effets sociaux sont déjà largement visibles. Le rôle d'une personne peut être *observateur* et donc engager un apprentissage par observation, il peut être celui d'être observé et promouvoir ce que l'on appelle un effet d'audience ou enfin, il peut être d'interagir de manière directe.

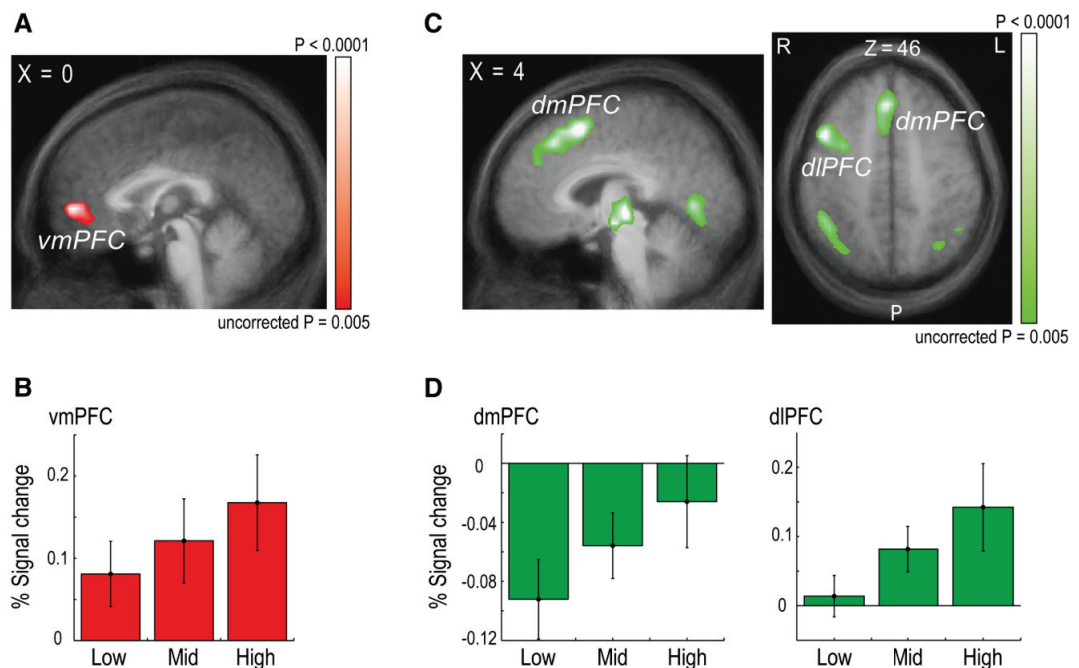
#### **1.2.1.1 Du point de vue d'un observateur**

Les *Homo sapiens*, mais aussi nombres d'autres espèces utilisent l'observation des autres individus (con-spécifiques ou non) pour naviguer dans leur environnement. L'une des formes les plus basiques d'apprentissage par observation est l'imitation. Elle apparaît tôt dans l'évolution, mais aussi dans le développement de l'individu. Elle fait appel à des processus sociaux distincts, notamment la théorie de l'esprit, l'empathie ou la compassion et la reconnaissance de la hiérarchie pour savoir qui imiter. L'imitation nécessite une capacité de renforcement viscérale, car aucune récompense n'est obtenue directement lors d'une observation d'un con-spécifique (Joiner et al., 2017). Plusieurs études récentes s'intéressent à l'apprentissage par observation. Particulièrement, Burke et al. (2010) montrent parmi les premiers qu'un tel apprentissage engage deux types

d'erreurs de prédiction : l'erreur de prédiction concernant l'action observée ainsi que l'erreur de prédiction du résultat observé de l'action. Ces travaux ont été repris et complétés, notamment dans une étude de Suzuki et al. (2012) en imagerie fonctionnelle par résonance magnétique (IRMf) qui montre qu'en observant quelqu'un jouer à un jeu (voir figure 1.7), le cerveau de l'observateur simule grâce à son réseau neural de l'évaluation et celui social (vmPFC, STG, gyrus cingulaire) la probabilité de récompense au moment du choix ainsi que la prédiction d'erreur de récompense au moment du résultat (vmPFC) (voir figure 1.8), comme s'il jouait lui-même. Comme Burke et al. (2010), Suzuki et al. (2012) a montré que ces calculs ne suffisent pas à reproduire les comportements observés expérimentalement. Un processus supplémentaire entre en jeu, celui de la confrontation entre la prédiction du participant et son observation des choix fait par l'autre joueur. On appelle cette confrontation l'erreur de prédiction d'action. Ce processus permet d'exploiter les divergences entre la prédiction d'action et l'action effectivement réalisée qui peuvent notamment être dues à la vitesse d'apprentissage différente entre l'observateur et l'observé ou à des différences de préférences au risque dans cette tâche. Cette prédiction d'erreur d'action est trouvée encodée dans le dmPFC ainsi que dans le dlPFC droit (voir figure 1.8) et le gyrus angulaire bilatéral.



**Figure 1.7** – Tâche expérimentale. La tâche est composée de deux conditions, la tâche *autrui* où le participant observe les choix d'un autre et la tâche contrôle où le participant fait des choix lui-même. Dans la tâche contrôle (où le participant joue lui-même) comme dans la tâche *autrui* (où le participant observe), la tâche contient 4 phases : Stimulus, réponse, intervalle inter-stimuli (ISI) et récompense. Dans les deux tâches, le participant doit choisir entre deux stimuli représentant des fractales. La réponse du participant est encadrée en gris durant l'ISI dans la condition contrôle. La réponse de l'autre dans la condition *autrui* est encadrée en rouge. Dans tous les cas, la bonne réponse apparaît au centre de l'écran. Les stimuli étaient récompensés avec une probabilité respective  $p$  et  $1 - p$ .  $p = 0.25$  ou  $p = 0.75$  en fonction des blocs. La magnitude de la récompense était inscrite sur les fractales. *Extrait de Suzuki et al. (2012).*



**Figure 1.8** – Activité neurale corrélée avec l’erreur de prédiction simulé de la récompense ou de l’erreur de la prédiction de l’action de l’autre lors de la condition *autrui* (où le participant observe les choix d’un autre). **(A)** Activité corrélant avec la magnitude de l’erreur de prédiction de récompense simulée au moment de la récompense (vmPFC).  $p < 0.005$  pour des questions de visualisation. **(B)** Moyenne du pourcentage de changement du signal BOLD dans le vmPFC lorsque la magnitude de l’erreur de prédiction est faible, moyenne ou haute (33<sup>e</sup>, 66<sup>e</sup> et 100<sup>e</sup> centile) 7-9 secondes après l’apparition de la récompense (résultat validé de manière croisée). **(C)** Activité corrélant avec la magnitude de l’erreur de prédiction de l’action simulé au moment de la récompense.  $p < 0.005$  pour l’affichage. **(D)** Moyenne du pourcentage de changement du signal BOLD dans le dmPFC et dlPFC lorsque la magnitude de l’erreur de prédiction est faible, moyenne et haute (33<sup>e</sup>, 66<sup>e</sup> et 100<sup>e</sup> centile) 7-9 secondes après l’apparition de la récompense (validée de manière croisée). *Extrait de Suzuki et al. (2012).*

Il est aussi possible que ce ne soit pas des régions entières qui soient spécifiques à l’apprentissage par observation, mais aussi des populations ou sous populations de neurones<sup>17</sup>. On voit ainsi dans l’étude de Hill et al. (2016) avec une tâche légèrement différente (voir figure 1.9) que les neurones dans le cortex cingulaire antérieur rostral

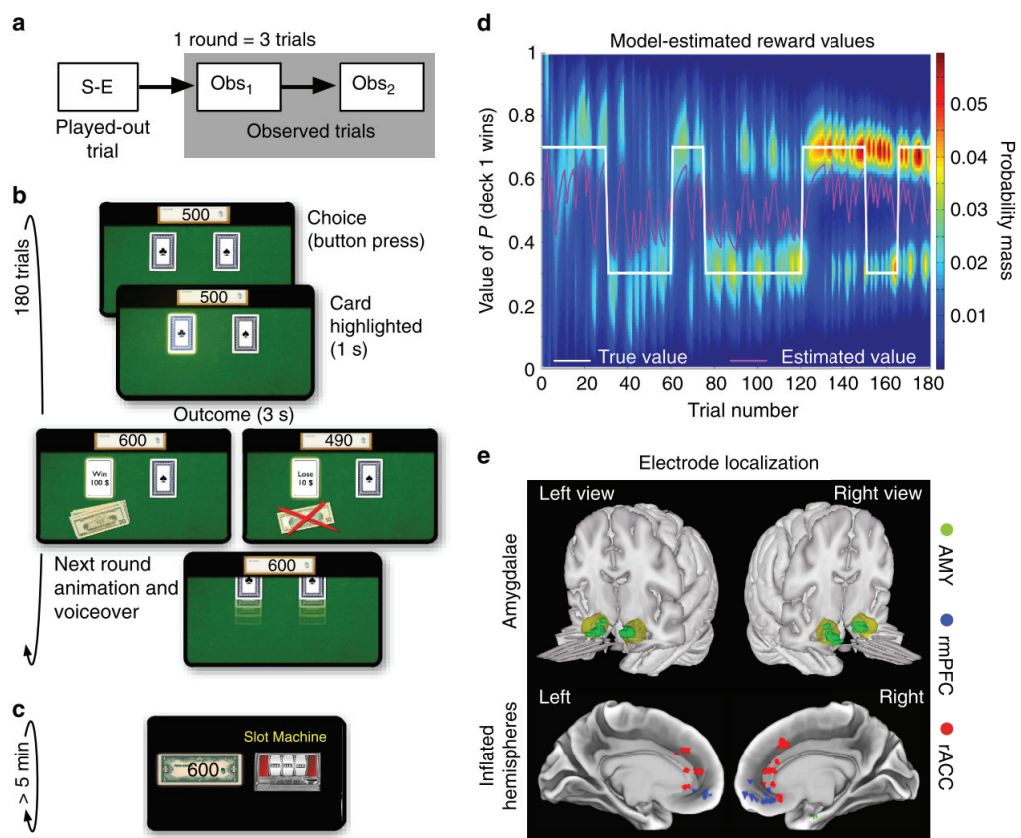
<sup>17</sup>Pour observer des populations de neurones localement il est possible d’utiliser l’électroencéphalographie intracrânienne. Ces techniques utilisent des antennes directement à l’intérieur du système nerveux central pour capter les variations du champ électrique. La résolution spatiale est grande (précision d’une sphère d’un rayon d’un millimètre environ, voir jusqu’au neurone dans certains cas) ainsi que la précision temporelle (micro seconde environ, souvent limitée par la puissance de calcul des ordinateurs)

(rACC), l'amygdale (AMY) et le cortex préfrontal rostromédial (rmPFC) encodent la récompense quand une personne joue elle-même. Ceux du rACC sont encore davantage recrutés lorsque l'on observe une autre personne jouer. On constate aussi qu'alors que les sous-populations de neurones dans l'AMY et le rmPFC qui encode la magnitude d'une récompense observée (pour autrui) encode aussi la magnitude d'une récompense expérimentée soi-même et l'encode dans la même direction, ceux du rACC encodent la récompense pour soi dans un sens opposé. Ce qui montre que la distinction entre le soi et l'autre pourrait entre autres être fait par des sous-populations de neurones dans le rACC. Cette équipe a aussi montré qu'une sous-population du rACC encoderais l'erreur de prédiction de récompense<sup>18</sup> mais uniquement celle lors de récompense observée et non lors de récompense expérimentée soi-même.

---

<sup>18</sup>Nous verrons dans la partie 1.3.1 qu'est ce qu'une erreur de prédiction plus en détail



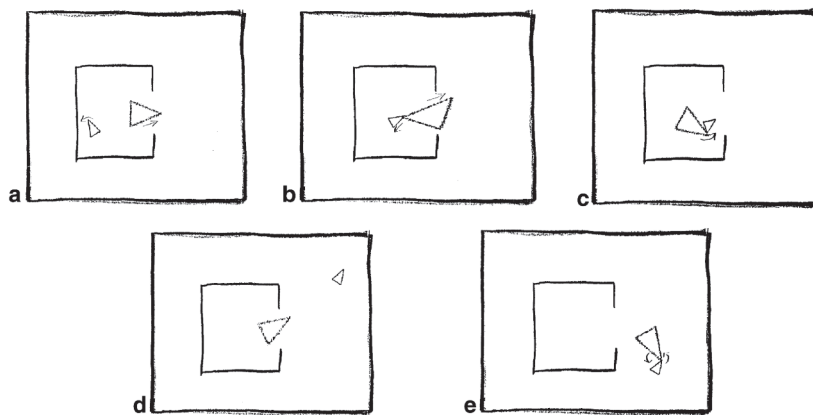


**Figure 1.9** – Tâche expérimentale, performance du modèle et localisation des électrodes. **(a)** Dans la tâche de jeu de cartes, les participants jouaient 12 parties de 5 cycles chacune. Chaque cycle consistait en 1 essai joué par le participant et deux essais observés. **(b)** Structure de chacun des 180 essais. Les essais joués et observés ont la même structure. **(c)** Écran du jeu de la machine à sous, auquel les participants jouaient à volonté pendant au moins 5 minutes. **(d)** Probabilité que le paquet de cartes 1 soit gagnant. En blanc la valeur réelle. La carte de chaleur représente la densité de probabilité, ou vraisemblance relative, pour chaque valeur de probabilité que le paquet numéro 1 soit récompensé<sup>19</sup>. En magenta la valeur moyenne de la distribution de probabilité qui est utilisée comme estimateur de la valeur attendue par le participant sur chaque essai. La probabilité réelle (70% ou 30%) que le paquet 1 soit gagnant est inversé au début de chaque partie (15 essais) avec une probabilité de 50%. Ici, le graphique représente la partie d'un participant. *Extrait de Hill et al. (2016).*

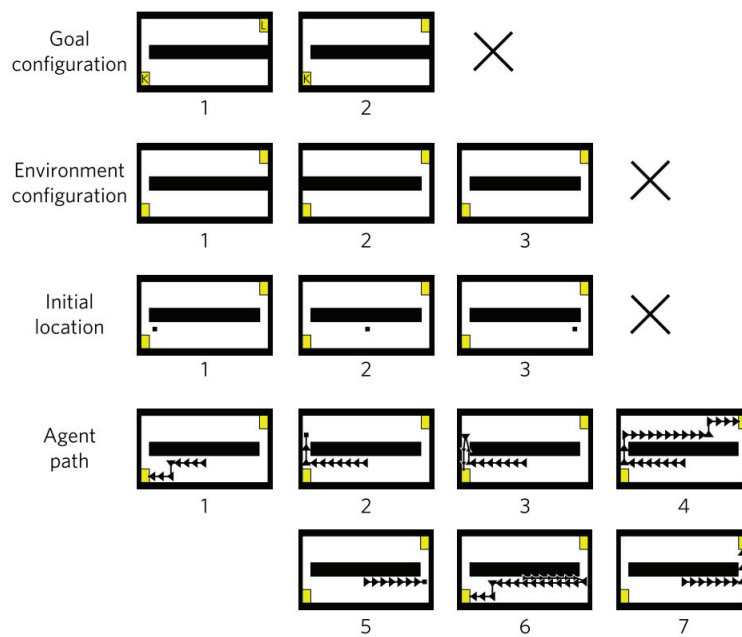
A force d'apprendre par observation, émerge d'une capacité à modéliser les états internes des autres, notamment les croyances. Ainsi, nous ne voyons plus les comportements des autres comme des simples mouvements, mais comme des actions intention-

<sup>19</sup>Ceci est lié au fait que le modèle utilisé est un modèle bayésien, qui donc maintient une densité de probabilité et non une simple valeur. Nous verrons ceci dans la partie 1.3.4

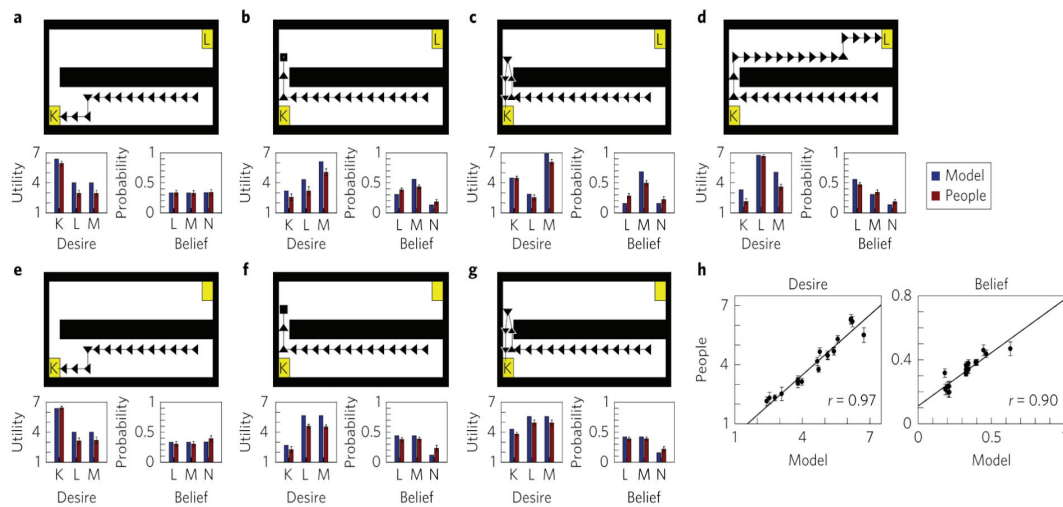
nelles, comme le résultat de plan dans le but d'accomplir des désirs étant donné des croyances (parfois incomplètes et fausses). Certaines expériences comme celle de Castelli et al. (2000) montrent que même dans un contexte non-social, en regardant des schémas dynamiques (voir figure 1.10) les humains parviennent à attribuer des croyances et des intentions à des flèches qui se déplacent. Dans leur papier, Baker et al. (2017) utilisent aussi des flèches en précisant aux participants qu'elles représentent des individus. Dans leurs travaux, ils parviennent à modéliser les prédictions des participants concernant les désirs et croyances de ces flèches (voir figure 1.12).



**Figure 1.10** – Cinq images extraites d'une des animations scénarisées (mère et enfant) : (a) La mère essaie d'inciter l'enfant à sortir. (b) L'enfant est réticent à sortir. (c) La mère pousse doucement l'enfant vers la porte. (d) L'enfant explore l'extérieur. (e) La mère et l'enfant jouent joyeusement ensemble. *Adapté de Castelli et al. (2000).*



**Figure 1.11** – Représentation des quatre facteurs variant dans la conception factorielle de la tâche de Baker et al. (2017). Les lignes noires représentent des murs bloquant mouvement et visibilité. Les carrés jaunes représentent des lieux où peuvent se garer des camions de restauration gratuite. 3 types de restaurants sont possibles : Coréen (K); Libanais (L) et Mexicain (M). L’agent observé est marqué par un triangle. L’observateur doit déterminer quel est le restaurant préféré de l’agent (K,L ou M) mais aussi quel était sa croyance initiale sur l’occupant de la place de parking lointaine (L, M ou rien (N)). *Extrait de Baker et al. (2017).*



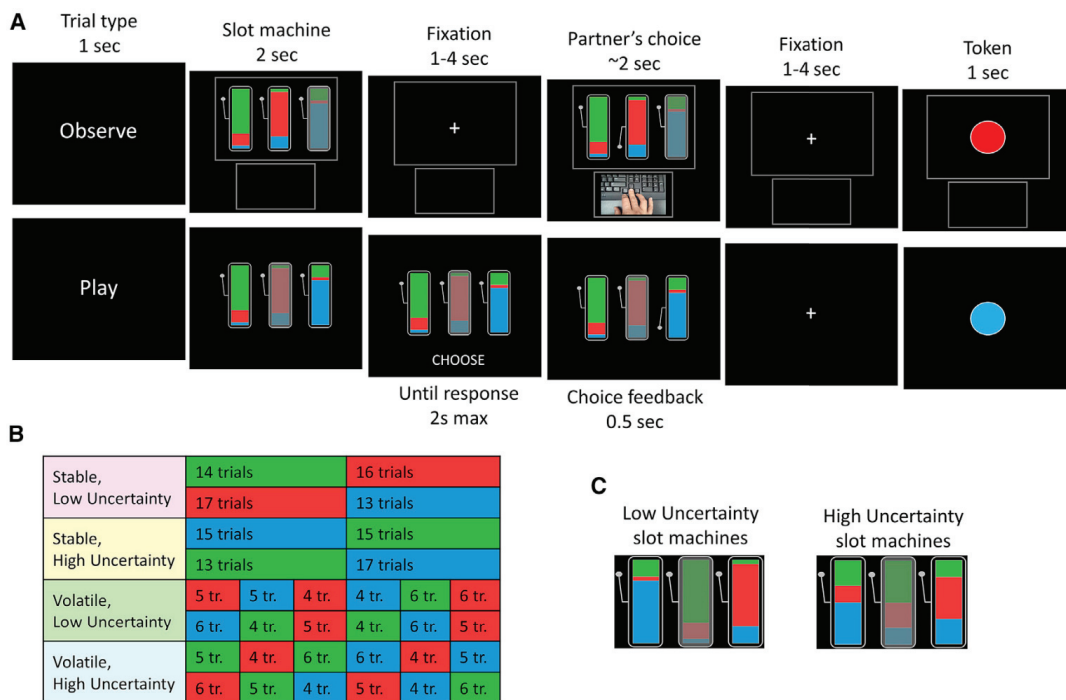
**Figure 1.12** – Tracés et corrélations (Coefficient de corrélation de Pearson  $r$ , pour  $n=21$ ) en utilisant les paramètres maximisant l’ajustement du modèle aux données ( $\beta = 2.5$ ). **(a-g)** Comparaison entre les données du modèle et celles des participants sur 7 scénarios clé (moyennes sur 16 participants, les barres d’erreurs représentent la somme des écarts à la moyenne). **(a-d)** Scénarios avec 2 camions présent. **(e-g)** Scénarios avec un unique camion présent. **(a,e)** L’agent va droit vers le premier camion. **(b-f)** Le chemin incomplet de l’agent se dirige derrière le bâtiment pour vérifier l’endroit éloigné. **(c,g)** L’agent retourne vers l’endroit le plus proche après avoir vérifié le plus lointain. **(d)** L’agent va vers l’endroit le plus loin après y avoir vérifié la présence d’un camion. **(h)** comparaison entre le modèle et la moyenne des participants ( $n=16$ ) concernant les préférences et croyances inférées sur les 7 scénarios (les barres d’erreur représentent la déviation standard au cours d’un essai) *Extrait de Baker et al. (2017)*.

Il est donc possible par observation d’inférer les états mentaux (croyances, buts, ...) d’autres personnes que l’on observe. Pour agir à notre tour, il vient alors la question du choix entre deux stratégies d’apprentissage par observation : imiter simplement une action ou émuler un nouveau choix depuis les buts et intentions inférés. Charpentier et al. (2020) tentent de répondre à cette question en proposant un jeu de machines à sous à des volontaires en alternant observation et action tout en faisant varier deux paramètres : l’incertitude et la volatilité concernant l’issue de chaque essai (voir figure 1.13). L’incertitude est liée aux probabilités qu’un jeton d’une certaine couleur sorte des machines à sous. La volatilité est liée au fait que la couleur du jeton qui a de la valeur change au cours du temps (très volatile si elle change souvent, peu volatile si elle change rarement). Le participant ne connaît pas quel est la couleur du jeton qui a le plus de valeur, mais il sait que la personne observée la connaît. Ainsi l’hypothèse

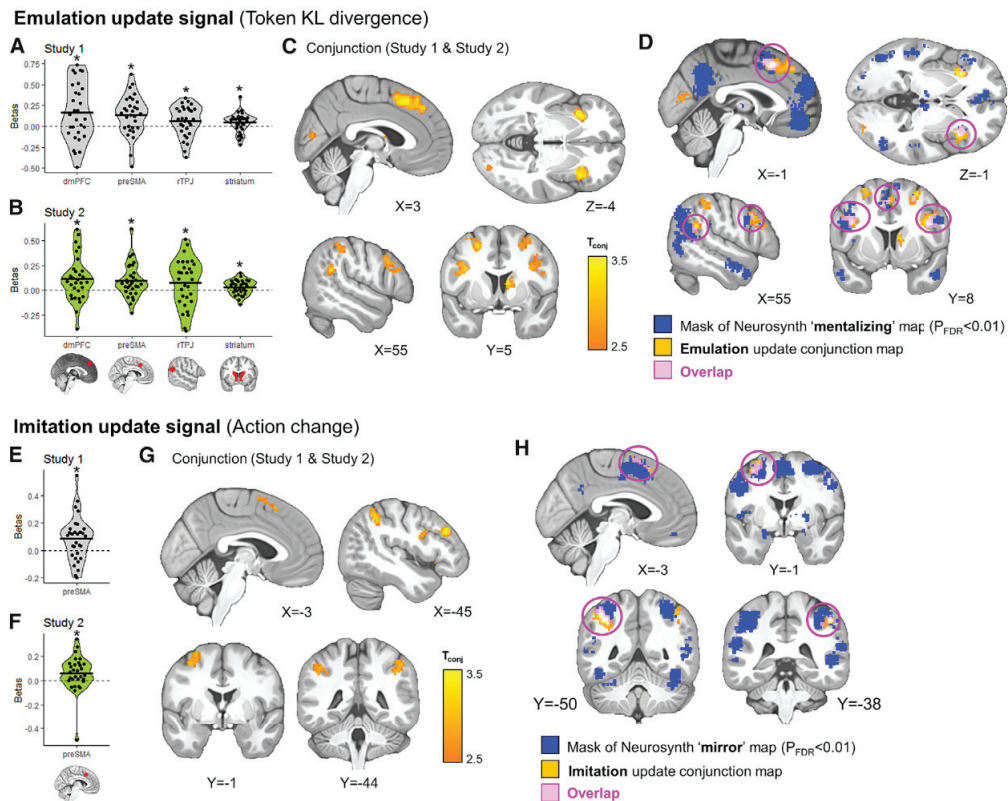
est que face à une grande volatilité concernant la couleur du jeton qui a la plus grande valeur, l'observé va être aussi volatile que la couleur du jeton puisqu'il est complètement informé. Ainsi, puisque l'observateur n'a pas connaissance du jeton qui a de la valeur, l'option d'imitation va être plus fiable pour lui que l'option d'émuler son propre choix. Inversement, si l'observé a un comportement plus stable, mais que les machines à sous sont plus imprévisibles, alors le participant devrait cette fois favoriser l'émulation. C'est en effet le résultat qu'ils trouvent grâce à un modèle par arbitrage<sup>20</sup>. Essentiellement, ce modèle consiste à faire l'hypothèse que deux processus fonctionnent simultanément dans le cerveau et qu'une région ferait l'arbitrage entre ces deux processus. De manière intéressante, ils retrouvent que le processus d'émulation est soutenu par un réseau sensible aux tâches de mentalisation alors que le processus d'imitation est implémenté par un réseau sensible à toutes les tâches impliquant les neurones miroirs (voir figure 1.14).

---

<sup>20</sup>L'arbitrage est issue d'une nouvelle théorie qui consiste à faire l'hypothèse que différents processus sont réalisés par différents experts dont le signal est ensuite intégré en fonction de leurs fiabilités respectives pour donner la décision finale.



**Figure 1.13** – Conception de la tâche de Charpentier et al. (2020). **(A)** Dans les essais d’observation (en haut), les participants voient le choix de machine à sous de l’agent observé. Les couleurs sur chaque machine représentent la probabilité relative qu’un jeton d’une couleur particulière va être délivré si l’agent choisit cette machine à sous. Dans chaque essai une des machines à sous est indisponible (grisé) mais les probabilités associées à cette machine reste visibles. Les participants savent qu’une unique couleur de jeton a de la valeur, mais ne savent pas lequel. Cependant, ils savent que l’agent observé possède toute l’information et joue de manière optimale. Le choix de l’agent est indiqué par une vidéo ainsi que par le bras de la machine à sous sélectionnée qui est alors descendu. **(B)** La tâche contient 8 blocs de 30 essais dans une conception factorielle 2 (stable/volatile) par 2 (faible/haute incertitude). La couleur des cases dans le tableau représente quel jeton (vert, rouge ou bleu) a de la valeur à ce moment (à l’insu du participant). L’ordre des blocs était contre-balançé entre les participants. Dans les blocs stables, seul un changement dans la couleur du jeton possédant une valeur à lieu. Dans les blocs volatils, 5 changements ont lieu. **(C)** Dans les blocs avec une faible incertitude la probabilité de distribution du jeton était de  $[0.75, 0.2, 0.05]$  rendant le calcul de la machine à sous plus facile que dans la condition de haute incertitude, pour laquelle la probabilité de distribution du jeton était  $[0.5, 0.3, 0.2]$  *Extrait de Charpentier et al. (2020).*



**Figure 1.14** – Signal de mise à jour de l'émulation et de l'imitation pendant l'observation. La divergence de Kullback-Leibler (mise à jour de l'émulation) ainsi que le changement dans l'action de l'agent observé par rapport à l'essai précédent (mise à jour de l'imitation) ont été ajoutés en tant que paramètre modulateur du signal BOLD au moment de la récompense. Les paramètres ont été mis en compétition pour expliquer la variance (i.e. qu'ils n'ont pas été orthogonalisés). (**A, B, E and F**) Dans la première étude et dans la réplication, la mise à jour du signal d'émulation a été significativement trouvée dans les régions d'intérêt du dmPFC, du preSMA, de la TPJ droite et le striatum dorsal (**A, B**). Une mise à jour significative du signal d'imitation a été trouvée dans une région d'intérêt du preSMA (**E, F**). (**C, G**) Une conjonction des résultats de niveau deux sur tout le cerveau de l'étude et de sa réplication montrent des groupes d'activation supplémentaires mettant à jours le signal d'émulation (**C**) ou de l'imitation (**G**). (Seuil à  $P < 0.0001$  non corrigé suivi d'une correction FWE à  $p < 0.05$ ). (**D, H**) Superposition des signaux d'émulation et d'imitation avec les cartes "mentalizing" (**D**) et "mirror" (**H**) de Neurosynth.org *Extrait de Charpentier et al. (2020)*.

### 1.2.1.2 Être observé

Lorsque nous sommes dans la situation d'être observé, on parle alors de facilitation sociale. Elle a aussi lieu lors d'interactions directes, mais ici nous nous intéressons au cas où une autre personne effectue une même tâche seul ou avec audience. Sous les

hypothèses de l'Homo-économus, l'observation par un autre ne devrait pas changer les décisions ou la balance coût bénéfice. Pourtant, il a déjà été montré que la présence d'une audience augmente l'éveil du participant lors d'une tâche complexe, qu'elle augmente la vitesse d'exécution des tâches simples et diminue celle des tâches compliquées. Concernant les performances, il semblerait que l'effet d'audience diminue la performance des tâches compliquées et augmente légèrement celle de tâches simples (Bond and Titus, 1983). Ces effets d'audience varient en fonction du stade de développement, des facteurs de personnalité, du contexte culturel et du diagnostic clinique, notamment l'autisme et les troubles anxieux (Hamilton and Lind, 2016). Ils sont décrits comme des effets liés aux capacités de théorie de l'esprit que nous étudierons à la fin de ce chapitre. Il a aussi déjà été vérifié dans des tâches de donation que l'approbation<sup>21</sup> sociale pouvait être intégré comme une récompense dans le réseau neural d'évaluation (Izuma et al., 2010), ce qui pourrait expliquer une générosité accrue lors de la présence d'une audience. Certaines études montrent que l'effet d'audience varie en fonction de l'expertise de cette audience (Henchy and Glass, 1968). Dans notre équipe, nous avons aussi voulu montrer que la variabilité interindividuelle de la subjectivité à l'effet d'audience était en partie due à la testostérone (Li et al., 2020), une hormone importante (parmi d'autres) dans la régulation des comportements sociaux (voir figure 1.15). Dans la première condition de cette tâche, les participants devaient accepter (ou refuser) de perdre un montant donné pour qu'une association "positive"<sup>22</sup> gagne une certaine somme d'argent. Dans l'autre condition, les participants pouvaient accepter ou rejeter une proposition de partage d'un montant entre eux et une association "négative". Il a d'abord été répliqué qu'avec une audience les participants acceptent davantage les échanges avec l'association positive qu'en privé. De même l'effet d'audience diminue l'acceptation des partages avec une organisation négative. L'équipe a aussi retrouvé le striatum<sup>23</sup> davantage impliqué

---

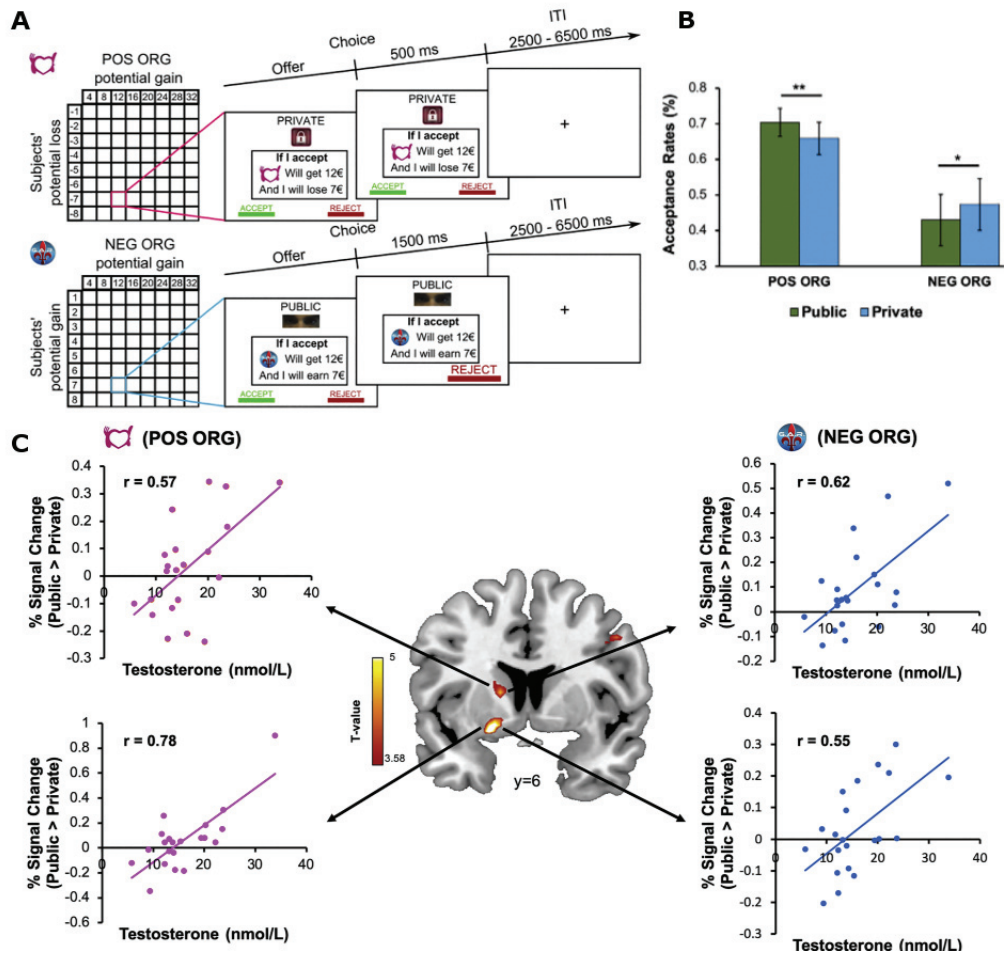
<sup>21</sup>A ne pas confondre avec le conformisme qui est le fait de se plier aux normes sociales, qui peut notamment avoir lieu même lorsque l'individu n'est pas observé. Park et al. (2017) ont dans notre équipe notamment étudié l'influence des choix de groupes de différentes tailles sur les choix d'un participant. L'approbation, elle, est le fait qu'un choix soit approuvé par un pair.

<sup>22</sup>La sélection des associations a fait l'objet d'un pilote comportemental

<sup>23</sup>Région connue pour son implication dans le circuit de la récompense et de l'évaluation d'un stimulus



dans la condition d'audience que seul. Le principal résultat nouveau est que le taux de testostérone des participants corrèle avec cette activité striatale lors des comportements pro-sociaux induits par l'audience <sup>24</sup>.



**Figure 1.15** – Principaux résultats de l'expérience de Li et al. (2020) sur l'effet d'audience. (A) Une conception factorielle 2 x 2 pour chaque sujet a été adoptée. Le participant devait accepter ou refuser une offre de partage d'un montant avec une association "négative" ou une offre de don à une association "positive" (facteur 1) en présence d'un public ou non (facteur 2). Les gains des associations variaient de 4€ en 4€ entre 4€ et 32€. Les gains et pertes des participants variaient entre 1€ et 8€. (B) Les résultats confirment un effet de l'audience vers ce qui semble socialement acceptable et recommandable. (C) Cette effet comportemental semble être entraîné par une activation supérieure du striatum ventral en public qu'en privé. Activation qui par ailleurs semble modulé par le taux de testostérone des participants. *Adapté depuis Li et al. (2020).*

<sup>24</sup>Ce qui est très intéressant car cette expérience donne un rôle pro-social à une hormone habituellement associée à des comportements d'agression.

### 1.2.1.3 Les interactions directes

Parfois, la situation est plus impliquante socialement qu'une simple observation, il peut y avoir interaction directe. Ainsi une interaction à lieu lorsque deux individus ou plus (d'une même espèce ou non) agissent l'un et l'autre par leur propre décision sur les états de l'autre. De nombreuses situations sont alors possibles et nous ne pourrions pas toutes les étudier ici. Parmi les situations possibles, nous avons notamment la distribution des ressources équitable ou non-équitable, les normes sociales et la morale, l'altruisme, le consensus, la confiance ou la rupture de confiance, le pardon, et bien d'autres encore. En neurosciences cognitive, pour étudier les interactions de groupe, nous utilisons souvent une tâche comportementale bien connue appelé le "jeu du bien commun". L'intérêt de cette tâche est qu'elle possède de nombreuses variantes permettant d'étudier l'impact sur la coopération de thèmes aussi varié que les différences social (Santos et al., 2008), la réputation (Milinski et al., 2002), la punition ou la récompense (Sigmund et al., 2001) ...<sup>25</sup> Le principe est celui qui suit.

Les participants sont inclus dans un groupe et ils vont effectuer ensemble un nombre de "périodes". À chaque période, tous les participants sont dotés d'un revenu  $w_i$  (la dotation peut varier en fonction de ce que l'on veut étudier e.g. l'inégalité, ou l'impact du moyen d'acquisition de la dotation). Chaque participant doit ensuite diviser son revenu en deux parties, une sur son compte privé  $x_i$  qui lui reviendra identique à la fin de la période multipliée par  $\alpha$ , et une sur un compte publique  $g_i$  dont la sommes de toute les contribution  $G$  sera multiplié par un facteur  $\beta$  pour être ensuite divisé et reversé à l'ensemble des participants. L'utilité comme elle a été définie précédemment est donc telle que :

$$u_i = \alpha * x_i + \frac{\beta}{\text{nombre de participants}} * G$$

Ainsi en fonction des paramètres  $\beta$ ,  $\alpha$  et du nombre de participants, il est plus ou moins intéressant pour les individus de coopérer.

---

<sup>25</sup>L'étude de ce jeu est si importante et intéressante car elle est une simplification de nombreuses situations sociales dont celle du consentement à l'impôt ou de l'engagement pour limiter le changement climatique par exemple.

Certaines variantes permettent notamment à chaque participant à la fin de chaque période d'endurer un coût pour infliger une perte d'argent (punition) à un participant au choix ou pour faire un don à un autre (récompense).

Une méta-analyse<sup>26</sup> a notamment permis d'élucider les paramètres psychologiques les plus importants pour augmenter la coopération (Zelmer, 2003). Cependant, les dernières études après 2003 ainsi que les jeux de bien public non-linéaire (je ne rentrerais pas dans le détail de la variante non linéaire) ne sont pas inclus dans l'étude. Tout de même, il a trouvé que le jeune âge, le marginal per capita (i.e.  $\frac{\beta}{\text{nombre de participants}}$ ; comment l'argent partagé avec le groupe fructifie et est retourné à chaque individu) et l'autorisation de communiquer ou non, sont les paramètres qui influent le plus sur la coopérativité (i.e. la tendance du groupe à mettre l'argent en commun plutôt que le garder pour eux). Suivie du cadre dans lequel est présenté le jeu (i.e. présenté de manière positive ou négative) et de si le même groupe interagit plusieurs fois ou non. Il a aussi montré que certains paramètres peuvent diminuer la coopérativité comme demander au participant ce que vont faire les autres avant de faire le partage de son revenu, l'habitude des participants avec cette tâche ou enfin l'attribution de revenus inégalitaires.<sup>27</sup>

Mon travail de thèse a notamment consisté à modéliser la prise de décision dans un tel contexte. Dans les travaux présentés dans le Chapitre (3) nous montrons comment l'attribution d'intention aux autres influe sur la décision de coopérer ou non du participant.

### 1.2.2 Quelles intentions derrière l'interaction

Dans un large groupe comme dans une interaction à deux, les interactions peuvent être assorties comme nous l'avons vu d'une intention. L'intention se place dans un continuum entre la coopération pure et la compétition pure. Par coopération pure,

---

<sup>26</sup>Une méta-analyse est une analyse d'analyses effectuées sur un même sujet. Elle permet notamment d'élucider les résultats spécifiques à certaines conditions expérimentales et les résultats plus généraux et robustes.

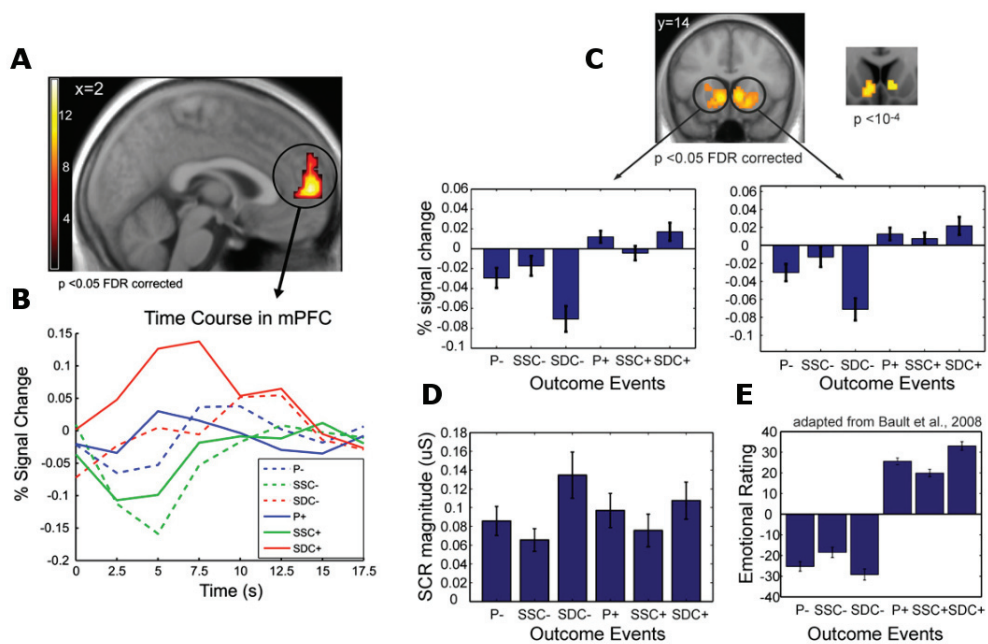
<sup>27</sup>Pour étudier les comportements de groupe, d'autres expériences hors de la théorie des jeux sont possible, comme avec cette étude (Goupil et al., 2020) qui étudie l'auto-coordination d'un large groupe de musique en improvisation et montre que la coordination peut émerger d'elle-même tout du moins quand les intentions sont compatibles et flexibles.

j'entends que les différentes parties qui coopèrent ont un but identique, rentre donc dans ce mode d'interaction la coordination. Par compétition pure, j'entends que les différentes parties ont des buts opposés ou incompatibles. La théorie des jeux nous permet d'ajuster les récompenses en fonction des états de l'interaction et de créer ainsi un continuum entre la coopération pure et la compétition pure en créant des états intermédiaires. Les modes d'interactions pouvant être dynamique au cours des interactions.

### **1.2.2.1 Les intentions compétitives**

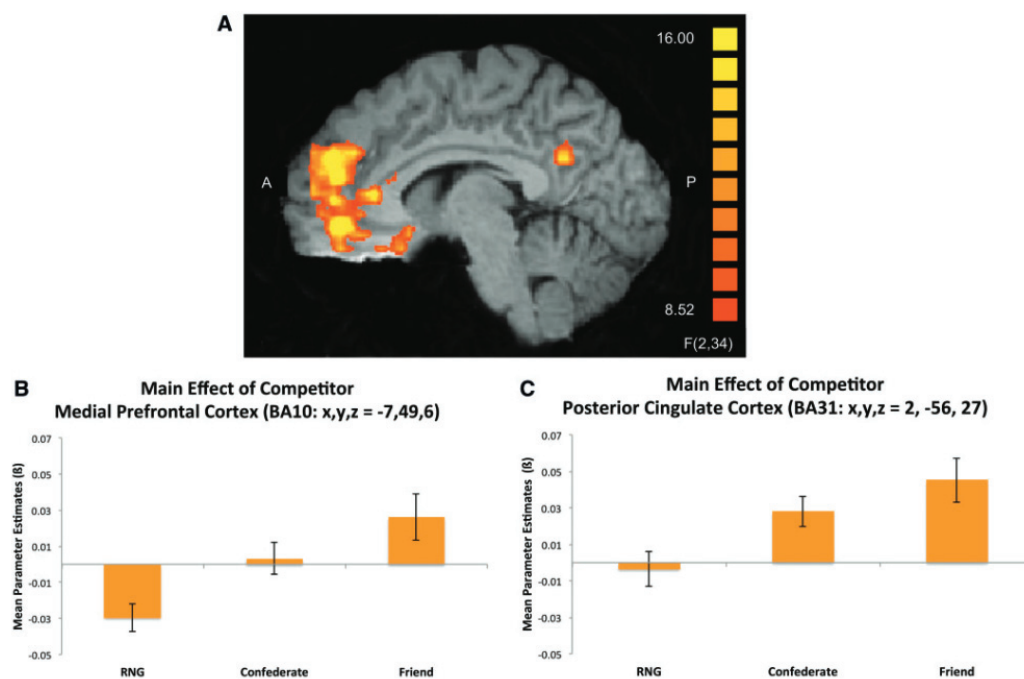
La compétition apparaît lorsque des ressources sont en quantité limitée et que deux individus ou plus veulent accéder à cette ressource qui peut aussi bien être primaire comme de la nourriture, des opportunités de reproduction, des abris ou l'occupation d'espace, que secondaire comme le prestige, l'argent ou le pouvoir. S'il n'y a pas de partage consensuel (i.e. coopération) des ressources, les individus se voient engagés dans une compétition sociale résultant en une victoire pour l'un et une défaite pour l'autre. Il y a donc nécessité d'une interdépendance et d'un conflit d'intérêt pour générer de la compétition. La victoire lors d'une compétition engagée avec un adversaire intelligent repose souvent sur la capacité d'inférer ses états mentaux pour prédire ses futures actions et l'influence de ses propres actions sur celles de son opposant. Ainsi, il n'existe pas une littérature propre à la compétition. Celle-ci se rattache souvent à celle de la dominance, celle de la théorie de l'esprit ou est discutée en opposition à la coopération par exemple. Toutefois, il est connu que si le contexte permet une comparaison entre ses propres récompenses et celles des autres cela engage les individus dans une compétition et les pousse à la prise de risque (Bault et al., 2011), même s'il n'y a pas de compétition directe pour les mêmes ressources (voir figure 1.16). Quelles sont les bases neurales de ce comportement observé ? Bault et al. (2011) ont montré que l'activité cérébrale dans le striatum ainsi que dans le mPFC et la TPJ est plus importante lorsque le participant fait mieux que son co-joueur dans la condition d'observation comparé à un même gain dans la condition où le participant joue seul. Inversement, cette activité est plus faible s'ils gagnent tous les deux. Ce résultat de compétition par comparaison sociale avait déjà été

montré dans une autre tâche où deux participants jouent à un même jeu perceptif dans lequel les récompenses varient en fonction de la réponse du participant (correcte ou non) et de la condition (récompense relative joueurA:JoueurB : 1:2 ou 1:1 ou 2:1). Ainsi Fließbach et al. (2007) ont prouvé que l'activation du striatum est notamment lié aux victoires et aux défaites, mais que cette activation est modulé par la récompense relative, c'est-à-dire que le striatum est plus activé dans la condition où le participant gagne plus que l'adversaire et inversement.



**Figure 1.16** – Résultats de Bault et al. (2011). **(A)** Activité du mPFC corrélé aux gains sociaux. Activité du mPFC discriminant entre les récompenses des 3 conditions (P=Privé, SSC = social, choix identique, SDC = social, choix différent) au moment où la récompense est révélée. Carte statistique du test de Fisher projeté sur le cerveau moyen du groupe. **(B)** Décours temporel dans le mPFC ( $x,y,z=0,54,9$ ) pour les 6 résultats possibles (- représente une défaite, + représente une victoire). Le mPFC est plus activé pour les gains sociaux que pour toutes les autres conditions. **(C)** Activité striatale qui encode à la fois la valence de la récompense et le contexte, au moment où la récompense est révélée. Cette vue coronale montre l'effet d'interaction entre la valence et le contexte. Le graphique bar montre le pourcentage de changement de signal ( $\pm$  écarts standards à la moyenne, SEM). Noyau caudé gauche ( $x,y,z = -9,9,-3$ ) et droit ( $9,12,-3$ ). **(D)** Conductance de la peau moyenne ( $\pm$  SEM) pour les six événements possible. Unité :  $\mu$ siemens. **(E)** Évaluation de l'émotion [sur une échelle de -50 (extrêmement négative) à +50 (extrêmement positive) en passant par 0 (ni positive ni négative)] faites par les 42 participants pour les 6 types de situations possibles. Réalisé dans une étude comportementale précédant l'étude par IRMf. *Adapté depuis Bault et al. (2011).*

Nous savons aussi que la personne qui interprète le rôle de l'opposant compte en terme de compétition. En effet Fareri and Delgado (2014) ont retrouvé que le striatum encode toute les récompenses, mais ils montrent en plus que le mPFC et le PCC encoderaient des données plus informative ou spécifique, comme la proximité sociale de l'opposant, ou comme le trouvent Hampton et al. (2008) les futures récompenses liées aux intentions et aux croyances de l'autre (voir figure 1.17).



**Figure 1.17** – Effet principal de l'adversaire au moment de la récompense. **(A)** L'ANOVA en mesure répétée sur le cerveau entier avec comme facteur 2 (*rôles*) \* 3 (*type d'adversaire*) \* 2 (*récompenses possibles*) révèle un effet principal de l'adversaire dans de nombreuses régions. **(B)** L'estimation des paramètres dans un groupe de voxel<sup>28</sup> centré en BA10 (x,y,z = -7,49,6) montre que cet effet est dirigé par une augmentation du signal lorsque l'adversaire est un ami comparé à un confédéré ou un algorithme. **(C)** Même résultats pour un groupe de voxel du cortex cingulaire postérieur à côté du cunéus (BA31). Les cartes d'activations sont d'abord seuillées à  $p < 0.001$  puis corrigé au niveau du groupe pour les faux positif et négatif à  $p < 0.05$ . Adapté depuis Fareri and Delgado (2014).

<sup>28</sup>Le voxel est comme le pixel en trois dimensions, c'est un volume alors que le pixel est en deux dimensions, c'est une surface.

### 1.2.2.2 Les intentions coopératives

Si la compétition est présente chez la quasi-totalité des espèces, la coopération est moins fréquente, notamment quand il s'agit de coopération basée sur un altruisme réciproque, qui est souvent la base de la coopération entre inconnus. Grâce à la théorie des jeux, il est possible de définir les jeux coopératifs dans un cadre formel.

|                 | Coopérer | Faire defection |
|-----------------|----------|-----------------|
| Coopérer        | R        | S               |
| Faire defection | T        | P               |

**Table 1.1** – Matrice de récompenses d'un jeu dans lequel deux stratégies sont possibles : coopérer ou faire défection. Les lignes indiquent les récompenses du joueur que l'on étudie. Extrait de (Rand and Nowak, 2013)

Pour qu'un jeu soit un dilemme coopératif il faut que (i) les deux coopérants soient plus récompensés que deux défections ( $R > P$ ), mais qu'il y ait une incitation à la défection qui peut soit être sous la forme de  $T > R$ , soit  $P > S$  soit finalement  $T > S$ . Dans le premier cas la meilleure stratégie est de faire défection face à un coopérateur, dans le deuxième cas il faut faire défection face à quelqu'un qui fait défection et dans le dernier cas, si l'un des deux seulement fait défection il vaut mieux être celui-ci. Dans le jeu le plus utilisé pour étudier la coopération entre deux individus, le dilemme du prisonnier, les trois cas sont réunis :  $T > R > P > S$ . Le jeu est le suivant : " Vous essayez de vous échapper d'une prison avec un autre prisonnier. Si vous coopérez vous êtes tous les deux libres. Si vous faites défection seul alors que l'autre coopère, vous serez récompensé. Si l'autre prisonnier fait défection seul il sera récompensé. Si vous faites tous les deux défections vous serez récompensés tous les deux mais dans une moindre mesure." Le jeu se matérialise par la matrice de gain ci-dessous. En général pour étudier la coopération à plus de deux individus c'est le jeu de bien public dont nous avons parlé précédemment qui est utilisé.

|          |                 | Joueur 2 |                 |
|----------|-----------------|----------|-----------------|
|          |                 | Coopérer | Faire defection |
| Joueur 1 | Coopérer        | \$2(2)   | \$0(3)          |
|          | Faire defection | \$3(0)   | \$1(1)          |

**Table 1.2** – Matrice de récompense du jeu du dilemme du prisonnier dans (Rilling et al., 2002). Entre parenthèses figurent les récompenses de l'autre joueur.

Ainsi Rilling et al. (2002) ont pu lors d'interactions répétées avec une même personne au travers de ce jeu commencer à déterminer les bases neurales de la coopération. Ils ont trouvé que les activations cérébrales ne dépendent pas uniquement du choix de l'individu observé, mais bien de l'interaction entre les choix des deux participants. Ainsi, ils ont trouvé que le striatum ventral et l'OFC étaient plus activés (au moment de la récompense) lors d'une coopération mutuelle ou défection mutuelle que lors d'une défection d'un unique joueur, qui pourtant est plus récompensant en terme monétaire pour le participant qui fait défection. Cependant, ces activations peuvent représenter une aversion à la inégalité. En comparant les essais où les participants ont tout deux coopéré avec la moyenne des activations lors des autres essais, ils constatent encore une activation des même deux régions. De plus, ils constatent que plus l'activation du striatum est grande chez les participants, plus ils vont avoir tendance à coopérer de nouveau à l'essai suivant. Il y a donc une notion de dynamique entre les essais. Ainsi, ils ont regardé ce qu'il se passe au niveau neural lorsqu'un participant choisi de coopérer après un essai où l'autre a coopéré comparé aux 3 autres type d'essais possible (défection après coopération, défection après défection ou coopération après défection). Ils ont trouvé une activation de l'ACC rostral et du striatum ventral. Ce qui laisse supposer que la coopération peut être maintenue par une forme d'altruisme. En effet, coopérer face à quelqu'un qui fait défection revient à endurer un coût. Cependant, cela fait vivre l'interaction sociale à l'autre comme coopérative et donc gratifiante, impulsant ainsi un acte de réciprocité. Cependant, les mécanismes sous-jacents restent méconnus à ce jour.

Par la suite, Rilling et al. (2008) a étudié les effets d'une non-réciprocité à une coopération en utilisant le même paradigme expérimental. Ainsi ils ont trouvé qu'une



non-réciprocité du joueur (2) face à une coopération du participant (1) était associée à une augmentation de l'activation de l'insula antérieure bilatérale comparé à une réciprocité. Ils constatent que ces mêmes aires répondent plus à une non-réciprocité qu'à une prise de risque non sociale non fructueuse (en effet, comme vu précédemment coopérer face à quelqu'un qui fait défection est vu comme une prise de risque, ils ont donc contrôlé cet effet en utilisant une condition non-sociale avec prise de risque). Ils ont de plus constaté que l'insula bilatérale, l'hippocampe gauche et le striatum ventral bilatéral du participant étaient plus activés lors d'une défection de l'autre si le participant avait coopéré au tour d'avant. Ces régions sont donc plus activées lorsque le choix de l'autre joueur est probablement perçu comme une attitude de non-réciprocité, non coopérative ou de trahison.

Ces mécanismes laissent à penser que l'évolution a modulé nos cerveaux de manière à favoriser la réciprocité directe ou indirecte (Rand and Nowak, 2013). En effet, Rand and Nowak (2013) proposent que la réciprocité promeuve la coopération par cinq mécanismes qui sont :

**La réciprocité directe** qui permet à un individu d'endurer un coût, lorsqu'il sait que d'autres interactions sont à venir, dans l'optique de gagner un geste de réciprocité futur. C'est dans ce cadre de réciprocité que s'inscrit la stratégie de coopération-réciprocité-pardon (CRP)<sup>29</sup>.

**La réciprocité indirecte** a lieu lorsqu'une personne effectue un acte de coopération alors qu'il est observé ou que cet acte de coopération est révélé et que les parties prenantes sont amenées à interagir de manière répétée. Ainsi, par exemple, la réputation participe à la réciprocité indirecte.

**La sélection des pairs ou parents** consiste en la reconnaissance d'un pair ou d'un parent et donc face à qui il convient d'adopter un comportement coopératif en conséquence.

---

<sup>29</sup>La stratégie est la suivante : D'abord je coopère avec un individu, ensuite par réciprocité, je donne à la hauteur de ce que je reçois (aide ou agression par exemple), puis s'il y a eu agression, il faut pardonner pour pouvoir de nouveau entrer dans un cycle de coopération

**La sélection spatiale** peut aussi favoriser la coopération. En effet, comme les individus interagissent avec d'autres proche d'eux, les coopérateurs peuvent former des groupes dans lesquels les coopérateurs sont plus susceptibles d'interagir avec d'autres coopérateurs, ainsi la coopération prévaut. La sélection spatiale peut représenter par exemple un réseau social, une zone géographique ou un phénotype.

**La sélection multi-niveau** intervient lorsque les interactions interindividuelles peuvent se décomposer en plusieurs niveaux<sup>30</sup>. Ainsi, la coopération à l'intérieur des groupes est favorisée, car elle permet à un groupe plus coopératif d'être meilleur dans la compétition inter-groupe qu'un groupe moins coopératif.

Nous avons vu que l'on peut utiliser le jeu de bien public pour étudier les interactions avec plus de deux individus, mais l'équipe a aussi utilisé ce paradigme expérimental pour étudier l'attribution d'intention au niveau du groupe (Park et al., 2019). Dans ce papier Park et al. (2019) ont montré que pour choisir de coopérer ou de faire cavalier seul les participants calculent d'abord la coopérativité du groupe (au vu des essais précédent) puis qu'ils en dérivent leur utilité individuelle (qualifiée aussi d'utilité de court terme au vu de la construction du modèle qui reproduit le mieux les données) ainsi que l'utilité du groupe (qualifié d'utilité de long terme, puisque la coopération de certains augmente la coopération des autres et maintient la cohésion pour les essais suivants). Ainsi, il a été démontré que la coopérativité du groupe est encodée dans les jonctions temporopariétales alors que la compétitivité du groupe activera plutôt l'ACC. L'utilité individuelle a été, elle retrouvée dans le vmPFC ainsi que dans l'insula droite. L'utilité du groupe, elle, a été retrouvée dans le cortex latéral préfrontale (IPFC) et dans le lobule intraparietal. Finalement, l'utilité individuelle et celle du groupe sont intégrées par l'ACC et le ventral IPFC afin de déterminer la stratégie à adopter. Il est ainsi intéressant de voir comment le conflit entre utilité individuel et utilité du groupe est intégré pour maintenir la coopération du groupe nécessaire pour maximiser ses propres profits.

---

<sup>30</sup>Au sein d'une nation en même temps qu'entre les nations par exemple

Afin de savoir avec qui coopérer ou avec qui entrer en compétition, il est nécessaire d'attribuer aux autres des aptitudes ou des capacités. Ces aptitudes peuvent être apprises par observation ou par interaction compétitive ou coopérative direct. De l'ordonnement d'individu le long de cet axe d'aptitude émergent une ou plusieurs hiérarchies chez l'Homo-sapiens comme chez de nombreuses autres espèces.

### 1.2.3 Émergence d'une hiérarchie et relation de dominance

Par hiérarchie de dominance, nous entendons une forme d'organisation dans laquelle il y a des dominants et des dominés, dans le cadre d'une compétition pour l'accès aux ressources. Dans une hiérarchie les dominants ont un accès privilégié aux ressources par la menace, la ruse, l'intimidation ou l'affichage de la force par exemple. La dominance étant ainsi défini comme la capacité à prévaloir en cas de conflit. La hiérarchie peut être variable au cours du temps au sein du même groupe, elle peut notamment être contextuelle (e.g. je domine au *babyfoot*, mais je suis dominé au jeu d'échecs), elle peut aussi être de prestige ou lié au physique. De nombreux travaux ont été faits sur le sujet au sein de notre équipe et qui ont notamment abouti à une revue de la littérature sur le sujet (Qu et al., 2017). Dans ce papier, l'équipe revoit les bases évolutives de la hiérarchie sociale. En effet, il a été montré que le leadership permet de coordonner les actions d'un groupe et qu'il permet de gagner du temps dans la prise de décision de groupe. Différents facteurs déterminent l'émergence théorique d'une hiérarchie, comme l'accès partielle à l'information pour certains membres du groupe, la personnalité des individus, des pouvoirs différenciés entre membres du groupe ou le contrôle des ressources par certains individus. Par conséquent, dans une société où l'accès aux ressources peut difficilement être contrôlé de manière monopolistique, où le partage des ressources est essentiel à la survie, où les individus forment des coalitions pour renverser les dominants et où il est facile de quitter le groupe auront des hiérarchies plus atténuées. Ainsi, les

disparités interindividuelles semblent être la source de l'émergence d'une hiérarchie verticale qui peut **parfois** être bénéfique à la fois pour le groupe et les individus<sup>31</sup>.

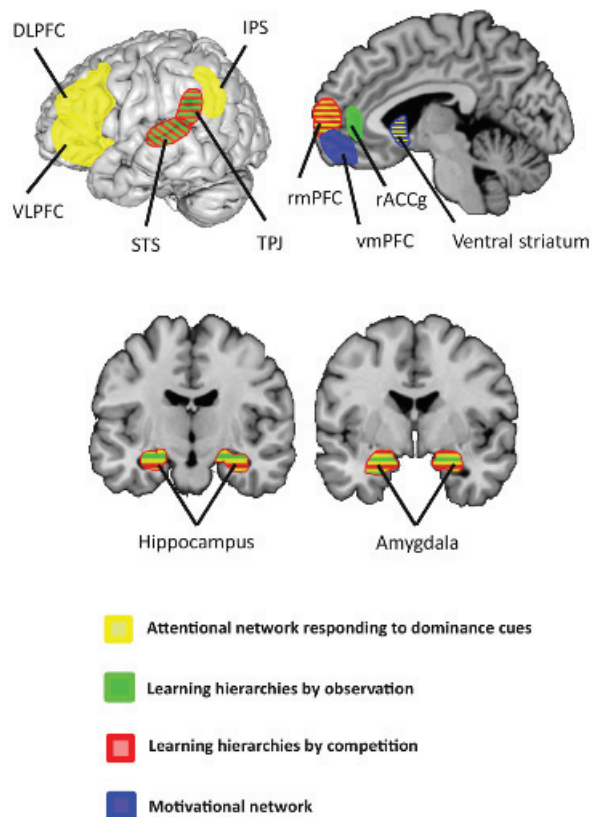
Une bonne connaissance de la hiérarchie d'un groupe permet une meilleure adaptation et navigation au sein de celui-ci et donc potentiellement un meilleur accès aux ressources. Notamment, elle est utile pour créer des alliances et éviter de perdre des ressources dans des conflits inutiles, mais elle est aussi importante pour les comportements d'imitation qui est un moyen primaire d'apprendre par observation. En effet, il a été montré que les individus imitent prioritairement les dominants (Clemson and Evans, 2012).

Cette hiérarchie peut être apprise par observation des indices visuels (taille, expression faciale, attribut physique...), par observation des interactions entre les membres d'un groupe <sup>32</sup> ou finalement par interaction dyadique directe. L'apprentissage de la dominance est connu pour mettre en jeu un réseau impliqué dans l'attention (e.i. amygdale, rmPFC et hippocampe, voir figure 1.18), un réseau motivationnel (motivation à gagné ou à éviter la perte ou les émotions négatives) avec l'engagement notamment du striatum ventral, mais aussi le réseau social avec la TPJ et le STS.

---

<sup>31</sup>En effet une hiérarchie peut dans certains cas être bénéfique pour un groupe, notamment en permettant une prise de décision rapide et flexible, ou en évitant des conflits menant à la perte de ressources individuels.

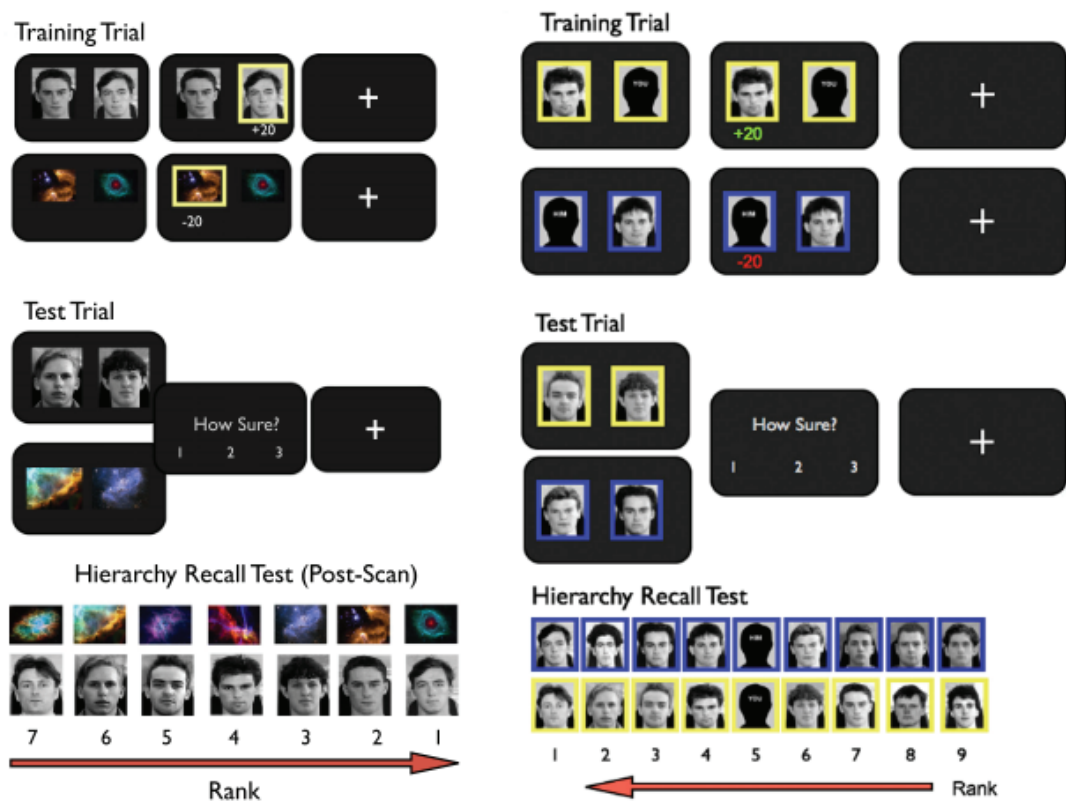
<sup>32</sup>Ce qui permettent d'éviter les coûts d'une confrontation directe



**Figure 1.18** – Principal réseau engagé dans les études de neuroimagerie lors de représentation de hiérarchie sociale basé sur la perception des rangs sociaux depuis des indices visuels (jaune), par observation (bleu) ou par compétition dyadique direct (rouge). En bleu, est représenté le réseau motivationnel classique. Les réseaux engagés sont composés de : (i) le réseau attentionnel répondant aux indices visuels de dominance, incluant les cortex pariéto-préfrontaux bilatéraux (jaune); (ii) un réseau engagé dans l'apprentissage des hiérarchies par observation composé de la TPJ, STS et du rACCg (vert); et (iii) un réseau reflétant l'apprentissage de la hiérarchie par compétition qui recrute le rmPFC (BA 10), et qui s'étend jusqu'au cortex préfrontal dorsomédian (rouge). Le quatrième réseau, le réseau motivationnel (bleu) composé du vmPFC et du striatum ventral est engagé dans l'apprentissage de ses propres actions et des récompenses. Les régions engagées dans plusieurs processus sont le striatum ventral, la TPJ, le STS, l'amygdale et l'hippocampe (lignes hachurées). Abréviations : DLPFC, cortex préfrontal dorsolatéral; IPS, sulcus interpariétal supérieur; rACCg, gyrus cingulaire antérieur rostral; rmPFC, cortex préfrontal rostromédial (BA 10); STS, sulcus temporal supérieur; TPJ, jonction temporo-pariétale; vLPFC, cortex préfrontal ventrolatéral; vmPFC, cortex préfrontal ventromédian. *Extrait de Qu et al. (2017)*

Kumaran et al. (2016, 2012) ont notamment participé à l'élaboration des résultats précédents. En effet, ils ont mis en place un paradigme expérimental pour étudier les comportements lors de l'apprentissage de la hiérarchie lors d'observation d'interactions

entre deux individus d'un même groupe ainsi que les bases neurales associées (voir figure 1.19). Leur protocole expérimental permet aussi de distinguer un apprentissage d'une hiérarchie non-sociale d'une hiérarchie sociale. Dans leur expérience, les participants étaient face à deux images (sociales ou non-sociales) et devaient apprendre par essais et erreurs lequel dominait l'autre, ceci par bloc de 6 essais. Les participants alternaient les blocs d'entraînement durant lesquels ils étaient face à des *objets* adjacents dans la hiérarchie pour lesquels ils avaient un retour sur leurs performances et les blocs d'évaluation pour lesquels les *objets* étaient non adjacents dans la hiérarchie, les participants devaient donc inférer la relation de dominance et ils n'avaient pas de retour sur leurs performances.



**Figure 1.19** – Tâche expérimentale de Kumaran et al. (2012, 2016). **(Haut.)** Exemple d’essai d’entraînement. À gauche dans l’expérience de 2012 impliquant une condition sociale (personne) et une condition non-sociale (galaxie). À droite dans l’expérience de 2016 avec une hiérarchie impliquant soi-même ou un ami. Dans l’ordre d’apparition des écrans, les participants voyaient 2 *objets* adjacents dans la hiérarchie. Le second écran apparaît quand le participant a sélectionné l’*objet* qui selon lui à le plus de pouvoir (en condition sociale) ou de minéraux (en condition non-sociale), il a alors un retour sur sa performance (-20 erreur ou +20 correct). **(milieu.)** Exemple d’un essai d’évaluation. Les participants étaient face à deux *objets* non-adjacents, ils devaient inférer lequel avait le plus haut rang et noter leur confiance en leur choix. Pas de retour sur leurs performances. **(Bas.)** Test de rappel de la hiérarchie (séance de débriefing): Les images des *objets* (personne ou galaxies) étaient montrées aux participants qui devaient les classer en fonction de leur ordre dans la hiérarchie. La position 1 donne le plus haut rang, la position 7 donne le plus bas rang. L’ordre des *objets* dans la hiérarchie était aléatoire pour chaque participant. *Adapté de Kumaran et al. (2012, 2016)*

Ainsi, dans leur première expérience, en 2012, ils ont trouvé un réseau commun à l’apprentissage de hiérarchie sociale et non sociale (i.e. hippocampe postérieur et le vmPFC) ainsi qu’un réseau spécifique aux hiérarchies sociales : l’amygdale et l’hippocampe antérieur (voir figure 1.20). De plus, ils ont trouvé que le volume de matière grise dans

l'amygdale prédit la capacité à inférer la hiérarchie de manière transitive (donc les performances dans les phases d'évaluation).

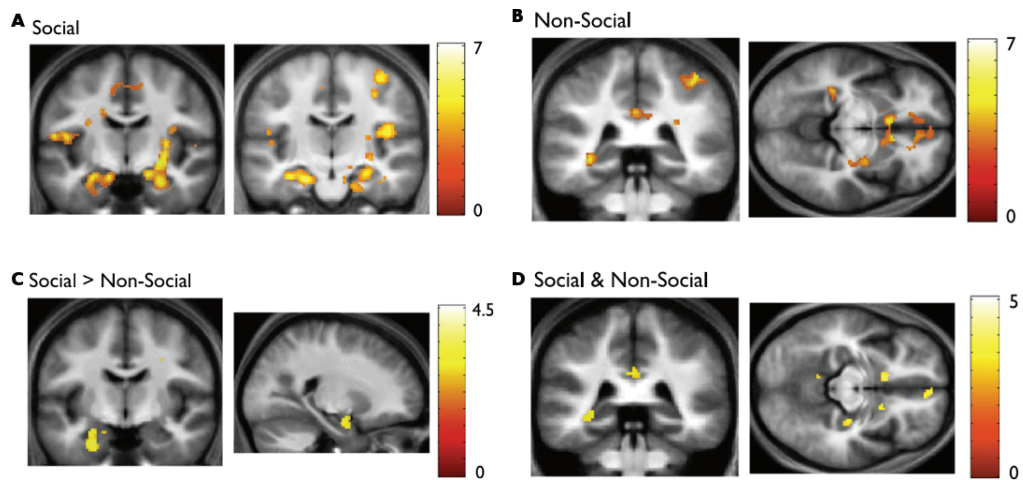
Dans la deuxième expérience en 2016, Kumaran et al. (2016) ont fait le choix d'une analyse basée sur des modèles. Ils comparent non plus une hiérarchie sociale avec une hiérarchie non-sociale, mais ils comparent une hiérarchie dont le participant fait partie intégrante avec une hiérarchie dont seulement un ami du participant fait partie (le paradigme est aussi expliqué figure 1.19). Ils ont tout d'abord trouvé que le meilleur modèle pour décrire le comportement des participants est un modèle Bayesian qui maintient une distribution de probabilité du niveau de dominances des *objets* et qui intègre notamment un oubli régulier des valeurs apprises d'un essai à l'autre. Ce modèle est le meilleur comparé à des modèles d'apprentissage par renforcement classique -non Bayesian- comme le RL-ELO<sup>33</sup> par exemple. Puis cherchant quelles régions cérébrales encodent la différence de valeur calculée par le modèle entre les deux *objets* présents, ils retrouvent l'implication de l'amygdale et de l'hippocampe<sup>34</sup> (voir figure 1.21). De plus, ils trouvent cette différence dans le vmPFC lors des essais en phase d'évaluation. En phase d'entraînement, ils trouvent le vmPFC plus engagé lors de la mise à jour des valeurs d'une hiérarchie impliquant soi-même (comparé à une hiérarchie impliquant un ami). Le vmPFC est aussi davantage engagé dans l'encodage de la valeur de *l'objet* choisi dans sa propre hiérarchie que dans celle de son ami. L'approche par les modèles est intéressante, car elle permet de confirmer l'implication des régions mises en évidence lors de la première expérience, mais en plus, elle permet de décrire leur rôle algorithmique (e.g. encodage de la valeur ou mise à jour de celle-ci)

---

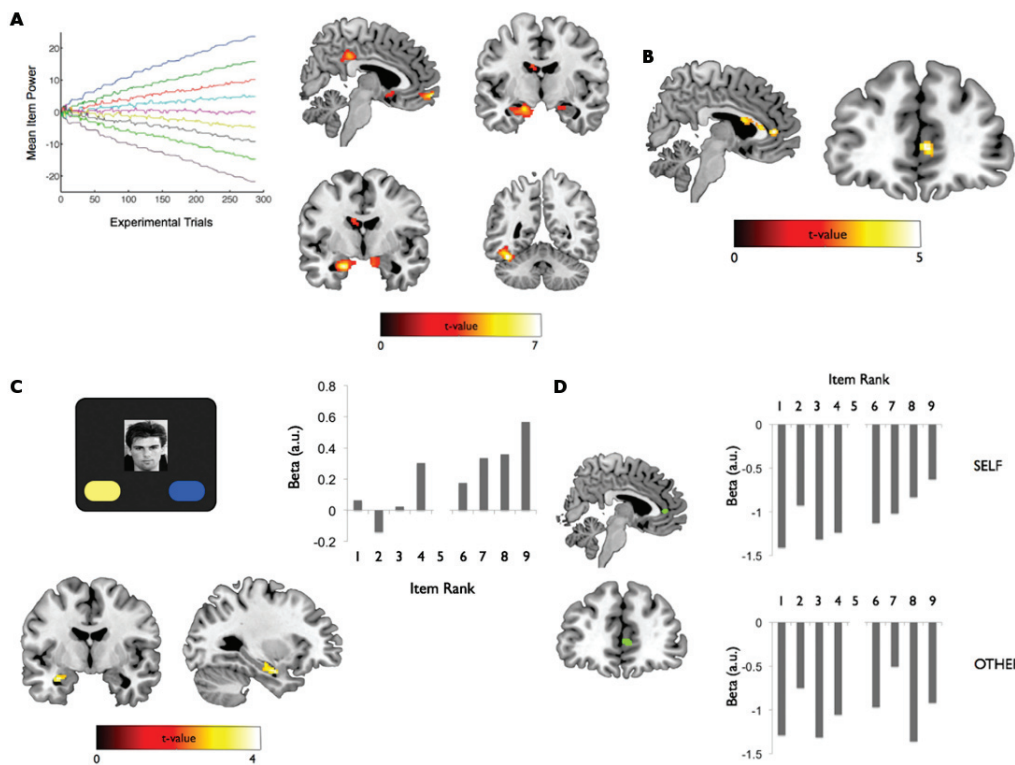
<sup>33</sup>C'est un modèle très connu notamment par son utilisation dans le classement international de jeu d'échec. Le principe est d'augmenter le score ELO d'un participant s'il gagne en fonction du score ELO de son adversaire. De même, le principe est de diminuer le score ELO après une défaite en fonction du score de son adversaire.

<sup>34</sup>Réplication des résultats de 2012 qui peut être rassurante quant à l'utilisation de la modélisation.





**Figure 1.20** – Résumé des résultats de Kumaran et al. (2012). **(A, B, C, D)** Phase d'apprentissage. **(A)** L'activité dans l'amygdale bilatérale ainsi que dans l'hippocampe antérieur bilatéral corrèle positivement avec l'index d'inférence dans le domaine social.  $p < 0.005$  non corrigé pour l'image affichée sur le cerveau moyen des participants. Significatif dans l'amygdale et l'hippocampe à  $p < 0.001$  non corrigé et corrigé au niveau du groupe à  $p < 0.05$ . **(B)** L'activité dans l'hippocampe postérieure gauche est significativement corrélée au score d'inférence en condition non-sociale (correction pour des petits volumes dans le vmPFC et l'hippocampe). **(C)** L'activité dans l'amygdale gauche/l'hippocampe antérieur et dans l'amygdale droite est davantage corrélée avec le score d'inférence en condition sociale comparé à la condition non-sociale. L'activité est significative lorsqu'elle est corrigée pour des petits volumes. **(D)** Résultat de l'analyse de conjonction. L'hippocampe gauche et le vmPFC corrèlent significativement avec le score d'inférence à la fois en condition sociale et en condition non-sociale (corrigé pour des petits volumes). *Adapté de Kumaran et al. (2012)*



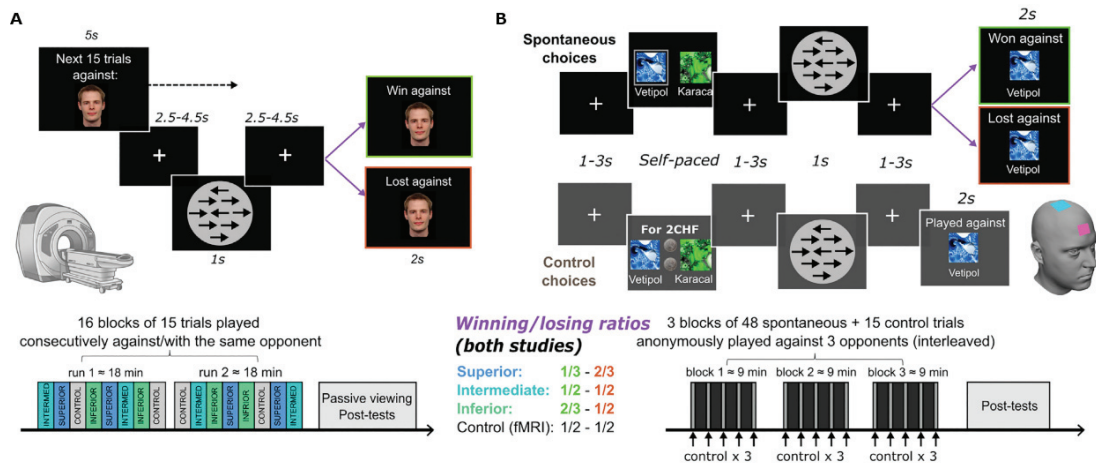
**Figure 1.21** – Résumé des résultats de Kumaran et al. (2016). **(A,B,C,D)** Analyse des images IRMf basée sur des modèles. **(A)** Activité neurale corrélant avec la valeur absolue de la différence de pouvoir entre les deux stimuli calculés par le modèle durant la phase d'apprentissage. Le pouvoir de chaque stimulus calculé par le modèle est à gauche. Les activations (à droite) de l'amygdale bilatérale, de l'hippocampe bilatéral du cingulaire postérieur ainsi qu'une région proche du fusiforme sont significatives lorsqu'elles sont corrigées au niveau du groupe. **(B)** Corrélation entre l'activité neural du mPFC et le pouvoir du stimuli sélectionné, plus forte durant l'apprentissage d'une hiérarchie impliquant soi-même qu'une hiérarchie impliquant un proche. Corrigé pour des petits volumes. **(C)** Phase de catégorisation : corrélation linéaire entre l'activité neurale de l'amygdale et de l'hippocampe antérieure et le rang des stimuli. Durant cette phase, les participants doivent dire à quelle hiérarchie (impliquant soi-même ou impliquant un proche) le stimulus appartient. L'activation est corrigée pour des petits volumes. Le graphique montre l'estimation des paramètres extraient du pique d'activation pour chaque stimulus et illustre la corrélation linéaire entre activation et rang. **(D)** Activité du mPFC corrélant linéairement avec le rang du stimulus observé durant la phase de catégorisation de la hiérarchie impliquant soi-même. Les graphiques illustrent la corrélation pour la condition impliquant soi-même et la non corrélation impliquant un proche (estimation des paramètres dans une région d'intérêt centrée sur l'activation dans le mPFC). Le rang 5 n'apparaît pas, car il représente soi, ou son proche en fonction des conditions. *Adapté de Kumaran et al. (2016)*

Les bases neurales d'apprentissage de la hiérarchie par observation sont donc assez bien connues, mais qu'en est-il pour un apprentissage impliquant une interaction directe

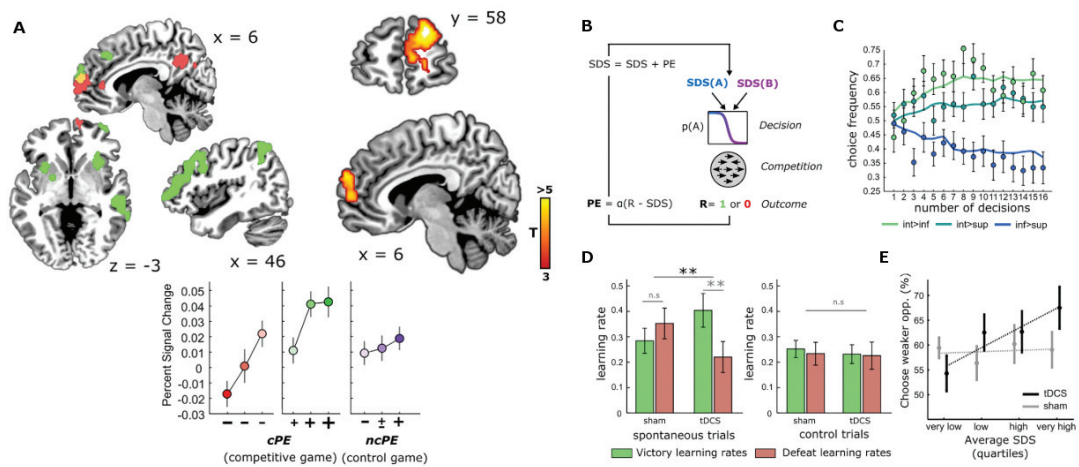
? Afin d'étudier l'apprentissage de l'attribution de capacité ou de prestige par compétition dyadique directe, Ligneul et al. (2016) ont utilisé un paradigme de compétition en duels répétés lors d'une tâche perceptive. Lors de ce jeu, les participants rencontraient quatre adversaires différents (Le quatrième étant un contrôle non-social de la tâche de perception). Dans chaque bloc, les adversaires étaient en fait non réels mais programmés pour être, soit supérieur (il gagnait deux tiers du temps), soit intermédiaire (50% de victoire), soit inférieur (il gagnait un tiers du temps). Dans la tâche contrôle, le taux de victoire était programmé à 50%. Dans chaque bloc, les participants rencontraient 2 fois chaque adversaire pour 15 essais consécutifs. L'ordre des adversaires était pseudo-randomisé (voir figure 1.22). Il a été démontré dans cette expérience que l'erreur de prédiction positive et négative lors d'une compétition n'est pas encodée dans les mêmes régions. Un chevauchement a cependant lieu dans le rmPFC qui encode les deux types d'erreur de prédiction, mais pas dans le même sens. Les prédictions d'erreur positives et négatives impliquent un changement de direction opposé dans le signal BOLD (négatif pour les prédictions erreurs négatives et positives pour les prédictions erreur positives, voir figure 1.23). Sachant que le signal BOLD de cette région corrèle avec l'erreur de prédiction, il peut être intéressant de vérifier la causalité<sup>35</sup> de l'implication de cette région dans l'apprentissage d'une hiérarchie par interaction dyadique. Ainsi, l'expérience a été reproduite (avec une variation mineure sur le nombre de blocs) avec comme facteur d'étude la stimulation transcrânienne à courant anodal direct ciblant le rmPFC. Les résultats confirment l'implication causale de cette région en augmentant la vitesse d'apprentissage par les victoires et la diminution de l'apprentissage par les défaites. De plus, il a été montré que la stimulation du rmPFC augmente l'influence de leur propre statut de dominance sur le choix des adversaires.

---

<sup>35</sup>En effet, une corrélation n'est pas une causalité.



**Figure 1.22** – Déroulement de l'expérience IRMf d'apprentissage d'une hiérarchie par interaction dyadique. **(A)** Durant 15 essais le participant joue contre le même opposant en compétition (ou dans la tâche contrôle). L'objectif était de déterminer le sens majoritaire vers lequel pointait les 46 flèches présentées au participant. Il était dit au participant que le premier à répondre correctement gagnait, cependant chaque opposant était associé à un taux de victoire (un tiers, la moitié ou deux tiers). **(B)** Déroulement de l'expérience en stimulation transcrânienne à courant direct. Les participants effectuaient une tâche similaire. La présentation des opposants était néanmoins différente. Ils leur étaient attribués un nom et une image. Le participant devait sélectionner l'individu contre lequel il souhaitait jouer (parmi deux joueurs proposés). Une condition contrôle permettait de distinguer les choix orientés dominance et les choix orientés récompense. La moitié des participants étaient réellement stimulés électriquement sur le rmPFC, les autres recevaient un simulacre de stimulation. *Adapté de Ligneul et al. (2016)*



**Figure 1.23** – Résultats of Ligneul et al. (2016). **(A)** Encodage de l’erreur de prédiction compétitive. En vert figure l’erreur de prédiction compétitive positive, en rouge la négative et en jaune la superposition des deux. Les activations passent la correction au niveau du groupe. À droite, figure la carte statistique pour la seule région encodant les erreurs de prédiction positive et négative. En bas, figure le pourcentage de changement de signal dans la région montrée en haut à droite en fonction de la taille et de la valence de l’erreur de prédiction. (cPE = erreur de prédiction compétitive, ncPE= erreur de prédiction non-compétitive). **(B)** Schémas du fonctionnement de l’apprentissage par renforcement. **(C)** Choix effectivement observés (points) et prédictions du modèle (traits pleins). **(D)** La stimulation transcrânienne module sélectivement la vitesse d’apprentissage dans la condition de compétition. Elle augmente la vitesse d’apprentissage à la suite d’une victoire et la diminue à la suite d’une défaite (\*\*  $p < 0.01$ ). **(E)** La stimulation transcrânienne module la probabilité de choisir l’opposant le plus faible en fonction du statut de dominance sociale du participant estimé à chaque essai. Sous stimulation, les participants alternent entre des périodes durant lesquelles ils challengent l’opposant le plus fort malgré plus de défaites (Statut de Dominance Social moyen bas) et de périodes où ils affirment leur statut en défiant l’opposant le plus faible (Statut de Dominance Social moyen haut) *Adapté de Ligneul et al. (2016)*

Ainsi, les processus engagés lors de l’apprentissage des capacités ou du rang social d’autrui diffèrent en fonction de la manière dont les connaissances sont apprises (par expérience directe ou par observation). Ce qui est possiblement en partie dû au fait qu’en interaction directe le statut social du participant soit directement mis en jeu.

Il est nécessaire pour bien naviguer dans un milieu social de connaître les compétences et capacités d’autrui, mais il est aussi nécessaire de leur attribuer des intentions ainsi que de prédire leurs décisions, leurs croyances et leurs états mentaux, c’est la théorie de l’esprit.

#### 1.2.4 Théorie de l'esprit

Comme vu précédemment, la capacité d'attribuer à autrui des états mentaux, des croyances, des préférences et des intentions s'appelle la théorie de l'esprit. Il est difficile de dire à quel âge commence la théorie de l'esprit et quelles espèces en "disposent" ou non tant ce champ théorique est vaste. Il serait plus adéquat de parler de profondeur ou de niveau dans la théorie de l'esprit. Elle contient notamment un pan lié aux émotions avec par exemple la compassion ou l'empathie, mais aussi un pan lié à la raison qui consiste à raisonner en faisant la distinction entre soi et l'autre.

Ainsi, dès le plus jeune âge, les enfants Homo sapiens parviennent à attribuer des préférences aux autres différentes des leurs, puis, plus tard, ils parviennent à leur attribuer des croyances différentes des leurs (Baillargeon et al., 2010). Ensuite, en combinant les désirs et croyances des autres, les enfants sont capables de différencier l'intentionnalité de l'accident (Meltzoff, 1995). La théorie de l'esprit n'est pas une spécificité de l'Homo sapiens, en effet, elle existe aussi chez de nombreux animaux non-humains, que ce soit au sein d'une même espèce (Horowitz, 2009; Bugnyar et al., 2016) ou entre espèces (Call and Tomasello, 1998). L'étude de la théorie de l'esprit a fait apparaître l'hypothèse selon laquelle elle émergerait de deux capacités : celle d'inhiber sa propre perspective ainsi que celle de raisonner sur les croyances des autres (Perner et al., 2002)

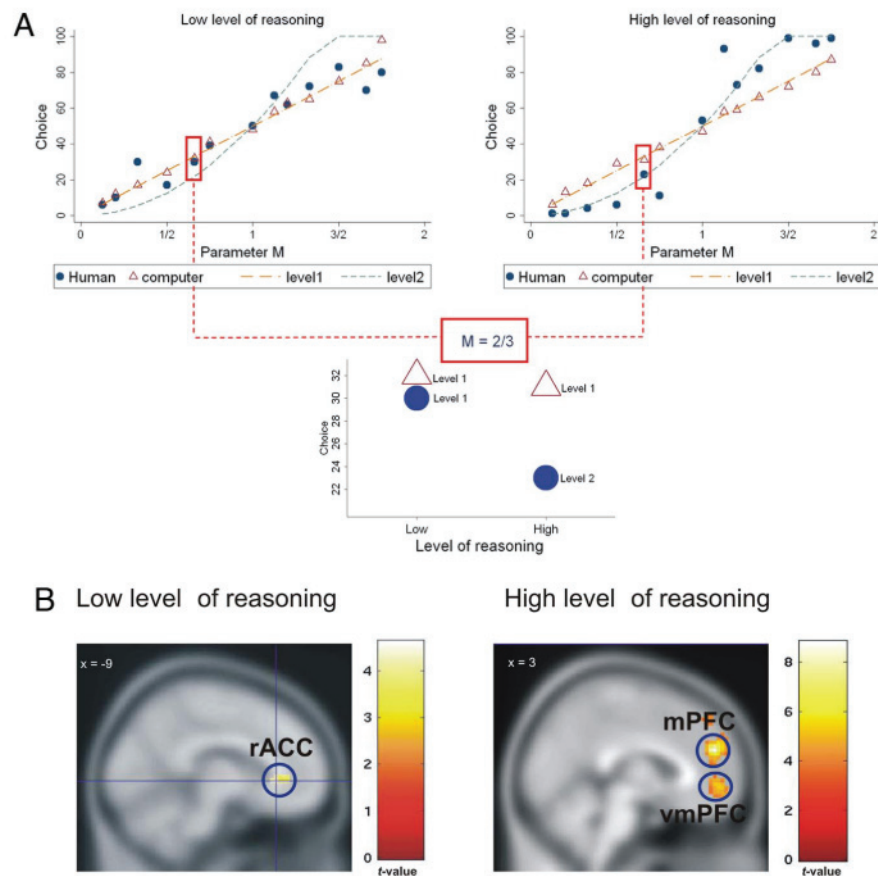
La profondeur de raisonnement est souvent utilisée comme mesure dans le cadre de la théorie des jeux. Cette mesure représente bien selon moi le côté rationnel de la théorie de l'esprit, mais omet radicalement le côté émotionnel. De plus, la mesure de la théorie de l'esprit ne prend souvent en compte que l'action qui résulte du raisonnement et non le raisonnement lui-même (car souvent inaccessible) qui peut être plus "profond" que l'action qui en résulte. Un des jeux les plus classiques pour étudier cette profondeur de raisonnement s'appelle le jeu du concours de beauté imaginé par Keynes, oui, celui connu pour ses analyses économiques. Il a remarqué que sur le marché boursier ce qui compte n'est pas la valeur intrinsèque des actions, mais bien ce que pensent les autres de ces actions. Il a donc fait l'analogie avec le jeu d'un magasin de l'époque où figuraient

100 femmes et dont le but était de sélectionner les 6 plus belles femmes. Le gagnant étant celui dont la sélection se rapproche le plus des photographies les plus choisies par l'ensemble des joueurs. Le but était alors non pas de sélectionner les plus belles femmes, mais bien de deviner ce qu'allait choisir les autres pour se rapprocher le plus possible du consensus. Ainsi, Keynes voulait montrer que ceux qui réussissent en bourse sont en fait ceux qui parviennent à anticiper la psychologie des foules. Ainsi, il en parle en ces termes (Keynes, 1942) : *" La technique du placement peut être comparée à ces concours organisés par les journaux où les participants ont à choisir les six plus jolis visages parmi une centaine de photographies, le prix étant attribué à celui dont les préférences s'approchent le plus de la sélection moyenne opérée par l'ensemble des concurrents. Chaque concurrent doit donc choisir non pas les visages qu'il juge lui-même les plus jolis, mais ceux qu'il estime les plus propres à obtenir le suffrage des autres concurrents, lesquels examinent tous le problème sous le même angle. Il ne s'agit pas pour chacun de choisir les visages qui, autant que chacun peut en juger, sont réellement les plus jolis ni même ceux que l'opinion moyenne considérera réellement comme tels. Au troisième degré où nous sommes déjà rendus, on emploie ses facultés à découvrir l'idée que l'opinion moyenne se fera à l'avance de son propre jugement. Et il y a des personnes, croyons-nous, qui vont jusqu'au quatrième ou au cinquième degré ou plus loin encore."* Il a donc été défini un équivalent mathématique avec un équilibre de Nash qui consiste à choisir un nombre entre 1 et 100. Dans cette version, le gagnant est celui qui trouve le chiffre le plus proche du nombre moyen choisi par les autres joueurs multiplié par un paramètre  $M$  choisi par l'expérimentateur. Ainsi si  $M = 1$  on se retrouve dans le cas qu'avait défini Keynes, si  $M < 1$ , par exemple  $M = \frac{2}{3}$  et que les participants ont choisi une moyenne de 60, le gagnant est celui qui aura choisi le nombre le plus proche de  $60 * \frac{2}{3}$  soit 40. En raisonnant de manière récursive un participant peut se dire que si tout le monde raisonne comme lui, il faudra qu'il choisisse un nombre proche de  $40 * \frac{2}{3}$  soit 27 par exemple. Ainsi, le raisonnement a été de calculer  $60 * \frac{2}{3} * \frac{2}{3}$  soit  $60 * (\frac{2}{3})^2$ . Ainsi, il est facile de calculer une profondeur de raisonnement en regardant combien de fois environ le participant a multiplié par le nombre  $M$ .

Cependant, comme je l'expliquais précédemment, ce qui est mesuré ici est plutôt la profondeur de raisonnement moyenne que le participant attribut au reste du groupe augmenté d'une itération de raisonnement supplémentaire. Cette mesure ne représente pas forcément directement la profondeur de raisonnement du participant lui-même, car il est impossible de différencier l'action du raisonnement.

Quoi qu'il en soit cela n'enlève rien aux résultats de Coricelli and Nagel (2009) qui ont demandé à des participants de faire cette expérience dans une IRM fonctionnelle. Ils ont trouvé qu'un raisonnement plus profond était associé à une activation mPFC/vmPFC alors qu'un raisonnement moins profond était associé à une activation du rACC (voir figure 1.24).


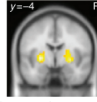

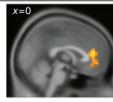

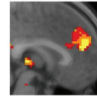




**Figure 1.24** – Comportement et activité cérébrale pour les basses et hautes profondeurs de théorie de l’esprit. **(A)** Résultat illustratif de deux participants pour tous les paramètres  $M$  contre un ordinateur (rouge) et contre un confédéré (bleu). Le participant de gauche représente les bas niveaux de raisonnement, celui de droite les hauts niveaux. La ligne rouge représente le modèle pour un raisonnement de profondeur 1 :  $50 * M$ , la ligne bleue un raisonnement de profondeur 2 :  $50 * M * M$ . Au total 7 participant ont été classifiés dans la catégorie basse profondeur et 7 dans la catégorie haute profondeur, 3 ont joué aléatoirement. **(B)** Résultats IRMf. Différence entre choisir un nombre contre un humain et choisir un nombre contre un ordinateur. Effet pour le groupe de basse profondeur d’esprit (gauche) et haute profondeur d’esprit (droite). Les activations passent la correction au groupe, mais sont affichées avec un seuil  $p < 0.001$ . *Extrait de Coricelli and Nagel (2009)*

Ces résultats ont été confirmés par Zhu et al. (2012) et Hampton et al. (2008) dans des tâches plus subtiles et dont les analyses des données IRMf étaient basées sur des modèles. Ces expériences supplémentaires ont permis de montrer l’implication du rACC dans les inférences de bas niveau et le mPFC pour les inférences d’un niveau supérieur.

Ainsi Griessinger and Coricelli (2015) ont résumé les bases neurales des comportements de plus en plus profonds en terme de théorie de l'esprit (voir figure 1.25).

| Learning mechanisms   | Levels of strategic thinking   | Neural correlates  |
|---|--|--|
| <b>Reinforcement learning (RL)</b><br>$V_{t+1}^a = V_t^a + \eta \delta_t$<br>(Sutton & Barto, 1998)                     | <b>Level zero k=0</b><br> | Striatal activity:<br>$\delta(t) = r(t) - V_a(t)$<br>RL Prediction error<br><br>(Zhu et al, 2012) |
| <b>Fictitious Play learning (FL)</b><br>$P_{t+1}^* = P_t^* + \eta \delta_t^p$<br>(Fudenberg et al, 1998)                | <b>Low level k=1</b><br>  | rACC<br>$\delta_t^p = P_t - P_t^*$<br>Belief Prediction error<br><br>(Zhu et al, 2012)            |
| <b>Influence learning (IL)</b><br>$P_{t+1}^* = P_t^* + \eta_1 \delta_t^p + \eta_2 \lambda_t^p$<br>(Hampton et al, 2008) | <b>High level k=2</b><br> | mPFC<br><br>(Hampton et al, 2008)   |

**Figure 1.25** – Penser et apprendre : calculs et corrélats neuronaux des différentes stratégies de penser et d'apprentissage. Le niveau 0 de profondeur d'esprit peut être associé à un algorithme classique d'apprentissage par renforcement. Le niveau 1 peut être associé à un algorithme d'apprentissage de jeu fictif. Enfin, le niveau deux peut être associé à un algorithme appelé apprentissage d'influence qui prend en compte comment l'action de l'autre est influencé par les nôtres. *Extrait de (Griessinger and Coricelli, 2015)*

De prime abord, on pourrait penser qu'avoir une profondeur d'esprit plus élevée conférerait un avantage évolutionnaire et donc la pression sélective devrait favoriser ce genre de phénotype. Pourtant, il a été montré que les humains ne sont pas si doués que l'on pourrait penser concernant la prédiction des actions des autres ou leurs intentions (Hedden and Zhang, 2002). Plusieurs explications sont possibles, notamment le coût énergétique impliqué par des calculs plus complexes. Pourtant, une autre hypothèse basée sur la théorie des jeux et l'analyse de modèles montre qu'avoir une trop grande profondeur de raisonnement par rapport aux autres est non seulement coûteux, mais n'apporte pas d'avantage évolutif lorsque l'on évolue au sein d'un groupe avec des profondeurs de raisonnement plus basses (Devaine et al., 2014b). En effet, le coût informationnel ne permet pas aux phénotypes de haute sophistication de complètement exploiter ceux de moins haute sophistication (en mode compétitif). De plus, dans la nature, les modes d'interactions ne sont pas uniquement compétitifs mais aussi

largement coopératifs. Cette même publication montre que la coopération favorise elle des niveaux de raisonnement plus bas permettant d'être plus prévisible (Devaine et al., 2014b).

### 1.3 La modélisation

Le cerveau est un organe très complexe interagissant avec de nombreux autres organes, produisant des réponses comportementales (ou émotionnelles) en réaction à des stimuli extérieurs ou intérieurs (des autres organes). Le principe de la modélisation est de mettre en équation les stimuli et les comportements afin de trouver les régularités qui les lient entre eux et qui restent vraies <sup>36</sup> au sein d'une population. Ainsi, la modélisation consiste à éliminer des sources de variations (perdre un peu d'information) au début pour décrire les principaux comportements puis complexifier (ou simplifier/modifier) pour expliquer de plus en plus de variance dans les données. J'apporte une grande importance à préciser que l'utilisation de modèles a de nombreux avantages, mais qu'elle induit un énorme biais théorique. J'entends par là que bien qu'un modèle semble fonctionner avec nos axiomes mathématiques rien ne nous dit qu'un autre ne pourrait pas fonctionner avec les mêmes axiomes, voire avec d'autres. Je me permets de citer Albert Einstein et Léopold Infeld pour expliciter ma pensée grâce à cette analogie sur la vérité objective qu'ils ont rendu célèbre : *"C'est en réalité tout notre système de conjectures qui doit être prouvé ou réfuté par l'expérience. Aucune de ces suppositions ne peut être isolée pour être examinées séparément. Dans le cas des planètes qui se meuvent autour du soleil, on trouve que le système de la mécanique est remarquablement opérant. Nous pouvons néanmoins imaginer un autre système, basé sur des suppositions différentes, qui soit opérant au même degré. Les concepts physiques sont des créations libres de l'esprit humain et ne sont pas, comme on*

---

<sup>36</sup>J'aimerais faire un aparté sur le vrai et le faux. Aujourd'hui, l'ensemble de la communauté fait souvent l'amalgame entre quelque chose de significatif au sens probabiliste du terme et quelque chose de vrai. Ce que j'entends par ici, c'est que d'abord, il n'est pas exclu d'avoir à faire à un artefact dans les probabilités, mais aussi qu'aujourd'hui la science ne repose plus sur un raisonnement booléen, c'est-à-dire vrai ou faux, mais sur un dégradé de plausibilité. Nous y reviendrons quand nous nous intéresserons aux modèles Bayesian. Je tiens aussi à préciser que la significativité est une chose, la taille d'effet en est une autre et ne doit pas être négligée non plus.

*pourrait le croire, uniquement déterminés par le monde extérieur. Dans l'effort que nous faisons pour comprendre le monde, nous ressemblons quelque peu à l'homme qui essaie de comprendre le mécanisme d'une montre fermée. Il voit le cadran et les aiguilles en mouvement, il entend le tic-tac, mais il n'a aucun moyen d'ouvrir le boîtier. S'il est ingénieux il pourra se former quelques images du mécanisme, qu'il rendra responsable de tout ce qu'il observe, mais il ne sera jamais sûr que son image soit la seule capable d'expliquer ses observations. Il ne sera jamais en état de comparer son image avec le mécanisme réel, et il ne peut même pas se représenter la possibilité ou la signification d'une telle comparaison. Mais le chercheur croit certainement qu'à mesure que ses connaissances s'accroîtront, son image de la réalité deviendra de plus en plus simple et expliquera des domaines de plus en plus étendus de ses impressions sensibles. Il pourra aussi croire à l'existence d'une limite idéale de la connaissance que l'esprit humain peut atteindre. Il pourra appeler cette limite idéale la vérité objective.".* Ainsi pour nous neuroscientifiques, une idée plaisante serait de trouver un modèle biologiquement plausible expliquant les observations allant des activations neuronales jusqu'aux comportements, voire jusqu'à la sociologie.

Ceci étant dit, nous pouvons nous pencher sur les utilisations possibles de la modélisation en neurosciences.

### **1.3.1 Qu'est ce qu'un modèle**

Nous définirons un modèle comme étant une équation (ou une série d'équations) qui transforme des entrées disponibles en sorties mesurables, les actions d'un participant par exemple. Prenons un exemple simple pour comprendre en quoi consiste un modèle d'apprentissage. On peut définir un modèle d'apprentissage de la valeur  $V$  d'une option associée à une récompense  $R$ , comme la différence entre la valeur à l'instant précédent la récompense ( $t$ ) et l'instant suivant la récompense ( $t + 1$ ), ainsi  $V_{t+1} - V_t = R$ . Cependant, comme nous l'avons vu dans la partie sur la dopamine et le système de récompense, ce n'est pas ce que nous observons dans les décharges dopaminergiques. En effet, quand on attend une récompense et qu'elle ne vient pas, il y a une chute des décharges des neurones dopaminergiques. Or dans l'équation précédente si  $R = 0$  la valeur n'est pas

modifiée à  $t+1$ . Le modèle n'est donc pas suffisamment bon. L'amélioration la plus courante consiste à utiliser la différence entre la récompense et l'attente qu'il y avait de cette récompense :  $V_{t+1} = V_t + (R - V_t)$ . Ainsi, lorsqu'il y a une attente, mais pas de récompense ( $R = 0$ ) la valeur diminue. Cependant, avec cette équation, en l'absence de récompense, la valeur de l'option est immédiatement remise à zéro, ce qui est un apprentissage trop rapide par rapport aux données observées. Ainsi, pour prendre en compte des comportements différents entre individus, comme la vitesse d'apprentissage, les modèles utilisent souvent ce que l'on appelle des "paramètres libres" afin de décrire la variabilité interindividuelle dans les données mesurables. Dans cette équation, la vitesse d'apprentissage est modélisée par le paramètre libre  $\alpha$ . Ainsi, l'équation que l'on appelle équation d'apprentissage par renforcement prend la forme de  $V_{t+1} = V_t + \alpha * (R - V_t)$ .  $(R - V_t)$  représente la différence entre la valeur attendue et la récompense, on l'appelle *erreur de prédiction*. Les paramètres libres sont ainsi "ajustés" pour reproduire au mieux les comportements des participants. L'ajustement de ces paramètres pour optimiser la reproduction des données observées par le modèle est par ailleurs un des enjeux principaux de la modélisation.

Les modèles ne dépendent pas toujours du temps comme précédemment. Par exemple dans notre publication <sup>37</sup> sur la corruption que vous trouverez en appendis, nous avons utilisé un modèle de Fehr-Schmidt, qui porte aussi le nom de modèle d'aversion à l'iniquité. Ce modèle attribue une valeur  $V$  à une option qui affecte un gain  $x_i$  à soi-même et un gain  $x_j$  à un autre individu. L'hypothèse est faite que dans cette tâche il n'y a pas d'apprentissage et que la valeur ne dépend pas du temps, mais uniquement de l'aversion à l'inégalité favorable ou celle défavorable. Ainsi, le modèle s'écrit  $V(x_i, x_j) = x_i - \alpha * \max(x_j - x_i, 0) - \beta * \max(x_i - x_j, 0)$ .  $\alpha$  et  $\beta$  sont les paramètres libres du modèle représentant respectivement l'aversion à l'inégalité favorable et celle défavorable.

Ensuite, une fois que nous avons attribué des valeurs à des actions, il faut transformer ces valeurs en probabilité d'action. Pour ceci, le plus courant est d'utiliser une

---

<sup>37</sup>Voir Appendis 1

fonction que l'on appelle *softmax*. Elle permet de transformer la valeur d'une action  $i$  en probabilité d'émettre cette action  $i$  parmi  $n$  actions.

$$\text{Softmax} : P(V_i) = \frac{e^{V_i}}{\sum_{k=0}^n e^{V_k}}$$

Il est aussi possible d'ajouter des paramètres libres dans cette opération. L'un est souvent appelé température inverse ( $\beta$ )<sup>38</sup> et permet de modéliser à quel point le participant est déterministe ou stochastique. L'autre paramètre souvent utilisé est le biais ( $C$ ) et représente la tendance à choisir davantage une action qu'une autre. L'équation devient alors :

$$\text{Softmax} : P(V_i) = \frac{e^{-\beta \cdot V_i + C}}{\sum_{k=0}^n e^{-\beta \cdot V_k + C}}$$

Avec les années de recherche, les chercheurs se sont aperçus que certains comportements étaient mieux décrits en utilisant une distribution de probabilité plutôt qu'une seule et unique valeur pour chaque option ou stimulus. Maintenir une distribution de probabilité permet de mieux décrire la manière dont l'Homo sapiens gère l'incertitude ainsi que la flexibilité que l'on observe dans ses comportements. Nous verrons dans le paragraphe dédié au cas bayésien que les modèles qui utilisent des distributions de probabilité sont souvent des modèles où l'opération centrale est une multiplication et non une addition comme dans les modèles présentés ci-dessus en exemple.

### 1.3.2 Utilisation de la modélisation en neurosciences

Plusieurs méthodologies différentes font appel à l'utilisation de la modélisation ou au contraire, la modélisation peut être utilisée suivant plusieurs méthodes. Afin de bien comprendre, il est important de rappeler la différence entre un raisonnement inductif et un raisonnement déductif. Dans un raisonnement déductif, des propriétés sont énoncées comme découlant d'axiome et de théorème, ainsi aucun nouveau résultat n'est produit, mais une propriété potentiellement vérifiable est émise. Tandis qu'un raisonnement inductif, utilise des observations souvent considérées comme vraies pour émettre une

<sup>38</sup>Il porte ce nom en rappel de l'origine de cette forme d'exponentielle qui est solution d'une équation différentielle. Cette équation vient de l'équilibre thermodynamique d'un système et pour lequel  $\beta = \frac{1}{kT}$  avec  $k$  la constante de Boltzmann et  $T$  la température. C'est pour ceci que  $\beta$  a gardé le nom de température inverse.

ou des hypothèses plausibles. Ainsi, la déduction nous donne des conclusions comme "Si A est vrai, B et C sont vrais" alors que l'induction donne des conclusions comme "Si B et C sont vrais, A est (fortement) plausible". Un raisonnement déductif ne peut être utilisé que dans un système axiomatique. Son utilisation en neurosciences permet principalement la construction d'axiomes plausibles ou réfutables. Quoi qu'il en soit, bien que l'utilisation de la modélisation puisse être très variée, 4 usages dominent la littérature : La simulation de données, l'estimation de paramètres libres, la comparaison de modèles et la production de variables latentes (Wilson and Collins, 2019). Plusieurs de ces analyses peuvent être faites sur les mêmes données ou le même paradigme.

**La comparaison de modèle** est la principale analyse de type inductif. En effet, lors d'inférences prédictives, les modèles permettent d'expliquer les données à posteriori, une fois celles-ci déjà observées. Plus précisément nous regardons la capacité des différents modèles de prévoir le comportement à un essai donné connaissant les essais précédents. En pratique, le chercheur définit un certain nombre de modèles (à priori) capable d'expliquer le comportement dans une tâche donnée, puis il récolte le comportement réel des participants (ce sont les observations considérées comme vraies), enfin, il compare la capacité de chacun de ses modèles à reproduire les données observées. Ainsi, on dit du modèle le mieux capable d'expliquer <sup>39</sup> les données observées qu'il est le modèle le plus plausible. Nous allons étudier un petit exemple qui nous permettra de nous familiariser avec l'utilisation des modèles. Ainsi, par exemple Devaine et al. (2014a) test dans une tâche simple si les participants ont un comportement faisant appel à de la théorie de l'esprit ou non ainsi que si les participants utilisent un raisonnement Bayesian ou non. Pour ce faire, elle a choisi une tâche qui est par ailleurs très simple, elle est parfois appelée "cache-cache". Le participant est face à deux options, si le participant choisit l'option aussi choisie par son opposant, il gagne, l'opposant perd. Si il choisit la mauvaise option, il perd et l'opposant gagne. L'opposant n'est en réalité pas un joueur réel, mais un algorithme.

---

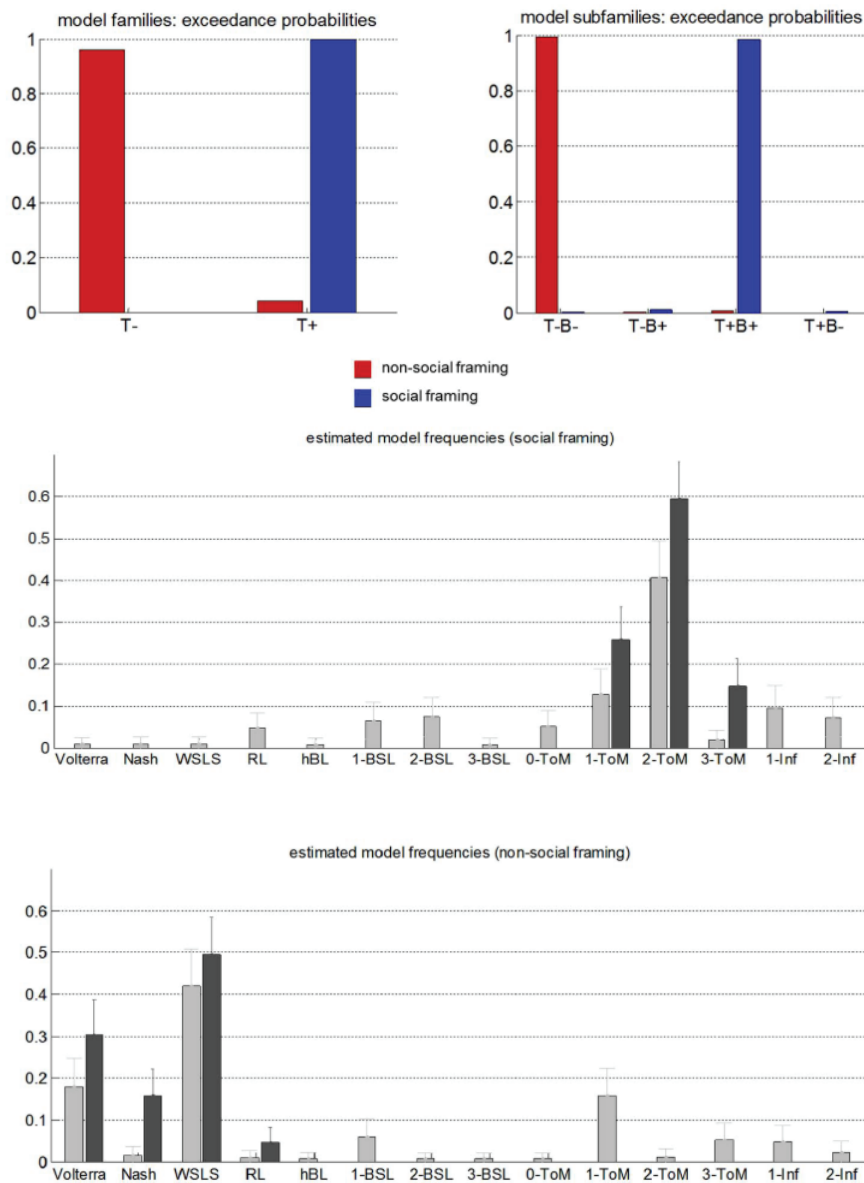
<sup>39</sup>Nous verrons plus loin quelles mesures nous pouvons utiliser pour déterminer quel est le modèle le mieux capable de reproduire des données observées.

|          |        | Joueur 2 |        |
|----------|--------|----------|--------|
|          |        | Gauche   | Droite |
| Joueur 1 | Gauche | \$1(0)   | \$0(1) |
|          | Droite | \$0(1)   | \$1(0) |

**Table 1.3** – Matrice de récompense du jeu du cache cache (Devaine et al., 2014a). Entre parenthèses figurent les récompenses de l’autre joueur, ici l’algorithme.

Le jeu était composé de deux fois soixante essais. La première fois, il leur a été dit que le jeu était social, la deuxième fois qu’il était non social, pourtant, ils jouaient au même jeu contre le même algorithme. Devaine et al. (2014a) ont sélectionné à priori 16 modèles pour tester leurs hypothèses. Les modèles étaient soit bayésiens soit non-bayésiens, mais aussi soit ils mentalisaient les stratégies des autres (théorie de l’esprit) soit non. Ainsi, ils ont optimisé les paramètres libres des modèles afin de maximiser leurs capacités respectives à reproduire les données observées. Ils ont ensuite comparé les modèles entre eux pour estimer la distribution des modèles dans la population. Ils ont ainsi trouvé une différence entre la condition sociale et la condition non-sociale, tant concernant le facteur bayésien que le facteur théorie de l’esprit. En effet, étant donné les observations il est probable qu’en contexte social les humains mentalisent de manière bayésienne alors qu’ils n’utilisent ni raisonnement bayésien ni mentalisation en contexte non-social. Pour le reste, ils trouvent qu’il est probable qu’il existe une distribution entre les différentes profondeurs de mentalisation au sein de la population (voir figure 1.26).





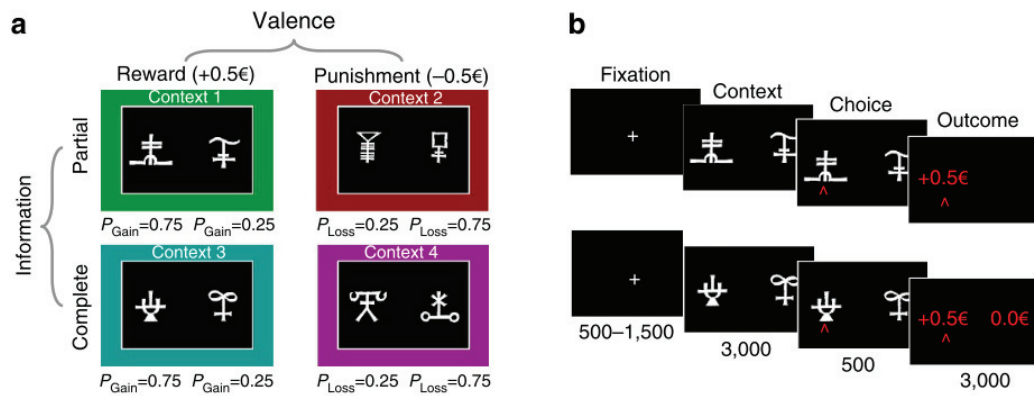
**Figure 1.26** – Comparaison bayésienne de modèles. **(Haut)** Probabilité de dépassement (i.e. probabilité qu’une famille de modèle soit plus fréquente que toutes les autres) des modèles qui ne mentalisent pas (T-) (gauche) en contexte social et non social. Probabilité de dépassement des modèles non bayésiens qui ne mentalisent pas (T-B-), qui mentalisent (T+B-), des modèles bayésiens qui ne mentalisent pas (T-B+) et qui mentalisent (T+B+). **(Milieu)** Distribution de sophistication de la théorie de l’esprit. Estimation de la fréquence des modèles en contexte social. En noir foncé, se trouve l’estimation si on restreint l’analyse à la famille gagnante (T+B+). Les barres d’erreur représentent l’erreur standard postérieure. **(Bas)** Distribution de sophistication de la théorie de l’esprit. Estimation de la fréquence des modèles en contexte non-social. En noir foncé, se trouve l’estimation si on restreint l’analyse à la famille gagnante (T-B-). Les barres d’erreur représentent l’erreur standard postérieure. *Adapté de Devaine et al. (2014a).*

**L'optimisation des paramètres** est une méthode de raisonnement déductif, tout du moins quand l'optimisation des paramètres n'est pas pour faire une comparaison de modèles, mais bien pour comparer les paramètres libres entre groupes ou participants. En effet, elle se base sur une hypothèse selon laquelle le comportement des participants dans une tâche spécifique peut être modélisé avec un même modèle. Ainsi, seuls les paramètres libres peuvent varier et expliquer les différences inter-individuelles. Cette méthode peut être utilisée notamment dans des tâches où un modèle fait consensus pour la description du comportement, après avoir démontrée par une sélection de modèles que l'ensemble de la population peut être décrite par un même modèle ou par exemple pour analyser l'impact d'une condition sur un facteur précis. Elle permet d'analyser l'impact de facteurs indépendants sur le comportement des participants en restreignant les hypothèses (seuls les paramètres libres peuvent être modulés par le facteur indépendant) tout en les rendant mesurables et interprétables. Ainsi, nous avons par exemple montré dans une publication en Appendix que la stimulation transcrânienne à courant continu (tDCS) du dlPFC droit (rdlPFC) augmente l'acceptation de la corruption. Pour mettre en lumière l'origine d'un tel comportement, nous avons utilisé le modèle de Fehr-Schmidt dont nous avons discuté précédemment. Ainsi, nous avons compris que la tDCS sur le rdlPFC diminue l'aversion à l'inégalité favorable, ce qui semble produire le comportement de corruptibilité observé. Ainsi, il est possible de faire des hypothèses sur le rôle du rdlPFC, notamment sur son rôle intégrateur d'information concernant les normes sociales.

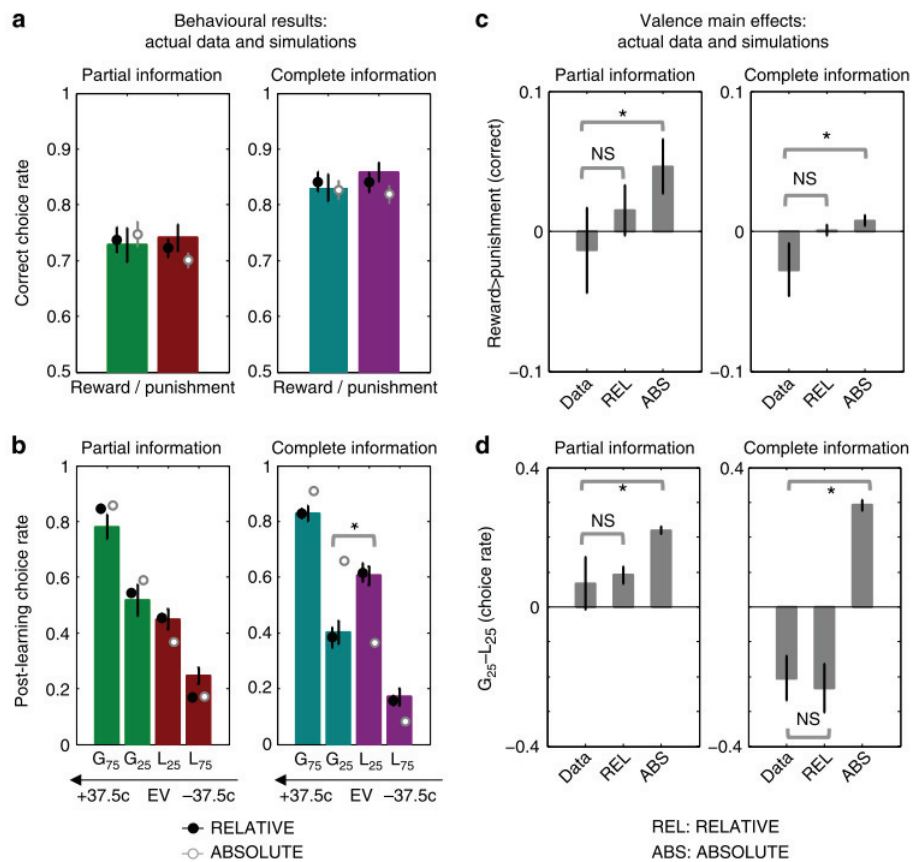
**La génération de données** est aussi un raisonnement déductif. Il permet notamment la falsification. Dans ce type d'analyse, il convient souvent de détecter un comportement spécifique des participants et de sélectionner (à priori ou après une sélection de modèles) un certain nombre de modèles dont on fait l'hypothèse qu'ils peuvent décrire le comportement des participants. Il faut ensuite sélectionner les paramètres libres avec lesquels nous voulons générer des données. Il est possible de les générer aléatoirement depuis une distribution de probabilité construite sur les distributions de paramètres estimés sur les participants (ou sur certains d'entre eux pour lesquels le modèle est suff-

isamment qualitatif). Cette distribution de paramètres est appelée hyper-distribution. Une autre possibilité est d'utiliser des paramètres trouvés dans d'autres expériences qui ont fait l'objet d'une publication. Il est enfin possible de fixer les paramètres à des valeurs qui font sens d'un point de vue théorique. Une fois les données générées (plus le nombre est important plus le résultat sera robuste), il convient d'effectuer exactement la même analyse qui a permis de mettre en évidence le comportement particulier des participants et de vérifier quel modèle est capable de le reproduire.

C'est notamment avec cette méthode que Palminteri et al. (2015) montrent que l'attribution de la valeur à une récompense dépend du contexte dans lequel évolue les participants. Pour appuyer leurs hypothèses, ils ont notamment utilisé la génération de données des modèles, candidats pour prouver leurs capacités ou non à reproduire les comportements observés. Ainsi, leur expérience avait 2 facteurs (voir figure 1.27) : la valence du résultat suite au choix du participant (récompense ou punition) et le contexte informationnel (les participants n'ont le résultat que du choix sélectionné - partiel, ou des choix sélectionnés et non sélectionnés - complets). Ils ont observé que le seul modèle qui parvient à reproduire les comportements est celui qui utilise un signal relatif (comparé à celui utilisant un signal absolu, qui est l'apprentissage par renforcement classique), c'est-à-dire un modèle qui modifie le point auquel le résultat est comparé en fonction du contexte (voir figure 1.28).



**Figure 1.27** – Conception expérimental de Palminteri et al. (2015). **(a)** Tâche d'apprentissage avec une conception à  $2 \times 2$  facteurs. Ainsi, il y a 4 différents contextes : récompense/partiel, punition/partiel, récompense/complet, punition/complet.  $P_{gain}$  = probabilité de gagné 0.5€;  $P_{defaite}$  = probabilité de perdre 0.5€. Les couleurs sont ajoutées à l'image pour des raisons illustratives, mais n'étaient pas présentes dans la tâche originale. **(b)** Écrans successifs d'un essai typique dans la condition récompense/partiel (en haut) et complet (en bas). Les durées sont données en millisecondes. *Adapté de Palminteri et al. (2015).*



**Figure 1.28** – Résultats comportementaux et de la simulation de modèles de Palminteri et al. (2015). **(a)** Choix corrects pendant le test d'apprentissage. **(b)** Taux de choix durant le test d'après-apprentissage.  $G_{75}$  et  $G_{25}$  : options associées respectivement à 75% et 25% de chance de gagner 0.5€.  $L_{75}$  et  $L_{25}$  : options associées respectivement à 75% et 25% de chance de perdre 0.5€. EV : valeur espérée absolue (Probabilité du résultat x Magnitude du résultat dans un seul et même essai. Les valeurs +37.5c et -37.5c correspondent respectivement aux options  $G_{75}$  et  $L_{75}$ . Dans **(a)** et **(b)** les barres de couleurs représentent les données des participants, les points noirs, les données générées par le modèle *RELATIF* et les points blancs, les données générées par le modèle *ABSOLU*. **(c)** Taux de choix correct pendant le test d'apprentissage dans la condition Récompense moins le taux de réponse correcte dans la condition punition. **(d)** Taux de choix de  $G_{25}$  moins  $L_{25}$  pendant la phase d'après-apprentissage. \* $P < 0.05$  t-test sur un échantillon; NS, non-significatif ( $N=28$ ). Les barres d'erreur représentent la somme des écarts à la moyenne. *Adapté de Palminteri et al. (2015).*

**La production de variables latentes**, est souvent une méthode par déduction puisque l'on considère ici que tout les participants sont bien décrits par un même modèle. Elle est souvent effectuée après une sélection de modèles. Les variables cachées, par exemple

la valeur de différentes options, ne sont pas directement observables dans les données comportementales, mais permettent selon la théorie de produire le comportement observé. Il est donc possible de vérifier la présence ou non de telles variables dans les données physiologiques et de réfuter la théorie ou de la renforcer. Ainsi, les variables latentes sont largement utilisées en imagerie fonctionnelle par résonance magnétique (IRMf), en électrophysiologie, ou avec des méthodes de pupilométrie par exemple. Cette méthode peut aussi être inductive, car il est possible de créer des variables latentes depuis plusieurs modèles et de regarder quelles sont celles qui expliquent le mieux les données physiologiques observées. O'Doherty et al. (2007) ont mis en place une méthode pour rechercher les variables latentes dans le signal d'IRMf et ont pu ainsi déterminer que l'erreur de prédiction de la récompense est encodée dans le striatum alors que l'évaluation de la valeur espérée d'une option serait plutôt encodée dans le vmPFC. Cette méthode est puissante et permet des inférences plus précises que l'imagerie classique, mais comme elle est déductive, elle implique de fortes hypothèses, voir d'importants biais dans le choix du modèle qui produit les variables latentes.

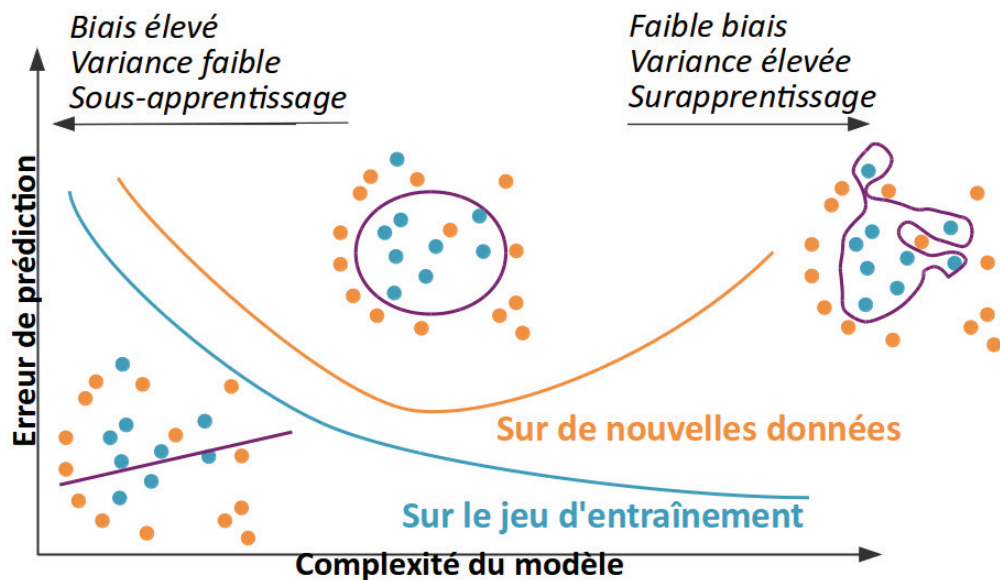
### **1.3.3 Sélectionner parmi les modèles**

Toute sélection rigoureuse implique des critères de sélection ou de diagnostic de l'ajustement des différents modèles aux données observées. Ces critères choisis, il convient de trouver les paramètres libres qui maximisent (ou minimisent en fonction de la mesure utilisée) le critère de sélection choisi. Ensuite, la sélection des modèles pourra se faire au niveau du participant ou du groupe, voir des deux.

#### **1.3.3.1 Les mesures utilisées**

Les mesures qu'il est possible d'utiliser sont nombreuses, parmi elles, il y en a qui prennent en compte le nombre de paramètres libres, d'autre non. Pourtant, pour comparer les modèles entre eux, il est important d'utiliser une mesure qui pénalise les modèles avec plus de paramètres. En effet avec plus de paramètres libres, il est plus facile de

rendre compte du comportement d'un jeu de données <sup>40</sup>, mais l'erreur sera plus grande lors de nouvelles observations. C'est ce que l'on appelle le sur-ajustement. L'utilisation de moins de paramètres libres permet une généralisation plus facile, ce que l'on souhaite en neurosciences : passer d'un groupe d'individus étudié à une population en général (voir figure 1.29).



**Figure 1.29** – Compromis entre complexité du modèle et erreur de prédiction. *Extrait du site de cours en ligne openclassrooms <https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4297218-comprenez-ce-qui-fait-un-bon-modele-d-apprentissage>*

Voici une liste non-exhaustive des métriques utilisables pour évaluer l'ajustement d'un modèle :

<sup>40</sup>Le théorème d'interpolation de Newton nous dit que tous  $k+1$  points peuvent être interpolés par un polynôme de degré  $k$ . Ce qui signifie qu'avec au maximum  $k$  paramètres libres, il est possible de représenter exactement  $k+1$  observations.

- Indépendant du nombre de paramètres :
  - **Similarité.** Elle représente le nombre de choix identique entre le modèle et le participant
  - **Coefficient de détermination ou  $R^2$ .** Il représente le pourcentage de variance des données observées expliqué par le modèle.
  - **Similarité balancé.** Elle a le même but que la similarité mais permet d'éviter qu'un modèle se biaise dans la direction du choix le plus récurant et d'être biaisé dans la qualité de l'ajustement du modèle (Brodersen et al., 2013).
  - **La log-vraisemblance.** Elle est le logarithme de la vraisemblance. La vraisemblance est la probabilité que les données observées soient générées par le modèle étudié. Concrètement, elle consiste en le logarithme du produit de la probabilité qu'a le modèle d'émettre le même choix que le participant chaque essai. Il est donc moins radical que la similarité qui est assimilée aux statistiques booléennes (vrai ou faux) alors que cette mesure est associée aux statistiques Bayésiennes (probabilité que le modèle soit vrai sachant les données).
  - **La validation croisée.** Elle ne prend pas en compte le nombre de paramètres, mais règle le problème de généralisation d'une manière différente. Elle sépare le jeu de données en deux parties. Les paramètres libres sont estimés sur une partie, puis la qualité de l'ajustement du modèle est mesurée sur la partie restante. L'opération est répétée plusieurs fois en changeant le sous-jeu de données sur lequel on teste la qualité de l'ajustement.
  
- Dépendant du nombre de paramètres :
  - **Critère d'information d'Akaike.** Augmente la log-vraisemblance du modèle d'une fois le nombre de paramètres. Le plus proche de 0 est le meilleur.
  - **Critère d'Information Bayésienne ou BIC.** Construit à partir de l'AIC, le BIC pénalise la vraisemblance en fonction du nombre de paramètre multiplié par



le logarithme du nombre d'observations. Ainsi plus les observations sont nombreuses, plus le BIC pénalise les paramètres supplémentaires et plus il est strict comparé à l'AIC.

- **L'énergie libre.** Pénalise la vraisemblance du modèle de la distance <sup>41</sup> entre la densité de probabilité postérieure des paramètres libres du modèle et une densité de probabilité quelconque sur les paramètres du modèle. Cette distance prend en compte la complexité du modèle. Nous y reviendrons quand nous aborderons l'optimisation des paramètres libres.
- **Autre méthodes.** Il existe d'autres méthodes pour corriger la vraisemblance du modèle, comme celles utilisant le gradient Hessien pour prendre en compte la forme de la courbe de vraisemblance ainsi qu'une correction concernant la probabilité d'observer les paramètres libres obtenus.

### 1.3.3.2 Optimisation des paramètres libres

Il convient avant d'optimiser les paramètres libres de choisir leurs domaines de définition (pour des raisons mathématiques ou d'hypothèses sur le comportement). Par exemple, on décide souvent que la température inverse ne peut pas être négative (sauf dans le cas de pathologie ou de comportement spécifique que l'on souhaite capturer). Pour fixer un paramètre positif, on peut par exemple en prendre l'exponentiel, pour le fixer entre 0 et 1 la sigmoïde, ...

L'optimisation des paramètres libres consiste souvent en un compromis entre vitesse d'exécution et précision des résultats. Les trois principales méthodes d'optimisation sont :

**L'optimisation en grille** est la méthode la plus simple et la plus intuitive. Elle consiste à définir un pas pour chaque paramètre et à tester les modèles pour toutes les combinaisons possibles puis de choisir la combinaison qui minimise le critère que l'on a choisi pour optimiser notre modèle. Le principal avantage est qu'il n'est que peu

---

<sup>41</sup>Plus exactement de la divergence de Kullback-Leibler qui mesure à quel point deux distributions de probabilité sont différentes.

probable de trouver un minimum local (et non global) de la fonction à optimiser si le pas des paramètres est suffisamment fin. L'inconvénient est que cette analyse est gourmande en calcul et donc en temps, surtout pour des modèles complexes.

**La méthode de Monte-Carlo par chaînes de Markov** est une méthode d'échantillonnage qui permet à partir d'un nombre suffisant d'échantillon de reconstruire la distribution de probabilités objectives. Pour ce faire, les échantillons subissent des transformations itératives qui ont pour loi stationnaire la distribution de probabilité des paramètres que l'on veut estimer. C'est-à-dire qu'à partir d'un certain nombre d'itérations (i.e. asymptotiquement) les échantillons reproduisent la distribution souhaitée. Plus le nombre d'itérations est grand, plus l'échantillon final est proche de la distribution souhaitée. L'avantage de cette méthode est sa potentielle précision concernant la génération des paramètres libres puisque leur convergence est garantie asymptotiquement. Cependant, il convient que la convergence soit suffisamment rapide puisque cette méthode est très gourmande en calcul et donc en temps. Le nombre d'itérations et le nombre de chaînages (i.e. un ensemble d'itération) sont déterminants pour la qualité de l'analyse. *hBayesDM* est par exemple une boîte à outils utilisable sur le logiciel R pour faire ce genre d'analyse.

**La méthode Bayésienne Variationnelle** est une méthode qui consiste à faire des approximations pour pouvoir calculer analytiquement quels paramètres minimisent l'énergie libre du modèle. Les approximations sont principalement faites concernant la fonction de densité de distribution des paramètres. Souvent, ces approximations ne permettent pas de calculer directement l'énergie libre, mais une borne inférieure de celle-ci. La première est issue de la physique moléculaire, c'est l'approximation des champs moyens (aussi appelée approximation du champ moléculaire) qui consiste à faire l'hypothèse que l'interaction entre les paramètres libres du modèle peuvent se réduire à leurs interactions moyennes. La deuxième est l'approximation de Laplace qui consiste à faire l'hypothèse que la fonction de distribution des paramètres libres peut se résumer avec ses deux moments principaux (moyenne et écart-type) sous forme d'une fonction gaussienne. L'avantage principal de cette méthode est la vitesse de l'analyse. Ces approximations permettent une convergence rapide vers un minimum du critère

d'évaluation utilisé : l'énergie libre. Cependant, les approximations peuvent avoir des effets secondaires, notamment le risque de se retrouver bloqué dans un minimum local. Pour éviter ce désagrément d'autres mesures de diagnostic existent, comme le pourcentage de variance expliqué ou la distribution des résidus. Pour éviter le problème du minimum local et non global, il est aussi possible de varier les a priori sur les paramètres libres ainsi que les variables du modèle et observer l'évolution de l'ajustement. Un autre risque est celui de la sur-confiance concernant les paramètres libres des modèles (i.e. sous-estimer leurs variances). Ceci est dû à l'approximation des champs moyens qui négligent les interactions conditionnelles. Pour éviter de tels problèmes, il convient de définir les paramètres libres d'une manière à les rendre séparables au sens de l'approximation des champs moyens. *VBA toolbox* est par exemple une boîte à outils utilisable sur le logiciel Matlab pour faire ce genre d'analyse.

### 1.3.3.3 Critère de sélection des modèles

Une fois les modèles optimisés selon le critère de notre choix, il est possible de les comparer entre eux et de sélectionner un modèle qui est vraisemblablement le plus représentatif du comportement observé, ou de déduire une distribution des modèles dans la population.

Certaines méthodes, surtout dans les premiers temps de la modélisation en neurosciences utilisent le calcul de la somme des BIC ou de la log-vraisemblance, rendant ainsi compte de la vraisemblance de chaque modèle sur l'ensemble des participants et de la complexité des modèles, mais il donne un poids relativement élevé aux mesures aberrantes comme un BIC très élevé ou très faible chez quelques participants dans le groupe.

Il est aussi possible d'utiliser la méthode fréquentiste sur le critère que l'on choisi pour déterminer quel modèle est le meilleur (i.e. T-test, F-test, correction pour comparaison multiple, test de Wilcoxon ...). Une autre possibilité est la méthode de sélection de modèles bayésienne (BMS) qui permet de traiter la fréquence des modèles dans la population comme une variable aléatoire dont la distribution de probabilité est une

distribution de Dirichlet. Cette distribution représente la fréquence de chaque modèle considéré. Un petit algorithme basé sur les inférences variationnelles et notamment l'approximation des champs moyens permet de faire évoluer la distribution de Dirichlet incrémentalement jusqu'à obtenir une estimation de la distribution des modèles sur la population.

Il peut aussi être intéressant d'effectuer une analyse de confusion pour prouver que le modèle le plus fréquent l'est parce qu'il représente bien les données et non parce que les autres modèles produisent des comportements non différentiables par les méthodes décrites précédemment. Pour ce faire, il s'agit de générer des jeux de données comme nous l'avons vu précédemment à partir de chacun des modèles testés. Ensuite, il faut ajuster chacun des modèles testés sur chaque jeu de données puis faire une comparaison et sélection de modèles. Si les modèles sont bien différentiables, la sélection de modèles doit aboutir sur une forte probabilité pour que le modèle le plus fréquent dans la sélection soit celui qui ait effectivement produit les données.

L'analyse d'identifiabilité est une autre analyse permettant de vérifier la qualité des modèles que l'on utilise. Elle permet cette fois non pas de vérifier que les modèles soient bien identifiables, mais de vérifier que les paramètres libres des modèles sont identifiables. Pour ceci, il convient de générer des données avec un modèle pour une série de paramètres libres choisis, d'estimer les données avec le modèle qui a permis de les générer puis de vérifier (avec une régression linéaire par exemple) que la variance des paramètres estimés est bien expliquée par la variance des paramètres choisis pour générer les données. Précisément, il faut vérifier que la variance de chaque paramètre estimé est bien expliquée par le paramètre générateur correspondant. S'il est expliqué par les autres paramètres générateurs, cela signifie qu'il y a une interaction entre les paramètres et donc qu'ils ne sont pas réellement identifiables. Dans ce cas, il convient de changer le modèle pour les rendre identifiables et d'être prudent sur l'interprétation des paramètres libres estimés.

De nombreuses analyses sont encore possibles sur les modèles comme la comparaison entre conditions ou entre groupes, l'utilisation de réseaux de neurones comme un outil

d'étude du cerveau (Cichy and Kaiser, 2019), les modèles hiérarchiques, les analyses par famille de modèles, les approximations de Laplace, les covariations de paramètres, les modèles dynamiques causaux, ...

### 1.3.4 Que sont les modèles et inférences bayésiennes ?

Nous aborderons ici uniquement les principes fondamentaux des analyses bayésiennes. Pour nous, le raisonnement bayésien intervient à au moins trois niveaux que nous déclinons : Au niveau du cerveau et des neurones, au niveau des modèles qui décrivent le comportement et enfin au niveau des analyses que nous effectuons. Le raisonnement bayésien est issu d'une formule simple :  $P(A|B) = \frac{P(B|A).P(A)}{P(B)}$ . avec  $P(A|B)$  la probabilité conditionnelle de A sachant B, aussi appelée probabilité à posteriori, après avoir confronté les à priori  $P(A)$  à la vraisemblance des données décrites par la probabilité conditionnelle de B sachant A  $P(B|A)$ . La probabilité de B,  $P(B)$ , est la probabilité d'observer les données de manière absolue<sup>42</sup>. Ainsi, la probabilité d'une hypothèse après l'observation des données (à posteriori) est proportionnelle à la probabilité à priori multipliée par la vraisemblance entre les données et l'hypothèse. Un des avantages de ce raisonnement est qu'il reste vrai pour toute probabilité donc aussi pour des densités de probabilités. La méthode bayésienne permet ainsi de manipuler des densités de probabilités et non une valeur scalaire (que ce soit pour les paramètres libres ou les variables latentes) et donc de manipuler une volatilité (ou incertitude) associée à la variable latente ou au paramètre libre. Voici un exemple concret issu du cours au collège de France sur le cerveau Bayésien du professeur et titulaire de la chaire de psychologie cognitive expérimentale Stanislas Dehaene : "Un de vos amis tousse beaucoup, que lui arrive-t-il ?" Plusieurs hypothèses s'offrent à vous, **H1** il a un cancer du poumon, **H2** il a une gastro-entérite, **H3** il a une rhume. L'**hypothèse une** a une forte vraisemblance puisqu'en effet, on tousse lorsque l'on a un cancer du poumon, cependant elle a une probabilité à priori faible puisqu'il y a peu de cancers du poumon dans la population.

---

<sup>42</sup>Concrètement cela se matérialise dans les modèles par une constante qui permet de garder le quotient entre 0 et 1. Dans ce type de raisonnement, comme  $P(B)$  est identique entre toutes les hypothèses, on ne le prend pas en compte dans le raisonnement.

L'**hypothèse deux** a une vraisemblance faible puisqu'on ne tousse pas forcément lorsque l'on a une gastro-entérite, cependant la probabilité à priori de **H2** est élevée puisqu'il est courant d'avoir une gastro-entérite. Concernant l'**hypothèse trois**, il est à la fois courant d'avoir un rhume (probabilité à priori élevée) et les données sont vraisemblables sous l'hypothèse (nous toussons lorsque nous avons un rhume), c'est donc cette hypothèse qui sera validée selon la méthode de raisonnement bayésienne.

Les illusions d'optique sont aussi un bon exemple (tiré du même cours au collègue de France) qui peut souvent être expliqué par un raisonnement Bayésien. En voici une figure 1.30. Vous devriez voir des ronds en volume à gauche et des ronds en creux à droite. Si je fais pivoter l'image de 180° comme dans la figure 1.31 qui suit<sup>43</sup>, vous devriez normalement encore voir sur la figure des ronds en volume à gauche et des ronds en creux à droite. Ce qui ne paraît pas normal puisque si les creux et les bosses n'ont pas changé dans le rectangle, les bosses devraient maintenant se trouver à droite. La théorie bayésienne explique ce phénomène. En effet, votre cerveau utilise ici le dégradé de gris pour donner du volume à l'image (supposant une ombre), cependant, vous ne savez pas d'où vient la lumière puisqu'il y a autant de dégradé dans un sens que dans l'autre sens, ainsi les vraisemblances qu'elle vienne du haut ou du bas sont égales. Or, dans la nature et dans votre quotidien, la lumière vient la majorité du temps du haut, donc la probabilité à priori que la lumière vienne du haut est plus grande que celle qu'elle vienne du bas. Ainsi à priori et vraisemblance combinée, la probabilité à posteriori que la lumière vienne du haut est plus grande, et donc votre cerveau interprète le dégradé du blanc vers le gris comme l'ombre d'une bosse et du gris vers le blanc comme l'ombre dans un trou.

---

<sup>43</sup>Oui oui, je l'ai bien retourné, tourné la page vous-même si vous ne me croyez pas !

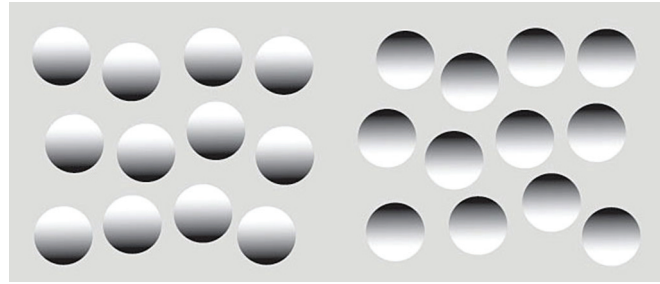


Figure 1.30 – Illusion d'optique 1

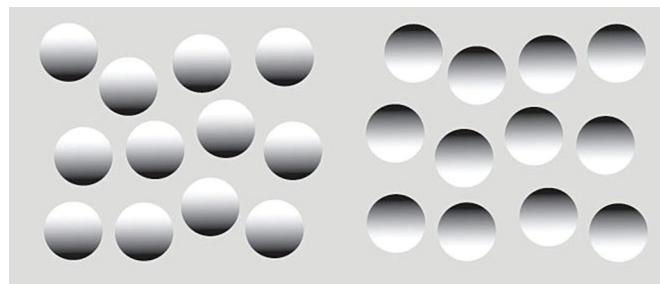


Figure 1.31 – Illusion d'optique 2

Il existe de nombreuses données comportementales et neurologiques permettant d'induire<sup>44</sup> que le comportement humain serait probablement bayésien, c'est-à-dire que le cerveau ne serait pas uniquement un système d'entrée-sortie mais ferait des prédictions et validerait ou non ces prédictions. Karl Friston, notamment, théorise mathématiquement le codage prédictif, qui s'inscrit selon lui dans le cadre de la minimisation de l'énergie libre (Ashburner and Friston, 2005; Friston, 2008, 2010; Friston et al., 2015, 2016, 2017). C'est ainsi selon lui que "les agents biologiques résistent à la tendance naturelle au désordre", "[qu']ils maintiennent leur état dans un environnement changeant ". Un modèle neuronal du codage prédictif a été proposé par Wacongne et al. (2012) et selon lequel les prédictions bayésiennes se propagent des couches neuronales supérieures vers les couches neuronales inférieures tandis que les erreurs de prédiction se propageraient dans l'autre sens. Ainsi, selon ses prédictions, les couches neuronales supérieures supprimeraient le signal attendu pour ne laisser remonter que les signaux d'erreur de prédiction.

<sup>44</sup>Dans le sens défini précédemment comme le contraire de déduire.

L'avantage évolutif des inférences bayésiennes est qu'elles sont optimales au sens où elles minimisent la mesure fréquentiste de l'erreur. En effet, les données étant manipulées en fonction de leur distribution, les inférences bayésiennes permettent d'appréhender les informations en fonction de leur incertitude. Que l'incertitude provienne des paramètres du modèle générateur du monde qui nous entoure ou du modèle générateur lui-même. La théorie bayésienne permet notamment d'expliquer les comportements lors de tâches où le temps est limité (Tavoni et al., 2019). Cependant, les inférences bayésiennes sont gourmandes en ressources de calculs et la puissance de calcul du cerveau est limitée. Pour résoudre ce problème, des chercheurs proposent des méthodes d'approximations des calculs bayésiens par des approches que nous avons abordé comme les méthodes variationnelles ou de Monte-Carlo (Tavoni et al., 2019) qui seraient aussi biologiquement probable.

Cependant, il convient de faire attention quand nous utilisons des méthodes bayésiennes ou des modèles bayésiens, car les formes et valeurs des à priori peuvent avoir un impact important sur les résultats. De plus comme toute fonction peut être décrite comme un couple d'une distribution d'à priori et d'une fonction de coût (ou de gain)<sup>45</sup>, il serait impossible de distinguer par l'expérience le modèle bayésien du modèle non bayésien (l'analyse de confusion serait une bonne manière de montrer que les deux modèles ne peuvent être distingués). Ce principe pose la question de la réfutabilité de l'hypothèse du cerveau bayésien qui est nécessaire afin d'en faire une théorie. Pour ceci, il est nécessaire de contraindre l'hypothèse Bayésienne, ce qui a notamment fait l'objet d'une thèse de doctorat à Grenoble (Diard, 2015).

## 1.4 Présentation des résultats obtenus

Le premier chapitre explore comment par l'observation d'interaction sociale dyadique, une personne est capable d'apprendre une hiérarchie complexe. Nous étudions dans ce chapitre les bases neurales liées à l'exploration d'une telle hiérarchie de 5 niveaux

---

<sup>45</sup>fonction d'optimisation que l'on cherche à minimiser (coût) ou maximiser (gain)



avec plusieurs individus par niveau. Ainsi, nous étudions par la même, comment les personnes font des inférences transitives concernant les rangs hiérarchiques de deux individus dont ils n'ont pas observé d'interaction dyadique direct.

Le deuxième chapitre étudie comment grâce à la stimulation transcrânienne, il est possible de vérifier les hypothèses obtenues grâce à l'imagerie par résonance magnétique fonctionnelle et comment nous pouvons moduler sélectivement l'apprentissage d'une hiérarchie sociale.

Dans le troisième chapitre, nous verrons comment il est possible d'apprendre une hiérarchie sociale et non-sociale par interaction directe. Nous étudierons quel est le rôle computationnel de la sérotonine dans la variabilité des comportements interindividuels dans ce genre d'apprentissage.

Enfin, dans un dernier chapitre, nous verrons comment lors d'interactions sociales minimales, les individus attribuent des intentions compétitives ou coopératives à autrui pour maximiser leurs bénéfices. Nous étudierons les mécanismes computationnels mis en jeu ainsi que leurs bases neurales.

## Bibliography

- Arias-Carrián, O., M. Stamelou, E. Murillo-Rodríguez, M. Menéndez-Gonzlez, and E. Pöppel (2010). Dopaminergic reward system: A short integrative review. *International Archives of Medicine* 3(1), 1–6.
- Ashburner, J. and K. J. Friston (2005). Unified segmentation. *NeuroImage* 26(3), 839–851.
- Baillargeon, R., R. M. Scott, and Z. He (2010). False-belief understanding in infants. *Trends in Cognitive Sciences* 14(3), 110–118.
- Baker, C. L., J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4), 0064.
- Bault, N., M. Joffily, A. Rustichini, and G. Coricelli (2011). Medial prefrontal cortex and striatum mediate the influence of social comparison on the decision process. *Proceedings of the National Academy of Sciences of the United States of America* 108(38), 16044–16049.
- Behrens, T. E., L. T. Hunt, and M. F. Rushworth (2009). The computation of social behavior. *Science* 324(5931), 1160–1164.
- Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences* 11(3), 118–125.
- Bond, C. F. and L. J. Titus (1983). Social facilitation : a meta-analysis of 241 studies . *Psychol Bull Social Facilitation : A Meta-Analysis of 241 Studies. Psychological Bulletin* 94(2), 265–292.
- Boot, N., M. Baas, S. van Gaal, R. Cools, and C. K. De Dreu (2017). Creative cognition and dopaminergic modulation of fronto-striatal networks: Integrative review and research agenda. *Neuroscience and Biobehavioral Reviews* 78, 13–23.

- Brodersen, K. H., J. Daunizeau, C. Mathys, J. R. Chumbley, J. M. Buhmann, and K. E. Stephan (2013). Variational Bayesian mixed-effects inference for classification studies. *NeuroImage* 76(May 2014), 345–361.
- Bugnyar, T., S. A. Reber, and C. Buckner (2016). Ravens attribute visual access to unseen competitors. *Nature Communications* 7, 3–8.
- Burke, C. J., P. N. Tobler, M. Baddeley, and W. Schultz (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences* 107(32), 14431–14436.
- Call, J. and M. Tomasello (1998). Distinguishing Intentional from Accidental Actions in Orangutans (*Pongo pygmaeus*), Chimpanzees (*Pan troglodytes*), and Human Children (*Homo sapiens*). *Journal of Comparative Psychology* 112(2), 192–206.
- Castelli, F., F. Happé, U. Frith, and C. Frith (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* 12(3), 314–325.
- Charpentier, C. J., K. Iigaya, and J. P. O’Doherty (2020). A Neuro-computational Account of Arbitration between Choice Imitation and Goal Emulation during Human Observational Learning. *Neuron* 106(4), 687–699.e7.
- Chi Yiu Yim and G. J. Mogenson (1980). Electrophysiological studies of neurons in the ventral tegmental area of tsai. *Brain Research* 181(2), 301–313.
- Cichy, R. M. and D. Kaiser (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences* 23(4), 305–317.
- Clemson, T. and T. S. Evans (2012). The emergence of leadership in social networks. *Physica A: Statistical Mechanics and its Applications* 391(4), 1434–1444.
- Cools, R. (2011). Dopaminergic control of the striatum for high-level cognition. *Current Opinion in Neurobiology* 21(3), 402–407.

- Coricelli, G. and R. Nagel (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences* 106(23), 9163–9168.
- Damasio, A. R. (1995). Toward a Neurobiology of Emotion and Feeling: Operational Concepts and Hypotheses. *The Neuroscientist* 1(1), 19–25.
- Damasio, Antonio, R. (DL 2003). *Spinoza avait raison : joie et tristesse, le cerveau des émotions / Antonio R. Damasio ; traduit de l'anglais par Jean-Luc Fidel*. Sciences. Paris: Odile Jacob.
- Damasio, Antonio, R. (DL 2010). *L'erreur de Descartes : la raison des émotions / Antonio R. Damasio ; traduit de l'anglais (États-Unis) par Marcel Blanc ([Nouvelle édition] ed.)*. Odile Jacob poches. Paris: Odile Jacob.
- D'Ardenne, K., S. M. McClure, L. E. Nystrom, and J. D. Cohen (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* 319(5867), 1264–1267.
- Devaine, M., G. Hollard, and J. Daunizeau (2014a). The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? *PLoS Computational Biology* 10(12).
- Devaine, M., G. Hollard, and J. Daunizeau (2014b). Theory of mind: Did evolution fool us? *PLoS ONE* 9(2).
- Diard, J. (2015). *Bayesian Algorithmic Modeling in Cognitive Science*. Ph. D. thesis, Université Grenoble Alpes.
- Edwards, W. (1954). The theory of decision making. *Psychological Bulletin* 51(4), 380–417.
- Fareri, D. S. and M. R. Delgado (2014). Differential reward responses during competition against in- and out-of-network others. *Social Cognitive and Affective Neuroscience* 9(4), 412–420.
- Fliessbach, K., B. Weber, P. Trautner, T. Dohmen, U. Sunde, C. E. Elger, and A. Falk (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science* 318(5854), 1305–1308.

- Fox, M. D., A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America* 102(27), 9673–9678.
- Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology* 4(11).
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11(2), 127–138.
- Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O’Doherty, and G. Pezzulo (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews* 68, 862–879.
- Friston, K., F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo (2015). Active inference and epistemic value. *Cognitive Neuroscience* 6(4), 187–224.
- Friston, K. J., R. Rosch, T. Parr, C. Price, and H. Bowman (2017). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews* 77(November 2016), 388–402.
- Gerfen, C. R. and D. J. Surmeier (2011). Modulation of striatal projection systems by dopamine. *Annual Review of Neuroscience* 34, 441–466.
- Goulden, N., A. Khusnulina, N. J. Davis, R. M. Bracewell, A. L. Bokde, J. P. McNulty, and P. G. Mullins (2014). The salience network is responsible for switching between the default mode network and the central executive network: Replication from DCM. *NeuroImage* 99, 180–190.
- Goupil, L., P. Saint-Germier, G. Rouvier, D. Schwarz, and C. Canonne (2020). Musical coordination in a large group without plans nor leaders. *Scientific Reports* 10(1), 1–14.
- Griessinger, T. and G. Coricelli (2015). The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences* 3, 73–79.

- Hamilton, A. F. d. C. and F. Lind (2016). Audience effects: what can they tell us about social neuroscience, theory of mind and autism? *Culture and Brain* 4(2), 159–177.
- Hampton, A. N., P. Bossaerts, and J. P. O’Doherty (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America* 105(18), 6741–6746.
- Hedden, T. and J. Zhang (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition* 85(1), 1–36.
- Henchy, T. and D. C. Glass (1968). Evaluation apprehension and the social facilitation of dominant and subordinate responses. *Journal of Personality and Social Psychology* 10(4), 446–454.
- Hill, M. R., E. D. Boorman, and I. Fried (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications* 7, 1–12.
- Horowitz, A. (2009). Attention to attention in domestic dog (*Canis familiaris*) dyadic play. *Animal Cognition* 12(1), 107–118.
- Izuma, K., D. N. Saito, and N. Sadato (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience* 22(4), 621–631.
- Joiner, J., M. Piva, C. Turrin, and S. W. C. Chang (2017). Social learning through prediction error in the brain. *npj Science of Learning* 2(1), 8.
- Keltne, D. and J. Haidt (2001). Social Functions of Emotions. In T. Mayne and G. A. Bonanno (Eds.), *Emotions and social behavior. Emotions: Current issues and future directions* (Guilford Press ed.), pp. 192–213.
- Keynes, J. M. (1942). *Théorie générale de l’emploi, de l’intérêt et de la monnaie*, traduit par Jean de Largentaye (Payot ed.).

- Kumaran, D., A. Banino, C. Blundell, D. Hassabis, and P. Dayan (2016). Computations Underlying Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing Self-Relevant Information. *Neuron* 92(5), 1135–1147.
- Kumaran, D., H. L. Melo, and E. Duzel (2012). The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. *Neuron* 76(3), 653–666.
- Li, Y., E. Météreau, I. Obeso, L. Butera, M. C. Villeval, and J. C. Dreher (2020). Endogenous testosterone is associated with increased striatal response to audience effects during prosocial choices. *Psychoneuroendocrinology* 122(September).
- Ligneul, R., I. Obeso, C. C. Ruff, and J. C. Dreher (2016). Dynamical Representation of Dominance Relationships in the Human Rostromedial Prefrontal Cortex. *Current Biology* 26(23), 3107–3115.
- Matsumoto, M. and O. Hikosaka (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459(7248), 837–841.
- Meltzoff, A. N. (1995). Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *developmental psychology*, 838–850.
- Mhatre V. Ho and Kelsey C. Martin, J.-A. L. (2012). Striatal Mechanisms Underlying Movement, Reinforcement, and Punishment. *Physiology* 23(1), 1–7.
- Milinski, M., D. Semmann, and H. J. Krambeck (2002). Reputation helps solve the ‘tragedy of the commons’. *Nature* 415(6870), 424–426.
- O’Doherty, J. P., A. Hampton, and H. Kim (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences* 1104, 35–53.
- Palminteri, S., M. Khamassi, M. Joffily, and G. Coricelli (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications* 6.

- Park, S. A., S. Goïame, D. A. O'Connor, and J.-C. Dreher (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biology* 15(6), e2001958.
- Park, S. A., M. Sestito, E. D. Boorman, and J. C. Dreher (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications* 10(1), 1–12.
- Pavlov, I. P. (1927). Conditioned reflexes. An investigation of the physiological activity of the cerebral cortex. Translated and edited by E. V. ANREP. *Conditioned reflexes. An investigation of the physiological activity of the cerebral cortex. Translated and edited by E. V. ANREP*, xv+427p. 18 fig.—xv+427p.
- Perner, J., B. Lang, and D. Kloo (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development* 73(3), 752–767.
- Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature reviews neuroscience* 9(Box 2), 148–158.
- Phelps, E. A., K. M. Lempert, and P. Sokol-Hessner (2014). Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience* 37, 263–287.
- Phillips, A. G., G. Vacca, and S. Ahn (2008). A top-down perspective on dopamine, motivation and memory. *Pharmacology Biochemistry and Behavior* 90(2), 236–249.
- Piaget, J. (1977). *Recherches sur l'abstraction réfléchissante*. Paris: Presses univ. de France.
- Qu, C., R. Ligneul, J. B. Van der Henst, and J. C. Dreher (2017). An Integrative Interdisciplinary Perspective on Social Dominance Hierarchies. *Trends in Cognitive Sciences* 21(11), 893–908.
- Rand, D. G. and M. A. Nowak (2013). Human cooperation. *Trends in Cognitive Sciences* 17(8), 413–425.



- Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory* 21(6), 64–99.
- Rilling, J. K., D. R. Goldsmith, A. L. Glenn, M. R. Jairam, H. A. Elfenbein, J. E. Dagenais, C. D. Murdock, and G. Pagnoni (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia* 46(5), 1256–1266.
- Rilling, J. K., D. A. Gutman, T. R. Zeh, and G. Pagnoni (2002). A Neural Basis for Social Cooperation. *Neuron* 35, 395–405.
- Santos, F. C., M. D. Santos, and J. M. Pacheco (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454(7201), 213–216.
- Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews* 95(3), 853–951.
- Schultz, W., P. Dayan, and P. R. Montague (1997). A Neural Substrate of Prediction and Reward. *Science* 275(5306), 1593–1599.
- Shirer, W. R., S. Ryali, E. Rykhlevskaia, V. Menon, and M. D. Greicius (2012). Decoding Subject-Driven Cognitive States with Whole-Brain Connectivity Patterns. *Cerebral Cortex* 3211(January).
- Sigmund, K., C. Hauert, and M. A. Nowak (2001). Reward and punishment. *Proceedings of the National Academy of Sciences of the United States of America* 98(19), 10757–10762.
- Suzuki, S., N. Harasawa, K. Ueno, J. L. Gardner, N. Ichinohe, M. Haruno, K. Cheng, and H. Nakahara (2012). Learning to Simulate Others' Decisions. *Neuron* 74(6), 1125–1137.
- Tavoni, G., V. Balasubramanian, and J. I. Gold (2019). What is optimal in optimal inference? *Current Opinion in Behavioral Sciences*, 11–126.
- Van Essen, D. C. (1997). A tension-based theory of morphogenesis and compact wiring in the central nervous system.

- von Neumann, J., O. Morgenstern, and A. Rubinstein (1944). *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press.
- Wacongne, C., J. P. Changeux, and S. Dehaene (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience* 32(11), 3665–3678.
- Wilson, R. C. and A. G. Collins (2019). Ten simple rules for the computational modeling of behavioral data. pp. 1–35.
- Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics* 6(3), 299–310.
- Zhu, L., K. E. Mathewson, and M. Hsu (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences of the United States of America* 109(5), 1419–1424.



## Chapter 2

# Neurocomputational mechanisms engaged in detecting cooperative and competitive intentions of others<sup>1</sup>

---

<sup>1</sup>Ce chapitre est un travail en collaboration avec Rémi Janet, Koosha Khalvati, Rajesh P.N.Rao. and Jean-Claude Dreher  
J'ai effectué l'ensemble de ces travaux avec l'aide éventuelle des différents collaborateurs.

# Neurocomputational mechanisms engaged in detecting cooperative and competitive intentions of others

R. Philippe <sup>1</sup>, R. Janet <sup>1</sup>, K Khalvati <sup>2</sup>, R.P.N. Rao <sup>2,3</sup>, D Lee<sup>4</sup>, JC. Dreher <sup>1\*</sup>

<sup>1</sup> CNRS-Institut des Sciences Cognitives Marc Jeannerod, UMR5229, Neuroeconomics, reward, and decision making laboratory. 67 Bd Pinel, 69675 Lyon, FRANCE

<sup>2</sup> Paul G. Allen School of Computer Science and Engineering, University of Washington, 185 Stevens Way, Seattle, WA 98195

<sup>3</sup> Center for Neurotechnology, University of Washington, Seattle, WA 98195

<sup>4</sup> The Zanvyl Krieger Mind/Brain Institute, Kavli Neuroscience Discovery Institute, Department of Neuroscience, Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA

\* Corresponding author

## **Abstract (150 words)**

Humans frequently interact with other agents whose intentions can fluctuate over time between competitive and cooperative strategies. How does the brain decide whether the others' intentions is cooperative or competitive when the nature of the interactions is not signaled? Here, we use fMRI and a task in which participants thought they were playing with a partner, who was in fact an algorithm that switched, without signaling, between cooperative and competitive strategies. We find that a neurocomputational mechanism underlying arbitration between competitive and cooperative experts outperform other learning models in predicting choice behavior. The ventral striatum and ventromedial prefrontal cortex tracked the reliability of this arbitration process. When attributing competitive intentions, these brain regions increased their coupling with a network that differentiated prediction error related to competitiveness *versus* cooperativeness. These findings provide a neurocomputational account of how the brain dynamically arbitrates between cooperative and competitive intentions when making adaptive social decisions.

## Introduction

During social interactions, humans are often uncertain whether others intend to compete or cooperate. The intentions of other agents can fluctuate over time, making it challenging to develop a behavioral strategy. A key question is to understand how the brain decides whether the other is using cooperative or competitive intentions during volatile situations in which the nature of the social interactions is not explicitly determined, such as when interacting with another individual alternating between competitive and cooperative strategies? This question is of importance since it lies at the heart of strategic social decision making<sup>1-9</sup>. In these types of situations, other agents can change behavior according to cooperative or competitive intentions. Pure cooperation is generally defined as involving a group of individuals working together to attain a common goal<sup>10,11</sup>. In contrast, pure competition involves one person attempting to outperform another in a zero-sum situation<sup>12</sup>. A number of theoretical accounts and experimental results demonstrate that the ability to mentalize, i.e. to simulate the other's belief about one's next course of action, is crucial for strategically sophisticated agents<sup>6,7,13,14</sup>. The neurocomputational mechanisms engaged in attributing intentions to others have been studied in situations in which participants are explicitly informed about the nature of the interactions, either in a collaborative context alone<sup>15-17</sup> or in a competitive context alone<sup>8,18-25</sup>. For example, during a cooperative game such as the coordination game, the best strategy is to try to choose one of two presented targets consistently. In contrast, in a competitive game such as the matching pennies<sup>19,25</sup>, the optimal strategy is to choose between two targets equally often and randomly across trials. If the identity of the game played is not known, the agent has to adjust his/her strategy based on repeated interactions with others and to infer cooperation/competition on the basis of observations. How the brain achieves such inference poses a unique computational problem because it not only requires the recursive representation of reciprocal beliefs about other's intentions, as in cooperative or competitive contexts alone, but it also requires to decide whether the other is competitive or cooperative to deploy a consecutive behavioral strategy.

Here, we thought to determine the neurocomputational mechanisms that underlie the inferences of whether another is competing or cooperating during volatile situations in which the nature of the interactions is not explicitly signaled. A recent computational account proposed that arbitration between strategies is determined by their predictive reliability, such that control over behavior is adaptively weighted toward the strategy with the most reliable prediction<sup>26</sup>. This approach has been tested successfully in the domains of instrumental or Pavlovian action

selection <sup>27</sup>, model-based and model-free learning <sup>28</sup> and learning by imitation or emulation <sup>29</sup>. Extending this concept of a mixture of experts to social interactions, we investigated whether the brain relies on distinct experts to compute the best choice between two possible intentions attributed to others (cooperation or competition) and then weight them by their relative reliability. We tested and compared these mixtures of models, dynamically attributing intentions to others with different classes of learning models: non-Bayesian vs Bayesian and non-mentalizing vs mentalizing (**see table 1**). This allowed us to identify the algorithms and brain mechanisms engaged with a key component of the estimation of other's intentions, i.e. whether a social partner is cooperating or competing. Below, we describe different classes of algorithms that have been developed to describe learning mechanisms engaged during strategic social interactions.

The majority of theoretical frameworks used to model feedback-dependent changes in decision making strategies, such as choice reinforcement and related Markov Decision Process (MDP) models, assume that optimal decisions can be determined from the observable events and variables by the decision makers. Clearly, these assumptions do not capture the reality and complexity of human social interactions because observable behaviors of other individuals provide only very partial information about their likely future behaviors. Moreover, model-free RL algorithms assume that values (utility or desirability of states and actions), change incrementally across trials. This assumption is incorrect when option values change abruptly, such as when the intention of the other shifts between cooperation and competition. These limitations explain why agents basing their behavior only on standard RL models can be exploited by opponents using more sophisticated algorithms<sup>6,30</sup>.

A more accurate account of strategic learning is based on a family of RL models which adds a mathematical term to the classical Temporal Difference (TD) algorithm to consider the other as an agent having their own policy, which can be influenced by oneself <sup>6,30,31</sup>. For example, fictitious play learning proposes a basic form of mentalizing by having a representation of the other's strategy. Influence models also consider that RL can be supplemented by a mentalizing term that represents how our actions influence those of others, and updated through a belief prediction error <sup>2,6,19,21,30,32-34</sup>. Such influence models formalize not only how players react to others' past choices, (first-order beliefs in Theory of Mind: ToM), but also how they anticipate the influence of their own choices on the others' behavior (i.e., mentalizing-related second-order beliefs). Other modeling approach of theory of mind used Bayesian algorithms to model inferences about the future actions of another, attempting to take their point of view and



to simulate their decision<sup>13,17,35</sup>. This strategy can be performed recursively so that participants make inferences concerning the others' inferences and so on. Such sophisticated approach is grounded in the theoretical framework of Partially Observable Markov Decision Processes (POMDPs)<sup>36</sup>. POMDPs provide a probabilistic framework for solving tasks involving action selection and decision making under uncertainty. Notably, this approach has recently been applied to strategic cooperation in groups<sup>36,37</sup>. These models, however, have mainly been limited to signaled cooperative or competitive tasks where the intentions of players do not change over a given period<sup>13,35,38,39</sup>.

Here, we explicitly tested the predictions of these different families of learning models against one another, investigating not only non-Bayesian vs Bayesian models and non-mentalizing vs mentalizing models, but also a mixtures of models deploying an arbitration process whereby the influence of attributing intentions to others is dynamically modulated depending on which type of intentions (i.e. cooperative vs competitive) is most suitable to guide behavior at a given time. We did so by using a novel model-based fMRI design (**Fig.1**) consisting of an iterative dyadic game in which participants were told that they would interact with another person via a computer. Unbeknownst to them, the other player was an artificial agent which switched between blocks of cooperative trials (matching pennies game) and blocks of competitive trials (hide and seek game) when playing a card matching game. Thus, the algorithm's goals were either the same than the ones of participants in the cooperative blocks or were orthogonal in competitive blocks. Participants remained uncertain with respect to the goals of their "partner" or "opponent", which alternated, without being signaled. This task allowed us to determine the algorithms used by the brain to recognize the "intentions" of others and to adopt appropriate strategies when the modes of interaction (cooperation vs competition) are unsignaled.

We found that the model (referred as mixed-influence model) accounting best for behavior was a mixture of influence models. Two expert systems work together to make strategic decision, one assessing competitive intentions and the other assessing cooperative intentions, a controller weighting between these experts according to their relative reliabilities. Each expert system use a classic RL algorithm complemented with a mentalizing term to infer another's actions. This mixed-influence model accounts for behavior observed in naturalistic environments in which the other's goal is often only partially congruent with our own, allowing for a continuous range of behavior between pure cooperation and pure competition. A brain network including the ventromedial prefrontal cortex (vmPFC) and the ventral striatum tracked the reliability signal

from the controller. This finding indicates that the mixed-influence model captures the higher-order structure of the mixed-intentions task (i.e., alternation between cooperation and competition). When comparing trials classified as competitive vs cooperative by the controller, we also identified a brain system engaged with an updating signal used for learning. Finally, when participants expected higher utility for choosing according to a competitive rather than a cooperative strategy, the vmPFC and the ventral striatum, tracked the intentions of others, and showed changes in functional connectivity with a brain system including the right temporo-parietal junction (rTPJ), dorsolateral prefrontal cortex (dlPFC) and the intra-parietal sulcus (IPS), which discriminate reward prediction error (PE) between classified modes of interaction. Together, these results provide a model-based account of the neurocomputational mechanisms guiding human strategic decisions during games in which the intentions of others fluctuate between cooperation and competition.

## Results

### Behavioral signature of tracking intentions

We assessed how participants used the history of previous interactions to make their choices. We used logistic regression to examine whether participants selected the same target as that from the previous trial (“Stay”) or chose the other target (“Switch”), depending on whether the previous three trials (at t-1, t-2 and t-3) had been won or lost, whether the previous decisions had been to Stay or Switch, and, whether the previous interactions from those trials indicated cooperation (see below). We also added sex, age and the number of trials as control variables. The cooperation was defined being indicated by a binomial variable, representing the interaction between the last action of the Artificial Agent (AA) and the participant’s own previous outcome (“Cooperativity signature”). Thus this variable was set at 1 if either the participant had won on the previous trial and the AA stayed on the same target for the next trial, or if the participant lost on the previous trial and the AA switched to the other target the trial just after. Otherwise the variable was set to 0. Indeed, from the perspective of the participant, if the AA is a cooperative partner, both players win at the same time and should then choose to keep the same target to be more predictable.

We found that the “Cooperativity signature” of the artificial agent’s predicted an increase in the “stay” probability of participants at t-1 and t-2 (Cooperativity signature  $t_{-1}$ : *estimate* = 0.05,  $p=0.026$ ; Cooperativity signature  $t_{-2}$ : *estimate* = 0.04,  $p=0.006$ ,  $\chi^2$  test **Fig.2.a**). This suggests the participants tracked whether the other agent was cooperating during the two previous trials (but not before). Participants used the outcome of the latest trial to make the next decision (staying or switching target) according to a win/stay, loose/switch strategy (victory $t_{-1}$ : *estimate* = 0.16,  $p<0.0005$ ,  $\chi^2$  test **Fig.2.a**).

### Computational models tracking intentions of the other agent

To elucidate the computations underlying strategic decision making, we compared the results of different computational models. These models were split into five classes (**see SI**). The first class of models, based on heuristics, are Win-Stay/Lose-Switch and Random Bias models. The other four classes of algorithms can be classified into non-Bayesian *versus* Bayesian model families along one dimension and mentalizing vs non-mentalizing model families along the other dimension. Thus, the second class of models includes non-Bayesian, non-mentalizing models represented by reinforcement learning (RL) models. The third class

represents Bayesian non-mentalizing models, including (1) a Hierarchical Gaussian Filter (HGF) which tracks the volatility of outcome <sup>40</sup>, (2) the  $k$ -Bayesian Sequence Learner which tracks the probability that one target will be selected by the AA after a history of specific length  $k$  and (3) the *active inference* model which minimizes the expected free energy <sup>41</sup>. The fourth class corresponds to Bayesian mentalizing models, which are the  $k$ -ToM models using recursive Bayesian inferences of depth  $k$  to predict the future choice of the AA. The fifth class of models contains non-Bayesian mentalizing models, namely the “influence models” which are RL models with an additional term representing how the actions of one player influence those of the other player. Each mentalizing model was tested using 3 versions: a competitive, a cooperative and a ‘mixed intentions’ version. The ‘mixed intentions’ version computes one decision value according to a competitive expert and another according to a cooperative expert and arbitrates between the two, based on the difference in their respective reliability (**see SI**).

Next, we performed a group-level random-effect Bayesian model selection on the models’ computed free energy, taking into account potential outliers and the number of free parameters <sup>42,43</sup>. We found that the ‘Mixed Intentions Influence Learning’ Model was the most frequent best fit across the population (**Fig.2.b**), demonstrating that subjects employed mentalizing-related computations in our mixed intentions task. This finding also indicates that arbitration between a cooperative and a competitive expert best explains most participants’ behavior, rather than either expert taken individually. Additionally, only the mixed-intention influence model, (and not the cooperative or the competitive one) succeeded to produce behavior, similar to participants, concerning the effect of the Cooperativity signature on the probability to stay (**see SI, Fig.2.c**). We conducted a logistic regression to understand how the mixed intention model explained differences in behavioral strategy to stay or switch target. This analysis included the reward prediction error at  $t-1$ , the valence of the arbitration between cooperative and competitive intention at time  $t$  ( $\text{sign}(\Delta)$ ; 1 for cooperative and -1 for competitive), and the interaction between these two variables. This analysis revealed a main effect of the valence of the arbitration (*valence of the arbitration<sub>t</sub>* : *estimate* = 0.24,  $p < 0.0005$ ,  $\chi^2$  test **Fig.2.d**) indicating that participant tend to stay more on the same target when they attributed cooperative intention to the other. Moreover, we found an interaction effect, i.e. participants did not integrate the prediction error in their strategy in the same way given the attributed intention (*valence of the arbitration<sub>t</sub> \* rPE<sub>t-1</sub>* : *estimate* = 0.20,  $p = 0.0227$ ,  $\chi^2$  test **Fig.2.d**). That is, higher negative prediction errors increased the probability that the participant stay on the same target when the controller attributes cooperative intentions compared to when it attributes

competitive intentions. In addition, we also performed another logistic regression analysis using the same variables and the actual mode of interaction (i.e. competitive vs cooperative), rather than the classified mode of interaction made by the controller. We did not find the same interaction effect when comparing actual competitive blocks and cooperative blocks ( $Block\ type_t * rPE_{t-1}: estimate = 0.003, p = 0.56, Block\ type_t: estimate = 0.03, p = 0.229,$  and  $PE_{t-1}: estimate = 0.14, p < 0.0005, \chi^2$  test; **see SI**), showing that only the classified intentions (but not competitive or cooperative block) had an effect on the use of prediction error.

Together, these analyses show that participants' behavior is best explained by the Mixed Intention Influence model when alternating between unsigned cooperative and competitive blocks. According to these findings, people use mentalization to update their beliefs about future chosen targets, and to dynamically arbitrate between the predicted intentions of the other agent to compete or cooperate (**Fig.3.a**).

### **Signed difference in reliability ( $\Delta$ ) is influenced by victory and the Cooperativity signature**

We reasoned that when facing an individual who can change his/her intentions to compete or cooperate over time, the brain may rely on distinct experts to compute the best choice based on these two possible intentions (i.e. cooperative or competitive), weighted by their relative reliabilities. We therefore built such an 'arbitrator' computation as a sigmoid function of the difference in reliability between the Cooperative and Competitive interactions ( $\Delta$ ), added to a bias ( $\delta$ ) that characterized each individual tendency to attribute competitive ( $\delta > 0$ ) or cooperative ( $\delta < 0$ ) intentions to others. To assess intentions of the others, participants only have access the outcomes of previous interactions, the choice (to stay or switch) of the artificial agent on previous trials, and the interaction between these two types of information.

We hypothesized that repeated social victories should favor the attribution of cooperative intentions because a series of victories suggests that both players are satisfied with the outcome and in such situation the other player (i.e. AA) has become more predictable, which is an important feature to build cooperation<sup>44</sup>. Moreover, the interaction between outcome and AA's choice (i.e. the tendency of the AA to "stay" after a participant wins or "switch" after a participant loses) should drive the arbitrator in favor of the cooperative mode because playing the same winning target for both players corresponds to the optimal Nash equilibrium of the cooperative game. To test this hypothesis, we regressed the signed difference in reliability on (1) the participant's last outcome, (2) AA choice to "stay" or to "switch", (3) the interaction between the

participant's outcome and the AA's choice to stay or switch (Cooperativity signature) over up to three retrospective trials. We found that the past two interactions between participant's outcome and AA's action (Cooperativity signature), the last outcome and switches by the AA at trial t-2 and t-3 explained the difference in reliability (*Cooperativity signature*  $t_{-1}$ : estimate =0.59,  $p<0.0005$ ; *Cooperativity signature*  $t_{-2}$ : estimate =0.06,  $p=0.037$ ; *Victory*: estimate =1.98,  $p<0.0005$ ; *switch*  $t_{-2}$ : estimate =-0.18,  $p<0.0005$ ; *switch*  $t_{-3}$ : estimate =-0.07,  $p<0.0005$ ),  $\chi^2$  test, **Fig.3.b**).

## Model-based fMRI analyses

First, we constructed a GLM (GLM1) to identify brain regions tracking the arbitration process (i.e.  $\Delta$ : signed reliability difference) between the two experts (one for cooperation the other for competition). We added the reliability difference  $\Delta$  as parametric regressor at the decision stage, and the expected reward for staying on the same target, as non-orthogonalized parametric regressors to allow them to compete for the variance. We added the reward prediction error as a parametric regressor at the outcome time and we controlled for the other's intention effect by adding  $\Delta$  as a non-orthogonalized regressor. The bilateral ventral striatum ( $x,y,z=9,12,0$  and  $x,y,z=-12,9,-6$ ), vmPFC ( $x,y,z=6,45,-8$ ), postcentral gyrus ( $x,y,z=-20,-44,48$ ), and middle cingulate cortex (MCC;  $x,y,z=11,-15,57$ ,  $p<0.05$  whole-brain family-wise error (FWE), **Fig.3.c and d**) tracked the difference in reliability between experts ( $\Delta$ ) at the decision time. Bilateral dorsal striatum (DS;  $x,y,z=17,6,-12$  and  $-14,3,-11$ ), bilateral orbitofrontal cortex (OFC;  $x,y,z=44,36,-14$  and  $-44,52,8$ ), posterior cingulate cortex (PCC;  $x,y,z=2,-34,38$ ), and bilateral angular gyrus ( $x,y,z=45,-30,46$  and  $-54,-62,39$ ) ( $p<0.05$  FWE, **Fig.4.a**) encoded the reward prediction error at the outcome time.

To investigate brain areas encoding the reward prediction error that were more engaged when the controller classified a trial as competitive vs cooperative, we tested another GLM (GLM2). Trial onsets were separated according to whether the value of the signed reliability difference  $\Delta$  added to the bias was positive or negative. If this value was  $\geq 0$ , the trial was classified as Competitive, and Cooperative otherwise. The computed expected reward for staying on the same target was used as a parametric regressor at the time of choice. We found that the right dlPFC ( $x,y,z=35,11,36$ ), the IPS region ( $x,y,z=50,-50,32$ ) and the right temporoparietal junction (rTPJ;  $x,y,z = 51,-50,33$ ,  $p<0.05$  FWE) were more engaged in encoding reward prediction error in trials classified as Competitive versus Cooperative ( $p<0.05$ , FWE **Fig.4.b and c**). This effect could not be explained by less variance in the PE regressor in trials

classified as Competitive trials compared to those classified as Cooperative, because we observed no difference in regressors' variance ( $p=0.57$ , Levene's test). No region was more engaged in trials classified as Cooperative compare to those classified as Competitive.

To further investigate the relationship between the behavior to stay after a trial classified as competitive vs cooperative and the BOLD signal, we conducted a logistic regression. Explanatory variable were the average of the weighted time series in the dlPFC and rTPJ/IPS region (see SI), the valence of the controller ( $\Delta$ ) and their interactions (Eq 1). This this signal was extracted at the time of the outcome. We found an interaction between the weighted time series and the valence of the arbitration (*valence of the arbitration<sub>t</sub> \* weighted time series<sub>t-1</sub>*:  $estimate = -3.419$ ,  $p = 0.0227$ ,  $\chi^2$  test). *Post hoc* test further revealed that this effect was driven by trial classified as competitive (*weighted time series<sub>t-1</sub> of trial classified as comptitive<sub>t-1</sub>*:  $estimate = -2.50$ ,  $p = 0.003$ ,  $\chi^2$  test). This results predict that activation of rdIPFC and rTPJ/IPS increases the probability to switch following a trial classified as competitive.

$$\text{logit}(P(\text{stay})) = \beta_0 + \beta_1 * \text{sign}(\Delta) + \beta_2 * \text{Time Series} + \beta_3 * \text{sign}(\Delta) * \text{Time Series} \quad (\text{Eq 1})$$

## Connectivity analysis

Finally, we performed a generalized psycho-physiological interaction (gPPI) connectivity analysis to understand the interactions between brain regions tracking the arbitration process (i.e.  $\Delta$ : reliability difference) for the cooperative and competitive experts and those more engaged with the PE when the controller attributes more competitive than cooperative intentions to the other (**see Online Methods**). We used the ventral striatum and vmPFC, encoding the controller, as seed regions (ROI extracted from the GLM1 striatal and vmPFC activity) in both competitive and cooperative modes (i.e. respectively  $\Delta < 0$  and  $\Delta > 0$ ) at the decision time. We found stronger functional connectivity between regions encoding the difference in reliability and the right dlPFC ( $x,y,z=38,34,34$ ), the left IPS region ( $x,y,z=-48;-44;58$ ) and the left TPJ ( $x,y,z=-42,-40,50$ ,  $p < 0.05$  FWE; **Fig.5.a**) at the decision time of trials classified as Competitive than those classified as cooperative. This result suggests that the dlPFC and IPS differentiate between PE for competitive *versus* cooperative situations and in updating the difference in reliability with respect to the intention of others.

## Discussion

To make a strategic decision when facing an individual with unknown and fluctuating intentions, it is necessary to make inferences as to whether we are in a competitive or cooperative situation. In the context of minimal information, for example when only the choices of other, but not their outcomes, are available, such inferences are much more difficult to make than when one is in a specific known setting (e.g., in a competitive game)<sup>19</sup>. Here, we provide evidence that the brain engages in dynamic tracking of another individual's cooperative/competitive intentions, despite having no explicit information regarding whether the situation is cooperative or competitive. We found that strategies of participants were mostly affected by the outcomes of previous interactions and by a “signature” of the other's cooperativity, i.e. the tendency of other (here the Artificial Agent or AA) to stay on the same target after the participant's victory. Comparison between computational models demonstrated that such behavior is best explained by a model in which choice is driven by a controller tracking the reliability difference between cooperative and competitive intentions. The fMRI results show that the neural computations of this controller are implemented in the ventral striatum and in the vmPFC. Thus, both behavior and brain imaging results can be accounted for by a model that includes a controller that allocates resources according to different experts' predictions. At the time of outcome, a common brain network, including the rostral anterior cingulate cortex (rACC), ventral striatum and lateral OFC encoded prediction error in trials classified as competitive or as cooperative. However, prediction error signals also depended on the classification of the current trial as Cooperative or Competitive as classified by the controller. That is, a distinct brain network, composed of the bilateral dlPFC, bilateral IPS regions and the rTPJ was more engaged for trials classified as competitive compared to those classified as cooperative. This latter brain network reflects a differential use of the outcome of the social interaction as a function of whether it is classified as competitive or cooperative (**Fig 2.d**).

Mentalizing processes are essential to correctly infer the strategy of others. This is true in the cooperative context, in which participants performed above chance, reflecting their ability to effectively infer the other's (i.e. AA) behavior. In the competitive context, participants performed below chance level, showing that the AA was able to predict their behavior and to exploit their previous choices/outcomes. The mixed influence model had the best ability to predict data and to generate very similar behavior to the participants. Each expert model is an expanded RL model, with a term accounting for one's choice influencing the choice of the other. Although only the influence term differed between the competitive and cooperative models, the



mixed influence model tracked intentions based on this second order mentalizing term by weighting the contribution of a cooperative and of a competitive expert. One key aspect of this mixed-influence model is that it captures higher order structures (fluctuations between cooperation and competition) during social interactions. In contrast, one important limitation of the classical RL model is that it does not exploit higher-order structures such as interdependencies between different stimuli, actions, and subsequent rewards. Previous studies demonstrated that models incorporating such structures can account for individual decision making in different situations <sup>45–48</sup>. Here, we demonstrate that the representation of abstract states, such as whether the other is cooperating or competing, can be extended to social decisions and underlies the ability to build strategies. To confirm that the mixed-influence model accounted more for neural activity in brain areas involved in social interactions, we formally compared the brain regions covarying more with the expected reward for staying on the same target, computed from the winning model, compared to the expected reward for staying on the same target computed from a simple RL model (**see SI**). The crucial difference between a simple RL model and the mixed-influence model is that in the former, only the value of the chosen option is updated and the valuation of the option that was not chosen does not change. In the latter, both the values of the chosen and unchosen options are updated to incorporate the knowledge that the current state has a given reliability to be cooperative or competitive. The controller weights the valuation produced according to the competitive or cooperative hypothesis, computed as a sigmoid of the difference in reliability of the two experts.

Activity of the ventral striatum and vmPFC increased as the cooperative prediction from the controller becomes more reliable than the competitive prediction, reflecting both the outcomes of the previous interactions (reliability difference modulated by last outcome) and other's "Cooperativity signature" over the last trials. Thus, these brain regions dynamically track the reliability difference between intentions classified as cooperative and competitive in a situation where the nature of the social interactions is implicit. Previous reports demonstrated a role of the ventral striatum when making cooperative choices alone, after a partner's cooperative choice in an explicit cooperation task, <sup>49</sup> and the attribution of intentions in a competitive context <sup>30</sup>. Our findings show that strategic social behavior can be explained by a Controller Theory in which cooperative/competitive social behavior results from the interaction of multiple systems, each proposing possible strategies for action <sup>26,28,29</sup>.

One strength of our computational approach was to assess and compare a large variety of competing models, (active inferences, recursive learning models k-TOM and a mixture of

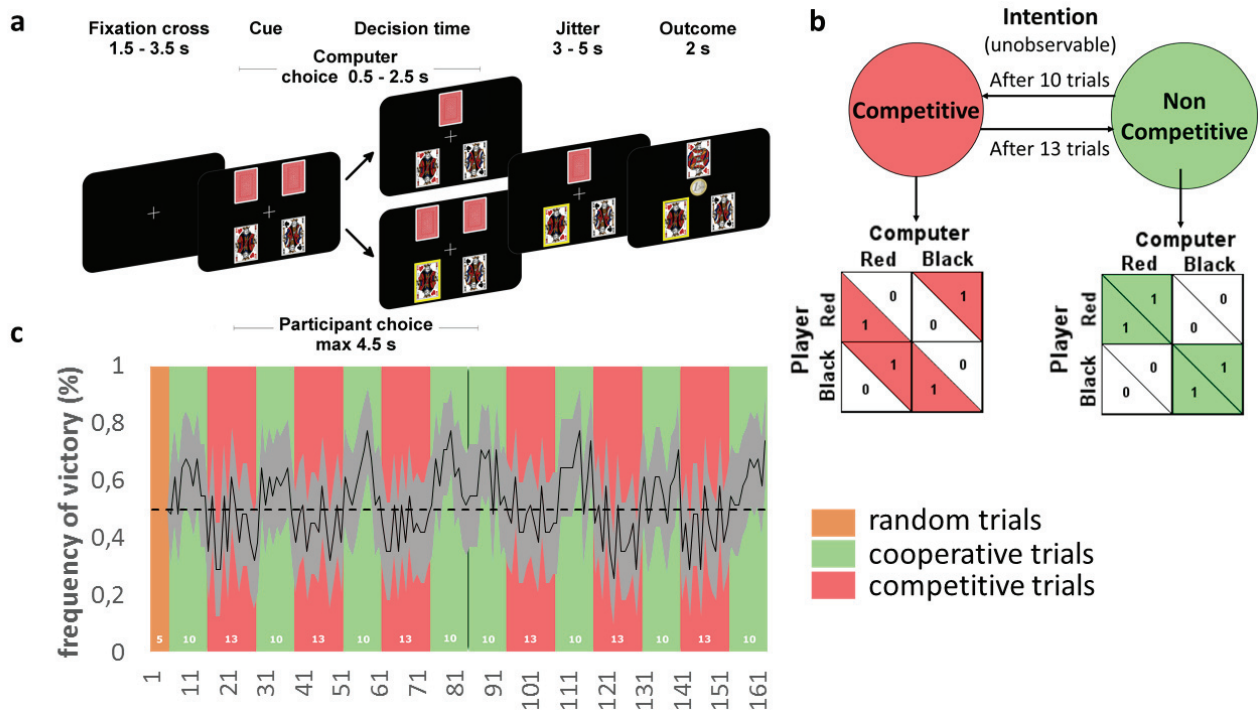
experts using k-TOM, influence models only for cooperative strategies or only for competitive strategies, a mixture of experts using influence models, fictitious learner, Bayesian Sequence learner, Hierarchical Gaussian Filter, Reinforcement Learning and Heuristic models), which had previously never been directly tested against each other (**see SI**). Our results agree with studies that concluded that social learning may be driven by non-specific reinforcement processes including a mentalizing term<sup>3,6,8,30,50</sup>. We demonstrate that when a task is not explicitly signaled as cooperative or competitive, this evokes the arbitration between strategies determined by the predictive reliability. Behavior is hence controlled by weighting towards the strategy with the most reliable prediction<sup>29</sup>. At first glance, it may be surprising to observe that the mixture of expert influence models performs better than mathematically more sophisticated models, such as POMDP and models mimicking different levels of sophistications of mentalizing (k-TOM). However, this is likely because in our setting the only information that can be integrated by participants is their own rewards and the history of the choices made by the other (i.e. AA). The nature of the social interaction is never explicitly signaled (participants are not told whether the other is cooperative or competitive), and the rewards of the other are not observed. This uncertainty could therefore result in the failure of POMDP models to reproduce human behavior, particularly when sudden flips occur between the AA strategies. This contrasts with previous neuroimaging studies that investigated learning of social interactions in either competitive or cooperative situations alone (matching-pennies or rock-paper-scissors games against computerized opponents)<sup>25,51</sup>. Our findings also broadly agree with a cognitive hierarchy of strategic learning mechanisms, proposing that distinct levels of strategic thinking correspond to different levels of sophistication of learning mechanisms, (in increasing order of complexity : reinforcement learning, fictitious play learning and influence learning)<sup>52</sup>. However, we propose a more general model based on a mixture of influence learning experts that function in parallel and are then compared with respect to their relative reliability

Competitive social interactions often emerge in situations where an agent's outcome depends on the choices of others, requiring the ability to infer the intentions of others<sup>6</sup>. In the context of our mixed intentions task, when participants attributed competitive, as opposed to cooperative intentions to others, the dlPFC and rTPJ/IPS specifically encoded a relative PE. This network has been broadly reported to be engaged with attentional processes and when inferring the intentions of others<sup>8,37,53</sup>. The strength of our computational account of theory of mind processes is to pinpoint that a key computation performed by this brain network is to compute a PE difference between trials that the controller classified as competitive *versus*

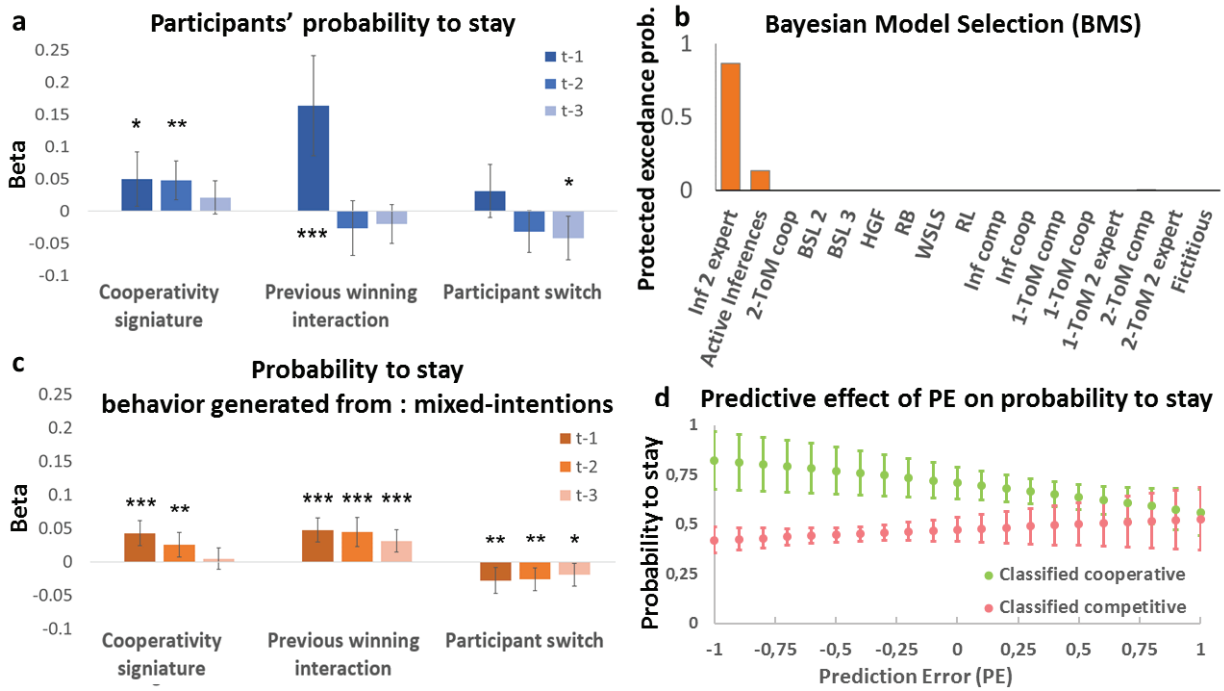
cooperative. This PE difference reflects a differentiation in the implementation and use of the outcome of the social interaction as a function of the classified interaction. Note that PE was not more volatile in trials when the competitive expert is more reliable than the cooperative expert, ruling out the possibility that the observed PE difference reflects higher PE volatility in competitive contexts. When comparing intentions classified as cooperative compared to competitive, participants tended to be more predictable, staying more on the same target after experiencing an unexpected social defeat (i.e. after higher negative PE) (**Fig 2d**). This behavior likely reflects a signal sent to the other to indicate one's willingness to stay on the same target, despite bearing the cost of staying on this target<sup>6,8,54</sup>. This is a key feature of successful coordination<sup>44</sup> in which agents who want to trigger reciprocity<sup>49</sup> are willing to incur a cost to promote cooperation from the other. Finally, we found higher functional connectivity between seed regions that encode the reliability difference of the controller (vmPFC and striatum) and brain regions more engaged in PE for trials classified as competitive vs cooperative (dlPFC, IPS) (**Fig. 5a**). This indicates that the brain system tracking the reliability of cooperation and competition of others communicates decision-relevant information provided by the PE, especially in the competitive context. Our data thus indicate that the ventral striatum/vmPFC integrate information from multiple brain areas including PE encoding areas by means of functional coupling to adapt behavior according to reliability signal (**Fig. 5.b**, BOLD signal predict switch in following trials classified as competitive).

Together, these findings provide a mechanistic framework for the neurocomputations underlying learning in strategic situations. Our mixed-intentions model may be useful in the field of computational neuropsychiatry to identify the specific computational components that are modified in theory of mind alterations, a key feature of autism spectrum disorder<sup>68</sup>.

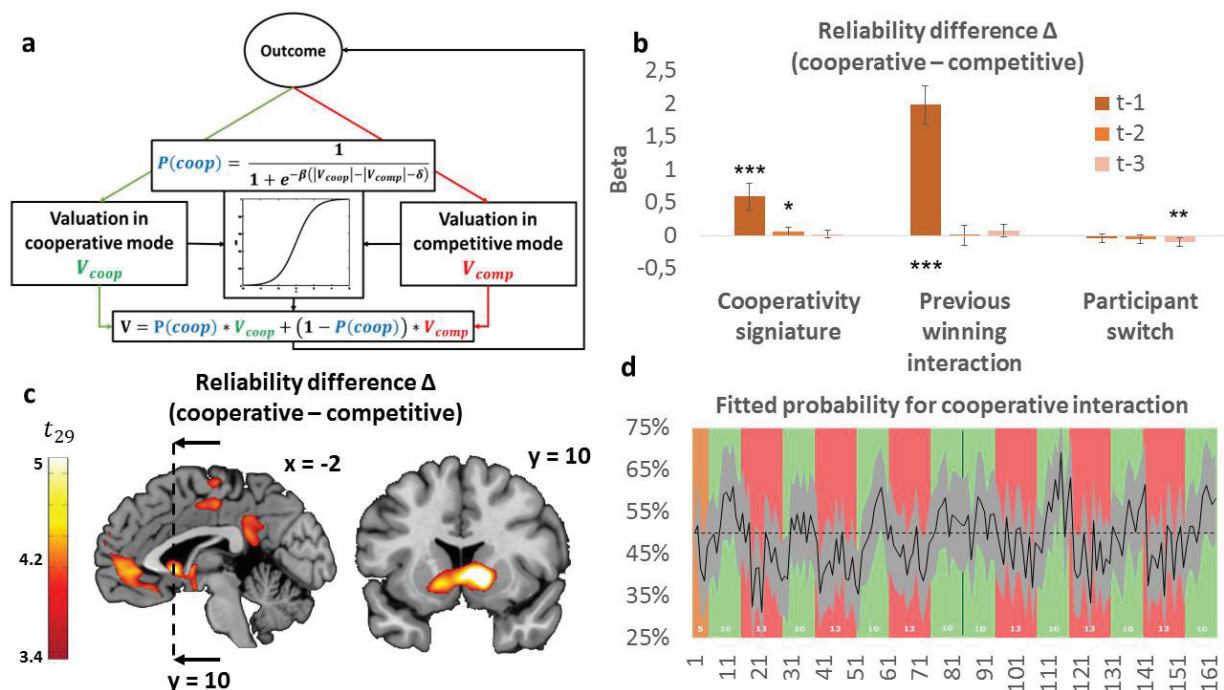
## Figures



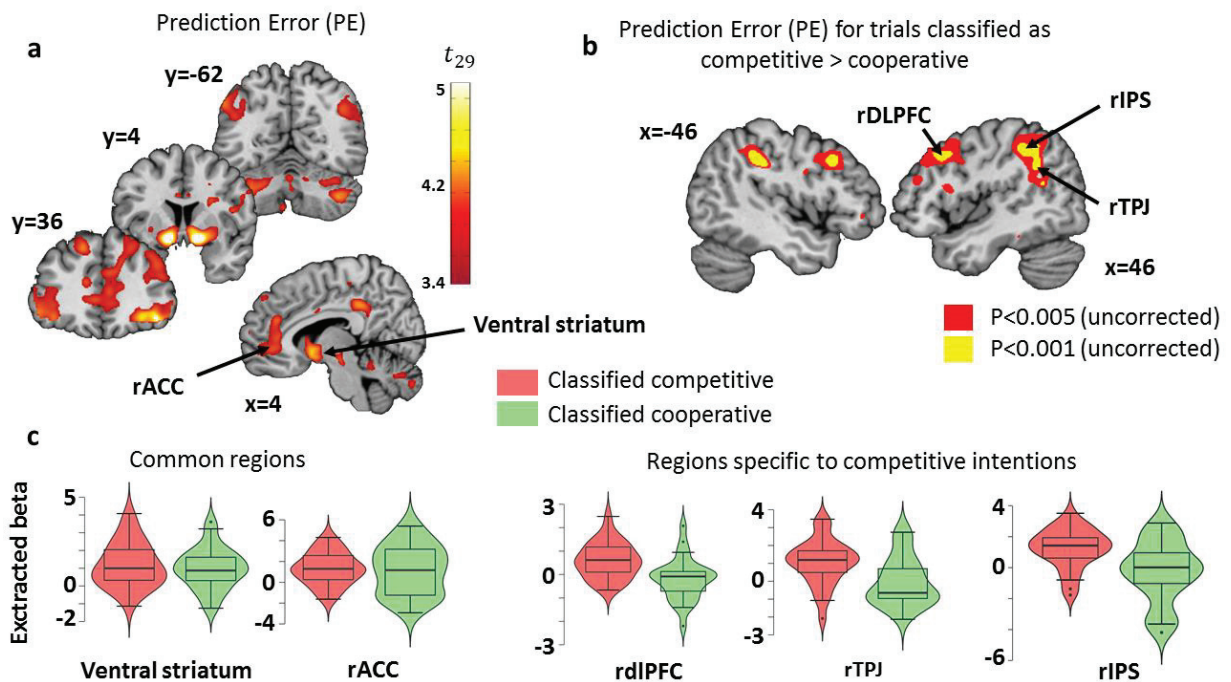
**Figure 1.** fMRI experiment. **a.** After a fixation cross, four cards were presented on the screen. The two cards shown on top of the screen represent the cards presented to the opponent/partner (i.e. Artificial agent), and not seen by the participant while the two kings (one black and one red) are the cards presented to the participant (shown in the bottom of the screen). The participants had to choose between these two cards. At the time of decision, the upper screen represents the display if the AA makes its choice first, while the lower screen shows how one card is highlighted with yellow border if the participant makes his choice first. Then a screen presents the participant's and Artificial Agent's choices together. Finally, at the time of outcome the participant wins if both he/she chooses the same card than the AA (here red king). **b.** Payoff matrix of the two types of block. **c.** Frequency of victory (black line) during competitive (red background) and cooperative (green background) blocks. The grey area represents the 95% confidence interval. The orange background represents 5 initial trials in which the AA played randomly for initialization purpose.



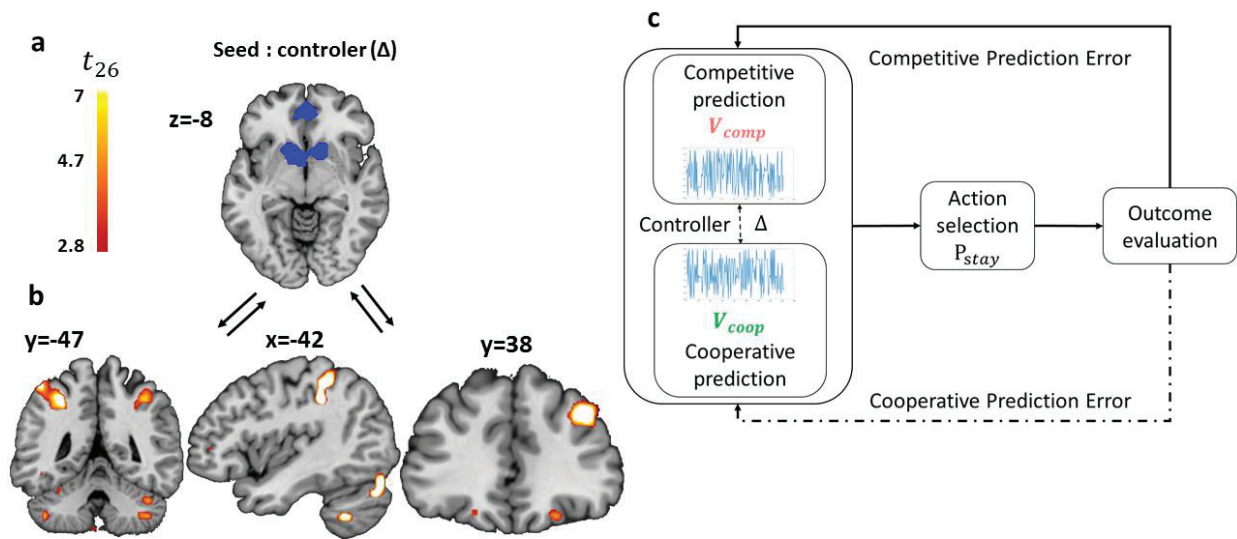
**Figure 2.** **a.** Model-free analysis. Random-effect logistic regression of the decision to stay after selecting a specific target with respect to the action of the artificial agent “Cooperativity signature” (i.e. participant wins then AA stay or participant loses then AA switch), the previous winning interaction (i.e. success or failure of past trials) and the choice to switch or stay, over the previous three trials. Error bars are the 95% confidence interval. **b.** Model comparisons based on Bayesian model selection. The protected exceedance probabilities indicate that the Mixed Intention Influence model (Inf 2 expert) explains decisions in the mixed intention task better than others: Active inference; k-ToM; Bayesian Sequence Learner (BSL); Hierarchical Gaussian Filter (HGF); Reinforcement Learning (RL); Heuristic models: Random Bias (RB); Win/Stay-Lose/Switch (WSLS). **c.** Model-free generative analysis. We generated one hundred sets of data using a free parameter from a normal distribution with mean and standard deviation calculated from the models fitted to the population, against the same artificial agent that participants played. We regressed the behavioral decision to stay after selection of a specific target on the previous trial depending on the interaction of the previous outcome and the action of the artificial agent (“Cooperativity signature”), the success or failure of up to three previous trials, and the action to switch or stay of the participant. Error bars are the 95% confidence interval (random-effect logistic regression). **d.** Marginal effect of the prediction error on the probability to stay on the same target in trials classified as Cooperative (green) and trials classified as Competitive (red). Error bars are the 95% confidence interval. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$  (random-effect logistic regression).



**Figure 3.** **a.** Scheme of the Mixed Intention Influence model. After the reception of an outcome, two influence models (one cooperative and the other competitive) compute a value for choosing one specific target (the black one). A controller uses the difference between the absolute value of the value of each expert (called reliability) to compute a probability that the other is cooperating. Then, the model weights the value of each expert according to the probability of being in cooperative and in competitive modes to produce a final decision value. **b.** Difference in reliability is influenced by the Cooperativity signature of the Artificial Agent (AA), specifically the interaction of the previous subject's outcome followed by the action of the artificial agent (Participant win then AA stay and Participant lose then AA switch), the latest outcome and the computer's switch at trial t-2 and t-3. Error bars are the 95% confident interval. **c.** BOLD signal in ventral striatum, mPFC and posterior cingulate cortex (PCC) ( $p < 0.05$  whole-brain family-wise error) are correlated with the difference in reliability,  $\Delta$ , of estimated competitive and cooperative intentions. **d.** Mean probability of the participant attempting to cooperate across all participants (black line) for the 163 trials. The initial orange area is the 5 random initializing trials, green areas are the cooperative blocks and red areas the competitive blocks. The grey area is the 95% confidence interval.



**Figure 4.** Correlations between BOLD activity and prediction error. **a.** Brain regions in which BOLD signal correlates with prediction errors for trials classified by the controller to be either competitive or cooperative. **b.** Brain regions in which BOLD activity correlates more with PE on trials estimated to be competitive compared to trials estimated to be cooperative. This network comprised dlPFC ( $x,y,z = 30,9,42$ ), IPS ( $x,y,z = 42,-47,42$ ) and the rTPJ ( $x,y,z = 51,-50,33$ ,  $p < 0.05$  whole-brain family-wise error). **c.** Beta value extracted for trials estimated to be either competitive or cooperative. Left: regions in the ventral striatum (left  $x,y,z = -14,3,-11$  + right  $x,y,z = 17,6,-12$ ) and rACC ( $x,y,z = 6,42,-3$ ) with increased activation in trials estimated to be either competitive or cooperative. Right: specific brain regions activated only when trials were classified as Competitive: dlPFC ( $x,y,z = 30,9,42$ ), IPS ( $x,y,z = 42,-47,42$ ) and rTPJ ( $x,y,z = 51,-50,33$ ) from 8 mm spheres centered on peak activation.



**Figure 5.** Neural mechanisms of arbitration between the attributions of competitive and cooperative intentions to the AA. **a.** The BOLD signal was extracted from seeds regions (mPFC and ventral striatum using GLM1) computing the reliability difference between cooperative and competitive intentions of others (in Blue). The reliability of the competitive and cooperative experts for each trial was calculated from the estimation of whether the AA would choose red or black. **b.** Brain regions in which BOLD signal correlated more with the BOLD signal from seed regions when comparing trials classified as competitive vs cooperative by the controller. The psychophysiological interaction effect shows higher functional coupling between seed regions and the left TPJ ( $x,y,z = -42,-40,50$ ), left IPS ( $x,y,z = -32, -48, 50$ ) and right dlPFC ( $x,y,z = 38, 34, 34$ ,  $p < 0.05$  FWE threshold at  $p < 0.001$ ). **c.** Scheme of the mixed intentions model that predicts attribution of competitive or cooperative intentions and strategic decisions during the mixed intentions task. In this model, the values of competitive and cooperative intentions are estimated by separate experts, and the controller computes the difference in their reliabilities. A strategic decision is selected (stay or switch target) given the inferred cooperative or competitive intentions of others. The inferred intention of other is then updated according to the outcome of the social interaction. The thick line represents stronger coupling and behavioral influence (see Fig 2d) between outcome evaluation and strategic decision making during competitive interactions.



## Online Methods

### Participants

32 participants (aged 20-40,  $M = 27$ ,  $SD = 5.1$  - 17 women) were recruited via a daily local newspaper and the University of Lyon 1 mailing list. All participants were screened to exclude those with medical conditions including psychological or physical illnesses or a history of head injury to prevent having confounding variables. They all provided informed consent and were paid a fixed amount. However, they were financially motivated in being told that they would be paid as a function of their decisions.

### Mixed intentions game

Participants performed a novel task comprising 163 trials in an MRI scanner. They were led to believe that they were interacting with another person via a computer interface, while in fact they were playing against an artificial agent (AA) managed by a computer program. Such simulated social interactions allowed us to investigate the dynamics and neural mechanisms arbitrating between multiple learning algorithms. Participants were faced with a screen containing four cards, two faces down (the other player's cards) and two faces up (their own cards). Participants were informed that to win, they had to choose the card of the same color as the one the other person was going to choose. Experimenters were careful not to specify whether the other was an adversary or a partner. Participants were told that they and the other player had to make their choices in four seconds (Figure 1.a). If the Artificial Agent (AA) played before the participant, only one of the two cards down remained face down on the playing field. If the participant chose first, only the selected card remained on the playing field. Then, once both had chosen either a red or black card, the chosen cards were revealed and the participant received a reward if the card colors matched, otherwise they received nothing. Participants were made to believe that their final payoff would be increased by 10ct for each win. No information about the other's payoff was given to the participants, they only knew that after an interaction, the other one saw the same screen but with its own outcome which could be different from theirs. The post-experiment debriefing revealed that 26 participants actually believed they were interacting with a real person. The other 5 participants answered all questions such as "what was your strategy?" by calling the algorithm "the other player", they finally expressed doubts only when answering the last question "Do you think the other player was a real person?".

Importantly, unbeknownst to the participants, the artificial agent alternated between Competitive and Cooperative trial blocks. During this mixed intentions task, the AA's strategy was determined by alternating 13 trials of a matching pennies (MP) task (competitive blocks), and 10 trials of a coordination game (cooperative blocks). The artificial agent algorithm was designed to predict the color that would be chosen by the participant on the basis of a probabilistic analysis of the two previous choices and outcomes (see Supp. Mat. for the algorithm). Here we defined a competitive choice, made by the AA, as choosing the card of the color the participant was expected not to play and a cooperative choice as choosing the card with the same color. Thus, the artificial agent exploited the bias of the participants in a stochastic way, i.e. the more predictable the participant was, the more the algorithm made correct competitive or cooperative choices (see SI). Participants were not informed of the switches between the two blocks (cooperative vs competitive), however their goal was always to choose the same color as that chosen by the other player (i.e. the AA).

The MP task is competitive, and the computer uses the record of the participant's choice and reward history to minimize the participant's payoff. Therefore, in this case the subject's optimal strategy during the MP task is to choose the two targets equally often and randomly across trials. During the coordination game, the AA tried to maximize the subject's payoff and in this case the subjects should try to choose one of the two targets consistently so that the computer can choose the same target as them. In some cases, this can be accomplished by a frequent switch to the target chosen by the computer. Since the participant is not informed of either the goals of the AA or the switches between blocks, they must adjust their strategy based on recent experience and infer cooperation/competition on the basis of their observations.

This task was designed to identify key components of the estimation of intentions regarding whether others are cooperating or competing. We took advantage of the fact that an individual's estimates as to whether they are engaged in a cooperative or competitive interaction can be assessed even when the individual is interacting with a computer program rather than another person. Transitions between the competitive and cooperative blocks were unsignaled, therefore subjects had to discover by trial and error the most successful strategy over consecutive blocks. This alternation between the two interaction modes functioned well because the participant's winning rate was significantly higher in cooperative (mean 60% std 1%) than in competitive (mean 44% std 1%) trials (paired t-test  $p < 10^{-4}$ ).

## Artificial agent

The AA calculated the probability  $p$  for the participant to select a particular target color based on the history of the two previous choices and their outcomes. Then to make the artificial agent behave more like a real person, this prediction was exploited in a probabilistic fashion (see SI). In the cooperative mode the AA chose the color card it predicted with probability  $p$ . In the competitive mode this color was chosen with probability  $1-p$ .

## Behavioral analysis

For the logistic regressions, we reported significant marginal effect of a given variable under the name “*estimate*” (for example: *Cooperativity signature*  $t_{-1}$ : *estimate*).

$$\text{Logistic regression : } \ln\left(\frac{P}{1-P}\right) = x_0 + x_1X_1 + x_2X_2 + \dots$$

$X_i$  represent independent variables and  $x_i$  represent the associated weights in the logistic regression.  $P$  represent the probability of a given event. The marginal effect of the variable  $X_1$  is defined as:

$$\widehat{y}_1 = \text{mean}(\text{logit}^{-1}(x_1))$$

The mean is computed across all observed data. Thus, the marginal effect called “*estimate*” can easily be interpreted as the discrete change of the dependent variable given a unitary change of an independent variable.

For the linear regressions, reported “*estimate*” represents  $x_i$  i.e. the regression coefficient. Indeed, in a linear regression, marginal effect of a variable is equal to the estimated coefficient.

## Models

To test for a dynamic tracking of implicit intention we compared 14 models, 9 involved theory of mind (*Inf,k-ToM*), the others were to control for other possible strategies. The influence models (*Inf*) rely on a Taylor expanded reinforcement learning<sup>69</sup> to take into account the influence of one’s own strategy on the strategy of the other. *k-ToM* models also take into account the influence of one’s own strategy on the other but in a Bayesian fashion<sup>13,35</sup>. These two models were adapted in their cooperative and competitive versions. Moreover, we constructed an adaptation of these two models (*Inf,k-ToM*) in which an arbitrator weights the cooperative and competitive versions according to their reliability before making the decision. Finally, because

k-ToM is a recursive model (“I think that you think that...), we included k-ToM of depth one and two for each version.

To control for strategies that did not include theory of mind we added 5 other models including two Bayesian inference types (*HGF* and *BSL*). The Hierarchical Gaussian Filter (*HGF*)<sup>40,70</sup> basically tracks the external volatility of the artificial agent choices in a Bayesian hierarchical way. The Bayesian Sequences Learner (*BSL*) strategy relies on Bayesian inference given past sequences of choices. In a model free analysis, we found that participants tended to use the past 2 choices to make their next choice, so we used sequences of depths 2 and 3. Finally, we added two non-Mentalizing non Bayesian models, a reinforcement learning model (RL) and a model based on the heuristic Win/stay – Lose/Switch that we observed in the model free analysis.

The Bayesian Model Selection (BMS) was performed using the VBA toolbox (Variational Bayesian Analysis) in a random effect analysis relying on the free energy as the lower bound of model evidence. We use protected Exceedance Probability measurements (pEP)<sup>42</sup> to select the model which is used most frequently in our population.

| <i>Model</i>                      | <i>Mentalizing</i>   | <i>Bayesian</i>      | <i>Mixed intentions</i> |
|-----------------------------------|----------------------|----------------------|-------------------------|
| <i>Influence</i>                  | +<br>(coop and comp) | -                    | -                       |
| <i>Fictitious</i>                 | +                    | -                    | -                       |
| <i>Influence mixed intentions</i> | +                    | -                    | +                       |
| <i>k-ToM</i>                      | +<br>(coop and comp) | +<br>(depth 1 and 2) | -                       |
| <i>k-ToM mixed intentions</i>     | +                    | +                    | +                       |
| <i>Active Inferences</i>          | -                    | +                    | -                       |
| <i>HGF</i>                        | -                    | +                    | -                       |
| <i>BSL</i>                        | -                    | +<br>(depth 2 and 3) | -                       |
| <i>RL</i>                         | -                    | -                    | -                       |
| <i>Wst/Lsw</i>                    | -                    | -                    | -                       |

**Table 1.** Classification of models according to 3 categories. The first depends on the ability of the model to mentalize, the second depends on whether the model is a Bayesian model, and the third concerns models that could be used with a mixture of experts.

## **fMRI data acquisition**

MRI acquisitions were performed on a 3 Tesla scanner using EPI BOLD sequences and T1 sequences at high resolution. Scans were performed in a Siemens Magnetom Prisma scanner HealthCare at CERMEP Bron (single-shot EPI, TR / TE = 1600/30, flip angle 75°, multiband acquisition (accelerator factor of 2), in an ascending interleaved manner with slices interlaced 2.40 mm thickness, FOV = 210 mm. We also use the iPAT mode with an accelerator factor of 2 and the GRAPPA method reconstruction. The number of volumes acquired varied given the time the participant took to make their decisions. The first acquisition was made after stabilization of the signal (3 TR). Whole-brain high-resolution T1-weighted structural scans (0.8 x 0.8 x 0.8 mm) were acquired for each subject, co-registered with their mean EPI images and averaged across subjects to permit anatomical localization of functional activations at the group level. Field map scans were acquired to obtain magnetization values that were used to correct for field inhomogeneity.

## **fMRI data analysis**

Image analysis was performed using SPM12 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK, [fil.ion.ucl.ac.uk/spm/software/spm12/](http://fil.ion.ucl.ac.uk/spm/software/spm12/)). Time-series images were registered in a 3D space to minimize any effect that could result from participant head-motion. Once DICOMs were imported, functional scans were realigned to the first volume, corrected for slice timing and unwarped to correct for geometric distortions. Inhomogeneous distortions-related correction maps were created using the phase of non-EPI gradient echo images measured at two echo times (5.20 ms for the first echo and 7.66 ms for the second). Finally, in order to perform group and individual comparisons, they were co-registered with structural maps and spatially normalized into the standard Montreal Neurological Institute (MNI) atlas space using the DARTEL method. Then we ran ARTrepair to deweight scans that could include movement artefacts <sup>71</sup>.

We ran general linear models (GLMs) analyses to identify which brain regions encoded: (a) one's belief that one is interacting in a cooperative or in a competitive situation ( $\Delta$ ); (b) the reward prediction error (PE) after such cooperative or competitive interactions; (c) the PE difference between the trials classified as cooperative vs competitive. In every GLM, an event was defined as a stick function. The participant's button press and the AA's selection of target were defined as onset of no interest in all GLMs. For all GLMs, missing trials were modeled with four events (cue, participant's button press, AA's choice and outcome) as separate onsets

without additional parametric regressors. Head movement parameters were added as parametric regressors of no interest to account for motion-related noise. Because the behavioral analysis showed that the bias towards competitive interaction affects the strategy of participants, we added the competitive bias ( $\delta$ ) as a covariate at the second level analysis in all GLMs.

Specifically, in GLM1, there were 4 onsets, including the time of the cue presentation (cards on screen), participant's button press, AA's choice and the feedback time. Parametric regressors were the difference in reliability  $\Delta$  and the expected reward for staying at the time of the cue onset and the reward prediction error (PE) at the feedback time, as well as  $\Delta$ , to control for the effect of the believed intention of the other on the PE brain encoding. For each GLM, we turned off the serial orthogonalization function of regressors to allow it to compete for the variance.

In a second GLM (GLM2), we separated trials given the sign of  $\Delta - \delta$  (positive or negative) to identify brain regions specifically engaged in cooperative or competitive mental states ( $\delta$  is a free parameter capturing the participant's bias toward competitive intent).  $\Delta$  refers to the difference in reliability of cooperative and competitive prediction and  $\delta$  is the competitive bias. For this GLM, there were 6 onsets, including the cue for trials classified as cooperative or competitive, participant's button press, AA's choice and the feedback time for trials classified as cooperative or competitive. For trials classified as cooperative or competitive, parametric modulators were: the difference in reliability  $\Delta$  and the expected reward for staying on the same target at the time of the cue and the PE and  $\Delta$  at the time of feedback. Three participants who always attributed the same intention to the AA were not included in GLM2.

To test the additional hypothesis that brain activation observed for classified intentions (in **Fig 4b**) is also present in competitive vs cooperative blocks, we conducted two more GLMs. The first, GLM3 is similar to GLM2 in the sense that we separated trials into two categories (cooperative and competitive), but the differentiation was made using the real mode of interaction of the AA rather than the classification made by the controller. Other onsets and parametric regressors were left unchanged.

Finally, we ran a last GLM to check that the results observed in GLM2 were not simply due to the effect of volatility of the rewarded target. This GLM (GLM4) is similar to GLM2, i.e. trials were classified according to the sign of  $\Delta - \delta$ . The only difference was that we added the actual probability that the AA would choose the same target as the previous trial as a parametric regressor at both the time of the cue and at the outcome.

We computed one sample t-tests with contrasts for main effect of  $\Delta$  in GLM1 and effect of PE at the outcome time. Then we computed the contrast between competitive and cooperative PE regressors in GLM2, GLM3 and GLM4. Finally, we computed a paired t-test between this contrast derived from GLM2 and GLM3 to formally show that activation coming from the difference between classified trials are significantly higher than those coming from the difference between the actual modes of interaction.

Reported brain areas show a significant activity at the threshold of  $p < 0.05$ , whole brain family-wise error (FWE), corrected for multiple comparisons at the cluster level (threshold at  $P < 0.001$  uncorrected).

### **Psychophysiological interaction (PPI) analysis**

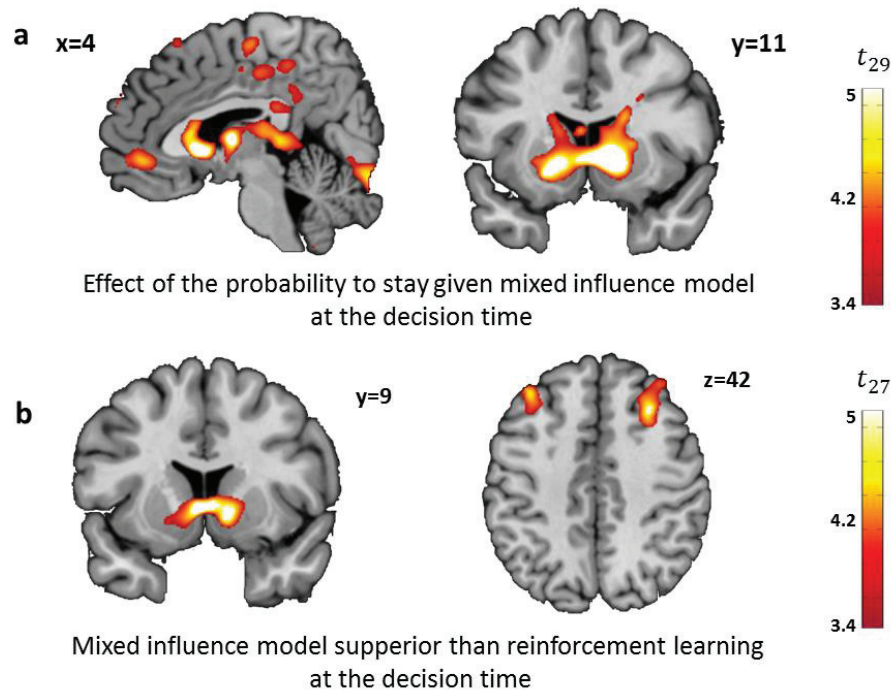
We defined the attribution of cooperative or competitive intentions at the time of decision making as the psychological factor. Thus, we were able to investigate the difference in functional connectivity when making a decision under cooperative or competitive intent. For this PPI analysis, we focused on decision time and on functional connectivity between regions encoding the others' intentions and all other voxels. Thus, for the physiological factor we took the BOLD signal of the striatal region elicited in GLM1 as encoding the intention of others. Otherwise, we used same regressor parameters and onsets as GLM2.

Reported brain areas show a significant activity at the threshold of  $p < 0.05$ , whole brain family-wise error (FWE) corrected for multiple comparisons at the cluster level (threshold at  $P < 0.001$  uncorrected).



## Supplementary Information

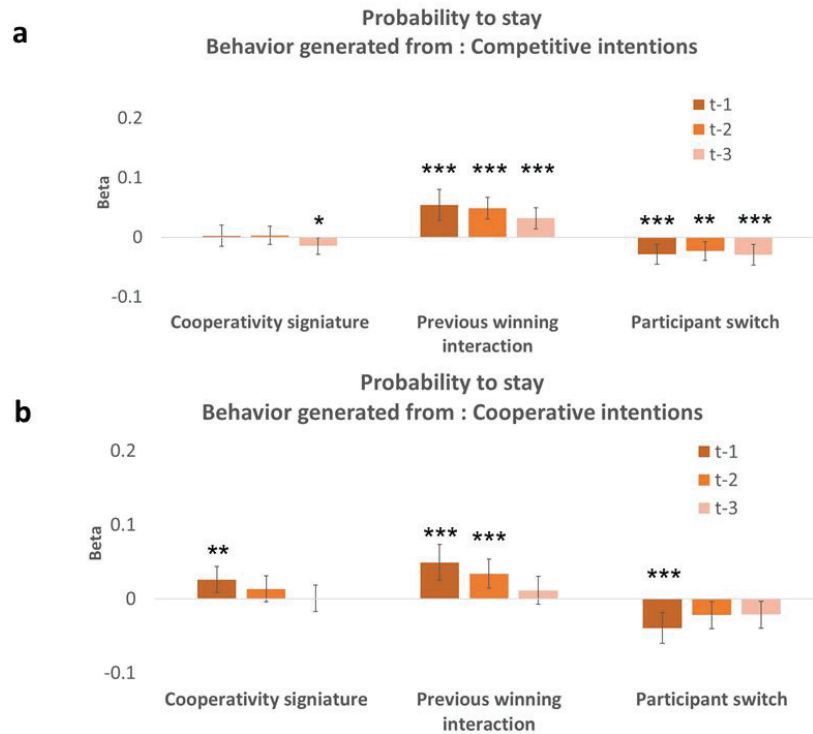
### Supp. Figures



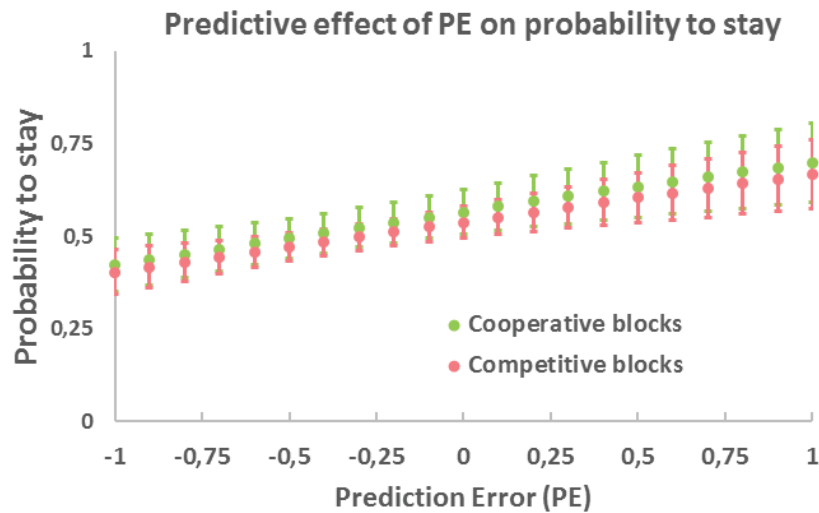
**Extended data figure 1.a.** Neural correlates of the expected reward for staying on the same target as the previous trial, computed by the mixed influence model. (Significant ventral striatum correlation  $x,y,z=14,11,-2$ ,  $p<0.05$  FWE corrected threshold at  $p<0.001$ ) **b.** Ventral Striatum ( $x,y,z=6,12,0$ ), bilateral DLPFC ( $x,y,z=-36, 33, 44$  and  $x,y,z= 30,24,42$ ) and MTG ( $x,y,z=65,-56,-8$ ,  $p<0.05$  FWE corrected threshold at  $p<0.001$ ) are best explained by the expected reward for staying of the mixed influence model than the expected reward for staying of a reinforcement learning model.

We searched for brain regions computing the expected value for staying on the same target for the mixed influence model and for classic reinforcement learning to compare for higher order inferences. To do this, we ran two GLMs (GLM5 and GLM6) containing the expected reward for staying on the same target as the only parametric regressor at the time of choice. In GLM5, the expected reward for staying was computed with the mixed influence model whereas in GLM6 it was computed with a reinforcement learning model. We found that the ventral striatum coded the expected reward for staying on the same target positively, as computed by the mixed influence model ( $x,y,z = 14, 11,-2$ ;  $p < 0.05$  whole-brain FWE corrected at the cluster level, threshold at  $p<0.001$ , see supplementary information). Comparison of the neural

correlates of the expected reward of the two models with a paired t-test revealed that Ventral Striatum ( $x,y,z=6,12,0$ ), bilateral dIPFC ( $x,y,z=-36, 33, 44$  and  $x,y,z= 30,24,42$ ) and MTG ( $x,y,z=65,-56,-8$ ,  $p<0.05$  FWE corrected threshold at  $p<0.001$ ) were encoding higher order features of the task.



**Extended data figure 2.** Model-free generative analysis. We generated one hundred sets of data using free parameters from a normal distribution with mean and standard deviation calculated from the "Influence models" in competitive (a) and cooperative (b) mode, fitted to the population. We generated a data set against the same artificial agent that participants played. We regressed the interaction of the previous outcome and action of the artificial agent (I win – AA / I lose – AA switch), the behavioral decision to stay after selecting a specific target at the previous trial based on the success or failure of the previous trial (Win) and the action to switch or stay of the artificial agent (Switch) in previous trials up to three trials back. Error bars are the 95% confident interval. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$  (random-effect logistic regression).



**Extended data figure 3.** Marginal effect of the prediction error on the probability to stay on the same target in Cooperative blocks (green) and in Competitive blocks (red). Error bars are the 95% confidence interval. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$  (random-effect logistic regression).

## Artificial agent

The artificial agent (AA) selected its target according to the probability for the player to choose a specific color after a given history. It stored the frequency that the participant chose each target after each possible history of four elements composed by two choices and two outcomes (**see table S1**). We call the probability of the player choosing the black card  $P_{black}$ . In competitive trials, the AA will choose the black card with probability  $1 - P_{black}$ , while in cooperative trials, it will choose the black card with probability  $P_{black}$ . A cooperative choice of the AA is defined as a AA choice following the most likely target chosen by the participant. Thus, even in competitive trials, the AA can make a cooperative choice. Since the algorithm needs to be initialized, we arbitrarily defined the first five trials as random (the AA plays the black target with probability 0.5). The possible combinations that are not encountered during these initialization trials, were assigned with a probability of choosing the black target of 0.5.

|                   |                         |                   |                         |
|-------------------|-------------------------|-------------------|-------------------------|
| H1: $BWBW$        | H2: $BWB\bar{W}$        | H3: $BWRW$        | H4: $BWR\bar{W}$        |
| H5: $B\bar{W}BW$  | H6: $B\bar{W}B\bar{W}$  | H7: $B\bar{W}RW$  | H8: $B\bar{W}R\bar{W}$  |
| H9: $RWBW$        | H10: $RWB\bar{W}$       | H11: $RWRW$       | H12: $RWR\bar{W}$       |
| H13: $R\bar{W}BW$ | H14: $R\bar{W}B\bar{W}$ | H15: $R\bar{W}RW$ | H16: $R\bar{W}R\bar{W}$ |

W = **Win**   B = **Black target chosen**  
 $\bar{W}$  = **Lose**   R = **Red target chosen**

**Table S1.** Exhaustive (16) possible histories  $H_i$  of outcomes (Win/Lose) and choices (Black chosen/red chosen) used by the algorithm to track the probability that the participant plays black.

### Weighted time series

To compute the logistic function on the probability to stay after a trial classified as competitive compare to a trial classified as cooperative, we extract the mean signal of dIPFC and TPJ which correlate more with PE during trial classified as competitive than those classified as cooperative (**GLM2, fig 4b and c**). We weight this time series by the convolution of feedback onset and hemodynamic function. We took as a predictor variable the mean of this weighted signal though the 2 scan before the theoretical peak of the hemodynamic function and the 2 scan following this theoretical peak.

### Mixed-influence model: complementary analyses

To determine whether the ‘Mixed Intentions Influence’ model could recover the observed behavioral Cooperativity signature (interaction between the participant’s outcome and the following choice by the AA to switch or not), we simulated one hundred behavioral sets using the influence model from each of the 3 versions (i.e. the competitive expert alone, the cooperative expert alone and the mixed intentions version arbitrating between the cooperative and competitive experts). Only the Mixed Intentions Influence Model was able to reproduce a similar effect of the Cooperativity signature on the probability to stay on the same target. This confirms the validity and specificity of the winning model (**Fig.2.c**) (for behavior generated by the cooperative expert alone or competitive expert alone, see supplementary information).

While our results suggest that the mixed intentions model best predicts and reproduces behavior of participants. It is possible that the tracking was not dynamic across trials but fixed throughout the entire experiment (for example with the same ratio of competitive/cooperative numbers of trial). To test this hypothesis, of a fixed arbitrator deciding between the two experts, we also tested a model using a static free parameter arbitrating between the two experts. The results of the Bayesian model selection still assigned the "mixed intentions" model to most participants, even after addition of this free parameter, demonstrating the importance of the dynamic aspect of the tracking. Finally, because only the second order term of the influence model differentiates the two modes of interactions, we tested the contribution of the mentalizing term to the fit this parameter was removed. This led to a decrease of 5.39% (95% confident interval [2.47;-8.31]) in the log-likelihood, indicating the importance of this mentalizing term.

An important parameter of the model is the competitive bias ( $\delta$ ), representing the participant's prior on the other's intentions. We observed a significant bias ( $\delta$ ) towards the competitive framework (mean=0.2247,  $p=0.0181$ ). This competitive bias, of the Influence Mixed Intention Model has an observable behavioral effect: a higher competitive bias increases the randomness of switching targets between two consecutive trials, therefore making participants behavior more unpredictable ( $\frac{d(\text{switch})}{d(\delta)} = 0.13$ , CI=[0.06; 0.20],  $p<0.0005$ ,  $\chi^2$  test when  $\delta$  is included in the logistic regression on the probability to stay).

## Supplementary fMRI analyses

It is possible that the observed DLPFC/IPS region activations are only due to the difference in behavior of the artificial agent between the competitive and cooperative blocks. To test this hypothesis, we conducted another GLM (GLM3), separating trials according to the cooperative blocks vs competitive blocks (and not according to classification of trials by the controller). We observed no difference in brain activation, even at a lower threshold ( $p>0.01$ ), indicating that the PE difference observed between the trials classified as competitive by the controller and the trials classified as cooperative were due to the effective tracking of attribution of intentions, and not to PE differences between the two (unsigned) types of trial blocks. Direct paired t-test comparison between PE for trials classified as competitive > cooperative (i.e.  $(\Delta>0)>(\Delta<0)$ ) and PE from the comparison between blocks of competitive and cooperative trials showed the engagement of the same regions (right angular gyrus  $x,y,z=48,-50,32$ , right dIPFC

x,y,z= 30,8,53 and left angular gyrus x,y,z=-33,-59,44 (p<0.05 FWE). The left dIPFC was also engaged at a lower threshold (x,y,z=-39,-2,26; p<0.01 FWE threshold at p<0.005).

Because the activities we observed could be due to more volatility in the rewarded target, we ran a fourth GLM (GLM4) to control for the volatility of the Artificial Agent. Indeed, the probability that the computer switches target was 11% higher in competitive than in cooperative trials (p<0.0001, 95%CI [8.8%; 14%]). However, when adding the trial by trial probability, that the artificial agent switches target as a non-orthogonalized regressor, the right dIPFC (x,y,z=30,9,42) and angular gyrus (x,y,z=51,-50,33) were still more positively correlated with PE in trials classified as competitive by the controller compared to those classified as cooperative. This result indicates that those activations are not due to the volatility of the competitive condition.

Finally, we searched for brain regions computing the expected value for staying on the same target. To do this, we ran a GLM (GLM5) containing the expected reward for staying on the same target as the only parametric regressor at the time of choice. We found that the ventral striatum coded positively the expected reward for staying on the same target (x,y,z=14, 11,-2; p < 0.05, FWE, **see SI**).

## Reinforcement learning

Reinforcement learning (RL) consisted of directly linking action or state and outcome to predict future rewards after performing a particular action or being in a particular state. In our experiment, we updated action value with the Rescola-Wagner rule:

$$V_{t+1}^a = V_t^a + \alpha * \delta_t$$

Eq 1

$$\delta_t = R_t - V_t^a$$

Eq 2

Where  $\alpha$  is the learning rate. The reward prediction error  $\delta_t$  is defined as the difference between the reward at trial  $t$ ,  $R_t$  and the expected value of the choice  $a$  at trial  $t$ ,  $V_t^a$ . Then the probability to choose action  $a$  is :

$$p^a = s(V^a - V^b)$$

Eq 3

With  $s(z) = 1/(1 + e^{-\beta z})$  the sigmoid function when  $\beta$  is a free parameter to capture the stochasticity of the participant's behavior (i.e. the exploration/exploitation trade-off). We defined the probability to stay as  $\begin{cases} p^{red} & \text{if previous choice was red} \\ p^{black} & \text{if previous choice was black} \end{cases}$ . We used the same definition of the probability to stay for other models.

### Fictitious play

In game theory, one can infer the probability that the other choose one particular action and choose one's own action to maximize one's expected reward. This model is called a first order fictitious play model. Thus, the opponent's probability  $p^*$  of choosing an action  $a$  is dynamically updated by tracking the choice history of the opponent:

$$p_{t+1}^* = p_t^* + \eta * \delta_t^p$$

Eq 4

$$\delta_t^p = P_t - p_t^*$$

Eq 5

Where  $\eta$  is the learning rate. The reward prediction error  $\delta_t^p$  is defined as the difference between the expected action of the opponent at trial  $t$ ,  $p_t^*$ , and the actual other's choice on trial  $t$ , ( $P_t = 1$ ) if the other's action is  $a$  and ( $P_t = 0$ ) if it is  $b$ . Then the probability to choose action  $a$  depends on the payoff matrix. In the competitive setting of our game we can derive the probability  $q$  that the participant chooses action  $a = \text{"Red card"}$  using the sigmoid function, the payoff matrix, and the probability that the other chooses action  $a = \text{"Red card"}$ :

$$p = s(2q^* - 1) \text{ Eq 6}$$

$p^*$  is the inferred probability that the other chooses  $a = \text{"Red card"}$ . Because the payoff matrix is the same for the participant in both competitive and cooperative trials, the mode of



interaction has no impact on the decision stage. However, considering the other's decision rule, it would be different in the competitive or cooperative modes:

$$\begin{aligned} q &= s(2p^* - 1) \quad \text{in cooperative mode} \\ q &= s(1 - 2p^*) \quad \text{in competitive mode} \end{aligned}$$

Eq 7

## Influence model

Another strategy could be to take into account how one's own actions influence the other's future actions. Thus, to compute the probability of updating of the other's strategy, we replaced update of opponent strategy (Eq. 4) in the player decision rule (Eq. 7). Then with a Taylor expansion taking  $\eta$  close to 0, we added the influence terms ( $\Delta p$  : influence update signal of the participant,  $\Delta q$  : influence update signal of the other):

$$\Delta q \approx + \eta 2\beta q_t(1 - q_t)(P_t - p_t^*)$$

Eq 8

$$\begin{aligned} \Delta p &\approx + \eta 2\beta p_t(1 - p_t)(Q_t - Q_t^*) \quad \text{in cooperative} \\ \Delta p &\approx - \eta 2\beta p_t(1 - p_t)(Q_t - Q_t^*) \quad \text{in competitive} \end{aligned}$$

Eq 9

Thus, in the competitive mode, there is only a sign difference between the term of influence of the two players which is not the case in the cooperative setting. A player can thus incorporate the influence of his/her action on the strategy of the other player:

$$\begin{aligned} p_{t+1}^* &= p_t^* + \eta_1(P_t - p_t^*) + \eta_2 2\beta p_t^*(1 - p_t^*)(Q_t - q_t^{**}) \quad \text{in cooperative} \\ p_{t+1}^* &= p_t^* + \eta_1(P_t - p_t^*) - \eta_2 2\beta p_t^*(1 - p_t^*)(Q_t - q_t^{**}) \quad \text{in competitive} \end{aligned}$$

Eq 10

$$q_{t+1}^* = q_t^* + \eta_1(Q_t - q_t^*) + \eta_2 k_1 2\beta q_t^*(1 - q_t^*)(P_t - p_t^{**})$$

Eq 11

*The Influence model update rules.  $p_t^*$  is the predicted opponent strategy.  $P_t$  is the opponent choice and then  $(P_t - p_t^*)$  is the action prediction error. The influence update is due to the  $(Q_t - q_t^{**})$  term. Thus  $Q_t$  is the player's own*

action and  $q_t^{**}$  the inferred probabilities that the opponent has of the player themselves (second-order beliefs). Thus, in the cooperative and competitive modes the influence will occur in the opposite directions. In the Mixed Intention Influence Model, we decline  $p_{t+1}^*$  in  $p_{t+1}^{coop*}$  and  $p_{t+1}^{comp*}$ .

To compute the  $p_t^{**}$  and  $q_t^{**}$ , we invert the decision function (Eq. 7):

$$\begin{aligned} q_t^{**} &= \frac{1}{2} - \frac{1}{2\beta} \ln\left(\frac{1-p^*}{p^*}\right) \\ p_t^{**} &= \frac{1}{2} + \frac{1}{2\beta} \ln\left(\frac{1-p^*}{p^*}\right) \quad \text{in competitive} \\ p_t^{**} &= \frac{1}{2} - \frac{1}{2\beta} \ln\left(\frac{1-q^*}{q^*}\right) \quad \text{in cooperative} \end{aligned}$$

Eq 12

We define the probability to stay as  $\begin{cases} p^{red} & \text{if previous choice was red} \\ p^{black} & \text{if previous choice was black} \end{cases}$

## k-ToM model

The k-ToM model is defined as in <sup>13</sup>. An economic game under game theory is defined by a utility table  $U(a^{self}, a^{other})$  to represent the payoff to players according to the actions of self, ( $a^{self}$ ) and the other player ( $a^{other}$ ). In our experiment, this utility table varies between competitive and cooperative blocks (see Figure 1.A.). Because participants make a binary choice,  $a^{self}$  and  $a^{other}$  take the value of  $a=0$  for the red option and  $a=1$  for the black option. According to Bayesian decision theory, agents try to maximize their expected value  $V = E[U(a^{self}, a^{other})]$ . We assume that agents use a softmax function as a decision rule:

$$P(a^{self} = 1) = s\left(\frac{V^1 - V^2}{\beta}\right)$$

Eq 13

$P(a^{self} = 1)$  is the probability that the agent choose option  $a^{self} = 1$ .  $s$  is the sigmoid function and  $\beta$  is a free parameter called inverse temperature and controls for the magnitude of behavioral noise. The value of each action depends on the probability of other's action with  $p^{other} = P(a^{other} = 1)$  and the utility table  $U(a^{self}, a^{other})$  :

$$V^i = p^{other} * U(a^{self} = i, a^{other} = 1) + (1 - p^{other}) * U(a^{self} = i, a^{other} = 0)$$

Eq 14

One key hypothesis of this model is that we consider that the other agent is itself a k-ToM agent. It means that the other agent has the same decision policy as equation 13. Thus, while the agent tracks  $p^{other}$ , the other track  $p^{self}$  to construct a recursive reasoning. This recursion induces distinct levels of ToM sophistication between the two agents, impacting how agents update their subjective prediction of  $p^{other}$ .<sup>13</sup> k-ToM agents are defined according to the way they update this prediction of  $p^{other}$  starting from 0-ToM. Definition of higher level of reasoning is based on the level 0, for which  $P(a^{other} = 1) = s(x_t^0)$ , with the log-odd  $x_t^0$  varying with a volatility  $\sigma^0$ . The updating rule for the hidden state  $x_t^0$  follows the Bayes-optimal probabilistic scheme :

$$q(x_{t+1}^0) \propto p(a_{t+1}^0 | x_{t+1}^0) \int q(x_t^0) p(x_{t+1}^0 | x_t^0) dx_t^0$$

Eq 15

With  $p(x_{t+1}^0 | x_t^0)$  the 0-ToM's prior belief on the volatility of the log-odd, and  $q(x_t^0) \equiv p(x_t^0 | a_{1:t}^{other})$ , the posterior belief about the log-odds  $x_t^0$  at trial t after the observation of all previous actions  $a^{other}$ . Thus, one can derive the 0-ToM's learning rule:

$$\hat{p}_{t+1}^{other} \approx s \left( \frac{\mu_t^0}{\sqrt{1 + \frac{3(\Sigma_t^0 + \sigma^0)}{\pi^2}}} \right)$$

Eq 16

$$\mu_t^0 \approx \mu_{t-1}^0 + \Sigma_t^0 (a_t^{other} - s(\mu_{t-1}^0))$$

Eq 17

$$\Sigma_t^0 \approx \frac{1}{\frac{1}{\Sigma_{t-1}^0 + \sigma^0} + s(\mu_{t-1}^0)(1-s(\mu_{t-1}^0))}$$

Eq 18

Where  $\mu_t^0$  is the approximate mean of 0-ToM posterior distribution of  $q(x^0)$  and  $\Sigma_t^0$  is it's approximate variance. Thus  $\mu_t^0$  is the estimate of the 0-ToM log-odds at trial t and  $\Sigma_t^0$  her subjective uncertainty about it.

A 1-ToM agent assumes that the other agent reasons with a 0 depth ToM. Thus, with the decision policy of a 0-ToM agent we can construct a 1-ToM agent. More specifically, in combining equation 13, 14 and 15, 1-ToM agent assumes that the probability for a 0-ToM agent to emit action  $a^{other} = 1$  is  $p^{other} = s \circ v^1(x_t^1)$  (we use the symbol  $\circ$  to refer to the composition of two functions defined as  $(g \circ f)(x) = g(f(x))$ ) with  $s$  the sigmoid function and  $v^1$  the value for 0-ToM agent to choose option 1 :

$$v^1(x_t^1) = \frac{p_t^{self} * \Delta U_t^1 + (1 - p_t^{self}) * \Delta U_t^0}{\beta_t}$$

Eq 19

With  $\Delta U_t^i = U(a^{self} = i, a^{other} = 1) - U(a^{self} = i, a^{other} = 0)$  which represents the incentive for the 1-ToM agent to choose option one if 1-ToM agent chooses option  $a^{self} = i$ .  $p_t^{self}$  for a 1-ToM agent is the same as  $p_t^{other}$  for 0-ToM agent, thus :

$$p_t^{self} \approx s \left( \frac{\mu_{t-1}^0}{\sqrt{1 + \frac{3(\Sigma_{t-1}^0 + \sigma_t^0)}{\pi^2}}} \right)$$

Eq 20

To let the 1-ToM agent eventually learn how 0-ToM agent learns about herself, and act in consequence, the 1-ToM agent assumes that hidden states  $x_t^1$  vary across trials with volatility  $\sigma^1$ , which leads to a meta learning rule similar to equation 16, 17, 18.

In a more general fashion, an agent of depth  $k \geq 2$  considers that the other agent is a lower sophistication  $\kappa$ -ToM agent ( $\kappa \leq k$ ), but this sophistication has to be learned in addition to the hidden states  $x^k$  that track the opponent's learning and decision making. Thus a  $k$ -ToM agent continuously tracks all possible other's sophistication levels and it's associate action probability  $p^{other, \kappa} = s \circ v^\kappa(x^k)$  that he will choose  $a^{other} = 1$ .

$$p_t^{other} = \sum_{l < \kappa} \lambda_t^{k, \kappa} * p_t^{other, \kappa}$$

Eq 21

$$p_t^{other} \approx s \circ \tilde{v}^\kappa(\mu_{t-1}^{k,\kappa}, \Sigma_{t-1}^{k,\kappa})$$

Eq 22

$$\lambda_t^{k,\kappa} \approx \left[ \frac{\lambda_{t-1}^{k,\kappa} * p_t^{other,\kappa}}{\sum_{\kappa' < \kappa} \lambda_{t-1}^{k,\kappa'} * p_t^{other,\kappa'}} \right]^{a_t^{other}} \left[ \frac{\lambda_{t-1}^{k,\kappa} * (1 - p_t^{other,\kappa})}{\sum_{\kappa' < \kappa} \lambda_{t-1}^{k,\kappa'} * (1 - p_t^{other,\kappa'})} \right]^{1 - a_t^{other}}$$

Eq 23

$$\mu_t^{k,\kappa} \approx \mu_{t-1}^{k,\kappa} + \lambda_t^k \sum_t^{k,\kappa} W_{t-1}^\kappa (a_t^{other} - s \circ v^\kappa(\mu_{t-1}^{k,\kappa}))$$

Eq 24

$$\Sigma_l^{k,\kappa} \approx \left[ (\Sigma_{l-1}^{k,\kappa} + \sigma^k)^{-1} + s' \circ v^k(\mu_{l-1}^{k,\kappa}) \lambda_l^\kappa W_{l-1}^\kappa{}^T W_{l-1}^\kappa \right]^{-1}$$

Eq 25

Where  $\lambda_t^k$  is k-ToM's probability that her opponent is  $\kappa$ -ToM,  $W^\kappa$  is the gradient of  $v^\kappa$  with respect to the hidden states  $x^\kappa$ . Here,  $v^k$  is obtained by the recursive injections of equation 5 in equation 1, as we have already done to obtain equation 4.  $\tilde{v}^\kappa$  is defined in terms of the expectation operator :  $E[s \circ v^\kappa(\mu_{t-1}^{k,\kappa}, \Sigma_{t-1}^{k,\kappa})] = s \circ \tilde{v}^\kappa(\mu_{t-1}^{k,\kappa}, \Sigma_{t-1}^{k,\kappa})$ . Equation 3 and 5 have been estimated using a Variational approach to approximate Bayesian inference.

## Two experts model

For models using the payoff matrix to update hidden states, making a difference between the cooperative and competitive modes (i.e. k-ToM and the Influence model), we fitted three models in different settings: competitive, cooperative or mixed intentions. For the mixed intention setting, we ran the competitive and cooperative models in parallel, generating a probability that the other will choose option  $a$  for each possible mode of interactions,  $P_{comp}^a$  and  $P_{coop}^a$ . We then transformed the probability with the sigmoid function to have values ranging from  $-\infty$  to  $+\infty$ . Because we have a binomial choice configuration,  $V^a = -V^b$  in both competitive and cooperative settings. Thus, as  $V_i^a$  and  $V_i^b$  get close to zero, uncertainty for  $i$  the other's intention, increases. We defined the reliability of the intention  $i$  as the absolute value of  $V_i^a$  and the

probability that the other intention is cooperative as the sigmoid function of the difference in reliability between the two modes :

$$P_{coop}^t = \frac{1}{1 + e^{\beta(|v_{coop}^t| - |v_{comp}^t| - \delta)}}$$

Eq 26

where  $\beta$  is the inverse temperature controlling for the stochasticity of the mode of interaction and  $\delta$  is the bias towards cooperative mode. Then, with  $P_{comp}^{a,t}$  and  $P_{coop}^{a,t}$  we computed the decision value given the competitive and cooperative payoff matrix,  $DV_{comp}^a$  and  $DV_{coop}^a$  respectively and weighted them by the probability of the corresponding mode of interaction to compute the total decision value :

$$DV^t = P_{coop}^t * DV_{coop}^a + (1 - P_{coop}^t) * DV_{comp}^a$$

Eq 27

The sigmoid function  $s$  generated the probability of selecting choice  $a$  at trial  $t$ :

$$p^{a,t} = s(DV^t)$$

Eq 28

Finally, the reward prediction error was defined as the reward at trial  $t$  for action  $a$ :

$$PE = R^{a,t} - p^{a,t}$$

Eq 29

## Active inference model

For this model based on <sup>64</sup>, we adopted the partially observable Markov decision process (POMDP) framework which is a way of describing transitions among states under the hypothesis that probability of the next state depends only on the current state. The partially observed aspect of the Markovian process means that states are not directly observable and have to be inferred through a set of (noisy) observations.

Active inferences are composed of a tuple (P, Q, R, S, A, U,  $\Omega$ ) :

- $\Omega$  is a finite set of possible observations
- A is a finite set of possible action
- S is a finite set of hidden states
- U is a finite set of control states
- R is the *generative process* over observation  $\tilde{o} \in \Omega$ , hidden states  $\tilde{s} \in S$ , and action  $\tilde{a} \in A$

$$R(\tilde{o}, \tilde{s}, \tilde{a}) = Pr(\{o_0, \dots, o_T\} = \tilde{o}, \{s_0, \dots, s_T\} = \tilde{s}, \{a_0, \dots, a_{T-1}\} = \tilde{a})$$

- P is the *generative model* over observation  $\tilde{o} \in \Omega$ , hidden states  $\tilde{s} \in S$ , and control states  $\tilde{u} \in U$

$$P(\tilde{o}, \tilde{s}, \tilde{u} | m) = Pr(\{o_0, \dots, o_T\} = \tilde{o}, \{s_0, \dots, s_T\} = \tilde{s}, \{u_0, \dots, u_T\} = \tilde{u}) \quad \text{with}$$

parameters  $\theta$ .

- Q is the *approximate posterior* over hidden and control states

$$Q(\tilde{s}, \tilde{u}) = Pr(\{s_0, \dots, s_T\} = \tilde{s}, \{u_0, \dots, u_T\} = \tilde{u}) \quad \text{with parameters or expectation } (\hat{s}, \hat{\pi}), \text{ where } \pi \in \{1, \dots, K\} \text{ is a policy that indexes a sequence of control states } \tilde{u} | \pi = (u_t, \dots, u_T | \pi)$$

Firstly, *generative process* describes the transition probabilities among hidden states which generate observations. Transition probabilities depend on actions which are sampled from *approximate posterior* belief about control states. Belief is formed using the *generative model* (denoted by  $m$ ) of how observation are generated by hidden states. The *Generative model* encodes belief and hidden states of the agent in term of expectation.

The active inference model assumes that both action and expectation minimize the free energy of observations. That is, expectation minimizes free energy and expectation of control states prescribes action in each trial.

$$(\hat{s}, \hat{\pi}) = \operatorname{argmin} F(\tilde{o}, \hat{s}, \hat{\pi})$$

Eq 30

$$Pr(a_t = u_t) = Q(u_t | \hat{\pi}^*)$$

Eq 31

With :

$$\begin{aligned} F(\hat{o}, \hat{s}, \hat{\pi}) &= E_Q[-\ln P(\tilde{o}, \tilde{s}, \tilde{u}|m)] - H[Q(\tilde{s}, \tilde{u})] \\ &= -\ln P(\tilde{o}|m) + D_{KL}[Q(\tilde{s}, \tilde{u}) || P(\tilde{s}, \tilde{u}|\tilde{o})] \end{aligned}$$

Eq 32

The generative mode could be defined as three marginal distributions:

$$P(\tilde{o}, \tilde{s}, \tilde{u}|m) = P(\tilde{o}|\tilde{s}) P(\tilde{s}|\tilde{u}) P(\tilde{u}|m)$$

Eq 33

Thus, heuristically, the decision consists firstly of figuring out which state is the most likely by optimizing its expectation according to free energy and the generative model. Then, after optimizing its posterior beliefs, an action is sampled from the posterior probability distribution over the control state. The environment generates a new observation given the selected action using the generative process and a new decision cycle begins.

For our experimental design, we have 8 possible observations:

$\Omega = \{ \text{"Previous black loose"}; \text{"Previous black win"}; \text{"Previous red loose"}; \text{"Previous red win"}; \text{"Current black win"}; \text{"Current black loose"}; \text{"Current red win"}; \text{"Current red lose"} \}$

We defined 20 hidden states:

$S = \{ \text{"Previous choice"} \times \text{"previous reward"} \times \text{"current correct answer"} \times \text{"current mode of interaction"} ; \text{"Current black win"}; \text{"Current black loose"}; \text{"Current red win"}; \text{"Current red lose"} \}$

The finite set of action is  $A = \{ \text{"Choose red"}; \text{"Choose black"} \}$  bringing the agent from the 16 first possible hidden states to their corresponding 4 last hidden states which are  $\{ \text{"Current black win"}; \text{"Current black loose"}; \text{"Current red win"}; \text{"Current red lose"} \}$ .

Log prior preferences over the observed states are  $C = [ 0 ; 1 ; 0 ; 1 ; 2 ; -1 ; 2 ; -1 ]$  meaning that the agent prefers to observe, in decreasing order, a current victory, then a previous victory, a previous defeat and finally a current defeat.



For each trial, prior beliefs about hidden states are equally spread between the four hidden states composed by {"Previous choice" x "previous reward"} leaving unknown the "current mode of interaction" and the "current good answer".

To allow the agent to learn about the hidden state "current mode of interaction" we add concentration parameters about observation. This is prior about what hidden state leads to what observation, and can be viewed as the number of occurrences encountered in the past. We arbitrarily set this number to 2 for being in "cooperative" mode, when observing a previous victory, and 1 for being in the hidden state "competitive". Inversely for a previous defeat, we set this parameter at 1 for being in the "cooperative" hidden state and 2 for being in the "competitive" state.

## Hierarchical Gaussian Filter

The Hierarchical Gaussian Filter was constructed to model the agent's learning under a volatile environment <sup>40,70</sup>. We are interested in a binary state  $x_1^t$  of the environment at time  $t$  (for convenience we will often omit the time index  $k$ ) :

$x_1^k \in \{ \text{"Red card is the good answer"} ; \text{"Black card is the good answer"} \}$  is causing sensory input  $u$ . Thus, we assume the following form of likelihood function:

$$p(u|x_1) = (u)^{x_1}(1 - u)^{1-x_1}$$

Eq 34

Because  $x_1$  is binary, it could be described by a single real number,  $x_2$ , the state at the next level of hierarchy. We then define the conditional prior density to map  $x_2$  to the probability  $x_1$  as a Bernnouilli law of parameter  $s(x_2)$ :

$$p(x_1|x_2) = s(x_2)^{x_1}(1 - s(x_2))^{1-x_1}$$

Eq 35

Where  $s(x) = \frac{1}{1+e^{-x}}$  is the sigmoid function. This prior density gives us  $x_1 = 1$  and  $x_1 = 0$  equally probable for  $x_2 = 0$  and for  $x_2 \rightarrow +\infty$  or  $-\infty$  we have respectively  $x_1 \rightarrow 1$  or  $0$ . Thus,  $x_2$  is an unbounded parameter of the probability that  $x_1 = 1$ . In our example, a higher  $x_2$

corresponds to a strong tendency for the red target to be the good answer. The only hypothesis on  $x_2$  is that it evolves with time as a Gaussian random walk.

$$p\left(x_2^{(k)} | x_2^{(k-1)}, x_3^{(t)}\right) = N\left(x_2^{(k)}; x_2^{(k-1)}, \exp\left(\kappa x_3^{(k)} + \omega\right)\right)$$

Eq 36

Where  $\omega$  and  $\kappa$  are two free parameters corresponding to the dispersion of the random walk.  $x_3^{(k)}$  represents the log-volatility of the environment, meaning the tendency of the red target to be the good answer, and follows a Gaussian random walk.  $\omega$  represents the volatility independent of the state  $x_3$ . Then we can apply the same approach to  $x_3$  as we do to  $x_2$ , and so forth, to add as many levels as desired. Here we stop at the fourth level introducing a new free parameter representing the volatility of  $x_3$  :

$$p\left(x_3^{(k)} | x_3^{(k-1)}, \vartheta\right) = N\left(x_3^{(k)}; x_3^{(k-1)}, \vartheta\right)$$

Eq 37

Then, using the Variational inversion method explained in <sup>40</sup>, we can inverse the model to update hidden states given a regular sensory entry  $u^{(k)}$ . The approximation of the inversion assumes Gaussian posteriors at all levels, with means  $\mu_i$  and precision (inverse of variance)  $\pi_i$  :

$$x_i^{(k)} | u^{(1)}, \dots, u^{(k)}, \chi \sim N\left(\mu_i^{(k)}, \left(\pi_i^{(k)}\right)^{-1}\right)$$

Eq 38

with  $\chi$  the set of all free parameters. Thus parameters  $\mu_i$  and  $\pi_i$  are the sufficient statistic, to be updated after each input  $u$  as follows:

$$\mu_i^{(k)} = \hat{\mu}_i^{(k)} + \frac{1}{2} \kappa_{i-1} v_{i-1}^{(k)} \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)}$$

Eq 39

$$\pi_i^{(k)} = \hat{\pi}_i^{(k)} + \frac{1}{2} \left( \kappa_{i-1} v_{i-1}^{(k)} \hat{\pi}_{i-1}^{(k)} \right)^2 \left( 1 + \left( 1 - \frac{1}{v_{i-1}^{(k)} \pi_{i-1}^{(k-1)}} \right) \delta_{i-1}^{(k)} \right)$$

Eq 40

With

$$v_i^{(k)} = \begin{cases} t^{(k)} \exp \left( \kappa_i \mu_{i+1}^{(k-1)} + \omega_i \right), & i = 1, \dots, n-1 \\ t^{(k)}, & i = n \end{cases}$$

Eq 41

$$\hat{\mu}_i^{(k)} = \mu_i^{(k-1)} \text{ by definition}$$

$$\hat{\pi}_i^{(k)} = \frac{1}{\sigma_i^{(k-1)} + v_i^{(k)}} \text{ by definition}$$

$$\delta_i^{(k)} = \frac{\sigma_i^{(k)} + \left( \mu_i^{(k)} - \hat{\mu}_i^{(k)} \right)^2}{\sigma_i^{(k-1)} + v_i^{(k)}} \text{ by definition}$$

## Bayesian sequence learner

The n-BSL (Bayesian sequence learner) is a model which tracks probabilities of a certain outcome “a” given the previous  $n$  outcomes as a Gaussian function:

$$P(a = \text{"Red target is the good answer"} | S_i) = N(\mu_i, \sigma_i)$$

With  $S_i$  the sequence of the  $n$  last outcomes “a”. For each observation at time  $t$ , the update is a Laplace-Kalman rule:

$$\sigma_i^{t+1} = \frac{1}{\frac{1}{\sigma_i^t + \Omega} + s(\mu_i^t) * (1 - s(\mu_i^t))}$$

Eq 42

$$\mu_i^{t+1} = \mu_i^t + (\sigma_i^{t+1} + \Omega) * (a^t - s(\mu_i^t))$$

Eq 43

With  $a^t = 1$  if "Red target is the good answer" and  $a^t = 0$  if "Black target is the good answer",  $s(x) = \frac{1}{1+e^{-x}}$ , and  $\Omega$  is a free parameter representing prior volatility.

## Win-Stay / Lose-Switch model

This model reproduces a heuristic behavior, precisely "I keep the same option if I just won, I switch if I just lost". To implement that, we use two pseudo Q-values,  $V^{stay} = 1$  for the action of stay and  $V^{switch} = -1$  for the action of switch. Then we use the sigmoid function to compute the probability of choosing the same option as the previous trial:

$$p^{stay} = s(V^{stay} - V^{switch})$$

Eq 44

## Acknowledgements

This research has benefited from the financial support of IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche and NSF in the CRCNS program to JCD (ANR n°16-NEUC-0003-01), and by CRCNS NIMH grant no. 5R01MH112166-03, NSF grant no. EEC-1028725, and a Templeton World Charity Foundation grant to RPNR.. We thank the CERMEP Staff for help during scanning and Pr Ed Derrington for proof reading the draft of the manuscript.

## Author contributions

J.-C.D., R.P., R. P.N.R, and D.L. developed the general concept, experiment and models. R.P. designed and programmed the task and ran the experiment under the supervision of J.-C.D. R.P. developed the model and implemented the algorithms under the supervision of J.-C.D., and analyzed the data in collaboration with J.-C.D. R.P. and J.-C.D. wrote the manuscript in collaboration with K.K., D.L and R.P.N.R.



## Chapter 3

# Modeling other minds: Bayesian inference explains human choices in group decision-making<sup>1</sup>

---

<sup>1</sup>Ce chapitre est un travail en collaboration avec Koosha Khalvati, Seongmin A. Park, Saghar Mirbagheri, Mariateresa Sestito, Jean-Claude Dreher and Rajesh P.N.Rao.  
Mon travail a essentiellement consisté à aider pour la modélisation et l'écriture du manuscrit.

## COGNITIVE NEUROSCIENCE

# Modeling other minds: Bayesian inference explains human choices in group decision-making

Koosha Khalvati<sup>1</sup>, Seongmin A. Park<sup>2,3</sup>, Saghar Mirbagheri<sup>4</sup>, Remi Philippe<sup>3</sup>, Mariateresa Sestito<sup>3</sup>, Jean-Claude Dreher<sup>3\*</sup>, Rajesh P. N. Rao<sup>1,5\*†</sup>

To make decisions in a social context, humans have to predict the behavior of others, an ability that is thought to rely on having a model of other minds known as “theory of mind.” Such a model becomes especially complex when the number of people one simultaneously interacts with is large and actions are anonymous. Here, we present results from a group decision-making task known as the volunteer’s dilemma and demonstrate that a Bayesian model based on partially observable Markov decision processes outperforms existing models in quantitatively predicting human behavior and outcomes of group interactions. Our results suggest that in decision-making tasks involving large groups with anonymous members, humans use Bayesian inference to model the “mind of the group,” making predictions of others’ decisions while also simulating the effects of their own actions on the group’s dynamics in the future.

## INTRODUCTION

The importance of social decision-making in human behavior has spawned a large body of research in social neuroscience and decision-making (1, 2). Human behavior relies heavily on predicting future states of the environment under uncertainty and choosing appropriate actions to achieve a goal. In a social context, the degree of uncertainty about the possible outcomes increases drastically as the behavior of others is much less predictable than the physics of the environment.

One approach to handling uncertainty in social settings is to act based on a belief about others. This approach includes inferring the consequences of one’s own behavior under uncertainty as opposed to “belief-free” models (3) that simply select the action that has been rewarding in the past, given current observations (4, 5). The difference between “belief-based” and belief-free models in social decision-making is closely related to “model-based” and “model-free” approaches (6, 7) in nonsocial decision-making but with a greater emphasis on uncertainty due to the greater unpredictability of human behavior in social tasks.

In belief-based decision-making, the subject learns a model of the environment, updates the model based on observations and rewards, and chooses actions based on a probabilistic “belief” about the current state of the world (5, 8, 9). As a result, the relationship of the current action with rewards received and current observations is indirect. Besides the history of rewards received and the current observation, the learned model can also include other factors such as potential future rewards and more general rules about the environment. Therefore, the belief-based (model-based) approach is more flexible than belief-free (model-free) decision-making (10, 11). However, belief-based decision-making requires more cognitive resources, for example, for simulation of future events. Thus, there is an inherent trade-off between the two types of approaches, and determining

which approach humans adopt for different situations is an important open area of research (12).

Several studies have presented evidence in favor of the belief-based approach by quantifying the similarity between probabilistic model-based methods and human behavior when the subject interacts with or reasons about another human (5, 13–18). Compared to reasoning about a single person, decision-making in a group with a large number of members can get complicated. On the one hand, having more group members disproportionately increases the cognitive cost of tracking minds compared to the cost of only tracking the reward history of each action given the current observations. On the other hand, consistent with the importance that human society places on group decisions, a belief-based approach might be the optimal strategy.

How does one extend a belief-based approach for reasoning about a single person to the case of decision-making within a large group? Group decision-making becomes even more challenging when the actions of others in the group are anonymous (e.g., voting as part of a jury) (19, 20). In such situations, reasoning about the state of mind of individual group members is not possible but the dynamics of group decisions do depend on each individual’s actions.

To investigate these complexities that arise in group decision-making, we focused on the volunteer’s dilemma task, wherein a few individuals endure some costs to benefit the whole group (21). Examples of the task include guarding duty, blood donation, and stepping forward to stop an act of violence in a public place (22). To mimic the volunteer’s dilemma in a laboratory setting, we used the thresholded binary version of a multiround public goods game (PGG) where the actions of each individual are hidden from others (21, 23).

Using an optimal Bayesian framework based on partially observable Markov decision processes (POMDPs) (24), we propose that in group decision-making, humans simulate the “mind of the group” by modeling an average group member’s mind when making their current choices. Our model incorporates prior knowledge, current observations, and a simulation of the future based on the current actions for modeling human decisions within a group. We compared our model to a model-free reinforcement learning approach based on the reward history of each action as well as a previous descriptive method for fitting human behavior in the PGG. Our model predicts

<sup>1</sup>Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. <sup>2</sup>Center for Mind and Brain, University of California, Davis, CA, USA. <sup>3</sup>Neuroeconomics Laboratory, Institut des Sciences Cognitives Marc Jeannerod, Lyon, France. <sup>4</sup>Department of Psychology, New York University, New York, NY, USA. <sup>5</sup>Center for Neurotechnology, University of Washington, Seattle, WA, USA. \*Joint senior authors. †Corresponding author. Email: rao@cs.washington.edu

human behavior significantly better than the model-free reinforcement learning and descriptive approaches. Furthermore, by leveraging the interpretable nature of our model, we are able to show a potential underlying computational mechanism for the group decision-making process.

**RESULTS**

**Human behavior in a binary PGG**

The participants were 29 adults (mean age, 22.97 years old ± 0.37; 14 women). We analyzed the behavioral data of 12 PGGs in which participants played 15 rounds of the game within the same group of *N* players (*N* = 5).

At the beginning of each round, 1 monetary unit (MU) was endowed (E) to each player. In each round, a player could choose between two options: contribute or free-ride. Contribution had a cost of *C* = 1 MU, implying that the player could choose between keeping their initial endowment or giving it up. In contrast to the classical PGG where the group reward is a linear function of total contributions (25), in our PGG, public goods were produced as a group reward (*G* = 2 MU to each player) if and only if at least *k* players each contributed 1 MU. *k* was set to two or four randomly for each session and conveyed to group members before the start of the session. The resultant amount after one round is therefore *E* - *C* + *G* = 2 MU for the contributor and *E* + *G* = 3 MU for the free-rider when public goods were produced (the round was a SUCCESS). On the other hand, the contributor has *E* - *C* = 0 MU and the free-rider has *E* = 1 MU when no public goods were produced (the round was a FAILURE).

Figure 1 depicts one round of the PGG task. After the subject acts, the total number of contributions, free-rides, and the overall outcome of the round is revealed (success or failure in securing the 2 MU group reward), but each individual player's actions remained unknown. In addition, as shown in the figure, the value of *k* for the current session was always presented on the screen to ensure that the subjects had it in mind when making decisions. Although subjects were told that they were playing with other humans, in reality, they were playing with a computer that generated the actions of all the other *N* - 1 = 4 players using an algorithm based on human data (see Methods). In each session, the subject played with a different group of players.

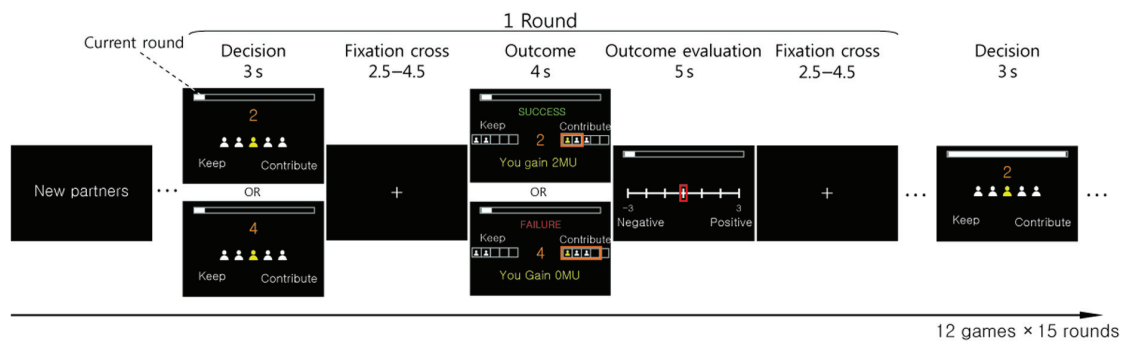
As shown in Fig. 2A, subjects contributed significantly more when the number of required volunteers was higher with an average contribution rate of 55% (SD = 0.31) for *k* = 4 in comparison to 33% (SD = 0.18) for *k* = 2 {two-tailed paired sample *t* test, *t*(28) = 3.94, *P* = 5.0 × 10<sup>-4</sup>, 95% confidence interval (CI) difference = [0.11,0.33]}. In addition, Fig. 2B shows that the probability of generating public good was significantly higher when *k* = 2 with a success rate of 87% (SD = 0.09) compared to 36% (SD = 0.29) when *k* = 4 {two-tailed paired sample *t* test, *t*(28) = 10.08, *P* = 8.0 × 10<sup>-11</sup>, 95% CI difference = [0.40,0.60]}. All but six of the subjects contributed more when *k* = 4 (Fig. 2C). Of these six players, five chose to free-ride more than 95% of the time. In addition, success rate was higher when *k* = 2 for all players (Fig. 2D).

The contribution rate of the subjects dropped during the course of the trial on average, especially for *k* = 4, but remained above zero. Figure 2E shows the average contribution rate across all subjects as a function of round number (1 to 15). We also compared the average contribution for the first five rounds with that for the last five rounds. For *k* = 4, the average contribution probability across all subjects for the first five rounds was 0.6 (SD = 0.20) and significantly higher than that for the last five rounds (average across subjects = 0.49, SD = 0.19) {two-tailed paired sample *t* test, *t*(28) = 3.65, *P* = 0.001, 95% CI difference = [0.05,0.17]}. For *k* = 2, the difference between the first five rounds (average = 0.53, SD = 0.32) and the last five rounds (average = 0.50, SD = 0.33) was insignificant {two-tailed paired sample *t* test, *t*(28) = 1.51, *P* = 0.14, 95% CI difference = [-0.01,0.06]}.

The average contribution probability did not change significantly as subjects played more games (Fig. 2F). In each condition, most of the players played at least five games (27 players for *k* = 2 and 26 for *k* = 4). For *k* = 2, in their first game, the average contribution rate of players was 0.37 (SD = 0.25), while in their fifth game, it was 0.30 (SD = 0.24) {two-tailed paired sample *t* test, *t*(26) = 1.34, *P* = 0.19, 95% CI difference = [-0.03,0.17]}. When *k* = 4, the average contribution rate was 0.57 (SD = 0.30) in the first game and 0.61 (SD = 0.35) in the fifth game {two-tailed paired sample *t* test, *t*(25) = -0.69, *P* = 0.50, 95% CI difference = [-0.16,0.08]}.

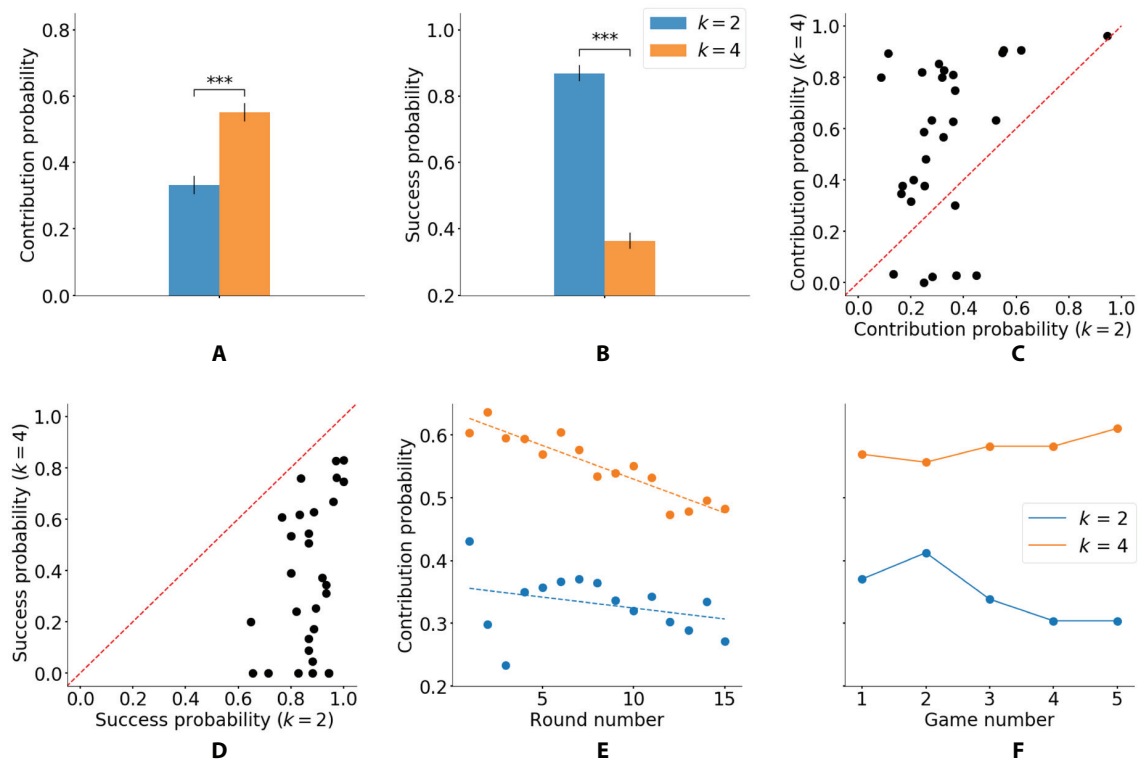
**Probabilistic model of theory of mind for the group in the PGG**

Consider one round of the PGG task. A player can be expected to choose an action (contribute or free-ride) based on the number of



**Fig. 1. Multiround PGG.** The figure depicts the sequence of computer screens a subject sees in one round of the PGG. The subject is assigned four other players as partners, and each round requires the subject to make a decision: Keep 1 MU (i.e., free-ride) or contribute 1 MU. The subject knows whether the threshold to generate public goods (reward of 2 MU for each player) is two or four contributions (from the five players). After the subject acts, the total number of contributions and overall outcome of the round (success or failure) are revealed.





**Fig. 2. Human behavior in the PGG Task.** (A) Average contribution probability across subjects is significantly higher when the task requires more volunteers ( $k$ ) to generate the group reward. (B) Average probability of success across all subjects in generating the group reward is significantly higher when  $k$  is lower. Error bars indicate within-subject SE (52). (C) Average probability of contribution for each subject for  $k = 2$  versus  $k = 4$ . Each point represents a subject. Subjects tend to contribute more often when the task requires more volunteers. (D) Average success rate for each subject was higher for  $k = 2$  versus  $k = 4$ . (E) Average probability of contribution across subjects decreases throughout a game, especially for  $k = 4$ . Dotted lines are linear functions showing this trend for each  $k$ . (F) Average contribution probability across subjects as a function of number of games played. The contribution probability does not change significantly as subjects play more games.

contributions they anticipate the others to make in that round. Because the actions of individual players remain unknown through the game, the only observable parameter is the total number of contributions. One can therefore model this situation using a single random variable  $\theta$ , denoting the average probability of contribution by any group member. With this definition, the total number of contributions by all the other members of the group can be expressed as a binomial distribution. Specifically, if  $\theta$  is the probability of contribution by each group member, the probability of observing  $m$  contributions from the  $N - 1$  others in a group of  $N$  people is

$$P(m | \theta) = \binom{N-1}{m} \theta^m (1 - \theta)^{N-1-m} \tag{1}$$

Using this probability, a player can calculate the expected number of contributions from the others, compare it with  $k$ , and decide whether to contribute or free-ride accordingly. For example, if  $\theta$  is very low, there is not a high probability of observing  $k - 1$  contributions by the others, implying that free-riding is the best strategy.

There are two important facts that make this decision-making more complex. First, the player does not know  $\theta$ .  $\theta$  must be estimated from the behavior of the group members. Second, other group members also have a theory of mind. Therefore, they can be expected to change their strategy based on the actions of others. Because of this ability in other group members, each player needs to simulate the effect of their action on the group’s behavior in the future.

To model the uncertainty in  $\theta$ , we assume that a probability distribution over  $\theta$  is maintained in the player’s mind, representing their belief about the cooperativeness of the group. Each player starts with an initial probability distribution, called the prior belief about  $\theta$ , and updates this belief over successive rounds based on the actions of the others. The prior belief may be based on the prior life experience of the player, or what they believe others would do through fictitious play (26). For example, the player may start with a prior belief that the group will be a cooperative one but change this belief after observing low numbers of contributions by the others. Such belief updates can be performed using Bayes’ rule to invert the probabilistic relationship between  $\theta$  and the number of contributions given by Eq. 1.

A suitable prior probability distribution for estimating the parameter  $\theta$  of a binomial distribution is the beta distribution, which is itself determined by two (hyper) parameters  $\alpha$  and  $\beta$

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ \text{Beta}(\alpha, \beta) : P(x | \alpha, \beta) &\propto x^{\alpha-1} (1 - x)^{\beta-1} \end{aligned} \tag{2}$$

Starting with a prior probability  $\text{Beta}(\alpha_1, \beta_1)$  for  $\theta$ , the player updates their belief about  $\theta$  after observing the number of contributions from the others in each round through Bayes’ rule. This updated belief is called the posterior probability of  $\theta$ . The posterior probability of  $\theta$  in each round serves as the prior for the next round.

In economics, the ability to infer the belief of others is sometimes called sophistication (27, 28). Here, we consider a simple form of sophistication: We assume that each player thinks other group members have the same model as themselves ( $\alpha$  and  $\beta$ ). This is justifiable due to computational efficiency and more importantly anonymity of players. As a result, with a prior of  $\text{Beta}(\alpha_t, \beta_t)$  after observing  $c$  contributions (including one's own when made) in round  $t$ , the posterior probability of  $\theta$  for the subject becomes  $\text{Beta}(\alpha_{t+1}, \beta_{t+1})$ , where  $\alpha_{t+1} = \alpha_t + c$  and  $\beta_{t+1} = \beta_t + N - c$ . Technically, this follows because the beta distribution is conjugate to the binomial distribution (29). Note that we include one's own action in the update of the belief because one's own action can change the future contribution level of the others.

Intuitively,  $\alpha$  represents the number of contributions made thus far, and  $\beta$  represents the number of free-rides.  $\alpha_1$  and  $\beta_1$  (that define prior belief) represent the player's a priori expectation about the relative number of contributions versus free-rides, respectively, before the session begins. For example, when  $\alpha_1$  is larger than  $\beta_1$ , the player starts the task with the belief that people will contribute more than free-ride. Large values of  $\alpha_1$  and  $\beta_1$  imply that the subject thinks that the average contribution probability will not change significantly after one round of the game when updated with the relatively small number  $c$  as above.

Decision making in the PGG task is also made complex by the fact that the actual cooperativeness of the group itself (not just the player's belief about it) may change from one round to the next: Players observe the contributions of the others and may change their own strategy for the next round. For example, players may start the game making contributions but change their strategy to free-riding if they observe a large number of contributions by the others. We model this phenomenon using a parameter  $0 \leq \gamma \leq 1$ , which serves as a decay rate: The prior probability for round  $t$  is modeled as  $\text{Beta}(\gamma\alpha_t, \gamma\beta_t)$ , which allows recent observations about the contributions of other players to be given more importance than observations from the more distant past. Thus, in a round with  $c$  total contributions (including the subject's own contribution when made), the subject's belief about the cooperativeness of the group as a whole changes from  $\text{Beta}(\alpha_t, \beta_t)$  to  $\text{Beta}(\alpha_{t+1}, \beta_{t+1})$  where  $\alpha_{t+1} = \gamma\alpha_t + c$  and  $\beta_{t+1} = \gamma\beta_t + N - c$ .

**Action selection**

How should a player decide whether to contribute or free-ride in each round? One possible strategy is to maximize the reward for the current round by calculating the expected number of contributions by the others based on the current belief. Using Eq. 1 and the prior probability distribution over  $\theta$ , the probability of seeing  $m$  contributions by the others when the belief about the cooperativeness of the group is  $\text{Beta}(\alpha, \beta)$  is given by

$$\begin{aligned}
 P(m | \alpha, \beta) &= \int_0^1 P(m | \theta) P(\theta | \alpha, \beta) d\theta \\
 &\propto \int_0^1 \binom{N-1}{m} \theta^m (1-\theta)^{N-1-m} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\
 &\propto \binom{N-1}{m} \int_0^1 \theta^{\alpha+m-1} (1-\theta)^{\beta+N-m-2} d\theta
 \end{aligned}
 \tag{3}$$

One can calculate the expected reward for the contribute versus free-ride actions in the current round based on the above equation. Maximizing this reward, however, is not the best strategy. As alluded

to earlier, the actions of each player can change the behavior of other group members in future rounds. Specifically, our model assumes that its own contribution in the current round increases the average contribution rate of the group in the future rounds. Equation 10 in Methods shows the exact assumptions of our model (with updates of  $\alpha_{t+1} = \gamma\alpha_t + c$  and  $\beta_{t+1} = \gamma\beta_t + N - c$  for its belief) about the dynamics of the actual (hidden) state of the environment. The optimal strategy therefore is to calculate the cooperativeness of the group through the end of the session and consider the reward over all future rounds in the session before selecting the current action. Thus, an optimal agent would contribute for two reasons. First, contributing could enable the group to reach at least  $k$  volunteers in the current round. Second, contributing encourages other members to contribute in future rounds. Specifically, a contribution by the subject increases the average contribution rate for the next round by increasing  $\alpha$  in the next round (see the transition function in Methods).

Long-term reward maximization (as discussed above) based on probabilistic inference of hidden state in an environment (here,  $\theta$ , the probability of contribution of group members) can be modeled using the framework of POMDPs (24). Further details can be found in Methods, but briefly, to maximize the total expected reward, our model starts from the last round, the reward is calculated for each action and state, and then the model steps back one time step to find the optimal action for each state in that round. This process is repeated in a recursive fashion. Figure 3A shows a schematic of the PGG experiment modeled using a POMDP, and Fig. 3B illustrates the mechanism of action selection in this model.

As an example of the POMDP model's ability to select actions for the PGG task, Fig. 4 (A and B) shows the best actions for a given round (here, round 9) as prescribed by the POMDP model for  $k = 2$  and  $k = 4$ , respectively (the number of minimum volunteers needed). The best actions are shown as a function of different belief states the subject may have, expressed in terms of the different values possible for belief parameters  $\alpha_t$  and  $\beta_t$ . This mapping from beliefs to actions is called a policy.

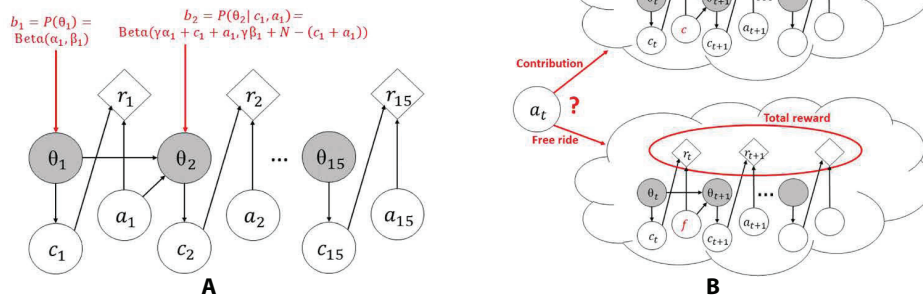
Our simulations using the POMDP model showed that considering a much longer horizon (e.g., 50 rounds) instead of just 15 rounds gave a better fit to the subjects' behavior, suggesting that human subjects may be inclined to use long horizons for group decision-making tasks (see Discussion). Such a long horizon for determining the optimal policy makes the model similar to an infinite horizon POMDP model (30). As a result, the optimal policy for all rounds in our model is very similar to the policy for round 9 shown in Fig. 4 (A and B).

In summary, the POMDP model performs two computations simultaneously. The first computation is probabilistic estimation of the (hidden) average contribution rate through belief updates. The average contribution rate changes during the course of the game as players interact with each other. The second computation involves selecting actions to influence this average contribution rate and to maximize total expected reward. This is the action selection component, which is performed by backward reasoning from the last round.

**POMDP model predicts human behavior in volunteer's dilemma task**

The POMDP model has three parameters,  $\alpha_1$ ,  $\beta_1$ , and  $\gamma$ , which determine the subject's actions and belief in each round. We fit these parameters to the subject's actions by minimizing the error, i.e., the difference between the POMDP model's predicted action and the

Downloaded from <http://advances.sciencemag.org/> on February 19, 2021



**Fig. 3. POMDP model of the multiround PGG.** (A) Model: The subject does not know the average probability of contribution of the group. The POMDP model assumes that the subject maintains a probability distribution (“belief,” denoted by  $b_t$ ) about the group’s average probability of contribution (denoted by  $\theta_t$ ) and updates this belief after observing the outcome  $c_t$  (contribution by others) in each round. (B) Action selection: The POMDP model chooses an action ( $a_t$ ) that maximizes the expected total reward ( $r_t$ ) across all rounds based on the current belief and the consequence of the action (contribution “ $c$ ” or free-ride “ $f$ ”) on group behavior in future rounds.

subject’s action in each round. The average percentage error across all rounds is then the percentage of rounds that the model incorrectly predicts (contribute instead of free-ride or vice versa). We defined accuracy as the percentage of the rounds that the model predicts correctly.

We also calculated the leave-one-out cross-validated (LOOCV) accuracy of our fits (29), where each “left out” data point is one whole game and the parameters were fit to the other 11 games of the subject. Note that our LOOCV accuracy is a prediction of the subject’s behavior in a game without any parameter tuning based on this game. In addition, while different rounds of each game are highly correlated, the games of each subject are independent from each other (given the parameters of that subject) as the other group members change in each game.

We found that the POMDP model had an average fitting accuracy across subjects of 84% (SD = 0.06), while the average LOOCV accuracy was 77% (SD = 0.08). Figure 5A compares the average fitting and LOOCV accuracies of the POMDP model with two other models. The first is a model-free reinforcement learning model known as Q-learning: Actions are chosen on the basis of their rewards in previous rounds (31), with the utility of group reward, initial values, and learning rate as free parameters (five parameters per subject; see Methods).

The average fitting accuracy of the Q-learning model was 79% (SD = 0.07), which is significantly worse than the POMDP model’s fitting accuracy given above {two-tailed paired  $t$  test,  $t(28) = -6.75$ ,  $P = 2.52 \times 10^{-7}$ , 95% CI difference =  $[-0.06, -0.03]$ }. In addition, the average LOOCV accuracy of the POMDP model was significantly higher than the average LOOCV accuracy of Q-learning, which was 73% (SD = 0.09) {two-tailed paired  $t$  test,  $t(28) = 2.20$ ,  $P = 0.037$ , 95% CI difference =  $[0.004, 0.08]$ }.

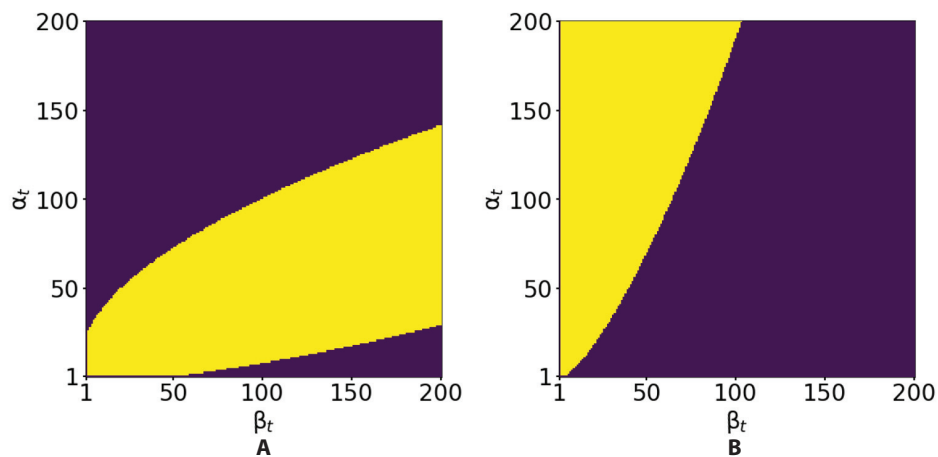
We additionally tested a previously explored descriptive model in the PGG literature known as the linear two-factor model (32), which predicts the current action of each player based on the player’s own action and contributions by the others in the previous round (this model has three free parameters per subject; see Methods). The average fitting accuracy of the two-factor model was 78% (SD = 0.09), which is significantly lower than the POMDP model’s fitting accuracy {two-tailed paired  $t$  test,  $t(28) = -4.86$ ,  $P = 4.1 \times 10^{-5}$ , 95% CI

difference =  $[-0.08, -0.03]$ }. Moreover, the LOOCV accuracy of the two-factor model was 47% (SD = 20), significantly lower than the POMDP model {two-tailed paired  $t$  test,  $t(28) = -7.61$ ,  $P = 2.7 \times 10^{-8}$ , 95% CI difference =  $[-0.38, -0.22]$ }. The main reason for this result, especially the lower LOOCV accuracy, is that group success also depends on the required number of volunteers ( $k$ ). This value is automatically incorporated in the POMDP’s calculation of expected reward. Also, reinforcement learning works directly with rewards and therefore does not need explicit knowledge of  $k$  (however, a separate parameter for each  $k$  is needed in the initial value function for Q-learning; see Methods). Given that the number of free parameters for the descriptive and model-free approaches is greater than or equal to the number of free parameters in the POMDP model, the higher accuracy of POMDP is notable in terms of model comparison.

We tested the POMDP model’s predictions of contribution probability for each subject for the two  $k$  values with experimental data (same data as in Fig. 2C; see Methods). As shown in Fig. 5 (B and C), the POMDP model’s predictions match the pattern of distribution of actual data from the experiments.

The POMDP model, when fit to a subject’s actions, can also explain other events during the PGG task in contrast to the other models described above. For example, based on Eq. 3 and the action chosen by the POMDP model, one can predict the subject’s belief about the probability of success in the current round. This prediction cannot be directly validated, but it can be compared to actual success. If we consider actual success as the ground truth, the average accuracy of the POMDP model’s prediction of success probability across subjects was 71% (SD = 0.07). Moreover, the predictions matched the pattern of success rate data from the experiment (Fig. 5, D and E). The other models presented above are not capable of making such a prediction.

The POMDP model’s predictions also match experimental data when the data points are binned on the basis of round of the game. The model correctly predicts a decrease in contribution for  $k = 4$  and lack of significant change in contribution rate on average for  $k = 2$  (Fig. 5F). Moreover, the model’s prediction of a subject’s belief about group success matches the actual data round by round (Fig. 5G). Further comparisons to other models, such as the interactive-POMDP model (33), are provided in the Supplementary Materials.



**Fig. 4. Optimal actions prescribed by the POMDP policy as a function of belief state.** Plot (A) shows the policy for  $k = 2$  and plot (B) for  $k = 4$ . The purple regions represent those belief states (defined by  $\alpha_t$  and  $\beta_t$ ) for which free-riding is the optimal action; the yellow regions represent belief states for which the optimal action is contributing. These plots confirm that the optimal policy depends highly on  $k$ , the number of required volunteers. For the two plots, the decay rate was 1 and  $t$  was 9.

### Distribution of POMDP parameters

We can gain insights into the subject's behavior by interpreting the parameters of our POMDP model in the context of the task. As alluded to above, the prior parameters  $\alpha_1$  and  $\beta_1$  represent the subject's prior expectations of contributions and free-rides, respectively. Therefore, the ratio  $\alpha_1/\beta_1$  characterizes the subject's expectation of contributions by group members, while the average of these parameters,  $(\alpha_1 + \beta_1)/2$ , indicates the weight the subject gives to prior experience with similar groups before the start of the game. The decay rate  $\gamma$  determines the weight given to past observations compared to new ones: The smaller the decay rate, the more weight the subject gives to new observations.

We examined the distribution of these parameter values for our subjects after fitting the POMDP model to their behavior (Fig. 6, A and B). The ratio  $\alpha_1/\beta_1$  was in the reasonable range of 0.5 to 2 for almost all subjects (Fig. 6C; in our algorithm, the ratio can be as high as 200 or as low as  $1/200$ ; see Methods). The value of  $(\alpha_1 + \beta_1)/2$  across subjects was mostly between 40 to 120 (Fig. 6D), suggesting that prior belief about groups did have a significant role in players' strategy, but it was not the only factor because observations over multiple rounds can still alter this initial belief. To confirm the effect of actions during the game, we performed a comparison with a POMDP model that does not update  $\alpha$  and  $\beta$  over time and only uses its prior. The accuracy of this modified POMDP model was 66% ( $SD = 0.17$ ), significantly lower than our original model {two-tailed paired  $t$  test,  $t(28) = -5.47$ ,  $P = 7.64 \times 10^{-6}$ , 95% CI difference =  $[-0.23, -0.11]$ }. The average  $\alpha_t$  and  $\beta_t$  for each of the 15 rounds, as well as distributions of their difference with the prior values  $\alpha_1$  and  $\beta_1$  are presented in the Supplementary Materials.

We also calculated the expected value of contribution by the others in the first round, which is between 0 and  $N - 1 = 4$ , based on the values of  $\alpha_1$  and  $\beta_1$  for the subjects. For almost all subjects, this expected value was between two and three (Fig. 6E).

In addition, we calculated each subject's prior belief about group success (probability of success in the first round) based on  $\alpha_1$ ,  $\beta_1$ , and the subject's POMDP policy in the first round. As group success depends on the required number of volunteers ( $k$ ), probability of success is different for  $k = 2$  and  $k = 4$  even with the same  $\alpha_1$  and

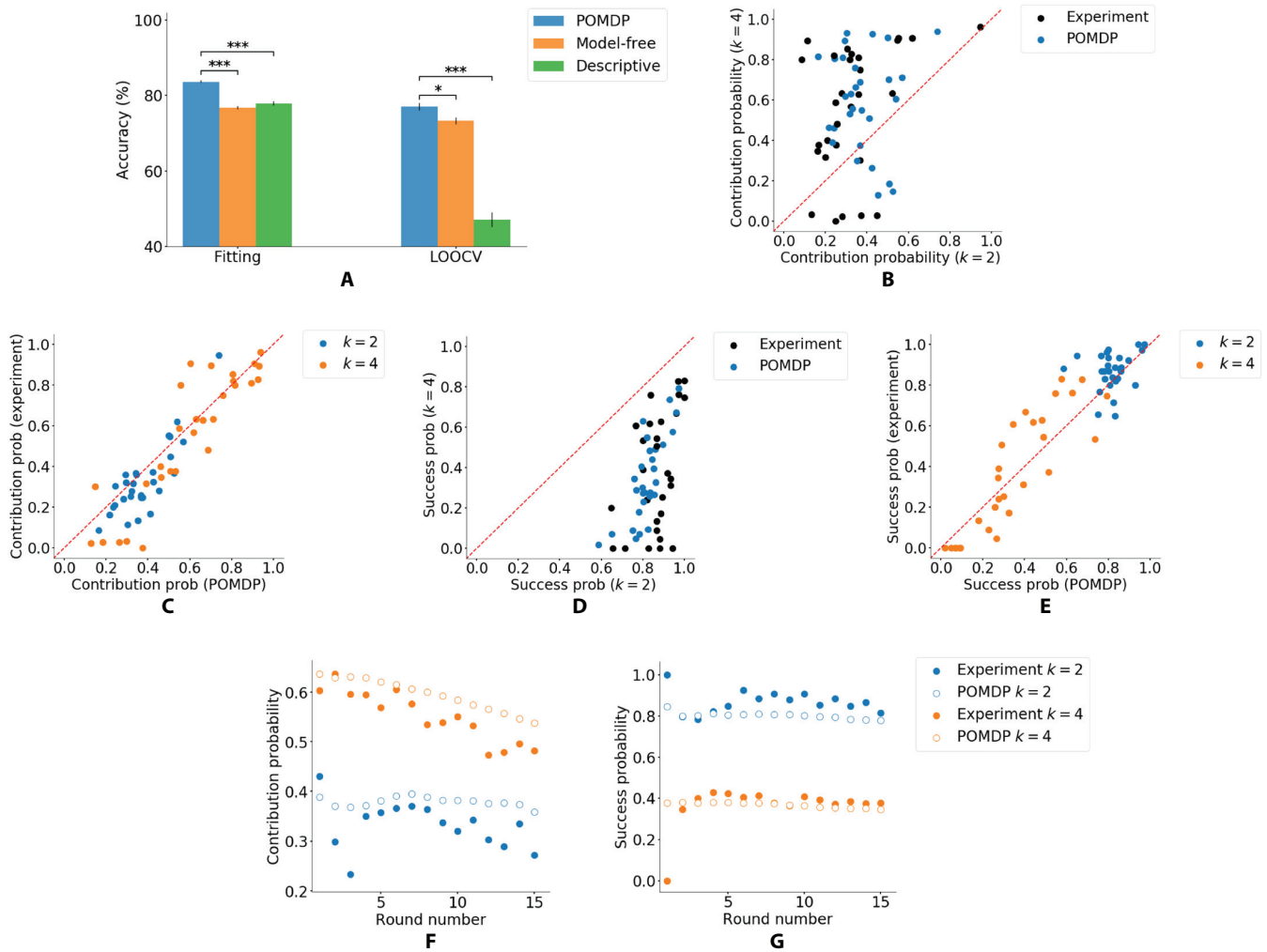
$\beta_1$ . Figure 6 (F and G) shows the distribution of this prior probability of success across all subjects for  $k = 2$  and  $k = 4$ . For  $k = 2$ , all subjects expected a high probability of success in the first round, whereas most of the subjects expected less than 60% chance for success when  $k = 4$ . While these beliefs cannot be directly validated, the results point to the importance of the required number of volunteers in shaping the subjects' behavior.

Additionally, the decay rate  $\gamma$ , which determines the weight accorded to the prior and previous observations compared to the most recent observation, was almost always above 0.95, with a mean of 0.93 and a median of 0.97 (Fig. 6H). Only three subjects had a decay rate less than 0.95 (not shown in the figure), suggesting that almost all subjects relied on observations made across multiple rounds when computing their beliefs rather than reasoning based solely on the current or most recent observations.

### DISCUSSION

We introduced a normative model based on POMDPs for explaining human behavior in a group decision-making task. Our model combines probabilistic reasoning about the group with long-term reward maximization by simulating the effect of each action on the future behavior of the group. The greater accuracy of our model in explaining and predicting the subjects' behavior compared to the other models suggests that humans make decisions in group settings by reasoning about the group as a whole. This mechanism is analogous to maintaining a theory of mind about another person, except that the theory of mind pertains to a group member on average.

This is the first time, to our knowledge, that a normative model has been proposed for a group decision-making task. Existing models to explain human behavior in the PGG, for example, are descriptive and do not provide insights into the computational mechanisms underlying the decisions (32). While the regression-based descriptive method we compared our POMDP model to can potentially be seen as a "learned" model-free approach to mapping observations to choice in the next round, our model was also able to outperform this method.

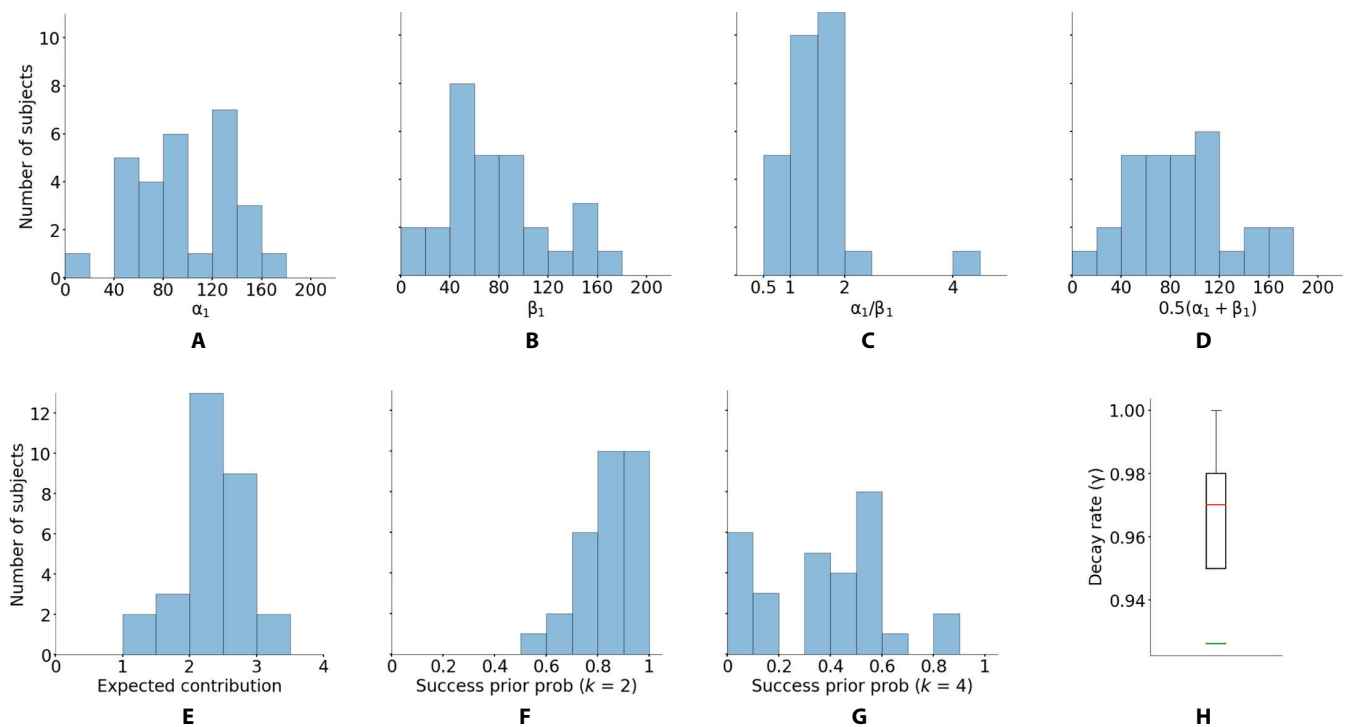


**Fig. 5. POMDP model's performance and predictions.** (A) Average fitting and LOOCV accuracy across all models. The POMDP model has significantly higher accuracy compared to the other models ( $*P < 0.05$  and  $***P < 0.001$ ). Error bars indicate within-subject SE (52). (B) POMDP model's prediction of a subject's probability of contribution compared to experimental data for the two  $k$  values [black circles: same data as in Fig. 2C]. (C) Same data as (B) but the POMDP model's prediction and the experimental data are shown for each  $k$  separately (blue for  $k = 2$  and orange for  $k = 4$ ). (D) POMDP model's prediction (blue circles) of a subject's belief about group success in each round (on average) compared to actual data (black circles, same data as in Fig. 2D). (E) Same data as (D), but the POMDP model's prediction and actual data are shown for each  $k$  separately (blue for  $k = 2$  and orange for  $k = 4$ ). (F) Same data as (B) and (C) but with the data points binned on the basis of round of the game. (G) Same data as (D) and (E) but with the data points binned based on round of the game.

In addition to providing a better fit and prediction of the subject's behavior, our model, when fit to the subject's actions, can predict success rate in each round without being explicitly trained for such predictions, in contrast to the other methods. In addition, as alluded to in Fig. 6 (C, D, and H), when fit to the subjects' actions, the parameters were all within a reasonable range, showing the importance of prior knowledge and multiple observations in decision-making. The POMDP model is normative and strictly constrained by probability theory and optimal control theory. The beta distribution is used because it is the conjugate prior of the binomial distribution (29) and not due to better fits compared to other distributions.

The POMDP policy aligns with our intuition about action selection in the volunteer's dilemma task. A player chooses to free-ride for two reasons: (i) when the cooperativeness of the group is low and therefore there is no benefit in contributing, and (ii) when the

player knows there are already enough volunteers and contributing leads to a waste of resources. The two purple areas of Fig. 4A represent these two conditions for  $k = 2$ . The upper left part represents large  $\alpha_t$  and small  $\beta_t$ , implying a high contribution rate, while the bottom right part represents small  $\alpha_t$  and large  $\beta_t$ , implying a low contribution rate. When  $k = 4$ , all but one of the five players must contribute for group success—this causes a significant difference in the optimal POMDP policy compared to the  $k = 2$  condition. As seen in Fig. 4B, there is only a single region of belief space for which free-riding is the best strategy, namely, when the player does not expect contributions by enough players (relatively large  $\beta_t$ ). On the other hand, as expected, this region is much larger compared to the same region for  $k = 2$  (see Fig. 4A). The POMDP model predicts that free-riding is not a viable action in the  $k = 4$  case (Fig. 4B) because not only does this action require all the other four players



**Fig. 6. Distribution of POMDP parameters across subjects.** (A) Histogram of  $\alpha_1$  across all subjects. (B) Histogram of  $\beta_1$  across all subjects. (C) Histogram of the ratio  $\alpha_1/\beta_1$  shows a value between 0.5 and 2 for almost all subjects. (D) Histogram of  $(\alpha_1 + \beta_1)/2$ . For most subjects, this value is between 40 and 120. (E) Histogram of prior belief  $\text{Beta}(\alpha_1, \beta_1)$  translated into expected contribution by the others in the first round. Note that the values, after fitting to the subjects' behavior, are mostly between 2 and 3. (F) When  $k = 2$ , all subjects expected a high probability of group success in the first round (before making any observations about the group). (G) When  $k = 4$ , almost all subjects assigned a chance of less than 60% to group success in the first round. (H) Box plot of decay rate  $\gamma$  across subjects shows that this value is almost always above 0.95. The median is 0.97 (orange line) and the mean is 0.93 (green line).

to contribute to generate the group reward in the current round but also such an action increases the chances that the group contribution will be lower in the next round, resulting in lesser expected reward in future rounds. The opposite situation can also occur especially when  $k = 2$ . A player may contribute not to gain the group reward in the current round but to encourage others to contribute in the next rounds. When an optimal player chooses free-riding due to low cooperativeness of the group, the estimated average contribution is so low that the group is not likely to get the group reward in the next rounds even with an increase in the average contribution due to the player's contribution. On the other hand, when an optimal player chooses to free-ride due to high cooperativeness of the group, the estimated average contribution rate is so high that the chance of success remains high in future rounds even with a decrease in average contribution rate due to the player free-riding in the current round.

In a game with a predetermined and known number of rounds, even if the player considers the future, one might expect the most rewarding action in the last rounds to be free-riding as there is little or no future to consider. However, our experimental data did not support this conclusion. Our model is able to explain these data using the hypothesis that subjects may use a longer horizon than the exact number of rounds in a game. Such a strategy provides a significant computational benefit by making the policies for different rounds similar to each other, avoiding recalculation of a policy for each single round. Recent studies in human decision-making have demonstrated that humans may use such minimal modifications of

model-based policies for efficiency (34, 35). More broadly, group decision-making occurs among groups of humans (and animals) that live together. Thus, any group decision-making involves practically an infinite horizon, i.e., there is always a future interaction even after the current task has ended, justifying the use of long horizons.

In the volunteer's dilemma, not only is the common goal not reached when there are not enough volunteers but also having more than the required number of volunteers leads to a waste of resources. As a result, an accurate prediction of others' intentions based on one's beliefs is crucial to make accurate decisions. This gives the model-based approach a huge advantage over model-free methods in terms of reward gathering, thus making it more beneficial for the brain to endure the extra cognitive cost. It is possible that in simpler tasks where the accurate prediction of minds is less crucial, the brain adopts a model-free approach.

Our model was based on the binomial and beta distributions for binary values due to the nature of the task, but it can be easily extended to the more general case of a discrete set of actions using multinomial and Dirichlet distributions (36). In addition, the model can be extended to multivariate states, e.g., when the players are no longer anonymous. In such cases, the belief can be modeled as a joint probability distribution over all parameters of the state. This, however, incurs a significant computational cost. An interesting area for future research is investigating whether, under some circumstances, humans model group members with similar behavior as one subgroup to reduce the number of minds one should reason about.

Our POMDP framework assumes that each subject starts with the same prior about average group member contribution probability at the beginning of each game. However, subjects might try to estimate this prior for a new group in the first few rounds, i.e., “explore” their new environment before seeking to maximize their reward (“exploit”) based on this prior (5). Such an “active inference” approach has been studied in two-person interactions (15, 16) and is an interesting direction of research in group decision-making.

Mimicking human behavior does not guarantee that a POMDP model (or any model) is being implemented in the brain. However, the POMDP model’s generalizability and the interpretability of its components, such as existence of a prior or simulation of the future, make it a useful tool for understanding the decision-making process.

The POMDP framework can model social tasks beyond economic decision-making, such as prediction of others’ intentions and actions in everyday situations (37). In these cases, we would need to modify the model’s definition of the state of other minds to include dimensions such as valence, competence, and social impact instead of propensity to contribute monetary units as in the PGG task (38).

The interpretability of the POMDP framework offers an opportunity to study the neurocognitive mechanisms of group decision-making in healthy and diseased brains. POMDPs and similar Bayesian models have previously proved useful in understanding neural responses in sensory decision-making (39–41) and in tasks involving interactions with a single individual (13, 17, 18). We believe that the POMDP model we have proposed can likewise prove useful in interpreting neural responses and data from neuroimaging studies of group decision-making tasks. In addition, the model can be used for Bayesian theory-driven investigations in the field of computational psychiatry (42). For example, theory of mind deficits are a key feature of autism spectrum disorder (43), but it is unclear what computational components are impaired and how they are affected. The POMDP model may provide a new avenue for computational studies of such neuropsychiatric disorders (44).

**METHODS**

**Experiment**

Thirty right-handed students at the University of Parma were recruited for this study. One of them aborted the experiment due to anxiety. Data from the other 29 participants were collected, analyzed, and reported. On the basis of self-reported questionnaires, none of the participants had a history of neurological or psychiatric disorders. This study was approved by the Institutional Review Board of the local ethics committee from Parma University (IRB no. A13-37030), which was carried out according to the ethical standards of the 2013 Declaration of Helsinki. All participants gave their informed written consent. As mentioned in Results, each subject played 14 sessions of the PGG (i.e., the volunteer’s dilemma), each containing 15 rounds. In the first two sessions, subjects received no feedback about the result of each round. However, in the following 12 sessions, social and monetary feedback were provided to the subject. The feedback included the number of contributors and free-riders, and the subject’s reward in that round. Each individual player’s action, however, remained unknown to the others. Therefore, individual players could not be tracked. We present analyses from the games with feedback.

In each round (see Fig. 1), the participant had to make a decision within 3 s by pressing a key; otherwise, the round was repeated. After the action selection (2.5 to 4 s), the outcome of the round was

shown to the subject for 4 s. Then, players evaluated the outcome of the round before the next round started. Subjects were told that they were playing with 19 other participants located in other rooms. Overall, 20 players were playing the PGG in four different groups simultaneously. These groups were randomly chosen by a computer at the beginning of each session. In reality, subjects were playing with a computer. In other words, a computer algorithm was generating all the actions of others for each subject. Each subject got a final monetary reward equal to the result of one PGG randomly selected by the computer at the end of the study.

In a PGG with  $N = 5$  players, we denote the action of player  $i$  in round  $t$  with the binary value of  $a_i^t$  ( $1 \leq i \leq N$ ), with  $a_i^t = 1$  representing contribution and  $a_i^t = 0$  representing free-riding. The human subject is assumed to be player 1. We define the average contribution rate of others  $\bar{a}_{2:N}^t = \frac{\sum_{i=2}^N a_i^t}{N-1}$  and generate each of the  $N - 1$  actions of others in round  $t$  using the following probabilistic function

$$\text{logit}(\bar{a}_{2:N}^t) = e_0 a_1^{t-1} + e_1 \left( \left( \frac{1 - K^{T-t+1}}{1 - K} \right)^{e_2} \bar{a}_{2:N}^{t-1} - K \right) \tag{4}$$

where  $K = k/N$ , in which  $k$  is the required number of contributors.

This model has three free parameters:  $e_0$ ,  $e_1$ , and  $e_2$ . These were obtained by fitting the above function to the actual actions of subjects in another PGG study (45), making this function a simulation of human behavior in the PGG task. Specifically, to generate the actions of others, we fixed  $e_2$  to 1 for all games.  $e_0$  was drawn randomly from the range of [0.15,0.35] for each game, and  $e_1$  was set to  $1 - e_0$ . This combination and the random sampling of  $e_0$  in each game simulated different response strategies for the others in each game, simulating new sets of group members. Higher values of  $e_0$  make the algorithm more likely to choose its next action based on the result of the group interaction in the previous round (especially the action of the subject). On the other hand, lower values of  $e_0$  make the algorithm more likely to stick to its previous action. For the first round of each game, we used the mean contribution rate of each subject as their fellow members’ decision.

**Markov decision processes**

A Markov decision process (MDP) is a tuple  $(S, A, T, R)$ , where  $S$  represents the set of states of the environment,  $A$  is the set of actions,  $T$  is the transition function  $S \times S \times A \rightarrow [0,1]$  that determines the probability of the next state given the current state and action, i.e.,  $T(s', s, a) = P(s' | s, a)$ , and  $R$  is the reward function  $S \times A \rightarrow R$  representing the reward associated with each state and action (30). In an MDP with horizon  $H$  (total number of performed actions), given the initial state  $s_1$ , the goal is to choose a sequence of actions that maximizes the total expected reward

$$\pi^* = \arg \max_{a_1, a_2, \dots, a_H} \sum_{t=1}^H E_{s_t} [R(s_t, a_t)] \tag{5}$$

This sequence, called the optimal policy, can be found using the technique of dynamic programming (30). For an MDP with time horizon  $H$ , the  $Q$  value, value function  $V$ , and action function  $U$  at the last time step  $t = H$  are defined as

$$\forall s \in S: \begin{cases} Q^H(s, a) \leftarrow R(s, a) \\ V^H(s) \leftarrow \max_a Q^H(s, a) \\ U^H(s) \leftarrow \arg \max_a Q^H(s, a) \end{cases} \tag{6}$$

For any  $t$  from 1 to  $H - 1$ , the value function  $V^t$  and action function  $U^t$  are defined recursively as

$$\begin{cases} Q^t(s, a) \leftarrow R(s, a) + \sum_{s' \in S} T(s', s, a) V^{t+1}(s') \\ V^t(s) \leftarrow \max_a Q^t(s, a) \\ U^t(s) \leftarrow \arg \max_a Q^t(s, a) \end{cases} \quad (7)$$

Starting from the initial state  $s_1$  at time 1, the action chosen by the optimal policy  $\pi^*$  at time step  $t$  is  $U^t(s_t)$ .

When the state of the environment is hidden, the MDP turns into a partially observable MDP (POMDP) where the state is estimated probabilistically from observations or measurements from sensors. Formally, a POMDP is defined as  $(S, A, Z, T, O, R)$ , where  $S, A, T,$  and  $R$  are defined as in the case of MDPs,  $Z$  is the set of possible observations, and  $O$  is the observation function  $Z \times S \rightarrow [0,1]$  that determines the probability of any observation  $z$  given a state  $s$ , i.e.,  $O(z, s) = P(z | s)$ . To find the optimal policy, the POMDP model uses the posterior probability of states, known as the belief state, where  $b_t(s) = P(s | z_1, a_1, z_2, \dots, a_{t-1})$ . Belief states can be computed recursively as follows

$$\forall s \in S: b_{t+1}(s) \propto O(z_t, s) \sum_{s' \in S} T(s, s', a_t) b_t(s') \quad (8)$$

If we define  $R(b_t, a_t)$  as the expected reward of  $a_t$ , i.e.,  $E_{s_t}[R(s_t, a_t)]$ , starting from initial belief state,  $b_1$ , the optimal policy for the POMDP is given by

$$\pi^* = \arg \max_{a_1, a_2, \dots, a_H} \sum_{t=1}^H E_{s_t}[R(b_t, a_t)] \quad (9)$$

A POMDP can be considered an MDP whose states are belief states. This belief state space, however, is exponentially larger than the underlying state space. Therefore, solving a POMDP optimally is computationally expensive, unless the belief state can be represented by a few parameters as in our case (30). For solving larger POMDP problems, various approximation and learning algorithms have been proposed. We refer the reader to the growing literature on this topic (46–48).

**POMDP for binary PGG**

The state of the environment is represented by the average cooperativeness of the group or, equivalently, the average probability  $\theta$  of contribution by a group member. Because  $\theta$  is not observable, the task is a POMDP, and one must maintain a probability distribution (belief) over  $\theta$ . The beta distribution, represented by two free parameters ( $\alpha$  and  $\beta$ ), is the conjugate prior for binomial distribution (29). Therefore, when performing Bayesian inference to obtain the belief state over  $\theta$ , combining the beta distribution as the prior belief and the binomial distribution as the likelihood results in another beta distribution as the posterior belief. Using the beta distribution for the belief state, our POMDP turns into an MDP with a two-dimensional state space represented by  $\alpha$  and  $\beta$ . Starting from an initial belief state  $\text{Beta}(\alpha_1, \beta_1)$  and with an additional free parameter  $\gamma$ , the next belief states are determined by the actions of all players at each round as described in Results. For the reward function, we used the monetary reward function of the PGG. Therefore, the elements of our new MDP derived from the PGG POMDP are as follows

- $S = (\alpha, \beta)$
- $A = \{c, f\}$

$$\begin{aligned} \bullet T(s', s, a) : & \begin{cases} P((\gamma\alpha + k' + 1, \gamma\beta + N - 1 - k') | (\alpha, \beta), c) = \binom{N-1}{k'} \frac{B(\gamma\alpha + k', \gamma\beta + N - 1 - k')}{B(\gamma\alpha, \gamma\beta)} \\ P((\gamma\alpha + k', \gamma\beta + N - k') | (\alpha, \beta), f) = \binom{N-1}{k'} \frac{B(\gamma\alpha + k', \gamma\beta + N - 1 - k')}{B(\gamma\alpha, \gamma\beta)} \end{cases} \\ \bullet R(s, a) : & \begin{cases} R((\alpha, \beta), c) = E - C + \sum_{k'=k-1}^N \binom{N-1}{k'} \frac{B(\alpha + k', \beta + N - 1 - k')}{B(\alpha, \beta)} G \\ R((\alpha, \beta), f) = E + \sum_{k'=k}^N \binom{N-1}{k'} \frac{B(\alpha + k', \beta + N - 1 - k')}{B(\alpha, \beta)} G \end{cases} \end{aligned}$$

$B(\alpha, \beta)$  is the normalizing constant:  $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta$ .

The POMDP model above assumes that the hidden state, i.e.  $\theta$ , is a random variable following a Bernoulli distribution, which changes with the actions of all players in each round. These actions serve as samples from this distribution, with  $\alpha_1$  and  $\beta_1$  being the initial samples. Also, the decay rate  $\gamma$  controls the weights of previous samples. Using maximum likelihood estimation, for any  $t$ ,  $\theta_t$  equals  $\alpha_t / (\alpha_t + \beta_t)$ . One can also estimate  $\theta$  in a recursive fashion

$$\theta_{t+1} \leftarrow \frac{1}{\gamma\alpha_t + \gamma\beta_t + N} \left( (\gamma\alpha_t + \gamma\beta_t)\theta_t + \sum_{i=1}^N a_i^t \right) \quad (10)$$

where  $a_i^t$  is the action of player  $i$  in round  $t$  ( $a_i^t = 1$  for contribution and 0 for free-ride).

According to the experiment, the time horizon should be 15 time steps. However, we found that a longer horizon ( $H = 50$ ) for all players provides a better fit to the subjects' data, potentially reflecting an intrinsic bias in humans for using longer horizons for social decision-making. For each subject, we found  $\alpha_1, \beta_1,$  and  $\gamma$  that made our POMDP's optimal policy fit the subject's actions as much as possible. For simplicity, we only considered integer values for states (integer  $\alpha$  and  $\beta$ ). The fitting process involved searching over integer values from 1 to 200 for  $\alpha_1$  and  $\beta_1$  and values between 0 and 1 with a precision of 0.01 (0.01, 0.02, ..., 0.99, 1.0) for  $\gamma$ . The fitting criterion was round-by-round accuracy. For consistency with the descriptive model, the first round was not included (despite the POMDP model's capability of predicting it). Because the utility value for public good for a subject can be higher than the monetary reward due to social or cultural reasons (49), we investigated the effect of higher values for the group reward  $G$  in the reward function of the POMDP. This, however, did not improve the fit. A preliminary version of the above model but without the  $\gamma$  parameter was presented in (50).

As specified above, the best action for each state in round  $t$  is  $U^t(s)$ . The probability of contribution (choice probability) can be calculated using a logit function:  $1 / (1 + \exp(z(Q^t(s, f) - Q^t(s, c)))$  (19). For each  $k$ , we used one free parameter  $z$  across all subjects to maximize the likelihood of contribution probability given the experimental data [implementation by scikit-learn (51)]. Note that the parameter  $z$  does not affect the accuracy of fits and predictions because it does not affect the action with the maximum expected total reward.

In round  $t$ , if the POMDP model selects the action "contribution," the probability of success can be calculated as  $\sum_{m=k-1}^{N-1} P(m | \alpha_t, \beta_t)$  (see Eq. 3). Otherwise, the probability of success is  $\sum_{m=k}^{N-1} P(m | \alpha_t, \beta_t)$ . This probability value was compared to the actual success and failure of each round to compute the accuracy of success prediction by the POMDP model.

**Model-free method: Q-learning**

We used Q-learning as our model-free approach. There are two  $Q$  values in the PGG task, one for each action, i.e.,  $Q(c)$  and  $Q(f)$  for



“contribute” and “free-ride,” respectively. At the beginning of each PGG,  $Q(c)$  and  $Q(f)$  are initialized to the expected reward for a subject for that action based on a free parameter  $p$ , which represents the prior probability of group success. As a result, we have

$$\begin{cases} Q^1(c) \leftarrow p(E - C + G) + (1 - p)(E - C) \\ Q^1(f) \leftarrow p(E + G) + (1 - p)E \end{cases} \quad (11)$$

We customized the utility function for each subject by making the group reward  $G$  a free parameter to account for possible prosocial intent (49). Moreover, as the probability of success is different for  $k = 2$  and  $k = 4$ , we used two separate parameters  $p_2$  and  $p_4$  instead of  $p$ , depending on the value of  $k$  in the PGG.

In each round of the game, the action with the maximum  $Q$  value was chosen. The  $Q$  value for that action was then updated on the basis of the subject’s action and group success/failure, with a learning rate  $\eta^t$ . This learning rate was a function of the round number, i.e.,  $\eta^t = \frac{1}{\lambda_0 + \lambda_1 t}$  where  $\lambda_0$  and  $\lambda_1$  are free parameters, and  $t$  is the number of the current round. Let the subject’s action in round  $t$  be  $a^t$ , the Q-learning model’s chosen action be  $\hat{a}^t$ , and the reward obtained be  $r^t$ . We have

$$1 \leq t \leq 15: \begin{cases} \hat{a}^t = \arg \max_{a \in \{c,f\}} Q^t(a) \\ Q^{t+1}(a^t) \leftarrow (1 - \eta^t) Q^t(a^t) + \eta^t r^t \end{cases} \quad (12)$$

For each subject, we searched for the values of  $\lambda_0$ ,  $\lambda_1$ , the group reward  $G$ , and the probability of group success  $p_2$  or  $p_4$  that maximize the round-by-round accuracy of the Q-learning model. Similar to the other models, the first round was not included in this fitting process.

**Descriptive model**

Our descriptive model was based on a logistic regression [implementation by scikit-learn (51)] that predicts the subject’s action in the current round based on their own previous action and the total number of contributions by the others in the previous round. As a result, this model has three free parameters (two features and a bias parameter). Let  $a_1^t$  be the subject’s action in round  $t$  and  $a_{2:N}^t$  be the actions of others in the same round. The subject’s predicted action in the next round  $t + 1$  is then given by

$$\hat{a}_1^{t+1} = \begin{cases} c & \kappa_0 + \kappa_1 a_1^t + \kappa_2 \left( \sum_{i=2}^N a_i^t \right) > 0 \\ f & \text{otherwise} \end{cases} \quad (13)$$

We used one separate regression model for each subject. As the model’s predicted action is based on the previous round’s actions, the subject’s action in the first round cannot be predicted by this model.

**Leave-one-out cross-validation**

For all three approaches, LOOCV was computed on the basis of the games played by each subject. For each subject, we set aside one game, fitted the parameters to the other 11 games, and computed the error of the model with fitted parameters on the game that was set aside. We repeated this for all games and reported the average of the 12 errors as LOOCV error for the subject.

**Static probability distribution and greedy strategy**

If a player does not consider the future and solely maximizes the expected reward in the current round (greedy strategy) or ignores the effect of an action on others, the optimal action is always free-

riding independent of the average probability of contribution by a group member. This is because free-riding always results in one unit more monetary reward (3 MU for success or 1 MU for failure) compared to contribution (2 or 0 MU), except in the case where the total number of contributions by others is exactly  $k - 1$ . In the latter case, choosing contribution yields one unit more reward (2 MU) compared to free-riding (1 MU). This means that the expected reward for free-riding is always more than that for contribution unless the probability of observing exactly  $k - 1$  contributions by others is greater than 0.5. We show that this is impossible for any value of  $\theta$ . First, note that the probability of exactly  $k - 1$  contributions from  $N - 1$  players is maximized when  $\theta = (k - 1)/(N - 1)$ . Next, for any  $\theta$ , the probability of  $k - 1$  contributions from  $N - 1$  players is

$$P(k - 1 | \theta) = \binom{N - 1}{k - 1} \theta^{k-1} (1 - \theta)^{N-k} \leq \binom{N - 1}{k - 1} \left( \frac{k - 1}{N - 1} \right)^{k-1} \left( \frac{N - k}{N - 1} \right)^{N-k} = 0.75^3 < 0.5 \quad (14)$$

for  $N = 5$  and for either  $k = 2$  or  $k = 4$ .

**SUPPLEMENTARY MATERIALS**

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/5/11/eaax8783/DC1>  
Supplementary Text

Fig. S1. Distribution and change in belief parameters over multiple rounds.  
Fig. S2. Data generated by the POMDP model compared to experimental data.

[View/request a protocol for this paper from Bio-protocol.](#)

**REFERENCES AND NOTES**

1. A. G. Sanfey, Social decision-making: Insights from game theory and neuroscience. *Science* **318**, 598–602 (2007).
2. J. Joiner, M. Piva, C. Turrin, S. W. C. Chang, Social learning through prediction error in the brain. *npj Sci. Learn.* **2**, 8 (2017).
3. D. Mookherjee, B. Sopher, Learning and decision costs in experimental constant sum games. *Games Econ. Behav.* **19**, 97–132 (1997).
4. C. F. Camerer, T. H. Ho, Experience-weighted attraction learning in normal form games. *Econometrica* **67**, 827–874 (1999).
5. K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O’Doherty, G. Pezzulo, Active inference and learning. *Neurosci. Biobehav. Rev.* **68**, 862–879 (2016).
6. P. Dayan, N. D. Daw, Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* **8**, 429–453 (2008).
7. N. D. Daw, P. Dayan, The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. B* **369**, 20130478 (2014).
8. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans’ choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
9. A. J. Culbreth, A. Westbrook, N. D. Daw, M. Botvinick, D. M. Barch, Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.* **125**, 777–787 (2016).
10. B. B. Doll, D. A. Simon, N. D. Daw, The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
11. R. J. Dolan, P. Dayan, Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
12. C. J. Charpentier, J. P. O’Doherty, The application of computational models to social neuroscience: Promises and pitfalls. *Soc. Neurosci.* **13**, 637–647 (2018).
13. W. Yoshida, B. Seymour, K. J. Friston, R. J. Dolan, Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* **30**, 10744–10751 (2010).
14. T. Xiang, D. Ray, T. Lohrenz, P. Dayan, P. R. Montague, Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Comput. Biol.* **8**, e1002841 (2012).
15. M. Moutoussis, P. Fearon, W. El-Derey, R. J. Dolan, K. J. Friston, Bayesian inferences about the self (and others): A review. *Conscious. Cogn.* **25**, 67–76 (2014).
16. M. Moutoussis, N. J. Trujillo-Barreto, W. El-Derey, R. J. Dolan, K. J. Friston, A formal model of interpersonal inference. *Front. Hum. Neurosci.* **8**, 160 (2014).
17. A. Hula, P. R. Montague, P. Dayan, Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Comput. Biol.* **11**, e1004254 (2015).
18. C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).

19. S. Suzuki, R. Adachi, S. Dunne, P. Bossaerts, J. P. O'Doherty, Neural mechanisms underlying human consensus decision-making. *Neuron* **86**, 591–602 (2015).
20. S. A. Park, S. Goïame, D. A. O'Connor, J.-C. Dreher, Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biol.* **15**, e2001958 (2017).
21. D. Diekmann, Volunteer's dilemma. *J. Confl. Resolut.* **29**, 605–610 (1985).
22. J. M. Darley, B. Latane, Bystander intervention in emergencies: Diffusion of responsibility. *J. Pers. Soc. Psychol.* **8**, 377–383 (1968).
23. M. Archetti, I. Scheuring, Coexistence of cooperation and defection in public goods games. *Evolution* **65**, 1140–1148 (2011).
24. L. P. Kaelbling, M. L. Littman, A. R. Cassandra, Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998).
25. E. Fehr, S. Gächter, Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
26. G. W. Brown, Iterative solution of games by fictitious play, in *Activity Analysis of Production and Allocation*, T. C. Koopmans, Ed. (Wiley, 1951), pp. 374–376.
27. M. Costa-Gomes, V. P. Crawford, B. Broseta, Cognition and behavior in normal-form games: An experimental study. *Econometrica* **69**, 1193–1235 (2001).
28. M. Devaine, G. Hollard, J. Daunizeau, Theory of mind: Did evolution fool us? *PLOS ONE* **9**, e87619 (2014).
29. K. P. Murphy, Machine learning: A probabilistic perspective, in *Adaptive Computation and Machine Learning* (MIT Press, 2012).
30. S. Thrun, W. Burgard, D. Fox, *Probabilistic Robotics* (MIT Press, 2005).
31. J. N. Tsitsiklis, Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16**, 185–202 (1994).
32. M. Wunder, S. Suri, D. J. Watts, Empirical agent based models of cooperation in public goods games, in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)* (ACM, 2013), pp. 891–908.
33. P. J. Gmytrasiewicz, P. Doshi, Interactive POMDPs: Properties and preliminary results, in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3* (IEEE Computer Society, 2004), pp. 1374–1375.
34. I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, S. J. Gershman, The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
35. E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Comput. Biol.* **13**, e1005768 (2017).
36. H. Attias, Planning by probabilistic inference, in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3 to 6 January 2003.
37. D. I. Tamir, M. A. Thornton, Modeling the predictive social mind. *Trends Cogn. Sci.* **22**, 201–212 (2018).
38. D. I. Tamir, M. A. Thornton, J. M. Contreras, J. P. Mitchell, Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 194–199 (2016).
39. R. P. N. Rao, Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
40. Y. Huang, R. P. N. Rao, Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. *PLOS ONE* **8**, e53344 (2013).
41. K. Khalvati, R. P. Rao, A Bayesian framework for modeling confidence in perceptual decision making, in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 7 to 12 December 2015, pp. 2413–2421.
42. Q. J. M. Huys, T. V. Maia, M. J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
43. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* **42**, 241–251 (2001).
44. P. Schwartenbeck, K. Friston, Computational phenotyping in psychiatry: A worked example. *eNeuro* **3**, ENEURO.0049-16.2016 (2016).
45. S. A. Park, S. Jeong, J. Jeong, TV programs that denounce unfair advantage impact women's sensitivity to defection in the public goods game. *Soc. Neurosci.* **8**, 568–582 (2013).
46. K. Khalvati, A. K. Mackworth, A fast pairwise heuristic for planning under uncertainty, in *Proceedings of The Twenty-Seventh AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2013), pp. 187–193.
47. G. Shani, J. Pineau, R. Kaplow, A survey of point-based POMDP solvers. *Auton. Agent. Multi-Agent Syst.* **27**, 1–51 (2013).
48. Y. Luo, H. Bai, D. Hsu, W. S. Lee, Importance sampling for online planning under uncertainty. *Int. J. Robot. Res.* **38**, 162–181 (2018).
49. E. Fehr, U. Fischbacher, S. Gächter, Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
50. K. Khalvati, S. A. Park, J.-C. Dreher, R. P. Rao, A probabilistic model of social decision making based on reward maximization, in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 5 to 10 December 2016, pp. 2901–2909.
51. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
52. D. Cousineau, Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutor. Quant. Methods Psychol.* **1**, 42–45 (2005).

#### Acknowledgments

**Funding:** This work was funded by the NSF-ANR Collaborative Research in Computational Neuroscience “CRCNS SOCIAL POMDP” n°16-NEUC to J.-C.D. and CRCNS NIMH grant no. 5R01MH112166-03, NSF grant no. EEC-1028725, and a Templeton World Charity Foundation grant to R.P.N.R. The experiments were performed within the framework of the Laboratory of Excellence “LABEX ANR-11-LABEX-0042” of Université de Lyon, attributed to J.-C.D., within the program “Investissements d’Avenir” (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). J.-C.D. was also funded by the IDEX University Lyon 1 (project INDEPTH). **Author contributions:** R.P.N.R. and J.-C.D. developed the general research concept. S.A.P. designed and programmed the task under the supervision of J.-C.D., and M.S. ran the experiment under the supervision of J.-C.D. K.K. developed the model under the supervision of R.P.N.R., implemented the algorithms, and analyzed the data in collaboration with R.P.N.R. S.M. interpreted the computational results in the context of social neuroscience. K.K. developed the reinforcement learning model after discussions with R.P. K.K., S.M., and R.P.N.R. wrote the manuscript in collaboration with S.A.P., R.P., and J.-C.D. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Both data and code are available upon request from the corresponding author. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 3 May 2019

Accepted 19 September 2019

Published 27 November 2019

10.1126/sciadv.aax8783

**Citation:** K. Khalvati, S. A. Park, S. Mirbagheri, R. Philippe, M. Sestito, J.-C. Dreher, R. P. N. Rao, Modeling other minds: Bayesian inference explains human choices in group decision-making. *Sci. Adv.* **5**, eaax8783 (2019).

## Supplementary Materials for

### **Modeling other minds: Bayesian inference explains human choices in group decision-making**

Koosha Khalvati, Seongmin A. Park, Saghar Mirbagheri, Remi Philippe, Mariateresa Sestito, Jean-Claude Dreher, Rajesh P. N. Rao\*

\*Corresponding author. Email: rao@cs.washington.edu

Published 27 November 2019, *Sci. Adv.* **5**, eaax8783 (2019)  
DOI: 10.1126/sciadv.aax8783

#### **This PDF file includes:**

Supplementary Text

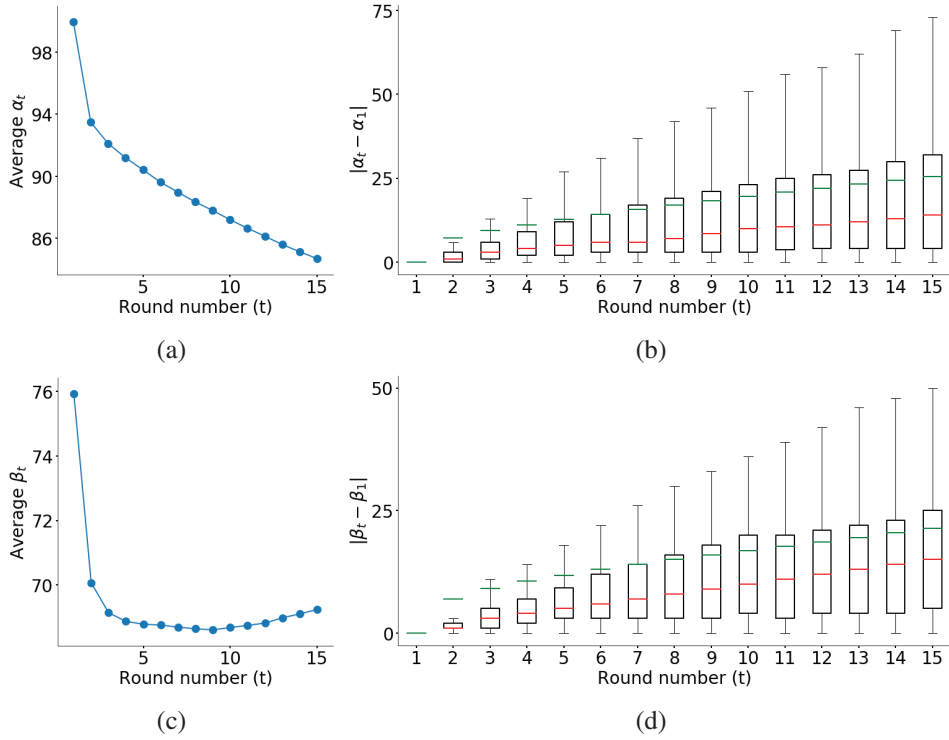
Fig. S1. Distribution and change in belief parameters over multiple rounds.

Fig. S2. Data generated by the POMDP model compared to experimental data.

## Supplementary Text

**Distribution of Belief Parameters in Each Round.** Our statistical analysis showed that despite large prior values, the actions of others during the game played an important role in determining the policy of our POMDP model (and the policies of subjects). Figure S1a shows the average  $\alpha_t$  across all games and subjects in each round. More importantly, Figure S1b shows the distribution of the difference between  $\alpha_t$  and its initial value, i.e.,  $(|\alpha_t - \alpha_1|)$ , for each round. Similarly, Figures S1c and S1d demonstrate the evolution of  $\beta_t$  over multiple rounds. As shown in these figures, the belief state parameters change quite drastically and this change increases as the game continues.

**Using the POMDP as a Generative Model** To further investigate the ability of the POMDP framework to model our experimental data in the Volunteer's Dilemma task, we performed a posterior predictive check by using the POMDP as a generative model of data, i.e., actions were sampled from  $Beta(\alpha_1, \beta_1)$  and their probability changed according to the dynamics of



**Fig. S1. Distribution and change in belief parameters over multiple rounds.** (a) Average  $\alpha_t$  across all games and subjects as a function of round number  $t$ . (b) Difference between  $\alpha_t$  and  $\alpha_1$  as a function of round number  $t$ . Red bars show the median and green bars represent the mean of distribution. Outliers are not shown. (c) Average  $\beta_t$  across all games and subjects as a function of round number  $t$ . (d) Difference between  $\beta_t$  and  $\beta_1$  as a function of round number  $t$ . Outliers are not shown.

the POMDP model (see equation 10 in Methods). Specifically, for each game, we sampled a  $\theta = \theta_1$  from the initial belief state of the subject, i.e.  $Beta(\alpha_1, \beta_1)$ , as the real initial state of the environment. In each round, contributions of others were generated based on the binomial distribution in equation 1 using the sampled  $\theta$  of that round. The next  $\theta$  were calculated based on  $\alpha$ ,  $\beta$ , and actions of that round as well as the decay rate, exactly as the POMDP model.

In the resulting synthesized data, the general patterns of both success rate and contribution probability (with  $z$  obtained from the actual experimental data) for the fitted subjects matched

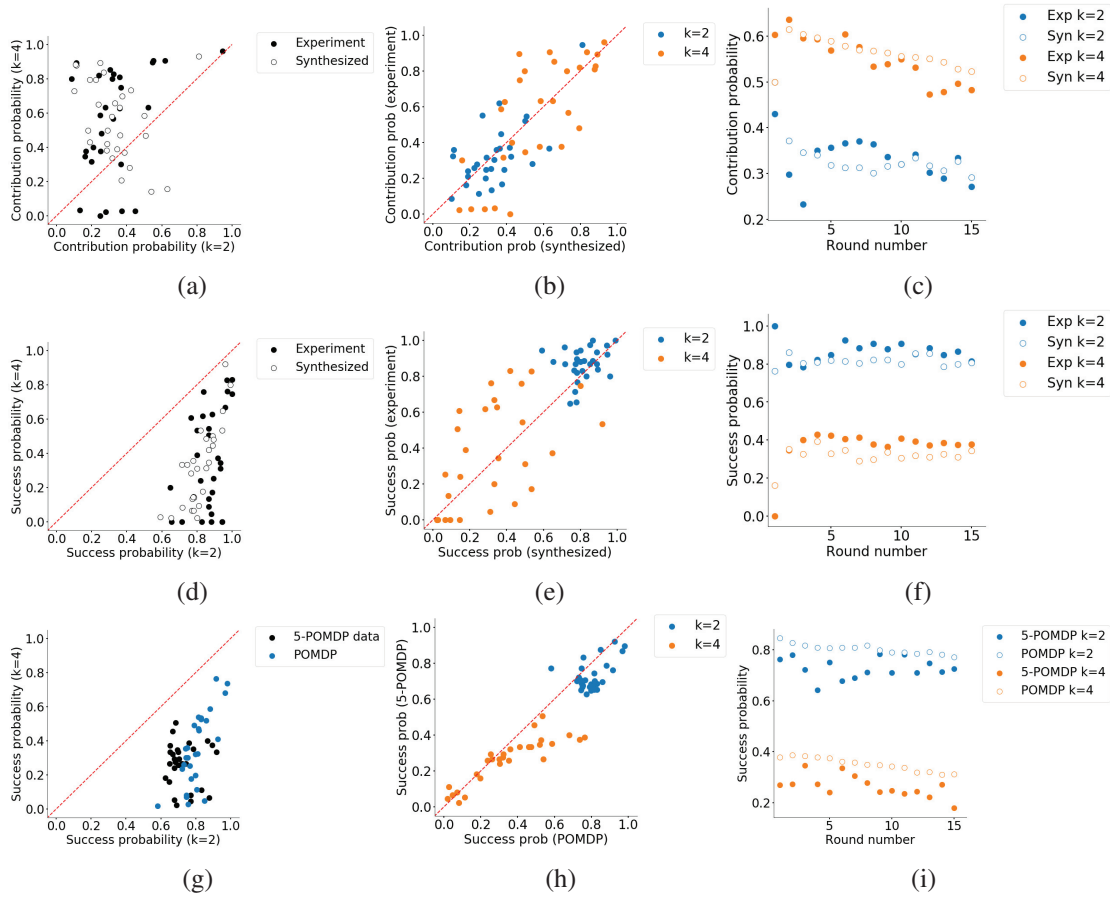
the experimental data of the subjects closely (Figures S2a-S2f). This result was robust to randomization - the same pattern was observed when the data was synthesized multiple times.

**Comparison with I-POMDP Model.** Our framework models the effect of the subject's actions on others by increasing the average contribution rate of the group by each contribution. To model higher levels of theory of mind, one can utilize an interactive-POMDP (I-POMDP) which assumes that the subject responds to  $N - 1$  policies generated by  $N - 1$  POMDPs, each modeling another individual. Each POMDP models the game with a separate set of  $\alpha_1$ ,  $\beta_1$ , and  $\gamma$  parameters. The subject however does not know the parameters of the others' models (here  $\alpha_1$ ,  $\beta_1$  and  $\gamma$ ). We tested a version of the I-POMDP model where the subject uses their own set of parameters for all members of the group (similar to our original POMDP model).

We found that our original POMDP, where the subject reasoned directly about the parameters of the group state, outperformed this I-POMDP model which had a fitting accuracy 73% with  $SD = .12$  (two-tailed paired t-test,  $t(28) = 4.91$ ,  $p = 3.53 \times 10^{-5}$ , 95% CI difference = [0.06, 0.14]). The better performance of our original POMDP over the I-POMDP model could be at least partly due to the computer algorithm used to mimic human players. To examine this potential issue, given that the later rounds are potentially more affected by the dynamics of the game, we compared the difference in fitting accuracy between the original POMDP model and the I-POMDP model for the first 7 fitted rounds of the game versus the last 7 rounds (the first round excluded). The difference in the fits for the first and last 7 rounds was not significant (two-tailed paired t-test,  $t(28) = -0.58$ ,  $p = 0.56$ , 95% CI difference = [-0.38, 0.21]).

**POMDP Model Capturing the Dynamics of Actions** We also investigated games where all group members are optimal agents to see if our POMDP model is capable of capturing the dynamics of actions by optimal agents. We created a dataset where each subject (simulated by POMDP) played with 4 POMDP agents in each of 12 games. The parameter sets of these 4

POMDPs were drawn from parameter sets fit to experimental data. In other words, this dataset captured subjects playing with other human subjects. We compared the predicted success by the (simulated) subject to actual success in the game, similar to what we did with the experimental data. The average accuracy of this prediction was 66% ( $SD = .07$ ). This accuracy was very robust across multiple runs of generated datasets. Figures S2g and S2h compare the actual success and the predicted success for the subject, similar to Figures 5e and 5d. Figure S2i shows that this match between the generated data and the model exists round by round.



**Fig. S2. Data generated by the POMDP model compared to experimental data.** (a) A subject's contribution probability in each round (on average) when the actions are generated based on the hidden state of the POMDP model (synthesized data, white circles) compared to experimental data (black circles, same data as in figure 2c). (b) Same data as (a) but comparing synthesized versus experimental contribution probability for each  $k$ . (c) Same data as (a) and (b) but with the data points binned based on round of the game. (d) Comparison of probability of group success in each round (on average) for the synthesized POMDP data compared to experimental data (black circles, same data as in figure 2d). (e) Same data as (d) but comparing synthesized versus experimental for each  $k$ . (f) Same data as (d) and (e) but with the data points binned based on round of the game. (g) Average probability of success for each subject in a generated data set where the actions come from 4 random POMDPs whose parameters were fit to experimental data from 4 random subjects (black circles) compared to the subject-fitted POMDP model's prediction of the success (blue circles) (h) Same data as (g) but comparing the generated data with the subject-fitted POMDP model's prediction for each  $k$ . (i) Same data as (g) and (h) but with the data points binned based on round of the game.





## Chapter 4

# Causal role of the medial prefrontal cortex in learning social hierarchies <sup>1</sup>

---

<sup>1</sup>Mon travail a consisté en l'analyse des données comportementales, la modélisation du comportement, l'interprétation de l'ensemble des résultats et la rédaction de la publication.

# Causal role of the medial prefrontal cortex in learning social hierarchies

Chen Qu<sup>1\*</sup>, Yulong Hang<sup>1\*</sup>, Rémi Philippe<sup>2,3\*</sup>, Edmund Derrington<sup>2,3</sup>, Philippe Jacquard<sup>4</sup> and Jean-Claude Dreher<sup>2,3</sup>

<sup>1</sup> Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, China; School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, China

<sup>2</sup> Laboratory of Neuroeconomics, Institut des Sciences Cognitives Marc Jeannerod, CNRS, Lyon, France

<sup>3</sup> Université Claude Bernard Lyon 1, Lyon, France

<sup>4</sup> EmLyon, Ecully, France

\* Equal contribution

\*Correspondence: [dreher@isc.cnrs.fr](mailto:dreher@isc.cnrs.fr)

## **Abstract**

Social hierarchy is a fundamental principle of social organization and an important attribute of social community stability and development in populations. Learning social hierarchies is critical to individuals' self-perception and orientation as well as adaptation to social organization. Little is known about the neurocomputational mechanisms supporting learning social hierarchies. Here, we demonstrate a causal role of the medial prefrontal cortex (mPFC), a core region in social cognition, when social hierarchy is learned. Using transcranial direct current stimulation (tDCS) over the mPFC we tested participants (n=128) while they completed a hierarchy learning task during. Anodal stimulation selectively impaired social hierarchy learning compared with a non-social learning task. A Bayesian model captured the effect of tDCS on social hierarchy learning better than a number of reinforcement learning models, indicating that this behavioral effect of anodal stimulation is specifically due to higher forgetfulness. Anodal stimulation also impaired transitive inferences, but only during early blocks when learning was not established. Our results offer a mechanistic account of the causal role of the mPFC in learning social hierarchy maps and provide causal evidence for the mPFC in computing forgetfulness during social hierarchy learning.

**Keyword:** Social hierarchy learning, Reinforcement learning, Bayesian inference, noninvasive brain stimulation, medial prefrontal cortex

## **Significance statement**

Social hierarchy learning affects social adaptation and organization. For individuals, social hierarchy knowledge facilitates self-perception and orientation of interactions in society. For society, social hierarchy is fundamental in maintaining the stability of social communities. Using transcranial direct current stimulation over the medial prefrontal

cortex, we conducted an experiment in which participants learned both social and non-social hierarchies through trial and error training and transitive inference testing. The mPFC-targeted stimulation selectively impaired the computation of uncertainties and performance of social hierarchy learning. This specialized functional region in the human brain seems to enable us to establish and update social hierarchy relationships to adjust our position in social groups. Interestingly, this process can be selectively modulated by external intervention.

## **Introduction**

We live in a social environment which is regulated by a variety of hierarchical competitions (Koski, Xie, and Olson 2015). Optimization of our social interactions requires us to perceive status cues and continuously update hierarchical relationships, by determining the power of others relative to ourselves, to make social judgments in daily life (Qu et al. 2017). Social hierarchy, as a group structure, exists widely in nature (Chiao et al. 2009, Wang et al. 2011), and is crucial to maintain the stability of populations and the health of individuals (Qu et al. 2017, Cheney and Seyfarth 2018, Sapolsky 2005). Animal studies, for example, show that fish can infer the social hierarchy of competitors by observation learning (Grosenick, Clement, and Fernald 2007), and adjust their size and growth rate according to their hierarchical position in the group (Buston 2003). Similarly, social hierarchy can affect human behavior (Cummins 2000), such as decision making (Santamaria-Garcia et al. 2014) and empathy (Feng et al. 2016), enabling people to choose favorable alliances in social competition and avoid potential conflicts. Failure to accurately perceive one's position in the social hierarchy can affect human health and increase the possibility of mental diseases such as externalizing disorders and social anxiety (Boyce 2004, Sapolsky 2005, Muscatell et al. 2012).

Individuals can assess hierarchy information in several ways (Qu et al. 2017), including by the perception of dominance-related cues, by observation learning (Kumaran et al. 2016, Kumaran, Melo, and Duzel 2012), and through competitive interactions, as is often the case in wildlife (Ligneul et al. 2016). Although assessing the strength of competitors by dominance cues, such as body postures, facial expressions and physical attributes (Marsh et al. 2009, Todorov et al. 2008, Zink 2008) is rapid and convenient, the information from such cues does not always coincide with the real hierarchy status. In addition, learning dominance relationships through direct dyadic competitive interactions, by experiencing successive victories or defeats

against competitors, is costly in terms of stress or physical injury, and is also time-consuming (Ligneul et al. 2016). Thus, learning social hierarchy by observation and without competitive interactions is an economic way to acquire social hierarchy knowledge.

The learning of social hierarchy engages the medial prefrontal cortex (mPFC), as well as the hippocampus and other structures (Kumaran et al. 2016, Kumaran, Melo, and Duzel 2012, Ligneul et al. 2016). Using model-based functional magnetic resonance imaging (fMRI), Kumaran et al. (2016) developed an observational hierarchy learning task that distinguished training and test phases to study the neural representations of the learning process (during training phase) and of transitive inferences (during test phase). A Bayesian inference scheme, called the Sequential Monte-Carlo (SMC), tracks social hierarchy learning better than Reinforcement Learning models (RL-ELO) (Kumaran et al. 2016). This SMC model captured behavior more effectively than RL models when participants are more uncertain about the relative power of individuals during the early phase of learning. The mPFC was selectively involved when updating hierarchies that include the participant (Apps and Sallet 2017), and represents self-related information (Joiner et al. 2017, Kelley et al. 2002, Rameson, Satpute, and Lieberman 2010, Webber 2011). In contrast, the hippocampus (and amygdala) appear to be involved in general social hierarchy learning (Kumaran, Melo, and Duzel 2012), independently of the presence of the participant in the hierarchy. More generally, recent models and experimental studies have proposed that the same neuronal representations that map space may be extended to a broad range of non-spatial problems in abstract cognitive space (Behrens et al., 2018; Garvert et al., 2017; Gershman and Niv, 2010; Niv, 2019; Schuck et al., 2016; Stachenfeld et al., 2017; Wang et al., 2018; Whittington et al., 2020; Wilson et al., 2014). These studies report that both the mPFC and the hippocampus are involved in non-spatial relational memory tasks allowing to make transitive inferences (Dusek and Eichenbaum, 1997; Whittington et al., 2020 ; Park et al. 2019).

However, it is unclear whether a specific component of this mPFC-hippocampus network (eg. specifically the mPFC) is causally necessary in two distinct processes needed to organize abstract relational information into a cognitive map: learning the rank (relationship between two items) and making transitive inferences to guide novel inferences. Second, it is unclear whether learning and transitive inference processes are performed in similar ways across domains (eg. social vs non social).

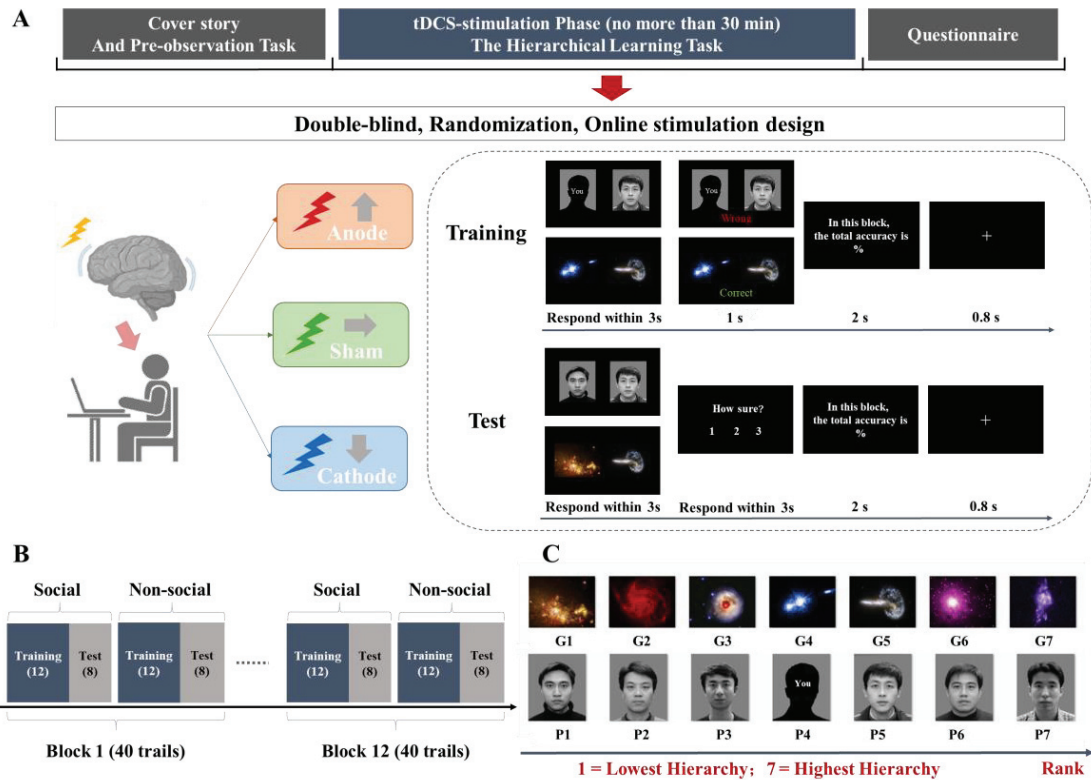
Previous research reporting an important role of the mPFC in learning social hierarchies by observation relied only on correlational fMRI approaches. However, it is unclear whether the mPFC causal evidence for this relationship, or any identification of what drives this covariation. Furthermore, although a previous study indicated the Bayesian Inference Scheme (SMC) captured social hierarchy learning better than the Reinforcement learning models (RL-ELO), it remains unclear whether this is restricted to social hierarchy learning or whether this would also be the case for a similar non-social hierarchy learning task. The combination of computational modeling and tDCS allowed us to explore four avenues: (i) to causally link the role of mPFC and the neural computations that support learning of hierarchy by observation (i.e. by trial and error); (ii) to establish whether the mPFC plays a causal role only for learning social hierarchy but not for non-social hierarchy; (iii) to examine whether Bayesian inference is a general feature of learning hierarchies (i.e. both non social and social) or is more specific to learning social hierarchies; (iv) to investigate whether the mPFC plays a causal role during learning hierarchies alone or also during the transitive inference processes.

Here, we combined transcranial direct current stimulation (tDCS) and computer modeling in 128 participants to explore the causal relationship between mPFC and hierarchical learning processes. tDCS is a noninvasive brain stimulation method that can modulate the neural excitability of specific brain regions using a low electrical current (Brunoni et al. 2011). This approach allowed us to explore the impact of mPFC-targeting brain stimulation on the behavior of learning hierarchies. We used a double-



blind sham-control, and online stimulation design, in which participants were randomly assigned to receive anodal (n=42), cathodal (n=42), or sham (n=44) stimulation over the mPFC. Participants were asked to imagine that recently they had joined a technology company that detected precious minerals in different galaxies. As a new member of the company, they were instructed to learn the hierarchical relationship between the staff members to help them adjust to work. In the meantime, they also needed to learn the mineral contents of the different galaxies to familiarize themselves with the company business (SI Appendix).

During tDCS brain stimulation, participants performed a Hierarchy Learning task which was developed, based on Kumaran et al. (2012). We added self-information in the social condition, including Training and Test phases in both social and non-social conditions (see Fig.1A and B). Human faces (matching to the participant's gender) were used in the social condition, whereas images of galaxies were used as the non-social condition (see Fig.1C). In the Training phase, participants were required to view pairs of hierarchically adjacent pictures (e.g., P4 versus P5, G4 versus G5; P=person and G=galaxy; P4 indicating the participant, "YOU"; SI Appendix). They indicated which picture they thought had a higher status (social) or more minerals (non-social), with the correct feedback for each trial. Thus, participants had to learn the relationship between the items, through trial and error, and thus update their hierarchy knowledge. Unlike the training phase, in the testing phase participants were required to use the hierarchy information learned in the training phase to make transitivity judgments concerning the hierarchical relationship between two non-adjacent entities, (e.g., P1 vs P5, G1 vs G5; SI Appendix), with no feedback provided, and also rate the confidence of their decisions from 1 (guess) to 3 (very sure).



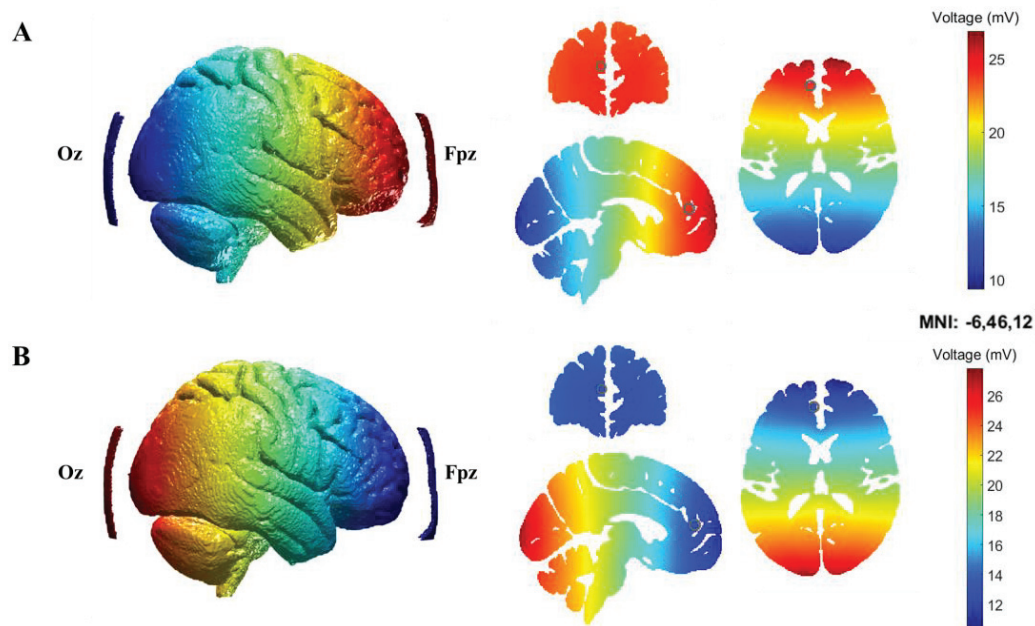
**Fig. 1 Experiment Design.** *A and B) Procedure.* First participants were given the cover story, then they completed an observation task to familiarize themselves with stimuli. The pictures presented in the observation task were the same as in the hierarchy learning task. Next, participants were randomly assigned to the Anode, Sham, or Cathode stimulation conditions, and instructed to perform the hierarchical learning task, including training trials and test trials. At the end of the experiment, participants were required to complete some questionnaires. **(A) Hierarchy Learning Task. Training phase:** There were 12 blocks of training trials, and each block included a 12 trial mini-block composed of the 6 training trial paired items (P1 vs P2, P2 vs P3, P3 vs P4, P4 vs P5, P5 vs P6, P6 vs P7) repeated twice. Participants were presented adjacent items of the hierarchy (e.g., P4 vs P5, G4 vs G5, where P4 = “You”, the participant, having rank equal to 4; and G4 = galaxy of rank equal to 4). The participants had to indicate the person they thought had higher status or the galaxy with more minerals. Through the correct feedback of their selection, they were able to learn the hierarchical relationships between the adjacent items. **Test phase:** There were 12 blocks of test trials one after each training block. Each included an 8 trial mini-block composed of the 8 test

trial paired items (P1 vs P4, P2 vs P4, P2 vs P5, P2 vs P6, P3 vs P5, P3 vs P6, P4 vs P6, P1 vs P7). Thus participants were required to view non-adjacent items in the hierarchy (e.g., P1 vs P5, G1 vs G5), infer which was the higher-ranked item, and rate their confidence of their selection—no feedback was provided. **(B) For the social condition**, there were 12 blocks including 12 training trials (6 adjacent paired items: P1 vs P2, P2 vs P3, P3 vs P4, P4 vs P5, P5 vs P6, P6 vs P7, each repeated twice) and then 8 test trials (8 non-adjacent paired items: P1 vs P7, P2 vs P6, P3 vs P5, P1 vs P4, P2 vs P4, P2 vs P5, P3 vs P6, P4 vs P6). **The non-social condition** was identical except that the pictures were of galaxies. **C) Stimuli.** Note that for the social condition female participants were presented with faces of women, whereas male participants were presented with male faces. For the non-social conditions, the galaxy pictures were the same for females and males (1 = Lowest position in the hierarchy, 7 = highest).

## Results

### Impact of mPFC targeted Brain Stimulation on Hierarchy Learning Process.

According to previous studies (Sellaro et al. 2015a, Casula et al. 2017, Liao et al. 2018, Guo et al. 2019, Wang et al. 2019, Sellaro et al. 2015b), we adopted the Fpz-Cz montage (EEG 10-20 system, both 70x50mm pad) with 1.5mA current as stimulation protocol. To ensure that the electrode montage effectively stimulated the mPFC, electrical potential simulations were performed using ROAST (Huang et al. 2018, Huang et al. 2019) with the MNI template brain. As illustrated in Fig.2A and B, the simulation shows that the voltage gradient spread through the prefrontal cortex and targeted mPFC (MNI: -6, 46, 12; from Kumaran et al. 2016).



**Fig. 2 Current flow simulation results.** **A) Anode stimulation.** Simulated voltage distribution over the prefrontal cortex (left), and in coronal, sagittal, and axial slices (right) using the anodal montage with the MNI template brain. The black circle shows the targeted mPFC coordinate from Kumaran et al. (2016) (MNI: -6, 46, 12). **B) Cathode stimulation.** Same as in A, but for the cathodal stimulation montage.

To investigate how brain stimulation modulated hierarchical learning we conducted panel logistic or linear regressions, depending on the form of dependent variable (binomial or continuous), on the population-average (generalized estimating equation, GEE). This analysis allowed us to observe the effect of stimulation at the level of the population, taking into account the effect of time. We used panel data of 480 trials clustered on each of 128 participants. The dependent variables were either the accuracy, the reaction time, or the confidence rating for each trial. The independent variables were the tDCS stimulation (Anode/Sham/Cathode), hierarchy condition (Social/Non-social), and block number (1-12). The percentage change effect was estimated via the marginal effect, which measures the percentage change of the

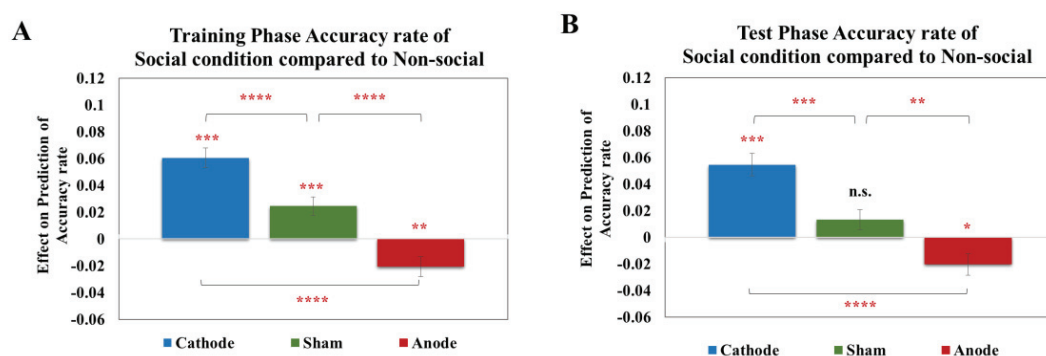
dependent variable versus a 1% change of the independent variable when holding all other independent variables constant.

We first focused on the impact of the mPFC-targeted brain stimulation regimes on hierarchy learning behavior between the Social and Non-Social conditions. During the training phase, under both Cathode and Sham stimulation, participants learned much better in the Social Condition relative to the non-social condition. This is indicated by the increased probability of accuracy in Social compared with Non-Social learning tasks [Accuracy rate Cathode Social>Non-social:  $\beta = 0.061$ , SE = 0.007,  $z = 8.27$ ,  $P < 0.001$ , 95% CI (0.046, 0.075); Sham Social>Non-social:  $\beta = 0.024$ , SE = 0.070,  $z = 3.52$ ,  $P < 0.001$ , 95% CI (0.011, 0.038); see Fig. 3A]. Moreover, cathodal stimulation increased social hierarchy learning compared to sham stimulation [Contrasts of average marginal effects Cathode > Sham in Training phase:  $\chi^2(1) = 12.68$ ,  $P < 0.001$ ]. On the contrary, anodal stimulation significantly decreased accuracy in the Social condition compared to Non-Social [Anode Social < Non-social:  $\beta = -0.021$ , SE = 0.007,  $z = -2.80$ ,  $P = 0.005$ , 95% CI (-0.035, -0.006); Contrasts of average marginal effects Anode < Sham in Training phase:  $\chi^2(1) = 19.82$ ,  $P < 0.0001$ ]. Furthermore, participants spent significantly more time learning the social hierarchy than the non-social when under anodal stimulation [Reaction Time Anode > Sham in Training phase:  $\chi^2(1) = 25.44$ ,  $P < 0.0001$ ; see SI Appendix], suggesting a specific impairment of social hierarchy learning under anodal stimulation.

Results in the test phase extended the observed effects of tDCS stimulation to transitive inferences. That is, under anodal stimulation participants' performance to infer social hierarchy was significantly worse than performance on the non-social hierarchy task [Anode Social<Non-social:  $\beta = -0.020$ , SE = 0.008,  $z = -2.53$ ,  $P < 0.05$ , 95% CI (-0.036, -0.005)] (Fig 3B). Under cathodal stimulation it was significantly better [Cathode Social>Non-social:  $\beta = 0.055$ , SE = 0.008,  $z = 6.49$ ,  $P < 0.001$ , 95% CI (0.038, 0.072); Contrasts of average marginal effects Anode < Sham in Test phase:  $\chi^2(1) = 9.25$ ,  $P < 0.005$ , Cathode > Sham in Test phase:  $\chi^2(1) = 13.39$ ,  $P < 0.0005$ ; see Fig.

3B]. However, under sham stimulation, unlike during training trials, there was no difference in performance accuracy between social and non-social hierarchy learning when making transitive inference [Sham:  $\beta = 0.013$ ,  $SE = 0.008$ ,  $z = 1.76$ ,  $P = 0.078$ , 95% CI (-0.001, 0.028); see Fig. 3A and B].

Overall, these results indicate that the mPFC-targeted brain stimulation has different impacts during social and non-social hierarchy learning. Cathodal stimulation improved social hierarchy learning whereas anodal stimulation impaired it. Without tDCS (sham condition), participants learned social hierarchies better than non-social hierarchies, but there was no significant effect on transitive inferences.



**Fig. 3 Effect of brain stimulation on hierarchy learning accuracy.** A) and B) tDCS modulation of Social condition compared to Non-Social in Training and Test phase accuracy. The blue bars indicate cathodal modulation of Social condition compared to Non-Social. The green bar shows the effect of sham. The red bar shows the effect of anodal stimulation. (\*indicates  $p < 0.05$ , \*\*indicates  $p < 0.005$ , \*\*\*indicates  $p < 0.001$ ; \*\*\*\*indicates  $p < 0.0001$ ; Error bars show SEM)

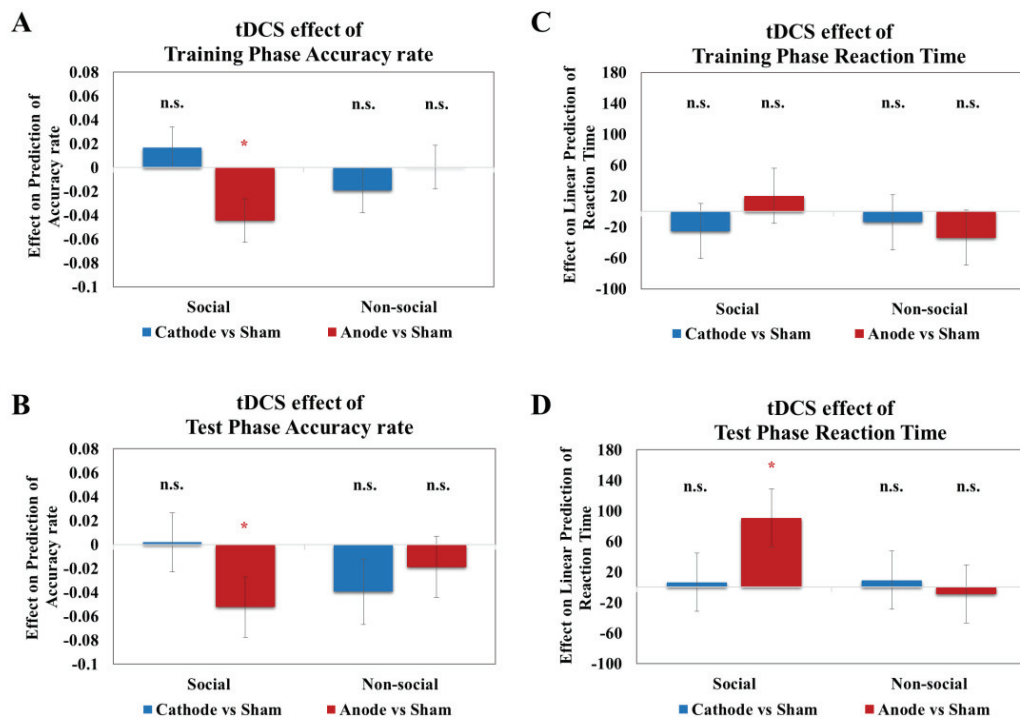
### Anode Stimulation Selectively Impairs Hierarchy Learning in the Social Condition

We estimated the marginal effect of learning a Social or Non-Social hierarchy under each specific mPFC-targeted stimulation compared to the Sham condition. This analysis was performed independently on performance in the Training and Test phases.

In line with prior findings, anodal stimulation resulted in lower accuracy during social hierarchy learning compared to non-social hierarchy learning in both the Training and the Test trials [Social condition Training Anode< Sham:  $\beta = -0.045$ , SE = 0.018,  $z = -2.45$ ,  $P = 0.014$ , 95% CI (-0.080, -0.009); Social condition Test Anode< Sham:  $\beta = -0.052$ , SE = 0.025,  $z = -2.07$ ,  $P = 0.038$ , 95% CI (-0.102, -0.003); see Fig. 4A and B]. Cathodal stimulation showed no significant effect on social hierarchy learning and there was no significant impact of either anodal or cathodal tDCS during the learning of non-social hierarchies.

### **Effect of tDCS on Reaction Times**

Brain stimulation showed a selective effect on the reaction time during the Test phase but not the Training phase and only in the Social condition, [Social condition Test Anode> Sham:  $\beta = 90.619$ , SE = 38.061,  $z = 2.38$ ,  $P = 0.017$ , 95% CI (16.021, 165.2161); see Fig. 4C and D]. These findings substantiate that the mPFC-targeted tDCS stimulation selectively modulated social hierarchy learning, but not the learning of non-social hierarchies, and that anodal stimulation specifically impairs performance in the social hierarchy learning condition.



**Fig. 4 Brain stimulation selectively modulates social hierarchy learning. A) and B) tDCS effect on Training and Test phase accuracy in Social and Non-Social conditions. The blue bar indicated cathode modulation compared to sham stimulation. The red bar is the effect of anodal stimulation compared to the sham. C) and D) Effect of tDCS on reaction times in the Training (C) and Test (D) phases during social and non-social hierarchy learning. (\*indicates  $p < 0.05$ , \*\*indicates  $p < 0.005$ , \*\*\*indicates  $p < 0.001$ ; \*\*\*\*indicates  $p < 0.0001$ ; Error bars show SEM)**

### tDCS impact on block to block learning performance

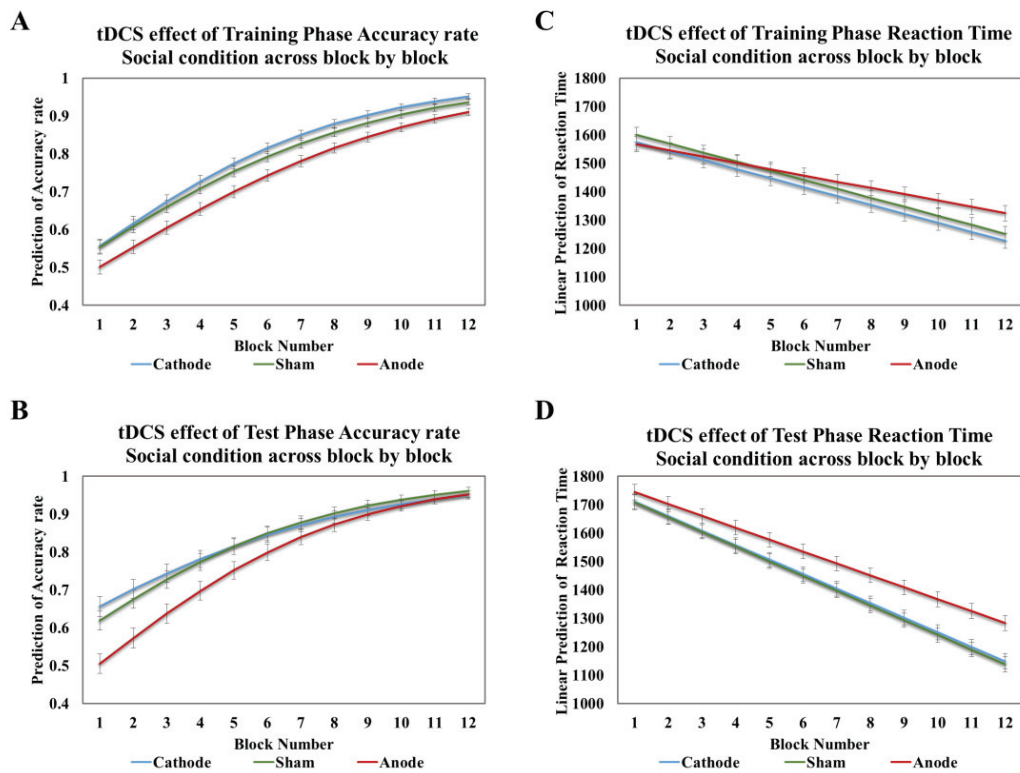
Next, we explored the effect of tDCS on the rate of learning social hierarchies from trial block to trial block. Interestingly there were contrasting effects of stimulation in the training and the test phases (Fig 5). First, there was a significant improvement in performance from block to block over consecutive blocks in both the training and the test phases [Training block:  $\beta = 0.035$ ,  $SE = 0.001$ ,  $z = 54.88$ ,  $P < 0.0001$ , 95% CI



(0.035, 0.037); Test block:  $\beta = 0.031$ ,  $SE = 0.001$ ,  $z = 35.78$ ,  $P < 0.0001$ , 95% CI (0.029, 0.032)] which confirms that the participants are indeed learning the social hierarchies, and are efficiently building on this learning to make successful transitive inferences as they progress in the blocks. There was no significant effect of tDCS on performance during the trial and error learning trials in the Training blocks [same slopes in the accuracy rate (block) estimated function; Cathode versus Sham:  $\chi^2(1) = 1.35$ ,  $P = 0.24$ , Anode versus Sham:  $\chi^2(1) = 0.21$ ,  $P = 0.65$ , see Fig. 5A]. However there was a specific effect of anodal stimulation, which increased the rate of acquisition of the ability to make transitive inferences from the learned trial and error trials when compared to either cathodal tDCS or sham [Test block Slope Anode > Sham:  $\beta = 0.007$ ,  $SE = 0.002$ ,  $\chi^2(1) = 10.72$ ,  $P = 0.0011$ , 95% CI (0.003, 0.011); Slope Cathode > Sham:  $\beta = -0.004$ ,  $SE = 0.002$ ,  $\chi^2(1) = 3.70$ ,  $P = 0.055$ , 95% CI (-0.008, 0.00008); see Fig. 5B]. These results show that mPFC stimulation did not affect the learning of adjacent stimuli over the successive blocks, but rather the ability to make transtansitive inferences. Interestingly, it appears that this effect is due to worse performance during the training phase under anodal stimulation rather than any adverse effect on the making of transitive inferences, which appears to improve and compensate for the lower performance on the training phase (see Fig. 4 and 5).

Anodal stimulation also resulted in a decrease in the rate of reduction of reaction times from trial block to trial block during both the Training and the Test phase compared to Sham whereas cathodal stimulation resulted in no significant change compared to Sham [Training: Slope anode > sham:  $\beta = 8.909$ ,  $SE = 1.558$ ,  $\chi^2(1) = 32.72$ ,  $P < 0.0001$ , 95% CI (5.856, 11.56); Slope cathode > sham:  $\beta = 4.566$ ,  $SE = 1.558$ ,  $\chi^2(1) = 8.59$ ,  $P = 0.0034$ , 95% CI (-0.008, 0.00008) see Fig. 5C; Test: Slope anode > sham:  $\beta = 12.12$ ,  $SE = 1.996$ ,  $\chi^2(1) = 36.88$ ,  $P < 0.0001$ , 95% CI (8.210, 16.03); Slope cathode > sham:  $\beta = 9.733$ ,  $SE = 1.996$ ,  $\chi^2(1) = 23.78$ ,  $P < 0.0001$ , 95% CI (5.821, 13.65), see Fig. 5D]. This suggests that anodal mPFC stimulation influence the social hierarchical knowledge updating and the making of transitive inferences

differently.



**Fig. 5 Anodal tDCS modulates social hierarchy learning from trial block to block.**

*Modulation of the evolution of the success rate, from trial block to trial block by tDCS over consecutive Training trial blocks (A) and consecutive Test trial blocks (B). Modulation of the evolution of Reaction Time shortening from trial block to trial block by anodal tDC. Anodal tDCS results in a reduction of the rate at which reaction time is reduced over consecutive trial blocks in both Training (C) and Test trial blocks (D). The accuracy and reaction times over consecutive training and test trials are indicated in blue (Cathode tCDS), Green (Sham tCDS) and red (Anode tCDS). Error bars show SEM.*

During Test trials, participants were also required to rate the confidence in their choices. The analysis of confidence ratings in the transitivity judgements showed that under cathode and sham stimulation, participants were more confident in their Social hierarchy than Non-Social hierarchy decisions [Cathode Social>Non-social:  $\beta = 0.129$ ,

SE = 0.012,  $z = 10.60$ ,  $P < 0.001$ , 95% CI (0.105, 0.153); Sham Social>Non-social:  $\beta = 0.155$ , SE = 0.013,  $z = 11.94$ ,  $P < 0.001$ , 95% CI (0.130, 0.181); see Fig.6A]. This was not the case for participants under Anodal stimulation [Anode Social vs Non-social:  $\beta = 0.022$ , SE = 0.127,  $z = 1.74$ ,  $P = 0.081$ , 95% CI (-0.003, 0.047)]. Moreover, in the Social compared to Non-Social conditions, participants under anodal stimulation were less confident in their judgements compared to those under sham stimulation. [Contrasts of average marginal effects Anode < Sham:  $\chi^2(1) = 53.86$ ,  $P < 0.0001$ ; see Fig.6A]. These results are consistent with the worse performance observed under anodal tDCS in the social condition compared to non-social, suggesting that participants were aware of their poorer performance in this condition.

### **mPFC-targeted Anode tDCS has different impacts on learning based on social ranks**

The social comparison theory posits that people are driven to compare themselves to others for accurate self-evaluations (Festinger, 1954). Specifically, people compare themselves to others in two opposite directions -downward and upward comparisons- that differ in motivations, comparison targets and consequences (Latane, 1966; Wills, 1981). Accordingly, we thought to split trials according to those involving higher hierarchical status (Trials included Social: P4, P5, P6, P7; Non-social: G4, G5, G6, G7) and lower hierarchical status (Trials included Social: P1, P2, P3, P4; Non-social: G1, G2, G3, G4) in the Training phase (SI Appendix). A previous study reported higher mPFC decreasing activity with greatest magnitude for the highest social rank (14). We thus predicted that anodal tDCS to the mPFC should preferentially disturb the learning of higher social ranks. Indeed, anodal stimulation is thought to depolarize neurons in the targeted area facilitating their excitation. Confirming this prediction, the tDCS-induced deficit in social hierarchy learning impinged asymmetrically on trials involving higher social ranks [Social higher status Anode<Sham:  $\beta = -0.055$ , SE = 0.020,  $z = -2.69$ ,  $P = 0.007$ , 95% CI (-0.095, -0.015); see Fig.S3A], with no significant effect on

trials involving lower social ranks [Social lower status Anode vs Sham:  $\beta = -0.030$ , SE = 0.021,  $z = -1.44$ ,  $P = 0.150$ , 95% CI (-0.070, 0.011); see Fig.S3C]. No similar asymmetrical effect was observed with respect to reaction times. It could be argued that this effect could be due to the participant's presence in the social hierarchy but not in the non social hierarchy. However, this effect cannot be accounted for by self-involvement or related factors because it remained robust when trial involving the participant (P4) were excluded [Social training Accuracy Anode<Sham:  $\beta = -0.053$ , SE = 0.020,  $z = -2.69$ ,  $P = 0.007$ , 95% CI (-0.091, -0.014); Social test Accuracy Anode<Sham:  $\beta = -0.047$ , SE = 0.026,  $z = -1.83$ ,  $P = 0.068$ , 95% CI (-0.097, 0.003); see SI Appendix].

These results implied that the mPFC updated social hierarchy information concerning the whole of one's own social group rather than only oneself. Moreover, as shown in SI Appendix Table S1, there was no significant difference in age, choice bias, the belief of cover story manipulation, the sensation of tDCS stimulation, and social dominance orientation among the three tDCS stimulation groups. Thus, any effect of the groups on social hierarchy learning behavior cannot be accounted for by the preexisting group differences or sensation of the current stimulus.

Taken together, these results suggest a causal role of mPFC in tracking the development of knowledge about a social, but not a non-social hierarchy. Anodal tDCS on the mPFC selectively impaired social but not non-social hierarchy learning performance. The anodal tDCS had distinct effects on the updating of social hierarchy knowledge and the making of transitivity inference judgments and was also specific to elements from hierarchy members of higher status. Finally, to tackle the mechanistic role of mPFC, we used model-based analyses to further understand its computational function.

**SMC-model scheme captures the social hierarchy learning behavior better than Bayesian inference models**

To better understand the computational processes engaged in hierarchy learning and the specific computational role supported by mPFC that is disrupted by anodal tDCS, we tested a number of RL and Bayesian inference models to fit the behavioral data. In total, we fitted six different models: four of which have been described recently (Kumaran et al. 2016) and two more RL models were developed to account for the asymmetrical learning after a victory or a defeat or asymmetric learning of relationships to hierarchical superiors and inferiors (SI Appendix).

For RL, we fitted a Rescorla Wagner (RW) model to test the hypothesis of a direct association between stimulus and outcome, but because of the symmetry of the hierarchies (except for the extremities), items were blocked to a null value with opposing updating values. We also tested an RL-ELO model, which is known to successfully learn hierarchies (e.g. in chess) by updating the item's value as a function of the value of the opponent's items, as well as the Value transfer model which add an indirect learning proportional to the winning item to the classical RW update. We derived two more models from the RL-ELO by adding an extra free parameter to allow the model to learn differently from a victory or a defeat, or, for higher and lower stimuli in the hierarchy. Finally, we included the Sequential Monte-Carlo (SMC) model, which maintains a probability distribution of value and implements forgetting (Doucet et al. 2000).

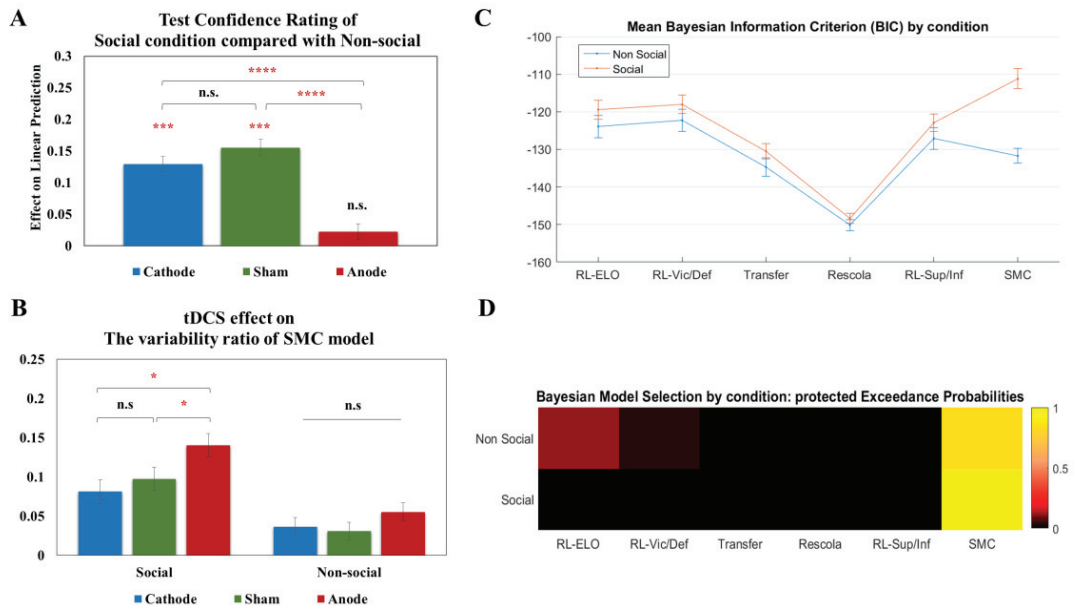
We first estimated the free parameters of each model for each participant in maximizing the likelihood of the model given the behavior in aggregate training and test trials. Then, we conducted a Variational Bayesian Model Selection for group study (BMS) to limit the outliers impact on the analysis. We used the BIC measure in the BMS to account for the number of free parameters in different models for the social and non-social conditions. The SMC model was the most frequent best model in our population in both social (protected exceedance probability,  $pEP = 1$ ) and non-social ( $pEP = 0.8338$ , see Fig.6D) hierarchy learning.  $pEP$  is the protected exceedance probability, which is the probability that one model is more frequent in the population

than the other, corrected by the Bayesian Omnibus Risk (Rigoux et al. 2014). Fig.6C indicates that the SMC model is far better than the other models in the social condition, however its superiority is far less marked in the non-social condition. Models comparison shows that, consistent with prior findings (Kumaran et al. 2016), the SMC model captures hierarchy learning behavior more effectively than the RL-ELO model and this is particularly true for the learning of social hierarchies.

### **Effect of tDCS on parameters from the Sequential Monte-Carlo model**

We ran a fixed effect analysis of the SMC model to determine how tDCS affects the distribution of the free parameters across the population. This allowed us to link mechanistic hypotheses about learning and making inferences to behavioral observations. The SMC model comprises two parameters, the Inverse Temperature and the Variability Ratio. The Inverse Temperature captures how participants exploit their knowledge about the hierarchy. The higher the Inverse Temperature, the more deterministic the participants become. The closer to zero the Inverse Temperature, the more participants tend to explore. The variability ratio captures the forgetfulness or the uncertainty in data learning. A higher value indicates more forgetfulness or increased uncertainty as captured by the drift of particles in the SMC model (SI Appendix). Higher values of Variability Ratio result in slower learning. The two-way ANOVA showed that the Variability Ratio was significantly modulated by the hierarchy conditions [ $F(1,125)=39.599$ ,  $P<0.001$ ,  $\eta^2=0.241$ ] and tDCS stimulation group [ $F(2,125)=4.600$ ,  $P<0.05$ ,  $\eta^2=0.06$ ]. We further examined whether tDCS affected both the Social and Non-Social conditions. Consistent with the behavioral results, during anodal stimulation, the Variability Ratio (uncertainty) of social hierarchy learning was significantly higher than the Sham [Anode>Sham: Mean difference=0.429, SD=0.021,  $P=0.044$ , see Fig.6B]. This explains the worse performance in the social hierarchy learning during both the Training and Test phases in the Anode condition. As for the behavioral results, there was no difference between the Cathode tDCS and Sham

conditions [Cathode<Sham: Mean difference=-0.016, SD=0.021, P=0.441; see Fig.6B].



**Fig. 6 Results of models comparison and effect of anodal stimulation on parameters from the SMC model.** A) tDCS modulation of Social condition compared to Non-Social in Test phase confidence rating. Bars indicate the proportional difference between confidence ratings in the Social Condition decisions expressed as a percentage of the Non-Social Condition decisions. The blue bar indicates cathodal tDCS, the green bar Sham, and the red bar anodal tDCS. B) Bayesian Model Selection (BMS) by conditions. Colormap represents the corrected probability that one model is more frequent than all others in social and non social conditions (pEP). C) Mean of Bayesian Information Criterion (BIC) by model across the population for Social (orange) and Non-Social (blue) condition D) Effect of tDCS on the SMC model's Variability Ratio. Sham: green, Cathode: blue, Anode: red; Social: orange. (\*indicates  $p < 0.05$ , \*\*indicates  $p < 0.005$ , \*\*\*indicates  $p < 0.001$ ; \*\*\*\*indicates  $p < 0.0001$ ; Error bars show SEM)

The analysis of Inverse Temperature shows a significant main effect of hierarchy condition [ $F_{(1,125)}=6.074$ ,  $p=0.015$ ,  $\eta^2=0.046$ ], such that the Social condition has a higher value of Inverse Temperature than the Non-Social condition. This reflects the

better performance in the Social Condition, because higher values of temperature indicate more exploitative behavior, and less explorative behavior. This finding suggests that participants exploit the low difference between noisy values of social hierarchy stimuli (faces) than in stimuli representing galaxies in the Non-Social condition. It explains why under Sham stimulation participants perform better in the social condition during the training phase. However, brain stimulation had no effect on this parameter. This could be explained by the fact that the purpose of SMC captures uncertainty in the process and often uncertainty is greater in social situations.

We collected two supplementary measures that could be used to compare the quality of models: Reaction Time (RT) in every trial and Confidence Rating in test trial. Usually we consider this data as related with the uncertainty during the decision represented by the entropy of choices in the model, which is a common measure of RL-ELO and SMC model in a trial by trial basis. To see which of the SMC and RL-ELO model better explain the RT in social and non-social we fit two linear regression making trial by trial entropy of SMC and RL-ELO model compete to explain the variance in the RT, one in social one in non-social. For both social and non-social SMC and RL-ELO model entropy explain the RT but the SMC model explain more variance of the RT (Social :  $\beta_{RL-ELO} = 53.8$ ;  $p=0.001$  and  $\beta_{SMC}=841$ ;  $p<0.0005$ , we test the difference of  $\beta$  with a chi2 test  $\text{Chi2}(1)=579$ ;  $p<0.00005$ ; Non-social :  $\beta_{RL-ELO} = 156$ ;  $p<0.0005$  and  $\beta_{SMC}=424$ ;  $p<0.0005$ , we test the difference of  $\beta$  with a chi2 test  $\text{Chi2}(1)=61.6$ ;  $p<0.00005$ ). We did the same analyze on the confident rating and found similar results for the social condition (Social :  $\beta_{RL-ELO} = -0.07$ ;  $p=0.001$  and  $\beta_{SMC} = -1.14$ ;  $p<0.0005$ , we test the difference of  $\beta$  with a chi2 test  $\text{Chi2}(1)=470$ ;  $p<0.00005$ , but an opposite result in non-social, meaning that the confidence rating in non-social was better explained by the entropy of the RL-ELO model (Non-social :  $\beta_{RL-ELO}=-0.70$ ;  $p<0.0005$  and  $\beta_{SMC} = -0.45$ ;  $p<0.0005$ , we test the difference of  $\beta$  with a chi2 test  $\text{Chi2}(1)=21.2$ ;  $p<0.00005$ ).



## Discussion

Using a combination of computational modeling with tDCS to perturbate mPFC functioning, our study examined the specific role of the mPFC in learning and making transitive inferences about social and non-social hierarchy relationships. Anodal stimulation over the mPFC selectively modulated social but not non-social hierarchy learning, which provides causal evidence to implicate the mPFC in the establishment of social hierarchy knowledge. Furthermore, we found that the Sequential Monte-Carlo model (SMC) captures this learning process better than Reinforcement learning models (RL-ELO), especially in the social condition, providing computational evidence that the mPFC is necessary to encode the forgetfulness with respect to the knowledge from which inferences are to be made.

The two-phase hierarchy learning task adopted in our study allowed us to effectively separate the updating and confirmation of hierarchical knowledge during the trial and error learning of the hierarchy in each Training block from the making of transitive hypotheses, on the basis of the acquired information, during the Test phase of each block. Anodal tDCS specifically reduced the accuracy of the building and updating of social hierarchy knowledge during the training phase (Fig 5A). Mechanically, because learning of the social ranks were not established during early blocks of the training phase, this also reduced the success of transitive inference during test phases in the anodal group. However, during later blocks of the test phase, while learning of the social hierarchy improved, the detrimental effect of anodal tDCS gradually decreased and disappeared. This shows that anodal tDCS does not disturb making of inferences for social hierarchy when learning of social ranks have improved over blocks. The apparent decreased in accuracy observed in early blocks of the test phase in the anodal group in social vs non social hierarchies is likely only the consequence of the impaired learning of social hierarchy during early training (Fig 3B). Thus, mPFC anodal stimulation disturbs learning of social hierarchy, but not the making of transitive inferences.

Our finding that anodal mPFC stimulation specifically disrupts the learning of social hierarchies, but leaves intact the learning of non-social hierarchies, indicates that the ability to learn these two types of hierarchies may rely on distinct cognitive processes. The fact that social hierarchy depends causally on the mPFC resonates with recent studies suggesting that task representations may differ across domains, such as the spatial and conceptual domains (Wu et al., 2020), or abstract vs naturalistic domains (Farashahi et al. 2020; Radulescu et al., 2021). Thus, the nature of the items themselves (here faces vs galaxies) may have influenced how they are learned because faces may be more easily learned. Confirming this hypothesis, differences in both learning accuracy and confidence were observed when directly comparing the sham group in the social and non-social conditions.

In addition to this intrinsic difference according to the nature of the stimuli presented, a key question is to understand the computations specifically impaired by anodal stimulation to learn social vs non-social hierarchies. Our computational modeling approach offers a mechanistic explanation of the observed mPFC anodal stimulation effect on learning social vs non social hierarchies. The mPFC is known to be engaged in social cognition, in learning social hierarchies and to encode the uncertainty in action-outcome learning (Alexander and Brown 2011, Behrens et al. 2007). Our results show that mPFC anodal stimulation plays on the forgetfulness specific to the training phase of the social condition. This results explains the reduced learning performance accuracy during early blocks of training. Performance accuracy progressively increases when the hierarchy knowledge is learned over blocks from the training phase. A key finding from this study is that the SMC mechanism is better able to capture behavior than the RL-ELO mechanism. One critical difference between the SMC and RL-ELO schemes precisely concerns uncertainty. The SMC model inherently models the uncertainty in the estimation of power while the RL-ELO maintain only a single scalar estimate of power for each individual (e.g., Niv et al., 2006; Schultz et al., 1997; although see Gershman, 2015). These models also differ in the nature of the

mechanism by which they update their estimates of the power of individuals within the hierarchy (see supplementary materials).

In addition, confidence increases as the amount of learned information increases (Peterson, Pitz, and Cognition 1988). The quantity and quality of information are represented in the SMC model by the variance of the probability distribution of the inferred variables (i.e. power). The SMC method captures the way that a participant can use information about the hierarchy that is acquired over the course of the experiment to make inferences about these powers. The SMC model offers a mechanistic explanation to the lower confidence level induced by anodal stimulation. Such tDCS effects generates higher evolution variance in the Gaussian random walk, which increases imperfect memory (i.e. forgetting).

The impact of the mPFC anodal stimulation on learning social hierarchy behavior may not be mediated by local activity alone, but by directed communication with other brain areas. For instance, the hippocampus has been reported to encode abstract general knowledge of relationships whatever the nature of the stimuli (spatial, abstract, social or non-social, ...) (Behrens et al., 2018 ; Park et al. 2019). Consistent with this proposal, the mPFC was found to selectively mediates the updating of knowledge about social hierarchy while domain-general coding of rank was observed in the hippocampus, even when the task did not require it (Kumaran 2016). Although another fMRI report did not find mPFC in social vs non-social (Kumaran et al., 2012), this was only a correlational fMRI study, which cannot account for the causal role of the mPFC. Functional coupling between the mPFC and the hippocampus has been shown to support social learning, which required individuals to update values and establish knowledge from prior information, and to incorporate new content (Schlichting, Mumford, and Preston 2015, Schlichting and Preston 2016).

Another important finding from our study is that mPFC stimulation left unimpaired transitive inferences processes, indicating that the mPFC is not causally necessary to make transitive inferences. Again, the hippocampus may be responsible for making

transitive inferences and establishing relationships between previously learned items, as reported by (Garvert et al., 2017; Whittington et al., 2018).

Our results show that anodal tDCS perturbs performance in an asymmetrical manner and preferentially impinges social comparison processes when individuals compare themselves to others above them in the hierarchy, while having no significant effect on the comparison with those ranked below them. The social comparison theory posits that people are driven to compare themselves to others for accurate self-evaluations (Festinger, 1954). Specifically, people compare themselves to others in two opposite directions—downward and upward comparisons—that differ in motivations, comparison targets and consequences (Latane, 1966; Wills, 1981). Upward comparison refers to comparing to those who are thought to rank higher. Upward comparison is most likely done to fulfill the motivation of challenging others and self-improvement. This type of social comparison invokes threat to the self (Brickman and Bulman, 1977) and provokes negative emotions such as envy (Chester et al., 2013; Jankowski and Takahashi, 2014; Tesser et al., 1988). In contrast, downward comparison is most likely done to fulfill the motivation of self-enhancement. Our findings that the mPFC is selectively engaged or focussed on social hierarchy individuals ranks higher to the oneself demonstrates that this region is causally necessary for upward comparison, demonstrating a causal relationship to previously correlational fMRI findings that mPFC distinguished between higher and lower ranks with respect to oneself (Kumaran, 2016), and engagement in different types of social valuation processes (Kim 2020, Lebreton et al. 2009).

Moreover, anodal mPFC disruption of learning of social hierarchies remains robust even when trials involving the participants themselves were excluded (SI Appendix), providing causal evidence that the mPFC not only updates the social hierarchy knowledge specifically related to oneself, but also generally supports the learning of social hierarchy information concerning others. This contrasts with a previous fMRI report showing that the mPFC selectively mediates the updating of knowledge about

one's own hierarchy, as opposed to that of another individual (Kumaran 2016).

Our tDCS approach established a causal relationship between mPFC and social hierarchy learning. Our modeling approach demonstrates that the SMC mechanism is better able to capture the behavioral effect of tDCS stimulation than the RL-ELO mechanism when learning social ranks by observation. This SMC model elucidates the computational principles underlying such impaired learning of social hierarchy learning (i.e. anodal effect on forgetfulness of the model). The maladaptive assessment of social dominance hierarchies is an important source of distress in social dysfunction disorders such as externalizing disorders and social anxiety (Sapolsky 2005, Johnson, Leedom, and Muhtadie 2012). Our current findings not only extend the understanding of the functions of mPFC and its involvement in the social learning process, but also suggest a possible novel avenue of treatment based on tDCS or related brain stimulation techniques to treat symptoms characterized by impaired social cognition (Sellaro, Nitsche, and Colzato 2016).

## **Materials and Methods**

### **Participants**

A total of 136 participants (67 males, 69 females) were recruited via online fliers and participated in the study after they gave informed written consent. All participants were right-handed, with no history of psychiatric or neurologic disorders, and were randomly assigned to receive anode, cathode, or sham stimulation over the medial prefrontal cortex (mPFC) while performing the hierarchy learning tasks. We set a threshold of 80% accuracy after the learning phase. Six participants did not reach this a priori threshold and were excluded from the analysis because they did not learn correctly both the social and non-social conditions in the training trials (accuracy rate of each block was lower than 2/3). In addition, one participant was excluded because he responded randomly, and another because the program was restarted twice. Thus, the data from 128 participants (males=63, females=65, mean age=19.90±0.145) were analyzed (anode=42, cathode=42, sham=44). The study was approved by the ethics committee of South China Normal University and all participants received 45 CNY after the task.

### **Stimuli**

Images of faces in social condition were selected from the CAS-PEAL Large-Scale Chinese Face Database (Gao et al., 2008). Silhouettes (2 faces, 1 female), were used to represent “You” referring to the participant in the experiment. Frontal images (12 neutral faces, 6 females, 6 males) identified fictive hierarchy members for the subsequent experiments. Images of galaxies were selected from a public astronomy website (<http://hubblesite.org/gallery/album/nbula>). All pictures were processed by Adobe Photoshop software to ensure grayscale and resolution were consistent. For the facial pictures, the hair and neck were preserved and the facial positions were as similar as possible. For female participants, the hierarchy was composed of pictures of females and for male participants, it was composed of pictures of males. This choice was adopted because men are known to be more competitive when facing men and

women are also more competitive when facing women. This also ensures attractive to opposite sex. The experiment program was written in E-prime 2.0 and presented on a 14-inch laptop.

### **Experiment Procedure**

With a double-blind and sham-control design, our study included three phases: The cover story and Pre-observation Task, the tDCS stimulation phase (the Hierarchical Learning Task), and Questionnaires (see Fig.1A).

#### **Cover story and Pre-observation phase**

Participants were asked to imagine that recently they had joined a technology company that detected precious minerals in different galaxies. Then they were instructed to observe the photos of staff members and galaxies related to the company. The observation task was to reduce the differential effects of extraneous stimuli during subsequent tasks. It consisted of three blocks, that is, each picture was randomly repeated three times. After that, as a new member of the company, participants were instructed to learn the power relationship between the staff members to help them adjust to work. In the meantime, they also needed to learn the mineral contents of the different galaxies to familiarize themselves with the company business.

#### **tDCS stimulation phase (Hierarchy Learning Task)**

During tDCS stimulation, participants were required to perform a Hierarchy Learning task, including Training and Test trials in both Social and Non-Social conditions. The condition presented first was consistent with the observation task and balanced pseudo-randomly among the participants. The sequences of paired pictures were randomized, as was the left or right location in which pictures were presented.

In each Training trial, participants were required to view a pair of adjacent hierarchical pictures (e.g., P4 versus P5, G4 versus G5; P=person and G=galaxy; P4

means "YOU"; see Fig.1A) and identify which picture they thought had a higher hierarchy (social) or more minerals (non-social), with the correct feedback for each trial. Each Training trial block was followed by a Test trial block in which no feedback was provided. For the Test trials, two non-adjacent hierarchical pictures were presented (e.g., P1 vs P5, G1 vs G5; see Fig.1A). Participants were required to make transitive inference judgments and rate the confidence of their decisions from 1 (guess) to 3 (very sure). For the Social condition, there were 12 blocks including 12 Training trials (6 adjacent paired items: P1 vs P2, P2 vs P3, P3 vs P4, P4 vs P5, P5 vs P6, P6 vs P7, each repeated twice) and then 8 Test trials (8 non-adjacent paired items: P1 vs P7, P2 vs P6, P3 vs P5, P1 vs P4, P2 vs P4, P2 vs P5, P3 vs P6, P4 vs P6). In both Training and Test trials, at the end of each block, they will have average accuracy feedback on their decisions. The Non-Social condition was identical except that the pictures were of galaxies rather than human faces.

### **Questionnaires**

After the Hierarchy learning task, participants were required to fill in the Social Dominance Orientation (SDO) scale, and then answer the post-questions about the task and tDCS stimulation rating: 1) the discomfort of electrode stimulation (range 1-5, 1 for none, 5 for very); 2) how much they believed that they were one of the members of the company (range 1-10, 1 for none, 10 for complete belief).

### **Brain stimulation and current modeling**

We applied NeuroConn transcranial direct current stimulation devices (NeuroConn, Ilmenau, Germany) for tDCS stimulation. According to previous studies (Sellaro et al. 2015a, Casula et al. 2017, Liao et al. 2018, Guo et al. 2019, Wang et al. 2019, Sellaro et al. 2015b), for anode stimulation, the center of the activated electrode was placed at Fpz (EEG 10-20 system), and the reference electrode was placed at Oz (EEG 10-20 system). We used gel to improve conductivity and reduce skin irritation. The



electrode sizes were both 5 cm x 7 cm (35 cm<sup>2</sup>) (see Fig.2). In all stimulation conditions, the current intensity was 1.5 mA, applied with a 30-second fade in and fade out at the beginning and the end of the stimulation. For both anode and cathode stimulation, the 1.5 mA stimulus lasted no more than 30 minutes (when participants completed the task in less than 30 minutes, the current was terminated earlier). For sham stimulation, the current only lasted 15 seconds. To account for possible delays in the onset of tDCS effects, participants were required to wait 2.5 minutes after the onset of stimulation to start the hierarchy learning task (For the sham condition, the current stimuli had ceased since 2.25 minutes).

To ensure that our electrode montage effectively stimulated the mPFC, current flow simulations were performed using ROAST (Huang et al. 2018, Huang et al. 2019) with the MNI152 template brain. Electrodes were simulated as pads, with a 70x50x3mm pad located over Fpz and Cz of standard 10-10 system locations. Tissue conductivities were set as white matter=0.11 S/m, gray matter=0.21 S/m, CSF=0.53 S/m, bone=0.02 S/m, and skin=0.90 S/m. For the anodal stimulation, 1.5mA was set as inward flowing current from the Fpz, and -1.5mA outward flowing current from the Cz. For the cathodal stimulation this was reversed.

### **Statistical analyses and Computational modeling.**

All the statistical analyses of behavioral data were conducted in STATA 14. The computational modeling was conducted in Matlab (r2015B—the MathWorks, Inc.) with the VBA toolbox (Variational Bayes Analysis). A detailed description of the computational modeling results is provided in the SI Appendix.

### **Acknowledgments**

This research has benefited from the financial support of IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d’Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université

de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche and NSF in the CRCNS program to JCD (ANR n°16-NEUC-0003-01), National Science Foundation of China to QC (31470995). We thank Siying Li, and Yaner Su for helpful assistance with data collection.

## References

- ALEXANDER, W. H., AND J. W. BROWN. 2011. "MEDIAL PREFRONTAL CORTEX AS AN ACTION-OUTCOME PREDICTOR." *NAT NEUROSCI* 14 (10):1338-44. DOI: 10.1038/NN.2921.
- APPS, M. A. J., AND J. SALLET. 2017. "SOCIAL LEARNING IN THE MEDIAL PREFRONTAL CORTEX." *TRENDS IN COGNITIVE SCIENCES* 21 (3):151-152. DOI: 10.1016/J.TICS.2017.01.008.
- BEHRENS, T. E., M. W. WOOLRICH, M. E. WALTON, AND M. F. RUSHWORTH. 2007. "LEARNING THE VALUE OF INFORMATION IN AN UNCERTAIN WORLD." *NAT NEUROSCI* 10 (9):1214-21. DOI: 10.1038/NN1954.
- BOYCE, W. T. 2004. "SOCIAL STRATIFICATION, HEALTH, AND VIOLENCE IN THE VERY YOUNG." *YOUTH VIOLENCE: SCIENTIFIC APPROACHES TO PREVENTION* 1036:47-68. DOI: 10.1196/ANNALS.1330.003.
- BRUNONI, A. R., J. AMADERA, B. BERBEL, M. S. VOLZ, B. G. RIZZERIO, AND F. FREGNI. 2011. "A SYSTEMATIC REVIEW ON REPORTING AND ASSESSMENT OF ADVERSE EFFECTS ASSOCIATED WITH TRANSCRANIAL DIRECT CURRENT STIMULATION." *INTERNATIONAL JOURNAL OF NEUROPSYCHOPHARMACOLOGY* 14 (8):1133-1145. DOI: 10.1017/S1461145710001690.
- BUSTON, P. 2003. "SOCIAL HIERARCHIES: SIZE AND GROWTH MODIFICATION IN CLOWNFISH." *NATURE* 424 (6945):145-146. DOI: 10.1038/424145A.
- CASULA, ELIAS P., GIULIA TESTA, PATRIZIA S. BISIACCHI, SARA MONTAGNESE, LORENZA CAREGARO, PIERO AMODIO, AND SAMI SCHIFF. 2017. "TRANSCRANIAL DIRECT CURRENT STIMULATION (TDCS) OF THE ANTERIOR PREFRONTAL CORTEX (APFC) MODULATES REINFORCEMENT LEARNING AND DECISION-MAKING UNDER UNCERTAINTY: A DOUBLE-BLIND CROSSOVER STUDY." *JOURNAL OF COGNITIVE ENHANCEMENT* 1 (3):318-326. DOI: 10.1007/s41465-017-0030-7.
- CHENEY, DOROTHY L, AND ROBERT M SEYFARTH. 2018. *HOW MONKEYS SEE THE WORLD: INSIDE THE MIND OF ANOTHER SPECIES*: UNIVERSITY OF CHICAGO PRESS.
- CHIAO, J. Y., T. HARADA, E. R. OBY, Z. LI, T. PARRISH, AND D. J. BRIDGE. 2009. "NEURAL

REPRESENTATIONS OF SOCIAL STATUS HIERARCHY IN HUMAN INFERIOR PARIETAL CORTEX." *NEUROPSYCHOLOGIA* 47 (2):354-363. DOI: 10.1016/J.NEUROPSYCHOLOGIA.2008.09.023.

CUMMINS, D. D. 2000. "HOW THE SOCIAL ENVIRONMENT SHAPED THE EVOLUTION OF MIND." *SYNTHESE* 122 (1-2):3-28. DOI: DOI 10.1023/A:1005263825428.

DOUCET, ARNAUD, SIMON GODSILL, CHRISTOPHE %J STATISTICS ANDRIEU, AND COMPUTING. 2000. "ON SEQUENTIAL MONTE CARLO SAMPLING METHODS FOR BAYESIAN FILTERING." 10 (3):197-208.

FENG, C. L., Z. H. LI, X. FENG, L. L. WANG, T. X. TIAN, AND Y. J. LUO. 2016. "SOCIAL HIERARCHY MODULATES NEURAL RESPONSES OF EMPATHY FOR PAIN." *SOCIAL COGNITIVE AND AFFECTIVE NEUROSCIENCE* 11 (3):485-495. DOI: 10.1093/SCAN/NSV135.

GARVERT, MONA M, RAYMOND J DOLAN, AND TIMOTHY E J BEHRENS. "A MAP OF ABSTRACT RELATIONAL KNOWLEDGE IN THE HUMAN HIPPOCAMPAL-ENTORHINAL CORTEX." EDITED BY LILA DAVACHI. *ELIFE* 6 (2017): E17086. [HTTPS://DOI.ORG/10.7554/ELIFE.17086](https://doi.org/10.7554/elife.17086).

GROSENICK, L., T. S. CLEMENT, AND R. D. FERNALD. 2007. "FISH CAN INFER SOCIAL RANK BY OBSERVATION ALONE." *NATURE* 445 (7126):429-432. DOI: 10.1038/NATURE05511.

GUO, W., J. SHI, X. LU, H. YE, AND J. LUO. 2019. "MODULATING THE ACTIVITY OF MPFC WITH TDCS ALTERS ENDOWMENT EFFECT." *FRONT BEHAV NEUROSCI* 13:211. DOI: 10.3389/FNBEH.2019.00211.

HUANG, Y., A. DATTA, M. BIKSON, AND L. C. PARRA. 2019. "REALISTIC VOLUMETRIC-APPROACH TO SIMULATE TRANSCRANIAL ELECTRIC STIMULATION-ROAST-A FULLY AUTOMATED OPEN-SOURCE PIPELINE." *J NEURAL ENG* 16 (5):056006. DOI: 10.1088/1741-2552/AB208D.

HUANG, YU, ABHISHEK DATTA, MAROM BIKSON, AND LUCAS C PARRA. 2018. "ROAST: AN OPEN-SOURCE, FULLY-AUTOMATED, REALISTIC VOLUMETRIC-APPROACH-BASED SIMULATOR FOR TES." 2018 40TH ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC).

JOHNSON, S. L., L. J. LEEDOM, AND L. MUHTADIE. 2012. "THE DOMINANCE BEHAVIORAL SYSTEM AND PSYCHOPATHOLOGY: EVIDENCE FROM SELF-REPORT, OBSERVATIONAL, AND BIOLOGICAL STUDIES." *PSYCHOLOGICAL BULLETIN* 138 (4):692-743. DOI: 10.1037/A0027503.

JOINER, JESSICA, MATTHEW PIVA, COURTNEY TURRIN, AND STEVE WC %J NPJ SCIENCE OF LEARNING CHANG. 2017. "SOCIAL LEARNING THROUGH PREDICTION ERROR IN THE BRAIN." 2 (1):8.

KELLEY, W. M., C. N. MACRAE, C. L. WYLAND, S. CAGLAR, S. INATI, AND T. F. HEATHERTON. 2002. "FINDING THE SELF? AN EVENT-RELATED FMRI STUDY." *JOURNAL OF COGNITIVE NEUROSCIENCE* 14 (5):785-794. DOI: DOI 10.1162/08989290260138672.

KIM, H. 2020. "STABILITY OR PLASTICITY? - A HIERARCHICAL ALLOSTATIC REGULATION MODEL OF MEDIAL PREFRONTAL CORTEX FUNCTION FOR SOCIAL VALUATION." *FRONT NEUROSCI* 14:281. DOI: 10.3389/FNINS.2020.00281.

KOSKI, J. E., H. L. XIE, AND I. R. OLSON. 2015. "UNDERSTANDING SOCIAL HIERARCHIES: THE NEURAL AND PSYCHOLOGICAL FOUNDATIONS OF STATUS PERCEPTION." *SOCIAL NEUROSCIENCE* 10 (5):527-550. DOI: 10.1080/17470919.2015.1013223.

KUMARAN, D., A. BANINO, C. BLUNDELL, D. HASSABIS, AND P. DAYAN. 2016. "COMPUTATIONS UNDERLYING SOCIAL HIERARCHY LEARNING: DISTINCT NEURAL MECHANISMS FOR UPDATING AND REPRESENTING SELF-RELEVANT INFORMATION." *NEURON* 92 (5):1135-1147. DOI: 10.1016/J.NEURON.2016.10.052.

KUMARAN, D., H. L. MELO, AND E. DUZEL. 2012. "THE EMERGENCE AND REPRESENTATION OF KNOWLEDGE ABOUT SOCIAL AND NONSOCIAL HIERARCHIES." *NEURON* 76 (3):653-666. DOI: 10.1016/J.NEURON.2012.09.035.

LEBRETON, M., S. JORGE, V. MICHEL, B. THIRION, AND M. PESSIGLIONE. 2009. "AN AUTOMATIC VALUATION SYSTEM IN THE HUMAN BRAIN: EVIDENCE FROM FUNCTIONAL NEUROIMAGING." *NEURON* 64 (3):431-9. DOI: 10.1016/J.NEURON.2009.09.040.

LIAO, C., S. WU, Y. J. LUO, Q. GUAN, AND F. CUI. 2018. "TRANSCRANIAL DIRECT CURRENT STIMULATION OF THE MEDIAL PREFRONTAL CORTEX MODULATES THE PROPENSITY

TO HELP IN COSTLY HELPING BEHAVIOR." NEUROSCIENCE LETTERS 674:54-59. DOI: 10.1016/J.NEULET.2018.03.027.

LIGNEUL, R., I. OBESO, C. C. RUFF, AND J. C. DREHER. 2016. "DYNAMICAL REPRESENTATION OF DOMINANCE RELATIONSHIPS IN THE HUMAN ROSTROMEDIAL PREFRONTAL CORTEX." CURRENT BIOLOGY 26 (23):3107-3115. DOI: 10.1016/J.CUB.2016.09.015.

MARSH, A. A., K. S. BLAIR, M. M. JONES, N. SOLIMAN, AND R. J. R. BLAIR. 2009. "DOMINANCE AND SUBMISSION: THE VENTROLATERAL PREFRONTAL CORTEX AND RESPONSES TO STATUS CUES." JOURNAL OF COGNITIVE NEUROSCIENCE 21 (4):713-724. DOI: DOI 10.1162/JOCN.2009.21052.

MUSCATELL, K. A., S. A. MORELLI, E. B. FALK, B. M. WAY, J. H. PFEIFER, A. D. GALINSKY, M. D. LIEBERMAN, M. DAPRETTO, AND N. I. EISENBERGER. 2012. "SOCIAL STATUS MODULATES NEURAL ACTIVITY IN THE MENTALIZING NETWORK." NEUROIMAGE 60 (3):1771-7. DOI: 10.1016/J.NEUROIMAGE.2012.01.080.

PETERSON, DANE K, GORDON F %J JOURNAL OF EXPERIMENTAL PSYCHOLOGY: LEARNING PITZ, MEMORY,, AND COGNITION. 1988. "CONFIDENCE, UNCERTAINTY, AND THE USE OF INFORMATION." 14 (1):85.

QU, C., R. LIGNEUL, J. B. VAN DER HENST, AND J. C. DREHER. 2017. "AN INTEGRATIVE INTERDISCIPLINARY PERSPECTIVE ON SOCIAL DOMINANCE HIERARCHIES." TRENDS IN COGNITIVE SCIENCES 21 (11):893-908. DOI: 10.1016/J.TICS.2017.08.004.

RAMESON, L. T., A. B. SATPUTE, AND M. D. LIEBERMAN. 2010. "THE NEURAL CORRELATES OF IMPLICIT AND EXPLICIT SELF-RELEVANT PROCESSING." NEUROIMAGE 50 (2):701-708. DOI: 10.1016/J.NEUROIMAGE.2009.12.098.

RIGOUX, L., K. E. STEPHAN, K. J. FRISTON, AND J. DAUNIZEAU. 2014. "BAYESIAN MODEL SELECTION FOR GROUP STUDIES - REVISITED." NEUROIMAGE 84:971-85. DOI: 10.1016/J.NEUROIMAGE.2013.08.065.

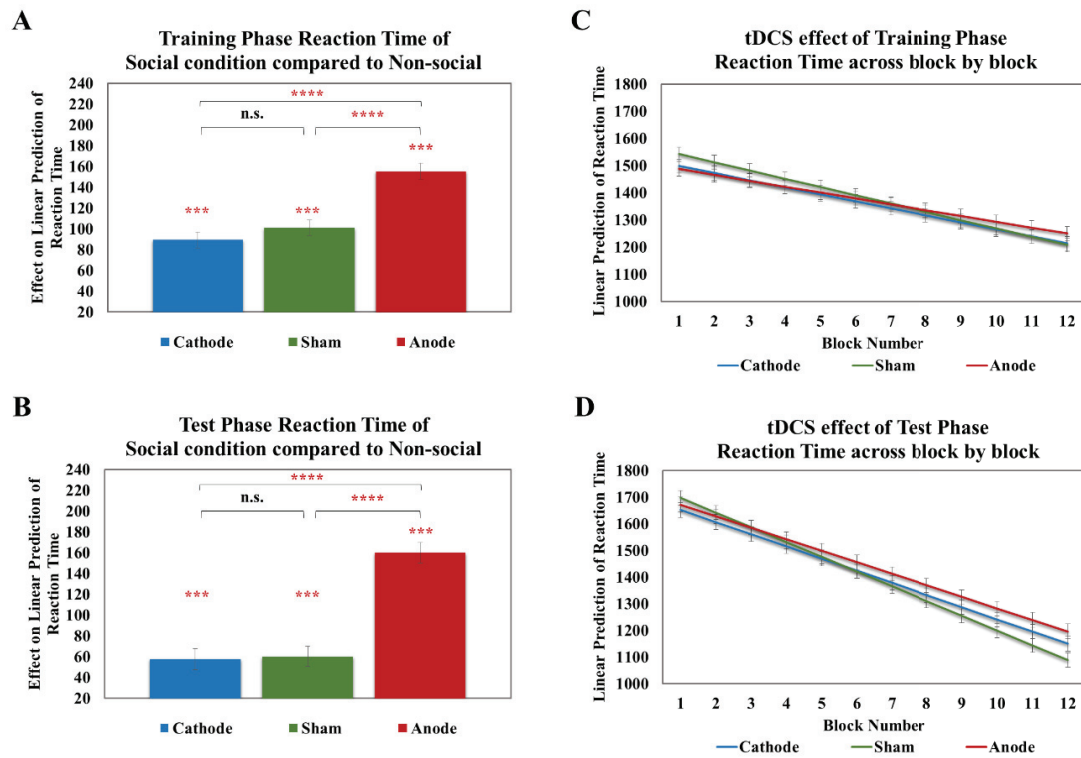
SANTAMARIA-GARCIA, H., M. PANNUNZI, A. AYNETO, G. DECO, AND N. SEBASTIAN-GALLES. 2014. "'IF YOU ARE GOOD, I GET BETTER' : THE ROLE OF SOCIAL HIERARCHY IN

## **Supplementary Materials for**

### **Modulating Social Hierarchy Learning in Human With Non-invasive brain stimulation**

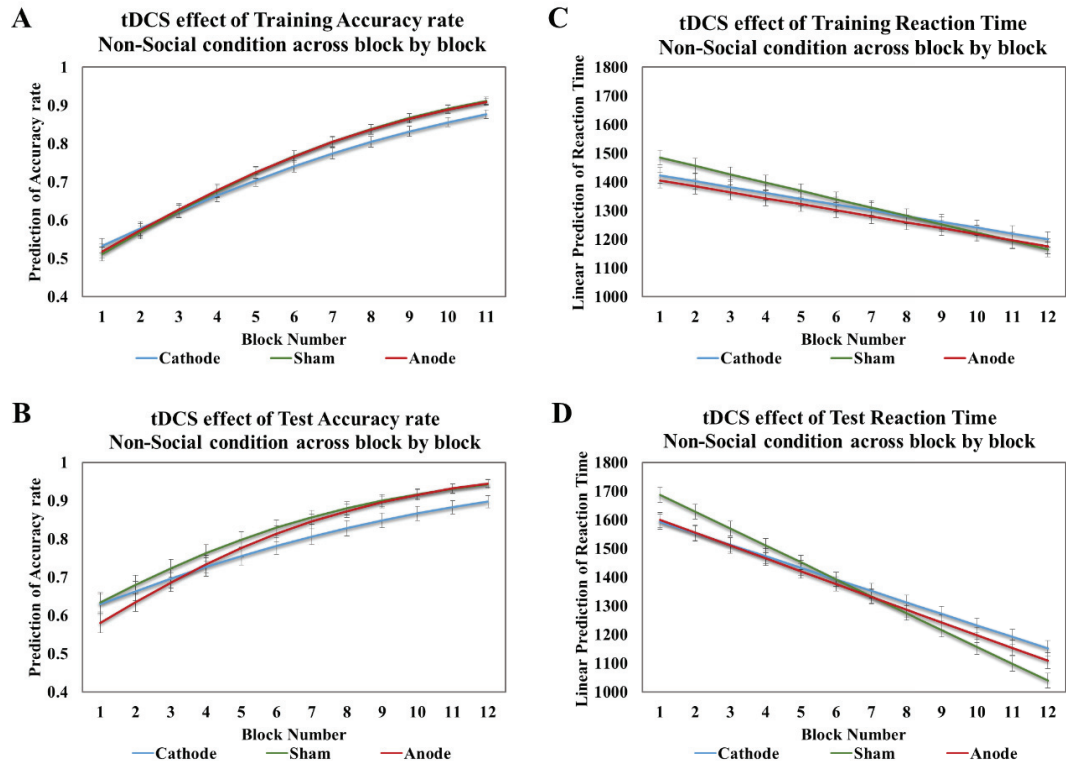
**This PDF file includes:**

Supplementary text  
Figures S1 to S3  
Tables S1  
SI References

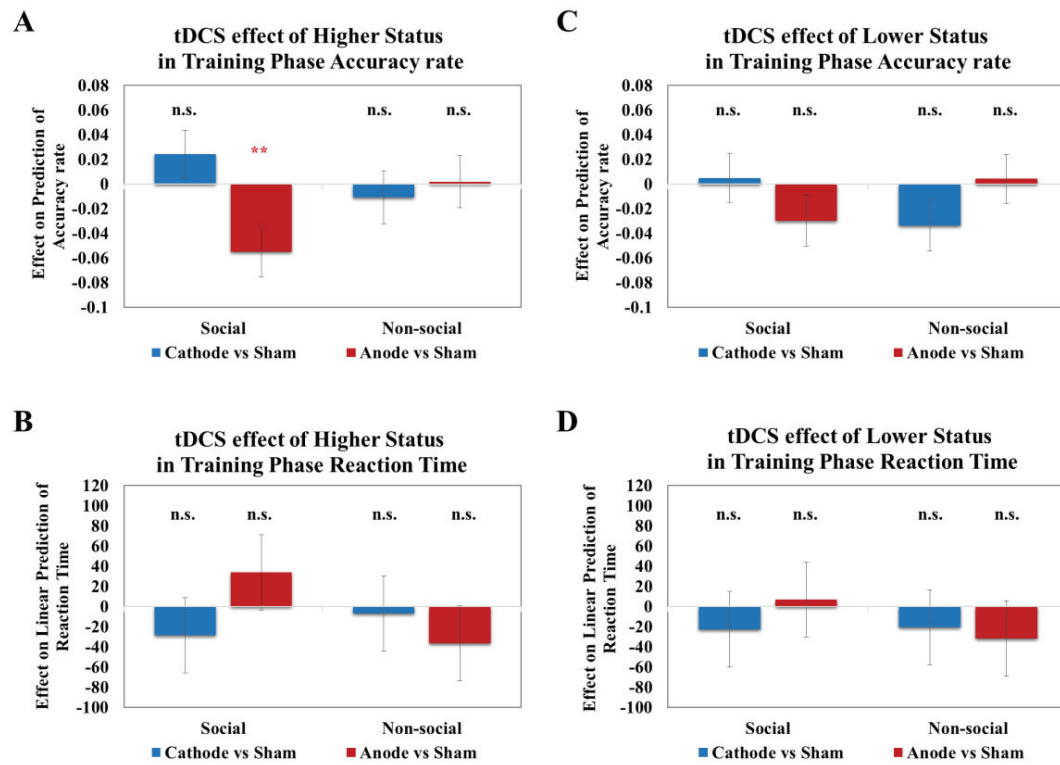


**Fig. S1. Brain stimulation effect of hierarchy learning in Reaction Time.** A) and B) *tDCS modulation of social condition compared to non-social in Training and Test phase reaction time.* The blue bar indicated cathode modulation of social condition compared to non-social. The green bar is sham effect and the red bar is Anode. C) and D) *tDCS effect across blocks in the Training and test phase reaction time.* Sham: green, Cathode: blue, Anode: red. (\*indicates  $p < 0.05$ , \*\*indicates  $p < 0.005$ , \*\*\*indicates  $p < 0.001$ ; \*\*\*\*indicates  $p < 0.0001$ ; Error bars show SEM)





**Fig. 5 tDCS effect of Non-social hierarchy learning across block by block. A) and B) tDCS modulation of non-social condition in Training and Test phase accuracy rate. The blue line indicated cathode effect. The green is sham effect and the red line is anode. C) and D) tDCS modulation of non-social condition in the Training and test phase reaction time. Same as accuracy rate. (Error bars show SEM)**



**Fig. 4 Brain stimulation effect of social hierarchy learning in higher and lower status.** A) and B) tDCS effect of Training phase accuracy rate and reaction time in higher status. The blue bar indicated cathode modulation compared to sham stimulation. The red bar is the anode effect compared to the sham. C) and D) tDCS effect of Training phase accuracy rate and reaction time in lower status. Same as the higher status. (\*indicates  $p < 0.05$ , \*\*indicates  $p < 0.005$ , \*\*\*indicates  $p < 0.001$ ; \*\*\*\*indicates  $p < 0.0001$ ; Error bars show SEM)

**Table S1. Demographic information, Questionnaires and Task performance compared among groups.**

|                                     | <b>Anode</b>  | <b>Sham</b>   | <b>Cathode</b>  | <b><i>p</i> value</b> |
|-------------------------------------|---|---|---|-----------------------|
| <b>Participants After Exclusion</b> | 42<br>(Males=21, Females=21)                                      | 44<br>(Males=22, Females=23)                                      | 42<br>(Males=21, Females=21)                                      | Non-sig.              |
| <b>Age</b>                          | 19.610±0.231  | 19.909±0.239  | 20.171±0.281  | Non-sig.              |
| <b>SDO</b>                          | 19.610±0.231  | 19.909±0.239  | 20.171±0.281  | Non-sig.              |
| <b>Truth Degree</b>                 | 5.452±0.350   | 5.545±0.337   | 5.976±0.352   | Non-sig.              |
| <b>Uncomfortable Rating score</b>   | 1.536±0.146   | 1.341±0.101   | 1.762±0.137   | Non-sig.              |
| <b>Choice Bias</b>                  | ❖ Training Trails:<br>▪ Left: 0.501±0.005<br>▪ Right: 0.499±0.005 | ❖ Training Trails:<br>▪ Left: 0.504±0.005<br>▪ Right: 0.496±0.005 | ❖ Training Trails:<br>▪ Left: 0.503±0.005<br>▪ Right: 0.497±0.005 | Non-sig.              |
|                                     | ❖ Test Trails:<br>▪ Left: 0.502±0.003<br>▪ Right: 0.498±0.003     | ❖ Test Trails:<br>▪ Left: 0.508±0.003<br>▪ Right: 0.492±0.003     | ❖ Test Trails:<br>▪ Left: 0.499±0.003<br>▪ Right: 0.501±0.003     | Non-sig.              |

## Computational Models

### Sequential Monte-Carlo (SMC) model.

The SMC model (Doucet et al., 2000; Kumaran et al., 2016) is a state-space inference model aimed at inferring the underlying state of an evolutionary dynamical system. In our particular case, the power of a set of individual faces and galaxies. During training trials, these hidden states diffuse according to a random walk pattern (with non-zero variance for models that allow for forgetting). The decision observation process is a sigmoid function of the power difference between the two presented individuals. The SMC method is a form of online Bayesian filter, relaxing the assumption for linear, Gaussian, Kalman filters that the probability density function (pdf) of the inferred variables (i.e. the power) is a normal distribution. This increases the flexibility of modeling multi-modal distributions (e.g. (Doucet et al., 2000)). In the SMC model, each particle (here, N=10 000) represents a set of values for the hidden state variable. A particle can thus be interpreted as representing a hypothesis about the rank of elements in the hierarchy. The population of particles is thus a multimodal pdf of the ranking (i.e. 7-dimensional given the 7 elements of the hierarchy). The particles are initialized with an equal weight, which is updated at each training trial according to the probability of the trial's outcome given the hypothesis about the ranking they represent. A particle resampling step ensures that the particle density is highest in the

regions of the (7-dimensional) space that are likely given the history of the observed data (i.e. have highest weights), by replacing conditionally unlikely particles (i.e. with low weights) with new, more appropriate particles.

**Prior model:** The state variable (called  $\vec{x}_0$  and denoting power) is initialized a normal distribution with a fixed initial variance  $\sigma_0^2 = 10$  (Eq. 1). The state process is described as a Gaussian random walk with the evolution parameter  $\sigma^2$  (Eq. 2 - defined as a free parameter optimized across subjects and a zero mean) which instantiates a form of imperfect memory (the reason why we called it the forgetting rate), letting to account for participant need around 110 trials to reach mastery of the task. The observation part of the model (Eq. 3) is a sigmoid function of the difference between each hypothetical power of the distribution of the two items presented in the current trial  $t$ . The sigmoid function is parametrized by a free parameter  $\beta$  letting to account for the deterministic/stochastic way participant exploit their knowledge. Here,  $a_t$  represent the item with current highest expected value and  $b_t$  with the lowest value.  $y_t = 1$  denotes the situation when the highest item valued item is the correct answer.

$$\vec{x}_0 \sim \mathcal{N}(\vec{0}, \sigma_0^2 \mathbf{I}) \quad (\text{Eq. 1})$$

$$\vec{x}_t | \vec{x}_{t-1} \sim \mathcal{N}(\vec{x}_{t-1}, \sigma^2 \mathbf{I}) \quad (\text{Eq. 2})$$

$$g(y_t = 1 | a_t, b_t, x_t) = \frac{1}{1 + e^{-\beta * (x_{a_t t} - x_{b_t t})}} \quad (\text{Eq. 3})$$

**Particles filter:** Let  $i$  index particles, of which there are  $N=10,000$ . Particles are initialized as samples from a normal distribution, with zero mean and variance  $\sigma_0^2$  (Eq. 4) each with equal normalized weight ( $\tilde{w}^{(i)}$ ) (Eq. 5). Because no feedback was provided in test trials, the following process of updating particles only occurred during training trials (Eq. 6). The

unnormalized weight ( $\tilde{w}_{t+1}^{(i)}$ ) of the particles are updated using the observation model and the normalized weight from previous trial (Eq. 7)

$$\vec{x}_0^{(i)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}) \quad (\text{Eq. 4})$$

$$\tilde{w}_0^{(i)} = \frac{1}{N} \quad (\text{Eq. 5})$$

$$\vec{x}_t^{(i)} \sim \mathcal{N}(\vec{x}_{t-1}^{(i)}, \sigma_0^2 \mathbf{I}) \quad (\text{Eq. 6})$$

$$w_{t+1}^{(i)} = g(y_t | a_t, b_t, \vec{x}_t^{(i)}) \tilde{w}_t^{(i)} \quad (\text{Eq. 7})$$

$N_{eff}$  determines the threshold for resampling in (Eq. 8)

$$N_{eff} = \frac{1}{\sum_{i=1}^N \tilde{w}_t^{(i)} \tilde{w}_t^{(i)}} < 0.25N \quad (\text{Eq. 8})$$

Then resample with replacement

$$a_t^{(i)} \sim \text{Categorical} \left( \left\{ \frac{w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}} \right\}_{i=1}^N \right) \quad (\text{Eq. 9})$$

$$\vec{x}_t^{(i)} = \vec{x}_t^{(a_t^{(i)})} \quad (\text{Eq. 10})$$

$$\tilde{w}_t^{(i)} = \frac{1}{N} \quad (\text{Eq. 11})$$

$$\text{Othserwise } \tilde{w}_t^{(i)} = w_t^{(i)} \quad (\text{Eq. 12})$$

The estimate of marginal likelihood is :

$$\hat{p}(y_{1:T}) = \prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N w_t^{(i)} \quad (\text{Eq. 13})$$

### RL models (RL-ELO).

In the RL-ELO model, rather than updating the value of items based on the difference between trial outcome and current value as in a Rescola-Wagner model, the value is updated as a function of the difference in current values between the two items :  $V_{L,t}$  &  $V_{R,t}$  (Indexed by their randomly assigned position Left and Right). This algorithm could be seen as a version of a policy gradient algorithm (Williams, 1992).

Initialize all ranks at  $V_0 = 0$

Free parameters:  $\alpha$  = learning rate;  $\beta$  = temperature

Training trial at time t with items on left and right sides of screen

Probability of choosing left item and right item:

$$p_{L,t} = \frac{1}{1+e^{-\beta(V_{L,t}-V_{R,t})}} \quad (\text{Eq. 1})$$

$$p_{R,t} = 1 - p_{L,t} \quad (\text{Eq. 2})$$

if L item is correct choice:

$$p_{win,t} = p_{L,t} \quad (\text{Eq. 3})$$

$$I_{L,t} = 1, I_{R,t} = -1 \quad (\text{Eq. 4})$$

if R item is correct choice:

$$p_{win,t} = p_{R,t} \quad (\text{Eq. 5})$$

$$I_{L,t} = -1, I_{R,t} = 1 \quad (\text{Eq. 6})$$

Update values of both items:

$$V_{L,t+1} = \alpha * I_{L,t} * (1 - p_{win,t}) + V_{L,t} \quad (\text{Eq. 7})$$

$$V_{R,t+1} = \alpha * I_{R,t} * (1 - p_{win,t}) + V_{R,t} \quad (\text{Eq. 8})$$

Please note that we examine two variant of the model. One incorporate two different alpha to test the hypothesis of a different learning given the outcome of the trial. Thus this model has one alpha to update values after a victory ( $\alpha_{win}$ ) and another to update values after a loss ( $\alpha_{loss}$ ). The other test the hypothesis of a different learning given the position of the pair in the hierarchy (higher or lower than the participant) also incorporate two different alpha, one is ( $\alpha_{superior}$ ) and the other is ( $\alpha_{inferior}$ ).

### Value transfer model

This model incorporates the standard update term from Rescorla Wagner, but also includes an indirect component: the incorrect item in a training trial has its value updated with a proportion (i.e. theta %) of the correct item.

Trial outcomes are +1 for correct choice, and -1 for incorrect choice.

3 free parameters:  $\alpha$  = learning rate;  $\beta$ =temperature;  $\theta$  = transfer factor.

Training trial at time t with items on left and right sides of screen

Probability of choosing left item, and right item:

$$p_{L,t} = \frac{1}{1+e^{-\beta(V_{L,t}-V_{R,t})}} \quad (\text{Eq. 1})$$

$$p_{R,t} = 1 - p_{L,t} \quad (\text{Eq. 2})$$

if L item is correct choice:

$$O_{L,t} = 1, O_{R,t} = -1 \quad (\text{Eq. 3})$$

$$I_{L,t} = 0, I_{R,t} = 1 \quad (\text{Eq. 4})$$

if R item is correct choice:

$$O_{L,t} = -1, O_{R,t} = 1 \quad (\text{Eq. 5})$$

$$I_{L,t} = 1, I_{R,t} = 0 \quad (\text{Eq. 6})$$

Calculate the direct component of update:

$$\delta V_{direct_{L,t}} = (O_{L,t} - V_{L,t}) * \alpha \quad (\text{Eq. 7})$$

$$\delta V_{direct_{R,t}} = (O_{R,t} - V_{R,t}) * \alpha \quad (\text{Eq. 8})$$

Calculate the indirect component of update:

$$\delta V_{indirect_{L,t}} = V_{R,t} * I_{L,t} * \theta \quad (\text{Eq. 9})$$

$$\delta V_{indirect_{R,t}} = V_{L,t} * I_{R,t} * \theta \quad (\text{Eq. 10})$$

Total update:

$$V_{L,t+1} = \delta V_{direct_{L,t}} + \delta V_{indirect_{L,t}} \quad (\text{Eq. 11})$$

$$V_{R,t+1} = \delta V_{direct_{R,t}} + \delta V_{indirect_{R,t}} \quad (\text{Eq. 12})$$

**Rescola Wagner.** As for Value transfer model, where theta parameter is set to zero.

**Computational model fitting and selection.** We quantified the fit of all models to participant's choice behavior during training and test trials. We used a maximum likelihood estimation procedure and optimized a separate set of parameters for each participant (Wimmer et al., 2012). Then the Bayesian Model Selection (BMS) was done using the VBA toolbox (Variational Bayesian Analysis) in a random effect analysis relying on the Bayesian Information Criteria (BIC) measure which penalizes more complex models. We use protected Exceedence Probability measurement (pEP) (Rigoux et al., 2013) to select the model which is used most frequently in our population.

## SI References

- Doucet, Arnaud, Simon Godsill, and Christophe Andrieu. “Methods for Bayesian Filtering.” *Statistics and Computing*, 2000, 197–208.  
<https://doi.org/10.1023/A:1008935410038>.
- Fersen, Lorenzo von, C. D.L. Wynne, Juan D. Delius, and J. E.R. Staddon. “Transitive Inference Formation in Pigeons.” *Journal of Experimental Psychology: Animal Behavior Processes* 17, no. 3 (1991): 334–41. <https://doi.org/10.1037/0097-7403.17.3.334>.
- Kumaran, Dharshan, Andrea Banino, Charles Blundell, Demis Hassabis, and Peter Dayan. “Computations Underlying Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing Self-Relevant Information.” *Neuron* 92, no. 5 (2016): 1135–47. <https://doi.org/10.1016/j.neuron.2016.10.052>.
- Rigoux, L., K. E. Stephan, K. J. Friston, and J. Daunizeau. “Bayesian Model Selection for Group Studies - Revisited.” *NeuroImage* 84 (2013): 971–85.  
<https://doi.org/10.1016/j.neuroimage.2013.08.065>.
- Williams, Ronald J. “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning.” *Machine Learning* 8, no. 3 (1992): 229–56.  
<https://doi.org/10.1007/BF00992696>.
- Wimmer, G. Elliott, Nathaniel D. Daw, and Daphna Shohamy. “Generalization of Value in Reinforcement Learning by Humans.” *European Journal of Neuroscience* 35, no. 7 (2012): 1092–1104. <https://doi.org/10.1111/j.1460-9568.2012.08017.x>.





## Chapter 5

# Regulation of social hierarchy learning by serotonin transporter availability <sup>1</sup>

---

<sup>1</sup>Mon travail a essentiellement consisté à la partie modélisation du projet ainsi qu'à l'interprétation de l'ensemble des résultats.

## **Regulation of social hierarchy learning by serotonin transporter availability**

R. Janet<sup>1</sup>, R. Ligneul<sup>2</sup>, A. Losecaat-Vermeer<sup>3</sup>, R. Philippe<sup>1</sup>, G Bellucci<sup>4</sup>, S Park<sup>5</sup>, P Jacquart<sup>6</sup>, JC Dreher<sup>1</sup>

<sup>1</sup>*CNRS-Institut de Sciences Cognitives Marc Jeannerod, UMR5229, Neuroeconomics, reward, and decision making laboratory and Lyon 1 University, France*

<sup>2</sup>*Champalimaud Neuroscience Program, Champalimaud Center for the Unknown, Lisbon, Portugal*

<sup>3</sup>*Neuropsychopharmacology and Biopsychology Unit, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Austria*

<sup>4</sup>*Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

<sup>5</sup>*Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Neuroscience Research Center, Berlin, Germany*

<sup>6</sup>*CNRS-GATE and Em-Lyon, Ecully, France*

## Short

Within hierarchical social groups learning each members' relative status sub serves adaptive behaviors. Although animal studies suggest that serotonin signaling plays a role in the establishment of social hierarchies, direct evidence in humans is still lacking. Here, combining a computational approach with simultaneous PET-fMRI acquisition in healthy humans, we investigated the link between serotonin transporter (SERT) availability and brain activity when learning one's rank in a group by competitive social interactions. Inter-individual differences in extracellular SERT in the dorsal raphe nucleus (DRN) covaried with the learning rates governing the updating of hierarchical status. Moreover, a negative relationship between SERT availability and the expected value of the social victories was observed in the ventral striatum. Consistent with the reinforcement learning theory, these results suggest that serotonin levels modulate the neural computations of the expected value of long-term social rewards.

## Long

Learning one's status in a group is a fundamental process in building social hierarchies. In animals, a substantial body of data indicates that serotonin activity in the raphe is tightly coupled with social rewards and the establishment of social hierarchies. In humans, although indirect evidence from pharmacological and clinical studies supports such an association, there has been no direct demonstration of a link between serotonergic transporter from the raphe nucleus and neural responses related to learning social hierarchies in healthy humans. Using a computational approach combined with a PET-3T fMRI scanner, we investigated the link between SERT availability and brain activity when learning from competitive social interactions. The results revealed a negative correlation between SERT availability and the learning rate. Moreover, a direct relationship between SERT availability and the expected value of the social victories was observed in the ventral striatum. These results suggest that serotonin levels modulate the neural computation of the expected value of long-term social rewards according to the reinforcement learning theory.

## Introduction

Dominance hierarchy is a strong evolutionary force because dominant individuals have better access to resources, including food and reproductive partners (Ellis, 1995; Sandi & Haller, 2015b; Sapolsky, 2004, 2005). A dominant status is attributed to the individual who wins competitive interactions against conspecifics (Qu et al., 2017; Raleigh et al., 1991; Wang et al., 2014; Zhou et al., 2018). Learning one's own rank during competitive dyadic interactions within a group is crucial to adapt behavior and avoid harmful social defeats. In animals, serotonin (5-HT) is tightly coupled with social rewards and the establishment of social hierarchies (Cohen et al., 2017; Li et al., 2016; Liu et al., 2014; Terranova et al., 2016; Sandi & Haller, 2015a). For example, in groups of velvet monkeys, enhancement or suppression of serotonin signaling can induce dominance or subordination respectively (Raleigh et al., 1991). Higher-ranked monkeys have more gray matter in the dorsal raphe nucleus (DRN) containing serotonergic neurons (Noonan et al., 2014). In both mice and monkeys, diverse types of DRN neurons are engaged by social defeats and post-defeat sensitization of these neurons decreases resilience to social defeat (Challis et al., 2013). In mice, the degree to which specific DRN neurons modulate behavior is predicted by social rank (Matthews et al., 2016).

In humans, vulnerability to anxiety and depression is increased by repeated social defeats (Johnson et al., 2012). Such experience can trigger maladaptive social avoidance and isolation, behavioral inhibition, and may even affect immune regulation in non-human primates (Johnson et al., 2012; Sandi & Haller, 2015a). Increasing serotonin biosynthesis through the administration of its precursor increases the frequency of dominance-related behaviors (Moskowitz et al., 2001). Although indirect evidence from preclinical, pharmacological and clinical studies suggests an association between the serotonergic transporter from the DRN and neural responses related to learning social hierarchies (Aan Het Rot et al., 2006; Moskowitz et al., 2001; Steenbergen et al., 2016), there has been no direct demonstration of such a link in humans.

Here, using reinforcement learning modeling and simultaneous PET-fMRI acquisition in the same subjects, we investigated the link between inter-individual brain activity during hierarchy learning and serotonin transporter (SERT) availability, which provides an indirect measure of serotonergic function. We induced an implicit dominance hierarchy with a competitive game that required subjects to choose between opponents of different strengths. Although the computational role of 5-HT in reinforcement learning (RL) has remained elusive (Crockett et al., 2009; Fonseca et al., 2015; Liu et al., 2014; Miyazaki et al., 2014; Seymour et al., 2012), a number of studies have associated 5-HT signaling with diverse rewards and punishments (J. Y. Cohen et al., 2015; Matias et al., 2017), including social rewards (Dölen et al., 2013; Li et al., 2016). Recent recordings from DRN 5-HT neurons revealed responses to both rewards and punishments, with modulations of tonic activity by context and phasic responses during reinforcer delivery, even when they are predicted (J. D. Cohen et al., 2017; Li et al., 2016; Liu et al., 2014). Moreover, a recent optogenetic study in mice suggested that the reinforcement learning rate may be under modulation of DRN 5-HT neurons (Iigaya et al., 2018). This learning rate determines how many trials over which reward histories are integrated to assess the value of actions that have been taken.

Based on these animal experiments, we reasoned that victories during social competitions may act as social rewards and that 5-HT levels may modulate the expected value of total social rewards over all successive steps. This variable, called the action value or expected value (Q-value), is commonly used in model-free RL to learn the (Q)uality of actions to take in a given state. Q-learning finds an optimal action-selection policy by maximizing the expected value of total (discounted) rewards. We thus tested the hypothesis that inter-individual differences in SERT

levels would covary with brain regions encoding the cumulative history of social expectation while subjects learned the ranks of opponents.

Confirming a previous fMRI study (Ligneul et al., 2016), we first observed that the medial prefrontal cortex (mPFC) and the ventral striatum tracked the expected value of social victories while competing with different opponents. Crucially, the learning rate associated with these competitive choices was increased in individuals with lower binding potential for 5-HT in the DRN. Moreover, a negative relationship between SERT levels and the expected value of social victories was observed in the ventral striatum. These findings provide a characterization of the interactions between DRN 5-HT function and the brain system engaged in social hierarchy learning in healthy humans.

## Results

### Behavioral results

Participants were led to believe they played in a group of four including themselves. Each trial, they were asked to choose an opponent among two others from the group, presented on the screen. Then, after competing with the selected opponent in a perceptual decision-making task, they received feedback from this competition, a victory or a defeat as outcomes stage. They were actually playing against a computer and the three opponents led to different reward probabilities (28% for the superior, 72% for the inferior and 50% for the intermediate opponent). Between two runs of the social hierarchy competition involving two distinct groups of opponents, participants were required to play a non-social hierarchy learning game. This comprised a slot machine selection task, where participants selected which slot machine to play among three distinct slot machines each with a reward probability equivalent to that of an opponent in the social hierarchy task (28%, 72% and 50% respectively). The design of this task was very similar to the social competition task. For each trial, participants had to select one of two slot machines presented on the screen, and then they received feedback of whether they had won or lost. During both tasks, participants were asked to rate their confidence of winning during specific trials (they had to rate four times each opponent).

To investigate the frequency of the choices made by participants during the tasks, a two-way repeated measure ANOVA including outcome reward probability conditions and task modalities as factors of interest was conducted. Results revealed a main effect of outcome Reward Probability ( $F_{(2,58)} = 42.81$ ;  $p < 0.001$ ) and an interaction effect between the Reward Probability and task modalities ( $F_{(2,58)} = 8.89$ ;  $p < 0.001$ ) (**Fig 1A**, right panel). *Post-hoc* analysis conducted on the social learning task revealed that participants selected the inferior opponent ( $M = 0.40$ ,  $SEM = 0.02$ ) more than the intermediate ( $M = 0.30$ , standard error of the mean,  $SEM = 0.01$ ;  $t_{(29)} = 3.74$ ,  $p = 0.004$ ) and the superior opponent ( $M = 0.29$ ,  $SEM = 0.02$ ;  $t_{(29)} = 3.33$ ,  $p = 0.04$ ). *Post hoc* tests conducted on the non-social learning task revealed that participants selected the easiest slot machine to win on ( $M = 0.48$ ,  $SEM = 0.08$ ) more frequently than the intermediate one ( $M = 0.33$ ,  $SEM = 0.06$ ;  $t_{(29)} = 4.68$ ,  $p < 0.000$ ) or the most difficult one ( $M = 0.18$ ,  $SEM = 0.03$ ;  $t_{(29)} = 8.82$ ,  $p < 0.001$ ) (**Fig 1B**, right panel). Also, participants selected the intermediate slot machine more frequently than the most difficult one ( $t_{(29)} = 5.18$ ,  $p < 0.001$ ).

We modeled the behavioral data using a reinforcement Q-learning algorithm (see supp. results). We compared 6 variants of the reinforcement learning scheme for the social competition task and 2 variants for the non-social learning task (see Computational modeling section). Bayesian model selection (BMS) for the social competitive task indicated that among these variants, the model with one learning rate and no updating according to performance was the

most likely model (called 'no accuracy monitoring'). The same BMS procedure conducted for the non-social task showed that the model with one learning rate was also the most likely model. Participants' choices and model predicted choices are shown on **Fig 2A**. We observed no significant difference in the alpha parameter for the social and non-social tasks ( $M = 0.38$ ,  $SEM = 0.41$ , social and  $M = 0.28$   $SEM = 0.42$ , non-social), however the non-parametric Wilcoxon test revealed that the beta parameter in the social learning task ( $M = 2.70$ ,  $SEM = 0.87$ ) was significantly lower than that estimated for the non-social task ( $z_{(29)} = -3.67$ ,  $M = 8.01$ ,  $SEM = 1.39$ ). This may reflect an increased tendency to explore options in the social context.

### **Dorsal Raphe Nucleus is linked to the social learning rate**

The Dorsal Raphe Nucleus (DRN) has the highest concentration of 5-HT neurons in the brain and is the main source of serotonin in the cortex and the basal ganglia and a good candidate for being at the origin of 5-HT regulation for learning social ranks. We therefore investigated the relationship between learning rate ( $\alpha$ ) and the non-displaceable Binding Potential ( $BP_{ND}$ ) of the SERT ligand [ $^{11}C$ ]-DASB, extracted from the DRN using the Automated Anatomical Labelling (AAL3) atlas (Rolls et al., 2020). A negative correlation was observed between the  $BP_{ND}$  in the DRN and learning rate, (alpha parameter), of social ranks ( $r = -0.366$ ,  $p = 0.046$ , two-tailed Spearman correlation) (**Fig 2B**). No such correlation was observed between  $BP_{ND}$ , DRN and learning rate in the non-social learning task ( $r = -0.190$ ,  $p = 0.315$ ). Additionally, a negative correlation was observed between the learning rate and the inverse temperature for both the social ( $r = -0.382$ ,  $p = 0.037$ ) and non-social tasks ( $r = -0.758$ ,  $p < 0.001$ ). No significant difference was observed between the correlation coefficients of the social task and non-social tasks ( $p = 0.25$ ).

We then estimated the brain SERT distribution using the average  $BP_{ND}$  (**Fig 2C**). Focusing on the DRN, we divided our sample using a median split procedure according to the level of SERT availability in this region. We formed two groups of 15 individuals, a low ( $M = 1.09$   $SEM = 0.07$ ) and a high  $BP_{ND}$  DRN group ( $M = 2.52$ ,  $SEM = 0.08$ ). This median split revealed a between group difference ( $p < 0.001$  Wilcoxon signed ranks test) (**Fig 2D, bar graphs**). Next, we computed a competitive index (cf. supplementary data) that reflected competitive choices and compared it between groups as the task progressed. A repeated measure ANOVA including the group (low vs high  $BP_{ND}$  DRN) and trial bins (1-6, in bins) revealed a group\*time interaction ( $F_{(1,5)} = 2.32$ ,  $p = 0.046$ ). *Post-hoc* test showed that the high  $BP_{ND}$  group made less competitive choices in the last bin ( $M = 0.34$ ,  $SEM = 0.06$ ) compared to the low  $BP_{ND}$  group ( $M = 0.52$ ,  $SEM = 0.06$ ;  $t_{(28)} = 2.27$ ,  $p = 0.031$ ) (**Fig 2D**).

### **fMRI analysis revealed the positive encoding of expected value of social victories during outcomes**

We first searched for brain regions encoding the expected value of social victories  $Q(t)$ , reflecting the dominance status of the opponent. To do so, we ran a general linear model (GLM1) using  $Q(t)$  as parametric regressor. A network of regions including the bilateral ventral striatum, the ventromedial prefrontal cortex (vmPFC), the posterior cingulate cortex and the posterior cingulate cortex coded positively for  $Q(t)$  (**Fig 3A**, table S1). Similar analysis conducted for the non-social task (GLM4) revealed that the vmPFC encodes the expected value of slot machine reward (**Fig 6A**, table S4). Activations are reported at a whole brain  $p < 0.05$ , FWE cluster corrected threshold, with an initial forming threshold of  $p < 0.001$  (**Fig 3A**, table S1).

As the expected value  $Q(t)$  can be decomposed into the previous expected value  $Q(t-1)$  and current violation of previous expectations  $PE(t)$ , we next sought to disentangle the neural underpinnings of these two components at the outcomes stage. Using GLM2, that includes as

parametric modulators  $Q(t-1)$  and  $PE(t)$ , we performed two one-sample t-tests at the group level, one for  $Q(t-1)$  and one for  $PE(t)$ . The bilateral striatum, the bilateral superior frontal gyrus, the vmPFC, the bilateral angular gyrus and the posterior middle cingulate cortex encoded  $PE(t)$  (**Fig 3B**, table S2) while the bilateral ventral striatum and the vmPFC encoded  $Q(t-1)$  (**Fig 3C**). Similarly, we investigated brain activity that varies parametrically with  $Q(t-1)$  and  $PE(t)$  in the non-social task (GLM5). The right ventral striatum, left medial PFC, superior frontal gyrus and posterior cingulate gyrus encoded  $PE(t)$  positively (**Fig 6B**) while the vmPFC encoded  $Q(t-1)$  positively (**Fig 6C**, table S4).

### **SERT DRN level correlates with ventral striatal expected value of social victories during outcomes**

Next, we sought to investigate the influence of 5-HT on learning social hierarchies. We therefore investigated the relationship between SERT level in the DRN and brain responses related to the expected value of social victories and to the Prediction error in the social dominance learning task. Our hypothesis was that the modulation of the learning rate in this social learning by SERT DRN levels should be reflected in the relationship between inter-individual differences in SERT DRN levels and brain regions encoding expected value of social victories or PE while subjects were learning the status of their opponents. We particularly focused on the ventral striatum because it has been reported to encode both expected value (Ito & Doya, 2015b, 2015a; Yamada et al., 2011) and PE (Diederer et al., 2016; Ottenheimer et al., 2020).

We first performed a correlation between  $BP_{ND}$  in DRN and BOLD signal extracted in the ventral striatum related to  $Q(t-1)$  and  $PE(t)$ , as estimated by GLM2. In the social task, a significant inverse correlation between the BOLD signal relative to  $Q(t-1)$  and  $BP_{ND}$  was observed in the DRN for the left VS ( $r=-0.383$ ,  $p=0.037$ ) and right VS ( $r=-0.392$ ,  $p=0.032$ ) (**Fig 4B**). No significant correlations were observed during the non-social task in the left VS ( $p=0.547$ ) or for the right VS ( $p=0.500$ ). Comparison of the correlation coefficients for the social and non-social tasks showed that they differed ( $p=0.016$ , right VS and  $p=0.023$ , left VS) and were significantly lower for the social task (right VS:  $r=-0.392$ , social task;  $r=0.128$ , non-social task; left VS:  $r=-0.383$ , social task,  $r=0.114$  non-social task).

We also investigated the relationship between the VS BOLD response in the social learning task and SERT availability not from the DRN but from the VS using a predefined ROI from the AAL3 atlas (**Fig 4C**). A negative correlation was observed between the VS BOLD signal related to  $Q(t-1)$  and the local SERT availability ( $r=-0.368$ ,  $p=0.045$  and  $r=-0.401$  and  $p=0.028$  for the left and right VS respectively). No correlation was observed between the local  $BP_{ND}$  and the BOLD signal related to  $PE(t)$  ( $p=0.118$  and  $p=0.098$  for the left and right VS, respectively) (**Fig 4C**).

Analysis performed for the non-social task revealed no correlations between BOLD signal and SERT availability in the right VS ( $p=0.610$ ) or in the left VS ( $p=0.085$ ). Comparison of the correlation coefficients obtained in the social and non-social task showed lower correlation coefficient for the social task (right VS:  $r=-0.401$ ; left VS:  $r=-0.368$ ) than the one for the non-social task (right VS:  $r=0.246$ ,  $p=0.004$ ; left VS:  $r=0.092$ ,  $p=0.047$ ).

### **SERT levels in the ventral striatum correlate with the BOLD signal related to social defeats**

A previous study reported a relative ventral striatum deactivation for social defeats compared to social victories (Ligneul et al., 2016). Results obtained in this study confirm this earlier report as ROI analyses conducted in the VS revealed a difference in the comparison [Victory>Defeat] ( $p < 0.001$  for both the left and right VS). Moreover, when directly comparing



social victories and social defeats, a significant BOLD response was observed in the right caudate, left putamen and bilateral middle temporal gyrus (**Fig 5**, table S3).

Next, we extracted ventral striatum BOLD signal related to the social defeats and social victories at outcome separately and investigated potential correlations between the BOLD signal and SERT level in the VS. The results revealed a positive correlation between the BOLD signal related to social defeats and SERT level in the left VS (Spearman correlation,  $r=0.511$ ,  $p=0.004$ ) and the right VS ( $r=0.485$ ,  $p=0.007$ ) (**Fig 5**). No correlation was found between SERT availability and BOLD signal related to social victories ( $p=0.571$  and  $p=0.982$  for the left and right VS respectively). We also investigated the differential correlation between local VS  $BP_{ND}$  in the contrast [Defeats > Victories]. The results revealed a negative correlation between the BOLD signal for the contrast [Defeats > Victories] and SERT availability in the right VS ( $r=0.410$ ,  $p=0.024$ ) and a corresponding tendency for the left VS ( $r=0.303$ ,  $p=0.072$ ).

Finally, we compared the correlation coefficients between the social and non-social task (i.e.  $r$  from  $BP_{ND}$  of the VS and the contrast [Defeats > Victories]). The correlation coefficient of the social task was significantly higher ( $r = 0.410$ ) than that for the non-social task ( $r=-0.198$ ) in the right VS ( $p=0.024$ ) and a similar trend towards significance was observed in the left VS ( $p=0.059$ ) ( $r=0.303$  for social task and  $r=-0.177$  for non-social task).

## Discussion

We investigated the link between serotonergic activity, reflected by the non-displaceable binding potential ( $BP_{ND}$ ) of [ $^{11}C$ ]-DASB to SERT, learning of social ranks and learning in a non-social context. Lower levels of SERT in the DRN were linked to higher learning rates during the social competition task, but no such relationship was observed in non-social context. When learning social ranks, activity in the ventral striatum and in the vmPFC correlated with the Prediction Error and the expected value of social victories, representing the total social victories over all successive steps. The expected value of social victories signal from the ventral striatum in the social learning task correlated negatively with the level of SERT availability in the DRN. Moreover, this relationship only occurred in the social learning context. This result indicates a direct relationship between 5-HT and the modulation of expected value of social victories signals, updating the expected values of victory, in a social context. Furthermore, individuals with lower SERT binding potential within the DRN showed higher levels of competitive behavior as the task progressed. These individuals also showed higher relative deactivation in the ventral striatum in response to social defeats, suggesting a link between extracellular serotonin levels and the ability to cope with social competition in relation to the neural response to repeated social defeats.

These results establish a link between SERT availability, learning social ranks and the neurocomputational mechanisms engaged in the integration of long-term social rewards. SERT availability measured using the  $BP_{ND}$  of the [ $^{11}C$ ]-DASB is proportional to the SERT density and affinity, which both contribute to serotonin clearance. Thus, low SERT availability may result in slower clearance of synaptic serotonin compared to when there is high SERT availability, but may also reflect a lower density of serotonergic synapses where SERT is expressed, an increased level of extracellular serotonin or lower SERT expression at nerve terminals.

Dorsal raphe nucleus 5-HT neurons may exert strong effects on a wide variety of behaviors, including social behavior (Kiser et al., 2012; Li et al., 2016; Tse & Bond, 2002; Watson et al., 2009), uncertainty (Miyazaki et al., 2018), punishment/rewards (Boureau & Dayan, 2011a; Lottem et al., 2018; Matias et al., 2017), inhibition of action (Boureau & Dayan, 2011a), patience (McDannald, 2015; Miyazaki et al., 2012, 2018; Miyazaki et al., 2014) and learning (Ligaya et al., 2018). Although apparently diverse, these behaviors are consistent with various aspects of our current findings.

### **Serotonin, social vs non-social rewards and unexpected uncertainty**

One strength of this study was to compare the SERT-BOLD relationship in social learning relative to non-social learning. This relationship between SERT (both in the DRN and the ventral striatum), and BOLD expected value-related striatal activity only occurred in the social context. The specificity of this relationship to the social context may be due to the fact that when decisions are made in such a context, the degree of uncertainty with respect to the possible outcomes increases dramatically, because the behavior of other individuals' is more difficult to predict than the outcome of a slot machine with fixed payoff probability (Khalvati et al., 2019; Park et al., 2019). A number of theoretical accounts have proposed that unexpected uncertainty (i.e. variability reflecting real changes in the environment) could be encoded by 5-HT (Soltani & Izquierdo, 2019; Yu & Dayan, 2005). Thus, differences in unexpected uncertainty between the social and non-social condition may be explained by the necessity of individuals to track the value of the opponent better, in order to update that opponent's value accurately for next trials. Confirming this difference between the social and non-social condition, direct assessment of both the modeled choice entropy and the temperature parameter (beta) were significantly different between conditions ( $p < 0.001$  Wilcoxon signed ranks test; for both the choice entropy and the beta parameter). This reflects higher exploratory choice behavior in the social as compared to the non-social condition

(Figure S4). Social decisions may also require more long-term computations of social expected value, in line with recent optogenetic results revealing that serotonin helps learning for long-term associations only, but not for short-term associations (Ligaya et al., 2018).

### **SERT level and ventral striatum encoding of expected value of social victories**

The expected value of social victories signal in the ventral striatum, modulated by SERT availability, can be interpreted as the cumulative prediction error that reflects the history of the participant's choices (**Fig 4**). This value of the selected option is updated, based on the previous expected value of social victories  $Q(t-1)$  and the current prediction error  $PE(t)$ . Such encoding of expected value of social victories is biologically relevant since this signal conveys information about the previous expected value of the options to update for future choices. Electrophysiological recordings have indicated that neurons in the ventral striatum encode such expected values in rats and non-human primates, (Gmaz et al., 2018; Kim et al., 2009; Lau & Glimcher, 2008; Strait et al., 2015). Our findings indicate that local ventral striatal computations of expected value of social victories are modulated by SERT availability, both locally in the ventral striatum and in projections from the DRN (**Fig 4**). This relationship between DRN SERT and BOLD-related ventral striatal activity is consistent with computational theories of serotonin (Daw et al., 2002; Luo et al., 2016), proposing that tonic serotonergic signal reflects the long-run average reward rate as an average RL algorithm (Boureau & Dayan, 2011a; Daw et al., 2002) or proposing that serotonin indicates how beneficial the current environment feels to the animal (Liu et al., 2014; Luo et al., 2016; Zhong et al., 2017). Generally, in reward RL algorithms, actions chosen to optimize the expected value optimize the long-term average reward received per time step, and not the cumulative reward received over a finite time window. More recently, the concept of beneficialness has been developed based on optogenetics and electrophysiological recordings from the DRN of freely behaving animals. It supports the theory that the firing rate of DRN serotonin neurons increases until the outcome is experienced, and is relative to the overall amount of reward earned during the previous trials (Liu et al., 2014; Zhong et al., 2017). The cumulated prediction error, which corresponds to a proxy for the expected value, reflects how much a particular option has been rewarded, and could relate to accumulated evidence of how beneficial an option is, based on previous experience and expected value. Thus, the link between SERT availability and striatal activity related to expected value of social victories establishes for the first time in humans a relationship between a computational role of serotonin (how beneficial the current environment appears to be to the animal) and local computations of expected value of social victories signal in the striatum.

### **Exploration or exploitation behavior?**

When participants were separated into two groups according to low or high SERT levels in the DRN (low and high  $BP_{ND}$  groups), we observed decreasing levels of competition as the task progressed in the high  $BP_{ND}$  group, who presumably have lower extracellular serotonin (we cannot rule out the others possible interpretations, leading to the same conclusion regarding the serotonin clearance) (**Fig 2D**). This behavior can be attributed to two underlying causes. First, recent recording of the serotonin level within the striatum showed that it may act as a protective signal that inhibits an over-reaction to negative outcomes (Moran et al., 2018). Thus, more competitive behavior observed in individuals with low SERT availability may be because they are less impacted by social defeats. Thus, such individuals are prepared to compete more, whereas individuals with high SERT availability tend to be more likely to select the weaker opponent, as they are more sensitive to social defeats. Alternatively, and non-exclusively, low  $BP_{ND}$  from/in the DRN may favor the persistence of a default choice to compete. This effect may be related to an

alternative interpretation of the classical role of serotonin in action inhibition or in waiting behavior (Crockett et al., 2012; Da Silva et al., 2018; Fonseca et al., 2015; Guitart-Masip et al., 2012). Indeed, a recent optogenetic study suggested that the reason that 5-HT stimulation favors patient waiting is not because it favors behavioral inhibition or passivity but because it favors persistence in a current behavior, even if it is active (Lottem et al., 2018). Here, the current behavior was to try to win the competitive social task, even after relative social status has been learned. Thus, lower  $BP_{ND}$  in the DRN, presumably resulting in higher extracellular 5-HT, may favor persistence in selecting the strongest opponent even when the alternative option (to play against the intermediate or lowest opponent) is more likely to lead to a social victory. Even though individuals with low  $BP_{ND}$  in the DRN learned the hierarchy faster than individuals with higher  $BP_{ND}$  in the DRN, they still consistently favored the more competitive option (i.e. they were more willing to challenge the strongest opponent).

Low SERT availability in the DRN was also associated with higher learning rate in the social context, which in turn was associated with lower inverse temperature (i.e. more random choices during the social task). Thus, participants with lower SERT level learned the hierarchy faster, but were less consistent in their choices during the task. This can be viewed as having a higher level of exploration or perhaps reflecting that they were willing to challenge the strongest opponents even after having learned the hierarchy. Having a high learning rate combined with low beta parameter in a stable environment (as is the case here, since winning probabilities were stable during the task) is not optimal to maximize benefits (Zhang et al., 2020). This suggests that individuals with lower SERT were suboptimal in their social behavior.

### **Punishments, social defeats and ventral striatum/SERT level relationship**

At 5-HT projections site, The ventral striatum reacted positively to victories and negatively to social defeats (**Fig 5**). More importantly, a positive correlation between the BOLD signal related to social defeats and SERT availability (low SERT levels presumably reflecting higher levels of extracellular 5-HT) was observed in the ventral striatum. Lower SERT availability is associated with larger differences in the victory *versus* defeat BOLD signal, suggesting that SERT availability modulates striatal activity related to the relative difference between defeats and victories. The fact that the ventral striatum responded in an asymmetric fashion to social defeats and victories, and that the relative striatal decrease in BOLD response to defeats was enhanced when there is lower SERT availability resonates with classical involvement of serotonin in coding punishment, often in asymmetric opposition to dopamine (DA) and rewards (Boureau & Dayan, 2011b; Daw et al., 2002; Dayan & Huys, 2008; Eldar et al., 2018; Michely, Eldar, Erdman, et al., 2020; Schultz et al., 1997). A classical theory is that 5-HT and DA play opposing roles, 5-HT being associated with punishments and DA with rewards. This view is consistent with our finding that the relative decrease in ventral striatum activity in response to social defeats was more pronounced when there was lower SERT availability, and perhaps therefore, higher levels of 5-HT. PE-related ventral striatal activity has often been associated with phasic DA release. Model-based fMRI was originally used to model DA-ergic neuronal responses with reinforcement learning (Schultz et al., 1997). However, it should be noted that the 5-HT-DA reward antagonism theory does not fit well with recent observations using optogenetic paradigms which show that rewards but not aversive stimuli excite DRN 5-HT neurons (Li et al., 2016; Liu et al., 2014; Wittmann et al., 2020; Zhong et al., 2017).

It should be noted that higher learning rates combined with increased competitiveness in low SERT individuals could result in enhanced negative neural reactions as reflected by negative striatal response to social defeats (**Fig 5**) but also in an enhanced striatal response for positive expectations of making successful future choices to win the social competition (positive

correlation with Q-value, **Fig 4**). We further tested whether the differential influence of 5-HT on brain responses to social victories and defeats could be related to a differential effect of 5-HT on learning rates from defeats and victories. However, this appears not to be the case because the modeling of learning with different learning rates for victories and defeats proved to be not the best fitting model (see methods and results section). It is even the case for both the low and high BP<sub>ND</sub> DRN group separately.

### **SERT, social defeats and psychosocial disorders**

Disrupted social behavior and social avoidance is a core clinical feature of many neuropathological disorders. In the social condition, individuals with lower SERT availability in the ventral striatum showed lower ventral striatal BOLD responses to social defeats compared to individuals with higher SERT availability (**Fig 5**). Depressed individuals are known to show blunted striatal responses to monetary rewards (Rappaport et al., 2020). Depression severity is also associated with diminished reward system activation (Satterthwaite et al., 2015). In contrast, up regulation of serotonin levels, through antidepressant medication or administration of precursors of serotonin biosynthesis, has been found to increase the frequency of dominance-related behaviors (Moskowitz et al., 2001). SERT is the target of Serotonin Selective Reuptake Inhibitors (SSRIs), commonly used against depression (Vaswani et al., 2003). There is a refractory period to SSRI treatment in two-thirds of patients. Higher pretreatment diencephalic SERT availability is correlated with the response to SSRI treatment four weeks later (Baudry et al., 2019; Kugaya et al., 2004). Moreover, individuals that carry a long variant of the promoter for SERT and show higher SERT levels, respond faster and better to SSRIs (Caspi et al., 2003; Keers et al., 2011; Porcelli et al., 2012; Ruhé et al., 2009; Serretti et al., 2007). These data, together with the fact that one of the most reliable animal models of depression, the chronic social defeat stress model (Knowland & Lim, 2018), suggest that higher SERT availability (both in the diencephalon and striatum) may confer resistance to social defeats in depression. Assessing SERT functions and SERT-BOLD relationships in response to social defeats could therefore be of great importance to predict the course of treatment in depression and to understand inter-individual differences in vulnerability to social stress that can result in subordination (Komori et al., 2019; van der Kooij & Sandi, 2015).

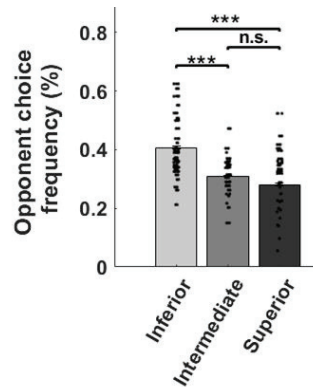
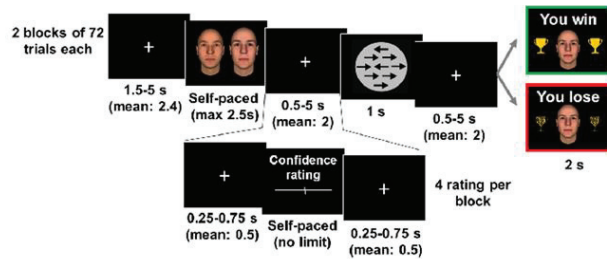
### **Time-scale of the relationship between ventral striatum activity and BP DRN**

The ventral striatum encoded both expected value of social victories of the opponent and PE at the time of the outcome. Yet, only ventral striatum activity related to the expected value of social victories (but not PE) correlated with SERT availability (both locally and from the DRN). This is consistent with the time-scales of both the PET measurement (i.e. one BPND value per subject for a given brain region), and the expected value of social victories that reflects the incorporation of future social victories over long periods. The 5-HT neuromodulatory system is known to participate in a variety of cognitive processes at different time scales, including slow time-scale cognitive processes such as motivation, mood and learning (Cools et al., 2011; Dayan, 2012; Eldar et al., 2018; Michely, Eldar, Martin, et al., 2020). Recent findings also reveal sub-second serotonin fluctuations which may be in opposition to dopamine signaling (Bang et al., 2020), and show positive transients to negative reward PE and negative transients to positive reward PE (Kishida et al., 2016; Moran et al., 2018). In humans, methods such as PET or pharmacological approaches are on the timescale of minutes and cannot resolve the sub-second computations believed to be supported by fast neuromodulation (Dayan, 2012). These different approaches are complementary as neuromodulators such as 5-HT can signal over more than one

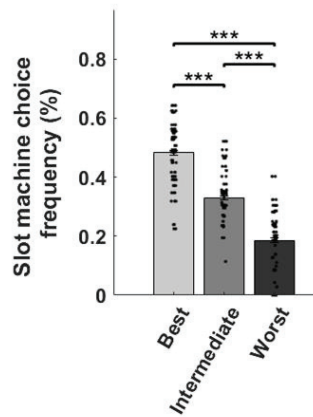
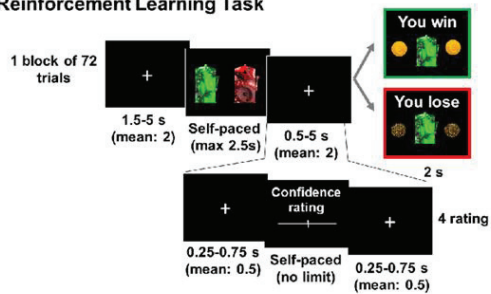
timescale, with partially separable tonic and phasic activity, and different receptor types that may be sensitive to the different timescales.

To conclude, during the learning of social dominance relationships through competition, levels of SERT availability impinge on social learning at the behavioral and neural levels. The level of SERT available in the DRN correlates with the learning rate, and local SERT availability in the ventral striatum is linked to both expected value of social victories and to defeat-related striatal BOLD responses. Inter-individual variations in 5-HT levels also affect confidence and competitiveness as individuals with low DRN SERT availability are less confident and more competitive than individuals with high SERT availability. Our simultaneous multimodal neuroimaging PET-fMRI approach reveals new direct relationships between the complex role of serotonin signaling and the neurocomputational basis of social learning during competitive interactions.

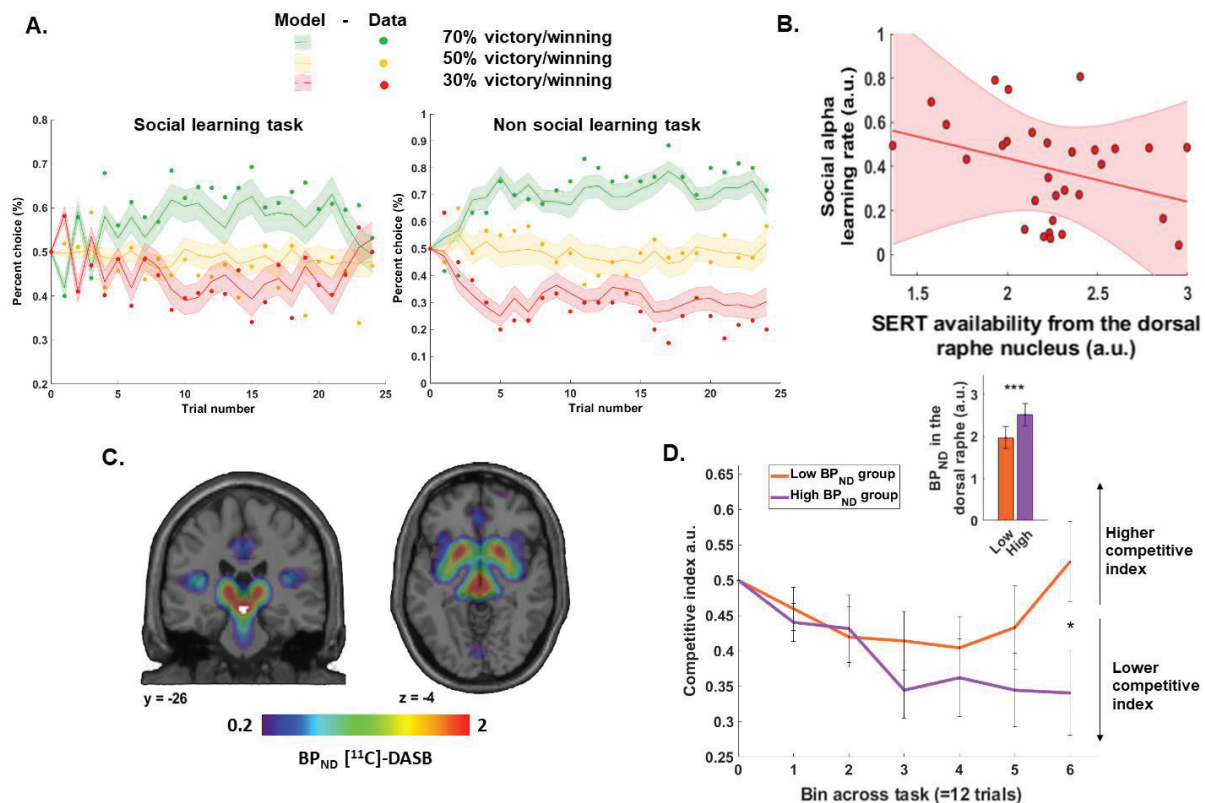
**A. Social Dominance Learning Task**



**B. Reinforcement Learning Task**

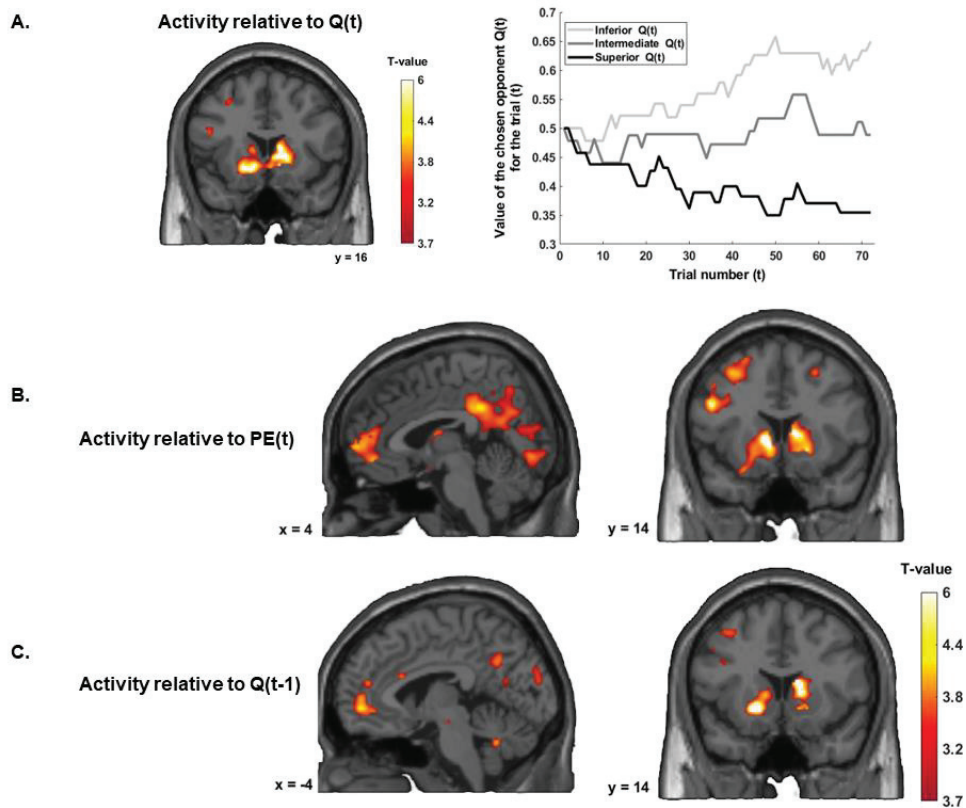


**Figure 1. Tasks and behavioral results.** **A.** Social Hierarchy Learning task (left). Participants were led to believe they were competing against one of three real opponents. Unbeknownst to them, the probability they would win was predefined at  $P=28\%$ ,  $50\%$  and  $72\%$  for the superior, intermediate and inferior opponents, respectively. In any one trial participants chose which one of two opponents they preferred to “compete” against. After competing in a perceptual decision-making task (circle with arrows), the outcome of the competition was delivered. For some trials (four per opponent), participants rated how confident they were of winning against the selected opponent. Right: bar graphs represent the frequency with which they selected each opponent. Participants preferred to select the opponent against whom they had more chance of winning. **B.** Reinforcement Learning paradigm. Similar to the Social hierarchy learning task, participants chose which one of two 2 slot machines from among 3 (winning probabilities:  $P=28\%$ ,  $50\%$  and  $72\%$  for the worst, intermediate and best chance to win, respectively) they preferred to bet on. This was followed by a outcome phase in which they were informed if they had won or lost. In some trials, participants estimated how confident they were of winning on the selected slot machine. Right: bar graphs represent the frequency that each slot machine was chosen. Participants preferred to select the slot machines on which they had more chance of winning. \*\*\* $p < 0.001$ .

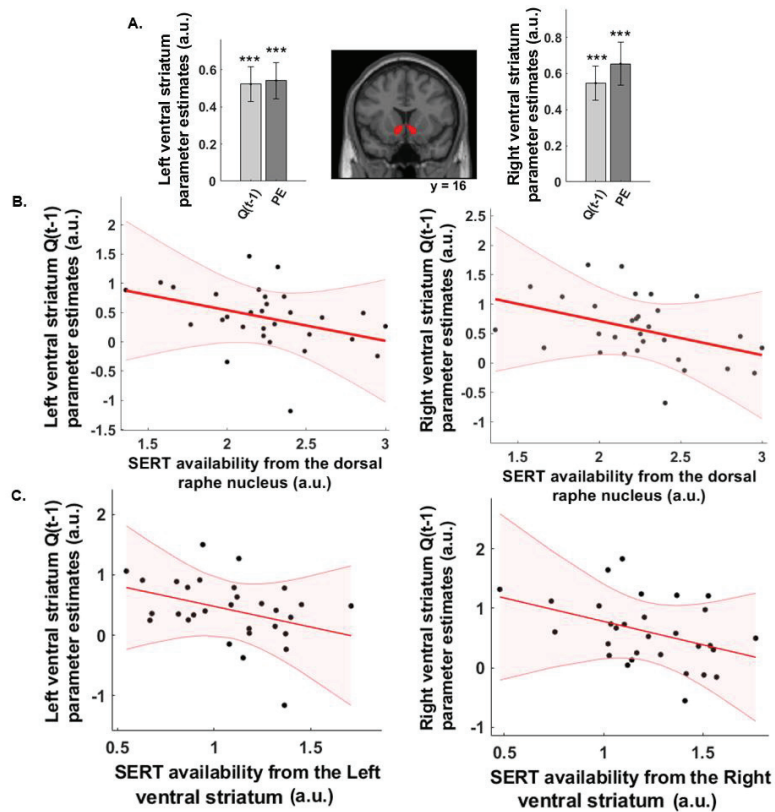


**Figure 2. Binding potential in the dorsal raphe nucleus modulates social learning and competitive behavior.** **A.** Left. Participants choice frequency during the social dominance learning task (dots) when facing the Inferior (green), Intermediate (orange) and Superior (red) opponents and model choice probability estimated by the RL algorithm. Right. Same illustration for the RL task. **B.** Negative correlation between the BP<sub>ND, DRN</sub> and participants' learning rate in the social task. No correlation was observed between BP<sub>ND, DRN</sub> and learning rate in the non-social task. **C.** Statistical map of the average BP<sub>ND</sub>. **D. Competitive choices in the High and Low Binding Potential groups.** Individuals with lower BP<sub>ND, DRN</sub> tended to increase their competitive choices, i.e. they chose to play against the stronger of the two opponents, in later trials. Interaction between trial bins and group (Low vs high BP<sub>ND</sub>) ( $F_{(1,5)}=2.32$ ,  $p = 0.046$ ). *Post hoc* tests conducted on the last bin revealed that the high BP<sub>ND</sub> group made less competitive choices in the last bin of the task ( $M=0.34$ ,  $SEM=0.059$ ) compared to the low BP<sub>ND</sub> group ( $M=0.52$ ,  $SEM=0.056$ ) ( $t_{(28)}=0.031$ ). The bar graphs show a between groups difference in BP<sub>ND, DRN</sub> level based on a median split of individuals. Errors bars represent SEM. \*\*\* $p < 0.001$ . BP<sub>ND</sub> = non-dissociable Binding Potential, DRN = Dorsal Raphe Nucleus.

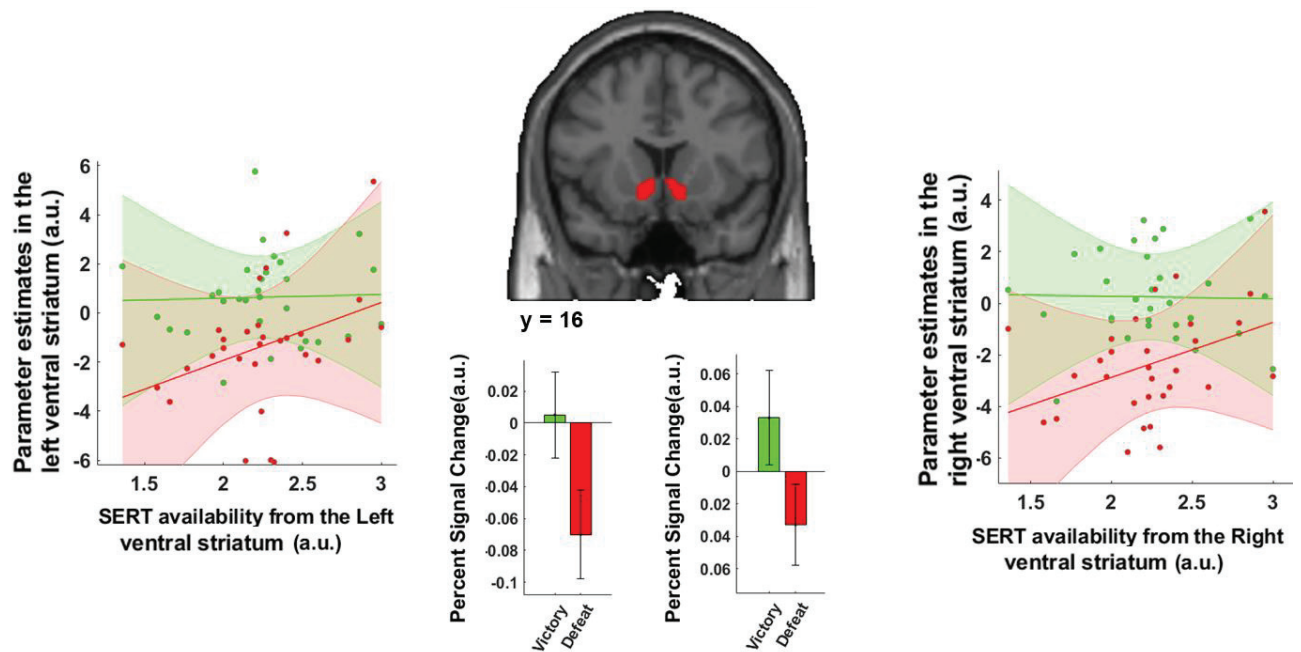




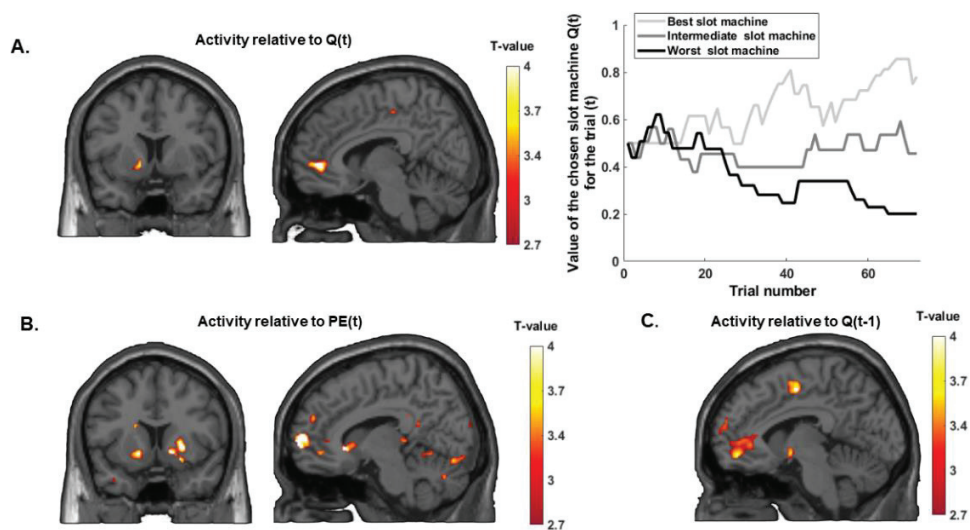
**Figure 3. Brain regions tracking the expected value of social victories  $Q(t)$  at the outcome of the competition.** **A.** The bilateral ventral striatum, vmPFC and right dlPFC encoded  $Q(t)$  during the outcomes, reflecting the updated expected value of social victories. The graph on the right side represents the evolution of  $Q(t)$  for a participant over the experiment. **B.** The prediction error  $PE(t)$  is encoded in the bilateral ventral striatum, vmPFC, bilateral superior frontal gyrus and posterior middle cingulate gyrus. **C.**  $Q(t-1)$  engages the bilateral ventral striatum, vmPFC and middle temporal gyrus. All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected. vmPFC = ventromedial Prefrontal Cortex, dlPFC = dorsolateral prefrontal cortex.



**Figure 4. Correlation between the signal of expected value of social victories in the ventral striatum and  $BP_{ND}$  from the dorsal raphe nucleus or ventral striatum. A.** Ventral striatum BOLD signal tracking both  $Q(t-1)$  and  $PE(t)$  at outcome. **B.** Negative correlation between ventral striatum activity tracking expected value of social victories  $Q(t-1)$  at the outcome stage and SERT availability in the DRN and **C.** with the SERT availability in the ventral striatum.



**Figure 5. Ventral striatum social defeat signal positively correlates with SERT availability in the ventral striatum.** Middle: statistical map of the contrast [Victory>Defeat] and BOLD signal extracted in the ventral striatum. Left and right. BOLD signal during social defeats outcomes positively correlates with SERT availability in the ventral striatum ( $p = 0.004$ ,  $r = 0.511$  and  $p = 0.007$ ,  $r = 0.485$  for the left and right VS).



**Figure 6: fMRI parametrical activity related to expected value and Prediction Error for the non-social learning task.** **A.** The ventromedial prefrontal cortex encodes the expected value of the slot machine at outcome for the non-social task. The graphs represent a participant Q-value for 3 slot machines. **B.** Parametric activation with the prediction error PE(t) observed in the right ventral striatum, the left medial prefrontal cortex, the superior frontal gyrus and the medial posterior cingulate gyrus. **C.** Parametric activation with Q(t-1) is observed in the vmPFC. All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected.

## References

- Aan Het Rot, M., Moskowitz, D. S., Pinard, G., & Young, S. N. (2006). Social behaviour and mood in everyday life: The effects of tryptophan in quarrelsome individuals. *Journal of Psychiatry and Neuroscience*, *31*(4), 253–262.
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. <https://doi.org/10.1016/j.neuroimage.2007.07.007>
- Bang, D., Kishida, K. T., Lohrenz, T., Tatter, S. B., Fleming, S. M., Montague, P. R., Bang, D., Kishida, K. T., Lohrenz, T., White, J. P., Laxton, A. W., & Tatter, S. B. (2020). Article Sub-second Dopamine and Serotonin Signaling in Human Striatum during Perceptual Decision-Making. *Neuron*, 1–12. <https://doi.org/10.1016/j.neuron.2020.09.015>
- Baudry, A., Pietri, M., Launay, J. M., Kellermann, O., & Schneider, B. (2019). Multifaceted regulations of the serotonin transporter: Impact on antidepressant response. *Frontiers in Neuroscience*, *13*(FEB), 1–13. <https://doi.org/10.3389/fnins.2019.00091>
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1960). *An Inventory for Measuring Depression The difficulties inherent in obtaining*. 561–571.
- Boureau, Y. L., & Dayan, P. (2011a). Opponency revisited: Competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, *36*(1), 74–97. <https://doi.org/10.1038/npp.2010.151>
- Boureau, Y. L., & Dayan, P. (2011b). Opponency revisited: Competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, *36*(1), 74–97. <https://doi.org/10.1038/npp.2010.151>
- Carver, C. S., & White, T. L. (2005). Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales. *Current History*, *104*(685), 380–389. <https://doi.org/10.1525/curh.2005.104.685.380>
- Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H. L., McClay, J., Mill, J., Martin, J., Braithwaite, A., & Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, *301*(5631), 386–389. <https://doi.org/10.1126/science.1083968>
- Challis, C., Boulden, J., Veerakumar, A., Espallergues, J., Vassoler, F. M., Christopher Pierce, R., Beck, S. G., & Berton, O. (2013). Raphe GABAergic neurons mediate the acquisition of avoidance after social defeat. *Journal of Neuroscience*, *33*(35), 13978–13988. <https://doi.org/10.1523/JNEUROSCI.2383-13.2013>
- Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, *20*(3), 304–313. <https://doi.org/10.1038/nn.4499>
- Cohen, J. Y., Amoroso, M. W., & Uchida, N. (2015). Serotonergic neurons signal reward and punishment on multiple timescales. *ELife*, *2015*(4), 1–25. <https://doi.org/10.7554/eLife.06346>
- Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology*, *36*(1), 98–113. <https://doi.org/10.1038/npp.2010.121>
- Crockett, M. J., Clark, L., Apergis-Schoute, A. M., Morein-Zamir, S., & Robbins, T. W. (2012). Serotonin modulates the effects of pavlovian aversive predictions on response vigor. *Neuropsychopharmacology*, *37*(10), 2244–2252. <https://doi.org/10.1038/npp.2012.75>
- Crockett, M. J., Clark, L., & Robbins, T. W. (2009). Reconciling the role of serotonin in behavioral inhibition and aversion: Acute tryptophan depletion abolishes punishment-induced inhibition in humans. *Journal of Neuroscience*, *29*(38), 11993–11999.

- <https://doi.org/10.1523/JNEUROSCI.2513-09.2009>
- Da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*(7691), 244–248. <https://doi.org/10.1038/nature25457>
- Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*, *10*(1). <https://doi.org/10.1371/journal.pcbi.1003441>
- Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4–6), 603–616. [https://doi.org/10.1016/S0893-6080\(02\)00052-7](https://doi.org/10.1016/S0893-6080(02)00052-7)
- Dayan, P. (2012). Twenty-Five Lessons from Computational Neuromodulation. *Neuron*, *76*(1), 240–256. <https://doi.org/10.1016/j.neuron.2012.09.027>
- Dayan, P., & Huys, Q. J. M. (2008). Serotonin, inhibition, and negative mood. *PLoS Computational Biology*, *4*(2). <https://doi.org/10.1371/journal.pcbi.0040004>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. <https://doi.org/10.1038/nn.3279>
- Diederer, K. M. M. J., Spencer, T., Vestergaard, M. D. D., Fletcher, P. C. C., & Schultz, W. (2016). Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. *Neuron*, *90*(5), 1127–1138. <https://doi.org/10.1016/j.neuron.2016.04.019>
- Dölen, G., Darvishzadeh, A., Huang, K. W., & Malenka, R. C. (2013). Social reward requires coordinated activity of accumbens oxytocin and 5HT. *Nature Communications*, *176*(12), 139–148. <https://doi.org/10.1016/j.physbeh.2017.03.040>
- Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of Reward Learning Signals Predicts Mood Fluctuations. *Current Biology*, *28*(9), 1433–1439.e7. <https://doi.org/10.1016/j.cub.2018.03.038>
- Ellis, L. (1995). Dominance and Reproductive Success Among Nonhuman Animals: A Cross-Species Comparison. *Ethology and Sociobiology*, *33*(3), 257–333.
- Fonseca, M. S., Murakami, M., & Mainen, Z. F. (2015). Activation of dorsal raphe serotonergic neurons promotes waiting but is not reinforcing. *Current Biology*, *25*(3), 306–315. <https://doi.org/10.1016/j.cub.2014.12.002>
- Gmaz, J. M., Carmichael, J. E., & van der Meer, M. A. A. (2018). Persistent coding of outcome-predictive cue features in the rat nucleus accumbens. *eLife*, *7*, 1–26. <https://doi.org/10.7554/eLife.37275>
- Goette, L., Bendahan, S., Thoresen, J., Hollis, F., & Sandi, C. (2015). Stress pulls us apart: Anxiety leads to differences in competitive confidence under stress. *Psychoneuroendocrinology*, *54*, 115–123. <https://doi.org/10.1016/j.psyneuen.2015.01.019>
- Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., & Hammers, A. (2008). Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, *40*(2), 672–684. <https://doi.org/10.1016/j.neuroimage.2007.11.034>
- Guitart-Masip, M., Chowdhury, R., Sharot, T., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Action controls dopaminergic enhancement of reward representations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(19), 7511–7516. <https://doi.org/10.1073/pnas.1202229109>
- Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., & Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, *19*(4), 224–247. <https://doi.org/10.1002/hbm.10123>

- Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F., & Dayan, P. (2018). An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, 9(1), 10–12. <https://doi.org/10.1038/s41467-018-04840-2>
- Ito, M., & Doya, K. (2015a). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *Journal of Neuroscience*, 35(8), 3499–3514. <https://doi.org/10.1523/JNEUROSCI.1962-14.2015>
- Ito, M., & Doya, K. (2015b). Parallel Representation of Value-Based and Finite State-Based Strategies in the Ventral and Dorsal Striatum. *PLoS Computational Biology*, 11(11), 1–25. <https://doi.org/10.1371/journal.pcbi.1004540>
- Johnson, S. L., Leedom, L. J., & Muhtadie, L. (2012). The dominance behavioral system and psychopathology: Evidence from self-report, observational, and biological studies. *Psychological Bulletin*, 138(4), 692–743. <https://doi.org/10.1037/a0027503>
- Keers, R., Uher, R., Huezo-Diaz, P., Smith, R., Jaffee, S., Rietschel, M., Henigsberg, N., Kozel, D., Mors, O., Maier, W., Zobel, A., Hauser, J., Souery, D., Placentino, A., Larsen, E. R., Dmitrzak-Weglaz, M., Gupta, B., Hoda, F., Craig, I., ... Aitchison, K. J. (2011). Interaction between serotonin transporter gene variants and life events predicts response to antidepressants in the GENDEP project. *Pharmacogenomics Journal*, 11(2), 138–145. <https://doi.org/10.1038/tpj.2010.14>
- Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J. C., & Rao, R. P. N. (2019). Bayesian Inference of Other Minds Explains Human Choices in Group Decision Making. *Science Advances*, November. <https://doi.org/10.1101/419515>
- Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience*, 29(47), 14701–14712. <https://doi.org/10.1523/JNEUROSCI.2728-09.2009>
- Kiser, D., Steemer, S. B., Branchi, I., & Homberg, J. R. (2012). The reciprocal interaction between serotonin and social behaviour. *Neuroscience and Biobehavioral Reviews*, 36(2), 786–798. <https://doi.org/10.1016/j.neubiorev.2011.12.009>
- Kish, S. J., Furukawa, Y., Chang, L. J., Tong, J., Ginovart, N., Wilson, A., Houle, S., & Meyer, J. H. (2005). Regional distribution of serotonin transporter protein in postmortem human brain: Is the cerebellum a SERT-free brain region? *Nuclear Medicine and Biology*, 32(2), 123–128. <https://doi.org/10.1016/j.nucmedbio.2004.10.001>
- Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., White, J. P., Ellis, T. L., Phillips, P. E. M., & Montague, P. R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1), 200–205. <https://doi.org/10.1073/pnas.1513619112>
- Knowland, D., & Lim, B. K. (2018). Circuit-based frameworks of depressive behaviors: The role of reward circuitry and beyond. *Pharmacology Biochemistry and Behavior*, 174, 42–52. <https://doi.org/10.1016/j.pbb.2017.12.010>
- Komori, T., Makinodan, M., & Kishimoto, T. (2019). Social status and modern-type depression: A review. *Brain and Behavior*, 9(12), 1–9. <https://doi.org/10.1002/brb3.1464>
- Kugaya, A., Sanacora, G., Staley, J. K., Malison, R. T., Bozkurt, A., Khan, S., Anand, A., Van Dyck, C. H., Baldwin, R. M., Seibyl, J. P., Charney, D., & Innis, R. B. (2004). Brain serotonin transporter availability predicts treatment response to selective serotonin reuptake inhibitors. *Biological Psychiatry*, 56(7), 497–502. <https://doi.org/10.1016/j.biopsych.2004.07.001>
- Kvaal, K., Ulstein, I., Nordhus, I. H., & Engedal, K. (2005). The Spielberger State-Trait Anxiety Inventory (STAI): The state scale in detecting mental disorders in geriatric patients.

- International Journal of Geriatric Psychiatry*, 20(7), 629–634.  
<https://doi.org/10.1002/gps.1330>
- Lammertsma, A. A., & Hume, S. P. (1996). Simplified reference tissue model for PET receptor studies. *NeuroImage*, 4(3), 153–158. <https://doi.org/10.1006/nimg.1996.0066>
- Lau, B., & Glimcher, P. W. (2008). Value Representations in the Primate Striatum during Matching Behavior. *Neuron*, 58(3), 451–463.  
<https://doi.org/10.1016/j.neuron.2008.02.021>
- Li, Y., Zhong, W., Wang, D., Feng, Q., Liu, Z., Zhou, J., Jia, C., Hu, F., Zeng, J., Guo, Q., Fu, L., & Luo, M. (2016). Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nature Communications*, 7. <https://doi.org/10.1038/ncomms10503>
- Ligneul, R., Obeso, I., Ruff, C. C., & Dreher, J.-C. (2016). Dynamical Representation of Dominance Relationships in the Human Rostromedial Prefrontal Cortex. *Current Biology*, 1–9. <https://doi.org/10.1016/j.cub.2016.09.015>
- Liu, Z., Zhou, J., Li, Y., Hu, F., Lu, Y., Ma, M., Feng, Q., Zhang, J. en, Wang, D., Zeng, J., Bao, J., Kim, J. Y., Chen, Z. F., ElMestikawy, S., & Luo, M. (2014). Dorsal raphe neurons signal reward through 5-HT and glutamate. *Neuron*, 81(6), 1360–1374.  
<https://doi.org/10.1016/j.neuron.2014.02.010>
- Lottem, E., Banerjee, D., Verтеchi, P., Sarra, D., Lohuis, M. O., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, 9(1), 1–12. <https://doi.org/10.1038/s41467-018-03438-y>
- Luo, M., Li, Y., & Zhong, W. (2016). Do dorsal raphe 5-HT neurons encode “beneficialness”? *Neurobiology of Learning and Memory*, 135(August), 40–49.  
<https://doi.org/10.1016/j.nlm.2016.08.008>
- Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37, 205–220. <https://doi.org/10.1146/annurev-neuro-071013-014017>
- Matias, S., Lottem, E., Dugué, G. P., & Mainen, Z. F. (2017). Activity patterns of serotonin neurons underlying cognitive flexibility. *ELife*, 6, 1–24. <https://doi.org/10.7554/eLife.20552>
- Matthew Brett, Jean-Luc Anton, Romain Valabregue, Jean-Baptiste Poline. Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage, Vol 16, No 2.
- Matthews, G. A., Nieh, E. H., Vander Weele, C. M., Halbert, S. A., Pradhan, R. V., Yosafat, A. S., Globber, G. F., Izadmehr, E. M., Thomas, R. E., Lacy, G. D., Wildes, C. P., Ungless, M. A., & Tye, K. M. (2016). Dorsal Raphe Dopamine Neurons Represent the Experience of Social Isolation. *Cell*, 164(4), 617–631. <https://doi.org/10.1016/j.cell.2015.12.040>
- McDannald, M. A. (2015). Serotonin: Waiting but not rewarding. *Current Biology*, 25(3), R103–R104. <https://doi.org/10.1016/j.cub.2014.12.019>
- Mérida, I., Reilhac, A., Redouté, J., Heckemann, R. A., Costes, N., & Hammers, A. (2017). Multi-atlas attenuation correction supports full quantification of static and dynamic brain PET data in PET-MR. *Physics in Medicine and Biology*, 62(7), 2834–2858.  
<https://doi.org/10.1088/1361-6560/aa5f6c>
- Michely, J., Eldar, E., Erdman, A., Martin, I. M., & Dolan, R. J. (2020). SSRIs modulate asymmetric learning from reward and punishment. *BioRxiv*, 2020.05.21.108266.
- Michely, J., Eldar, E., Martin, I. M., & Dolan, R. J. (2020). A mechanistic account of serotonin’s impact on mood. *Nature Communications*, 11(1). <https://doi.org/10.1038/s41467-020-16090-2>
- Miyazaki, K., Miyazaki, K. W., & Doya, K. (2012). The role of serotonin in the regulation of patience and impulsivity. *Molecular Neurobiology*, 45(2), 213–224.



- <https://doi.org/10.1007/s12035-012-8232-6>
- Miyazaki, K., Miyazaki, K. W., Yamanaka, A., Tokuda, T., Tanaka, K. F., & Doya, K. (2018). Reward probability and timing uncertainty alter the effect of dorsal raphe serotonin neurons on patience. *Nature Communications*, 9(1). <https://doi.org/10.1038/s41467-018-04496-y>
- Miyazaki, K. W., Miyazaki, K., Tanaka, K. F., Yamanaka, A., Takahashi, A., Tabuchi, S., & Doya, K. (2014). Optogenetic activation of dorsal raphe serotonin neurons enhances patience for future rewards. *Current Biology*, 24(17), 2033–2040. <https://doi.org/10.1016/j.cub.2014.07.041>
- Moran, R. J., Kishida, K. T., Lohrenz, T., Saez, I., Laxton, A. W., Witcher, M. R., Tatter, S. B., Ellis, T. L., Phillips, P. E., Dayan, P., & Montague, P. R. (2018). The Protective Action Encoding of Serotonin Transients in the Human Brain. *Neuropsychopharmacology*, 43(6), 1425–1435. <https://doi.org/10.1038/npp.2017.304>
- Moskowitz, D. ., G., P., D.C., Z., L., A., & S.N., Y. (2001). The effect of tryptophan on social interaction in everyday life: A placebo-controlled study. *Neuropsychopharmacology*, 25(2), 277–289. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed5&NEWS=N&AN=2001221285>
- Noonan, M. A. P., Sallet, J., Mars, R. B., Neubert, F. X., O'Reilly, J. X., Andersson, J. L., Mitchell, A. S., Bell, A. H., Miller, K. L., & Rushworth, M. F. S. (2014). A Neural Circuit Covarying with Social Hierarchy in Macaques. *PLoS Biology*, 12(9). <https://doi.org/10.1371/journal.pbio.1001940>
- Ottenheimer, D. J., Bari, B. A., Sutlief, E., Fraser, K. M., Kim, T. H., Richard, J. M., Cohen, J. Y., & Janak, P. H. (2020). A quantitative reward prediction error signal in the ventral pallidum. *Nature Neuroscience*, 23(10), 1267–1276. <https://doi.org/10.1038/s41593-020-0688-5>
- Park, S. A., Sestito, M., Boorman, E. D., & Dreher, J. C. (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications*, 10(1), 1–12. <https://doi.org/10.1038/s41467-019-12937-5>
- Porcelli, S., Fabbri, C., & Serretti, A. (2012). Meta-analysis of serotonin transporter gene promoter polymorphism (5-HTTLPR) association with antidepressant efficacy. *European Neuropsychopharmacology*, 22(4), 239–258. <https://doi.org/10.1016/j.euroneuro.2011.10.003>
- Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social Dominance Orientation: A Personality Variable Predicting Social and Political Attitudes. *Journal of Personality and Social Psychology*, 67(4), 741–763. <https://doi.org/10.1037/0022-3514.67.4.741>
- Qu, C., Ligneul, R., Van der Henst, J.-B., & Dreher, J.-C. (2017). An Integrative Interdisciplinary Perspective on Social Dominance Hierarchies. *Trends in Cognitive Sciences*, xx, 1–15. <https://doi.org/10.1016/j.tics.2017.08.004>
- Raleigh, M. J., McGuire, M. T., Brammer, G. L., Pollack, D. B., & Yuwiler, A. (1991). Serotonergic mechanisms promote dominance acquisition in adult male vervet monkeys. *Brain Research*, 559(2), 181–190. [https://doi.org/10.1016/0006-8993\(91\)90001-C](https://doi.org/10.1016/0006-8993(91)90001-C)
- Rappaport, B. I., Kandala, S., Luby, J. L., & Barch, D. M. (2020). Brain reward system dysfunction in adolescence: Current, cumulative, and developmental periods of depression. *American Journal of Psychiatry*, 177(8), 754–763. <https://doi.org/10.1176/appi.ajp.2019.19030281>
- Rathus, S. A. (1973). *Assertif Behavior Scale*.
- Reilhac, A., Merida, I., Irace, Z., Stephenson, M. C., Weekes, A. A., Chen, C., Totman, J. J., Townsend, D. W., Fayad, H., & Costes, N. (2018). Development of a dedicated rebinner with rigid motion correction for the mMR PET/MR Scanner, and Validation in a Large

- Cohort of 11C-PIB Scans. *Journal of Nuclear Medicine*, 59(11), 1761–1767.  
<https://doi.org/10.2967/jnumed.117.206375>
- Rolls, E. T., Huang, C. C., Lin, C. P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, 206(May 2019), 116189.  
<https://doi.org/10.1016/j.neuroimage.2019.116189>
- Ruhé, H. G., Ooteman, W., Booij, J., Michel, M. C., Moeton, M., Baas, F., & Schene, A. H. (2009). Serotonin transporter gene promoter polymorphisms modify the association between paroxetine serotonin transporter occupancy and clinical response in major depressive disorder. *Pharmacogenetics and Genomics*, 19(1), 67–76.  
<https://doi.org/10.1097/FPC.0b013e32831a6a3a>
- Sandi, C., & Haller, J. (2015a). Stress and the social brain: behavioural effects and neurobiological mechanisms. *Nature Reviews Neuroscience*, 16(5), 290–304.  
<https://doi.org/10.1038/nrn3918>
- Sandi, C., & Haller, J. (2015b). Stress and the social brain: Behavioural effects and neurobiological mechanisms. *Nature Reviews Neuroscience*, 16(5), 290–304.  
<https://doi.org/10.1038/nrn3918>
- Sapolsky, R. M. (2004). Social Status and Health in Humans and Other Animals. *Annual Review of Anthropology*, 33(1), 393–418.  
<https://doi.org/10.1146/annurev.anthro.33.070203.144000>
- Sapolsky, R. M. (2005). The influence of social hierarchy on primate health. *Science*, 308(5722), 648–652. <https://doi.org/10.1126/science.1106477>
- Satterthwaite, T. D., Kable, J. W., Vandekar, L., Katchmar, N., Bassett, D. S., Baldassano, C. F., Ruparel, K., Elliott, M. A., Sheline, Y. I., Gur, R. C., Gur, R. E., Davatzikos, C., Leibenluft, E., Thase, M. E., & Wolf, D. H. (2015). Common and Dissociable Dysfunction of the Reward System in Bipolar and Unipolar Depression. *Neuropsychopharmacology*, 40(9), 2258–2268. <https://doi.org/10.1038/npp.2015.75>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Serretti, A., Kato, M., De Ronchi, D., & Kinoshita, T. (2007). Meta-analysis of serotonin transporter gene promoter polymorphism (5-HTTLPR) association with selective serotonin reuptake inhibitor efficacy in depressed patients. *Molecular Psychiatry*, 12(3), 247–257.  
<https://doi.org/10.1038/sj.mp.4001926>
- Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, 32(17), 5833–5842. <https://doi.org/10.1523/jneurosci.0053-12.2012>
- Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nat Rev Neurosci*, 176(1), 139–148. <https://doi.org/10.1038/s41583-019-0180-y>. Adaptive
- Steenbergen, L., Jongkees, B. J., Sellaro, R., & Colzato, L. S. (2016). Tryptophan supplementation modulates social behavior: A review. *Neuroscience and Biobehavioral Reviews*, 64, 346–358. <https://doi.org/10.1016/j.neubiorev.2016.02.022>
- Strait, C. E., Sleezer, B. J., & Hayden, B. Y. (2015). Signatures of value comparison in ventral striatum neurons. *PLoS Biology*, 13(6), 1–22. <https://doi.org/10.1371/journal.pbio.1002173>
- Tse, W. S., & Bond, A. J. (2002). Serotonergic intervention affects both social dominance and affiliative behaviour. *Psychopharmacology*, 161(3), 324–330.  
<https://doi.org/10.1007/s00213-002-1049-7>
- Van der Kooij, M. A., & Sandi, C. (2015). The genetics of social hierarchies. *Current Opinion in Behavioral Sciences*, 2, 52–57. <https://doi.org/10.1016/j.cobeha.2014.09.001>

- Vaswani, M., Linda, F. K., & Ramesh, S. (2003). Role of selective serotonin reuptake inhibitors in psychiatric disorders: A comprehensive review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 27(1), 85–102. [https://doi.org/10.1016/S0278-5846\(02\)00338-X](https://doi.org/10.1016/S0278-5846(02)00338-X)
- Wang, F., Kessels, H. W., & Hu, H. (2014). The mouse that roared: Neural mechanisms of social hierarchy. *Trends in Neurosciences*, 37(11), 674–682. <https://doi.org/10.1016/j.tins.2014.07.005>
- Watson, K. K., Ghodasra, J. H., & Platt, M. L. (2009). Serotonin transporter genotype modulates social reward and punishment in rhesus macaques. *PLoS ONE*, 4(1). <https://doi.org/10.1371/journal.pone.0004156>
- Wittmann, M. K., Fouragnan, E., Folloni, D., Klein-Flügge, M. C., Chau, B. K. H., Khamassi, M., & Rushworth, M. F. S. (2020). Global reward state affects learning and activity in raphe nucleus and anterior insula in monkeys. *Nature Communications*, 11(1). <https://doi.org/10.1038/s41467-020-17343-w>
- Yamada, H., Inokawa, H., Matsumoto, N., Ueda, Y., & Kimura, M. (2011). Neuronal basis for evaluating selected action in the primate striatum. *European Journal of Neuroscience*, 34(3), 489–506. <https://doi.org/10.1111/j.1460-9568.2011.07771.x>
- Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4), 681–692. <https://doi.org/10.1016/j.neuron.2005.04.026>
- Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). *Using reinforcement learning models in social neuroscience: frameworks, pitfalls, and suggestions of best practices*. 1–5.
- Zhong, W., Li, Y., Feng, Q., & Luo, M. (2017). Learning and stress shape the reward response patterns of serotonin neurons. *Journal of Neuroscience*, 37(37), 8863–8875. <https://doi.org/10.1523/JNEUROSCI.1181-17.2017>
- Zhou, T., Sandi, C., & Hu, H. (2018). Advances in understanding neural mechanisms of social dominance. *Current Opinion in Neurobiology*, 49, 99–107. <https://doi.org/10.1016/j.conb.2018.01.006>

## STAR Methods

### Participants

Thirty-two healthy volunteers (only males; age range 19 to 32 years; and mean age (M)  $23.4 \pm$  (SD) 2.9) were recruited through a mailing list from the University of Claude Bernard Lyon 1. For inclusion in the study, participants were required to follow the following criteria: French-speaking, right-handed, no current medical treatment, no history of neurological or psychiatric disorders and no auditory, olfactory or visual deficits. Furthermore, volunteers were screened for general MRI counter-indications. A physician conducted medical examinations to follow inclusion criteria. Participants gave their written consent and received monetary compensation for the completion of the study. This study was approved by the Medical Ethics Committee (CPP Sud-Est IV, ID RCB: 2016-A01588-43).

### Scanning procedure and data collection

The stimuli were presented with a screen resolution of  $1280 \times 1024$  pixels, displayed at a visual angle of  $24 - 18^\circ$ , centered on the screen, and surrounded by a black background. The participants were asked to use their index and middle fingers of the same hand to answer by pressing a 4-button controller. Stimuli were presented, and the responses to the stimuli were collected using the Psychtoolbox toolbox Version 3 (PTB-3) on MATLAB (version 7.16.0, R2013a, Natick, Massachusetts: The MathWorks Inc).

Thirty-two subjects underwent functional MRI and PET scans simultaneously to specifically study serotonin transporter (SERT) binding using [ $^{11}\text{C}$ ]-DASB. Neuroimaging was performed at the CERMEP Center (Lyon, France). The subjects were positioned supine on the scanner beds, with their head held in place. Before the bolus injection, the anatomical image was acquired. Then after the bolus injection, the participant rested during 10 minutes before starting the task. The social hierarchy learning task commenced by a fake internet connection to a behavioral laboratory room (see procedure below) for the first social hierarchy competition. After a first block, participants were told to rest for 5 minutes, then they played the non-social learning task and after a second 5 minutes rest period they completed the second social hierarchy learning task.

### Experimental design

#### Social hierarchy learning Task

Participants were first trained outside of the scanner on a perceptual decision-making task. In this task a participant has to indicate within 1 s whether the majority of arrows point to the left or right direction on screen, using a left/right button press. This training period included (non-social) feedback based on trial performance: a red fixation cross indicated that the participant's decision was incorrect, a green cross indicated the participant's decision was correct. A yellow cross meant that the participant did not respond within one second.

The PET-fMRI experiment consisted of a social hierarchy learning task divided into two runs of approximately 15 min each, interleaved by a non-social learning task and followed by a post-task rating to assess learning. In the PET-fMRI Social Hierarchy Learning Task, participants were led to believe that they were competing against three other participants anonymously connected on-line (**Fig. 1A**). Participants were told that they had to interact in real time with the other individuals, and received the following instructions: "If both responses are correct, the fastest player wins. If one player gives an incorrect response, the accurate player wins. If both responses are incorrect, the slowest player wins. If one player doesn't respond, then he loses automatically." For each trial participants first had to select against which of the opponents,

between the two presented on the screen they were going to play against. Then they had to play the competitive perceptual decision-making task (for which they had been trained). Unbeknownst to the subjects, outcomes were manipulated to produce three different probabilities of winning called the Reward Probability (28%, 50%, or 72% of victories), depending on which of the 3 possible opponents they had chosen to play. Importantly, in the social hierarchy learning task, winning or losing against opponents was not associated with monetary incentives but only to social victories or social defeats. Subjects played 72 trials (24 trials per pair of opponents).

The task was composed of the following stages. First, participants had to choose one avatar that would be used to represent him henceforth during the social competition task. Following a short (fake) internet connection with the other participants, the task commenced. At the beginning of every trial participants were asked to choose against which opponent of the two proposed they wanted to compete. Participants were told that they could identify each opponent thanks to the first letter of their name and a neutral avatar. They played the competitive perceptual decision-making task according to the rules explained above. After each trial subjects received feedback concerning the outcome of the competition, which was externally determined according to the defined probability. After each trial a fixation cross was presented with a jittered duration lasting from 2 s to 5 s. In some trials (one at the beginning, one at the end and two in the middle of the task, resulting in four rating accordingly to the participant choice), participants were asked to indicate their confidence level with respect to their probability of winning against the selected opponent (**Fig. 1A**).

Finally, after scanning was completed, participants had to rate their opponents. This rating was composed of two stages. First, they were presented two of the opponents among three and they were required to indicate which had been better during the competition. Second, participants were asked to indicate the percentage of victories they estimated for each opponent. This allowed us to ascertain whether participants had learned the opponents' ranks/status.

### **Non-social learning task (slot machines)**

The non-social learning task was formally similar to the social hierarchy learning task, except that participants were not led to believe that they were in a social interaction and thus did not have to compete against opponents. Instead, participants were told to choose between 2 slot machines from a group of three possible slots machines on each trial. Each slot machine had a defined probability of allowing the participants to win (28%, 50% or 72%). After a fixation cross, participants received feedback about whether they won or lost based on the Reward Probability. As in the social hierarchy learning task, in some trials, participants were asked to indicate their confidence level with respect to the probability of winning with the slot machine they had selected on that trial (**Fig. 1B**).

### **Computational modeling (estimation and comparison procedures)**

#### **Reinforcement Learning model of the social hierarchy learning task**

To capture behavior, we used 6 variations of the Q-learning model. We compared them using Bayesian information criterion to select the model that best described the data using Bayesian group comparison (Daunizeau et al., 2014). All models were constructed based on a similar algorithm. As described above, the models assumed that the probability of choosing to compete with opponent  $i$  over another opponent  $j$  depends on the relative difference in the value of that opponent *versus* the other opponent (both being presented on the screen) (Equation 1). This relationship defines the softmax decision rule:

$$p(i) = \frac{\exp(\beta * Qi)}{\exp(\beta * Qi) + \exp(\beta * Qj)} \quad (\text{Eq. 1})$$

This equation defines the stochastic decision rule (softmax) that calculates the probability  $p(i)$  of choosing the opponent  $i$  given the other opponent  $j$ .  $\beta$  is the inverse temperature parameter. It is a free parameter that dictates to what extent the decision is deterministic relative to the  $Q$  expected value of social victories of the available options ( $\beta$  was constrained in the interval  $0$ - $+\infty$ ).

Then, according to the choice of the participant, the value of the opponent selected was updated following the exponential, recency-weighted average algorithm. The free parameter  $\beta$  is the inverse temperature parameter and dictates to what extent the decision is deterministic relative to the  $Q$  expected value of social victories of the available opponents  $i$  and  $j$ .  $\beta$  was constrained to be positive, and a  $\beta$  of  $0$  represented more exploratory behavior (random choice), whereas a high  $\beta$  denotes a highly exploitative behavior by which the participant prefers to compete with the weaker opponent.

$$Qi(t) = Qi(t - 1) + \alpha(R(t) - Qi(t - 1)) \quad (\text{Eq. 2})$$

$$Qi(t) = Qi(t - 1) + \alpha PE(t) \quad (\text{Eq. 3})$$

Equation 2 assumed that the value of the selected opponent is updated according to his previous value and the differences between the actual reward  $R(t)$  ( $R$  was arbitrarily set to  $1$  for a victory and  $0$  for a defeat) and his previous expected value of social victories  $Q(t-1)$ . This difference is modulated by the free parameter  $\alpha$ , which represents the learning rate of the model and was constrained between  $0$  and  $1$ . The prediction error  $PE(t)$  is the difference between the reward at time  $t$  from the current social value  $Qi(t-1)$  resulting from the ongoing competitive interaction (reward was arbitrary set to  $0$  for a defeat and  $1$  for a victory).  $Qi$  reflects the cumulative social reward prediction error and is called the expected value of social victories. This definition allows us to model that participants mostly preferred to compete against the inferior opponent after probing the other opponents' strengths to avoid social defeats. In fact, defeating an opponent will increase his relative expected value of social victories and losing will decrease his relative expected value of social victories.  $\alpha$  represents the subject's learning rate (between  $0$  and  $1$  and is assumed to be the same for defeats and victories). This allowed us to investigate the neural correlates of both the  $PE(t)$  and the  $Q(t-1)$  value of the selected option in the current trial. Note that  $Qi$  was updated even if the participant did not choose that opponent, because the participant might choose them as an adversary in a future trial, and update their value. There was an exception to this with three participants who systematically chose not to play against one of their three opponents during the entire first block of social competition. These blocks for these participants were therefore entirely removed from the analysis.

Because performance of the perceptual decision-making task could affect the updating of the expected value of social victories ( $Q(t)$ ), we constructed three different families of models. The first family, including RL3 and RL6 followed the updating rule as defined by Equation 2. The second family, including the RL1 and RL4, did not update the expected value of social victories ( $Q(t)$ ) after incorrect answers to the perceptual decision-making task, (ACC monitoring). The last family, including RL2 and RL5 models, did not update after incorrect answers and included a performance-weighting parameter  $\omega$  (Equation 4 and 5 for a victory and a defeat respectively) (ACC monitoring, weighted RT).

$$Q_i(t) = Q_i(t-1) + \omega * (1 - Perf) * (R(t) - Q_i(t-1)) + (1 - \omega) * \alpha * (R(t) - Q_i(t-1)) \quad (\text{Eq. 4})$$

$$Q_i(t) = Q_i(t-1) + \omega * Perf * (R(t) - Q_i(t-1)) + (1 - \omega) * \alpha * (R(t) - Q_i(t-1)) \quad (\text{Eq. 5})$$

The  $\omega$  is the performance weighting parameter, with a higher value of  $\omega$  reflecting a higher effect of the performance, Perf, on the prediction error. It represents how much a participant is sensitive to his own performance and updates the value of the opponent according to the participant's own performance. Perf is the normalized performance of the participant on the current trial and is computed as:

$$Perf = 1 - \frac{\log(RT_t) - \min(\log(RT_t))}{\max(\log(RT_t)) - \min(\log(RT_t))} \quad (\text{Eq supp. 6})$$

Note that the variant of the models used varies according to the expected value of social victories ( $Q(t)$ ). Note that we created two distinct families according to the learning rate for each model defined above (no updating, accuracy monitoring and accuracy monitoring weighted by the reaction time). A first group of models used the same alpha learning rates for victories and defeats and included RL1, RL2 and RL3. The remaining models, RL4, RL5 and RL6, updated the expected value of social victories ( $Q(t)$ ) according to two different alpha learning rates, one for victories and one for defeats. For the models with two alpha rates, the probability of choosing one opponent  $i$  to bet on, over another opponent  $j$  was defined with the same softmax decision rule (Eq supp 1). The difference concerns the value updating. Compared to the first model RL1 there are two different learning rates, one for winning and one for losing:

$$Q_i(t) = Q_i(t-1) + I * \alpha_{win} * PE(t) + (1 - I) * \alpha_{loss} * PE(t) \quad (\text{Eq supp. 7})$$

Where  $I = 1$  if the participant won and 0 if he lost. It assumed that participants learned differently after experiencing a victory or a defeat using a dual asymmetrical learning rate. Estimation of optimal parameters and goodness of fit were performed on the subject level using the Variational Bayes Approach (VBA) proposed by (Daunizeau et al., 2014) and implemented in a validated MATLAB toolbox.

We tested these 6 variants of RL models (table S6) for the social task. Alternative models were compared using Bayesian group comparison with the VBA Toolbox on MATLAB (Daunizeau et al., 2014). For each trial type, only the model presenting the highest exceedance probability using the Bayesian information criterion (BIC) was analyzed and the Log Likelihood (LL) (table S6 for further details).

### Reinforcement Learning model of the non-social learning task

In order to investigate the neural processes underlying learning, we used a Q learning model RL1, and estimated the learning rate and inverse temperature of each participant. This model assumed that the probability of choosing one slot machine  $i$  to bet, over another slot machine  $j$  is highly related to the difference in their internal expected values ( $Q_i$  and  $Q_j$ ).

$$p(i) = \frac{\exp(\beta * Q_i)}{\exp(\beta * Q_i) + \exp(\beta * Q_j)} \quad (\text{Eq supp. 8})$$

This equation defines the stochastic decision rule (softmax) that calculates the probability  $p(i)$  of choosing the slot machine  $i$  given the other  $j$ .  $\beta$  is the inverse temperature parameter. It is a free parameter that dictates to what extent the decision is deterministic relative to the values of the available options ( $\beta \in \mathbb{R}$ ).

Using this RL algorithm, we modeled the dynamic of the expected value of the slot machine  $i$  ( $Q_i$ ) and how it varies according to the feedback received during the outcome stage following the gamble at time  $t$ :

$$Q_i(t) = Q_i(t - 1) + \alpha PE \quad (\text{Eq supp. 9})$$

$R$  is the “reward” resulting from the on-going slot machine, and was arbitrarily set to 0 for a monetary loss and 1 for winning money.  $\alpha$  represents the learning rate of the model. It is constrained between 0 and 1 and is equal for loss and win.  $Q_i(t)$  represents the expected value of the chosen slot machine at the outcome and reflects the weighted cumulative prediction error. This procedure enables us to investigate the neural correlates of the chosen value at the outcome. It is important to note that during modeling  $Q$  was updated if the participant did not choose a slot machine.

To investigate learning in the non-social task in a similar way, we decided to use a similar model. Nevertheless, we also tested two variants of this reinforcement-learning scheme, including a single- versus a dual-learning rate applied to update the expected value ( $Q$ ) after winning and losing and we compared them using Bayesian group comparison (Daunizeau et al., 2014). For each trial type, only the model presenting the highest exceedance probability with the BIC criterion was analyzed. The results confirmed that the one alpha learning rate is the best model (Figure S5).

## **PET and MRI acquisition performed simultaneously on a Siemens Biograph mMR**

### **PET data acquisition**

PET data were acquired in list-mode, over 90 min. The acquisition started with the intravenous injection of a bolus of [ $^{11}\text{C}$ ]-DASB, a radiotracer that binds SERT. Mean [ $^{11}\text{C}$ ]-DASB injected activity (Mean = 268.3MBq, SEM = 7.3MBq). PET data were submitted to list mode motion correction (Reilhac et al., 2018), then re-binned into 24-time frames (variable length frames, 8 x 15s, 3x60s, 5x120s, 1x300s, 7x600s) for dynamic reconstruction. Images were reconstructed using OP-OSEM 3D incorporating the system point spread function using 3 iterations of 21 subsets. Sinograms were corrected for scatter, random, normalization and attenuation (Mérida et al., 2017). Reconstructions were performed with a zoom of 2 yielding a voxel size of 1.04x1.04x2.08 mm<sup>3</sup> in a matrix of 344 x 344 x 127 voxels. Gaussian post-reconstruction filtering (FWHM=2mm) was applied to PET images.

### **PET preprocessing and kinetic modeling**

Average PET image was computed for coregistration purposes. Anatomical T1 MPRAGE was coregistered (rigid transform) onto the average PET image. Regional labeling of the brain structure was performed with the Hammersmith 83 regions atlas (Gousias et al., 2008; Hammers et al., 2003). It allowed us to extract regional time activity curves based on the subject space, by coregistering the atlas on the subject space and performed extraction. Parametric images of non-displaceable binding potential ( $BP_{ND}$ ) were computed by applying the Simplified Tissue Reference Model (SRTM) (Lammertsma & Hume, 1996) and using cerebellar grey matter as a reference region assumed to be devoid of SERT transporters (Kish et al., 2005). PET images were then



spatially normalized into the standard Montreal Neurological Institute (MNI) atlas space using DARTEL (diffeomorphic anatomical registration through exponentiated lie algebra) toolbox procedure, using the T1 SPM template and resulting in voxels of 2 x 2 x 2 mm (Ashburner, 2007).

### **MRI data acquisition**

All functional MRI acquisitions were performed using EPI BOLD sequences. Functional scans were performed using the following parameters, single-shot EPI, TR / TE = 2400/34, flip angle 85 °, 52 axial slices interlaced, 2 mm thickness, 2 mm gap, FOV = 192x192x125. Volumes were collected, in an interleaved manner. The first acquisition was performed after stabilization of the signal. Anatomical MRI acquisition consisted of 3D sagittal T1-weighted sequences, repetition time = 2300 ms; echo time = 2.34 ms; flip angle = 8; field of view = 256 mm; voxel size = 1 x 1 x 1 mm<sup>3</sup>. The anatomical volume covered the entire brain using 256 adjacent slices of 1-mm thickness.

### **fMRI data preprocessing**

Image analysis was performed using SPM12 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK, [fil.ion.ucl.ac.uk/spm/software/spm12/](http://fil.ion.ucl.ac.uk/spm/software/spm12/)). Time-series images were registered in a 3D space to minimize any effect that could result from participant head-motion. Once DICOMs were imported, functional scans were corrected for slice timing, realigned to the first volume and corrected for motion displacement. Structural images were previously co-registered on the average Dynamic PET image computed. This procedure ensured that functional images were in the same space as PET images. Finally, to perform group and individual comparisons, EPI Images were co-registered with structural maps and spatially normalized into the standard Montreal Neurological Institute (MNI) atlas space using DARTEL toolbox procedure, using the T1 SPM template (Ashburner, 2007). Images were then spatially smoothed with an 8 mm isotropic full-width at half-maximum (FWHM) Gaussian kernel using the standard procedures in SPM12. Note that all the following general linear models (GLM) described also included motion parameters as regressors of no interest, and two session constants (representing our four runs and accounting for its effect) were added for GLM of the social task.

### **Encoding of the expected value of social victories obtained by choosing a particular opponent**

A first GLM (GLM1), allowed us to investigate the brain regions encoding the expected value of social victories ( $Q(t)$ ) (i.e. choice value) computed by summing  $\alpha PE(t)$  and  $Q(t-1)$ . This chosen value represents the expected value of the option during the next trial that will guide the choice for future decisions. To do so, GLM1 included one categorical boxcar regressor of interest representing the outcomes phase (victory or defeat) with a fixed duration of 2 s.  $Q(t)$  was added as a parametric regressor to this categorical onset. This parametric modulator was previously normalized using the Fisher z-score transformation. In addition to this regressor of interest, GLM1 also included three others regressors. The first denoted the choice stage which was parametrically modulated by the difficulty of the choice (computed as  $1 - [Q(i)(t-1) \text{ chosen opponent} - Q(j)(t-1) \text{ unchosen opponent}]$ ) and the decision reaction time, previously normalized using the Fisher z-score transformation. It was modeled as a boxcar function with the duration of the choice. The second regressor represented the confidence rating. It was modeled as a boxcar function with the duration of the rating. The last regressors represented the perceptual decision competition and was parametrically modulated by the accuracy and the reaction time. It was modeled as a boxcar function with the duration of the RT to respond to the perceptual decision competition. In addition, two regressors of no-interest were included and denoted both the miss

choice and the miss competition as separate regressors (representing both the no choice and the no response for the PDM).

### **Dissociating brain representations of the expected value of social victories and prediction error**

To investigate the relative variance explained by both the prediction error  $PE(t)$  and the  $Q(t-1)$  for updating the new expected value of social victory  $Q(t)$ , we created GLM2. GLM2 included one categorical boxcar regressor of interest representing the outcome phase with a fixed duration of 2 s. The expected value of social victories  $Q(t-1)$  and the competitive prediction errors  $PE(t)$  computed by the reinforcement learning algorithm were respectively added as parametric modulators for outcomes. The orthogonalization procedure was disabled to give equal "weight" to each of the parametric modulators ( $Q(t-1)$  and  $PE(t)$ ) related to the outcome phase and to let them compete to explain the variance. These parametric modulators were previously normalized using a z-score transformation. In addition to this regressor of interest and as for the GLM1, GLM2 also included three others regressors. The first represented the choice onsets which were parametrically modulated by the difficulty of the choice (computed as  $1-[Q(i) \text{ opponent} - Q(j) \text{ unchosen opponent}]$ ) and the decision reaction time, modeled as a boxcar function with the duration of the choice. The second regressor denoted the confidence rating. It was modeled as a boxcar function with the duration of the rating. The last regressor denoted the perceptual decision competition, parametrically modulated by the accuracy and the reaction time. It was modeled with a boxcar function with the duration of the perceptual competition. In addition, two regressors were included to denote both miss choice and miss competition as separate regressors at the end of the GLM2. Note that before entering the  $Q(t-1)$  and the  $PE(t)$  into GLM2, we controlled for the correlation between the two parameters. Results revealed inverse correlations between  $Q(t-1)$  and the  $PE(t)$  at the group level (mean  $p = 0.094$ , mean  $r = -0.368 \pm 0.03 \text{ SEM}$ ).

### **Social victories and defeats**

Another GLM (GLM3) was created to investigate the differences between victories and defeats in the social competition task. GLM3 included two regressors of interests, one denoted victories, the other defeats. These regressors were modeled using a boxcar function, with a fixed duration of 2 s. They were parametrically modulated by their respective prediction error. These parametric modulators were previously normalized using the Fisher z-score transformation. GLM3 also included regressors of no-interest to control for the effect of other stages of the task. The first regressors of no-interest denoted the choice stage. It was modeled using a boxcar function with a duration of the choice, and was parametrically modulated by the difficulty of the choice, the value of the chosen opponent for this trial, and the reaction time. The second regressor of no-interest denoted the confidence rating. It was modeled as a boxcar function, with the duration of the rating. The last regressor denoted the perceptual decision competition, parametrically modulated by the accuracy and the reaction time, it was modeled with a boxcar function with the duration of the competition. Finally, two last regressors of no-interest were included to denote both miss choice and miss competition as separate regressors at the end of GLM3. We computed the contrast [Victory>Defeat] at the single subject level and performed a one-sample t-test at the group level to reveal regions that are differently activated by victories compared to defeats.

### **Non-social learning task**

GLM4 was constructed for the non-social reinforcement learning task. It was built similarly to GLM1 except that there was no regressor encoding competition. The same procedure was used for the parametric modulators. First,  $Q(t-1)$  was normalized using the Fisher transformation,

and then entered it as modulator of the outcomes categorical regressor. Similarly, GLM5 was constructed for the non-social reinforcement learning task. GLM5 was constructed in the same way as GLM2 except that there was no regressor encoding competition in this task. The same procedure was used for the parametric modulators. First,  $Q(t-1)$  and  $PE(t)$  was normalized using the Fisher transformation, and then entered as modulators of the outcomes categorical regressor. The orthogonalization procedure was disabled to give equal "weight" to the parametric modulators and allow them to compete to explain the variance.

All GLM models included a high-pass filter to remove low-frequency artifacts from the data (cut-off = 128 s) as well as a run-specific intercept and 6 motion parameters estimated from the realignment step, in order to covary out potential movement-related artifacts in the BOLD signal. Temporal autocorrelation was modeled using an AR(1) process. Regressors of interest were convolved with the canonical hemodynamic response function (HRF) using a boxcar lasting the duration of the visual stimulus associated with each regressor.

### **ROIs definition**

We decided to study the relationship between the SERT level and the BOLD signal related to both the expected value  $Q(t)$  and the prediction error  $PE(t)$ . As we were particularly interested in the VS, we defined two ROIs using an anatomical definition of the nucleus accumbens based on MNI space from the Automated Anatomical Labeling atlas (Rolls et al., 2020). The search volume was defined by a ROI of the left VS and another of the right VS. We also used the DRN definition from the same atlas to extract the SERT level from the DRN. Before any extraction, ROIs were resampled in SPM12 to match the size of the voxels and images of the MRI and PET acquisitions (already co-registered together). Extraction of the BOLD signal was conducted on the single level subject estimated signal within the ROIs defined and using MarsBaR toolbox for MATLAB (Brett, Anton, Valabregue, & Poline, 2002).

### **Behavioral scales**

At the end of the experiment, participants completed a series of questionnaires aimed at assessing different aspects of personality. To assess anxious temperament, they completed the Spielberger trait and state anxiety scale (Y-T and Y-S version) (Kvaal et al., 2005). A distinction is made between the trait (YT), which is a general temperament, and the state (YS), which is more variable over time and corresponds to the person's current temperament. To measure depressive temperament, participants completed the BECK scale (Beck et al., 1960). Participants also completed the BIS-BAS questionnaire (Carver & White, 2005), which assesses two general motivational systems underlying behavior: the behavioral inhibition scale (BIS) and the behavioral approach system (BAS). Finally, a scale to assess the self-assertiveness and social orientation of individuals was also completed. Social assertiveness was assessed using the assertiveness questionnaire (Rathus, 1973) and social orientation was assessed using the social dominance orientation scale (Pratto et al., 1994). All demographic data are summarized in the supplementary data (**table S5**).

### **Behavioral analysis**

For the behavioral analysis, we excluded trials where a participant did not make a choice at the time of selection of the opponent or of the slot machine. These trials represent less than 5% of all trials and the results remain similar even when they are included in the analyses. Such trials were considered as missed choices.

All statistical analyses were performed using SPSS v21.0 (SPSS Inc., Chicago, IL, USA). Normal distribution was assessed with a Shapiro-Wilk test and histograms plots. If data distribution was not normal, we performed a Friedman test, otherwise a repeated measure ANOVA was conducted. Then, we ensured that homoscedasticity of variances was respected using a Mauchly test. If not, we applied a Greenhouse-Geisser correction to our ANOVA. For multiple comparisons, a *post-hoc* comparison (with Bonferroni correction) was conducted according to the previous test used.

Concerning the correlation analyses, a Shapiro-Wilk test and histogram plots were used to assess the normal distribution. If data were normal, we performed a Pearson correlation, otherwise, Spearman correlation was conducted to investigate the correlation between the SERT level from the DRN and the VS bold signal related to the chosen value and prediction error. In addition, correlation coefficients from the social competition task and from the non-social task were compared using a single side test of correlation comparison (Eid et al., 2011).

For the mixed effect linear regression explaining the confidence rating, we included several explanatory variables: the trial number, the reward probability (represented by the opponent or slot machine selected), the reaction time of the choice, the BP<sub>ND</sub> of the DRN, the block number and the task condition (social and non-social). The trial was coded as the trial number when the confidence rating was requested during the task (approximated with a continuous variable). The reward probability categories were coded for both social and non-social task as 3 = superior or worst, 2 = intermediate or middle and 1 = inferior opponent or best (nominal variable). The logarithm of the reaction time was for each decision (opponent or slot machine) selection ( $0.23 \pm 0.37$ s, continuous variable). The BP<sub>ND</sub> was the average level of SERT within the DRN for each participant (continuous variable). The block number was coded as 1 for the first block of social competition and the block of non-social learning and 2 for the second block of social competition (nominal variable). Finally, the task condition was coded as 1 for the social competition task and 2 for the non-social task (nominal variable).

### **Computation of the competitive index**

The competitive index was based on the opponents presented on the screen. For each trial, a competitive value of 1 was assigned to the stronger opponent, based on the predefined strength, and a competitive value of 0 for the weaker. Then, the overall proportion of competitive choices was computed by summing the competitive value in each trial divided by the number of trials played by participants. A value close to 1 represents a highly competitive index, whereas a value close to 0 reflects a non-competitive index.

## Supplementary Materials

### Supplementary results

#### Reaction times

Concerning the reaction time for the opponent selection, we first z-scored all reaction times and then performed a one-way repeated measures ANOVA by pooling results from the two blocks of competition. The results revealed significant differences in the speed of opponent selection ( $F_{(2,52)} = 13.35$ ;  $p < 0.001$ ) (Figures S1B). The *post-hoc* tests revealed that participants were faster when selecting inferior opponents ( $M = -0.12$ ,  $SEM = 0.03$ ) compared to the intermediate ( $M = 0.13$ ,  $SEM = 0.04$ ;  $t_{(31)} = -4.35$ ,  $p < 0.001$ ) and the superior opponent ( $M = 0.17$ ,  $SEM = 0.06$ ;  $t_{(31)} = -3.95$ ,  $p < 0.001$ ). No significant differences were observed between the reaction times to select the intermediate and the superior opponent (paired  $t_{(31)} = -1.29$ ,  $p = 0.205$ ).

Concerning reaction times for the perceptual competition, we first z-scored all the reaction times and then performed a one-way Repeated measured ANOVA by pooling results from the two blocks of competition. The results revealed no significant difference in the speed of decision during the competition ( $F_{(2,52)} = 0.12$ ;  $p = 0.94$ ) (Figures S1C).

#### Selection of opponents in the social hierarchy learning task: control for block effects

During the two independent competition tasks, participants were told to choose with which opponents they wanted to compete with in each trial. The two-way repeated measures ANOVA with a Greenhouse-Geisser correction, including opponent and competition blocks as factors of interest, revealed significant differences in the choice frequency of the participants according to the opponent  $F_{(2,52)} = 7.16$ ;  $p = 0.005$ ) (**Fig S1A**). The *post-hoc* analysis revealed that participants selected the inferior opponent more ( $M = 0.40$ ,  $SEM = 0.02$ ) compared to the intermediate ( $M = 0.29$ ,  $SEM = 0.01$ ,  $t_{(1,29)} = 3.33$ ;  $p = 0.001$ ) and the superior opponent ( $M = 0.27$ ,  $SEM = 0.01$ ;  $t_{(1,29)} = 3.75$ ;  $p = 0.004$ ). No significant differences were revealed between the choice frequency for the intermediate and the superior opponent ( $t_{(1,29)} = 1.12$ ;  $p = 0.22$ ). There was no significant interaction effect between block and opponent types ( $F_{(2,52)} = 0.28$ ;  $p = 0.75$ ) (Figures S1A).

We also performed a similar analysis on RT for the opponent selection, adding the block as a factor of interest into a two-way repeated measure ANOVA. The Greenhouse-Geisser corrected results from this ANOVA, revealed a significant effect of the opponent ( $F_{(2,58)} = 14.76$ ;  $p < 0.001$ ). Post hoc tests shown that during the first block of competition, participants were faster in selecting the inferior opponent ( $M = -0.17$ ,  $SEM = 0.04$ ) compared to the intermediate ( $M = 0.11$ ,  $SEM = 0.06$ ;  $t_{(29)} = -3.56$ ,  $p < 0.001$ ) and the superior opponent ( $M = 0.25$ ,  $SEM = 0.06$ ;  $t_{(31)} = -4.57$ ,  $p < 0.001$ ). They were also faster in selecting the intermediate opponent compare to the superior ( $t_{(29)} = -2.29$ ,  $p = 0.029$ ). During the second block of competition, participants were faster in selecting the inferior opponent ( $M = -0.08$ ,  $SEM = 0.04$ ) compared to the intermediate ( $M = 0.16$ ,  $SEM = 0.2$ ; paired  $t_{(29)} = -3.95$ ,  $p < 0.001$ ). No significant differences were observed between the reaction time for selecting the intermediate opponent compared to the superior opponent during the second block. No effect of block was revealed ( $F_{(2,62)} = 0.073$ ;  $p = 0.930$ ), yet we observed a significant interaction effect between the opponent and the block ( $F_{(2,58)} = 5.32$ ;  $p = 0.01$ ) (**Fig S1B**).

Also, statistical analysis of the reaction times at the perceptual decision making task (PDM) revealed no significant opponent effect ( $F_{(2,58)} = 0.08$ ;  $p = 0.92$ ), and no block effect across the three different opponents ( $F_{(2,58)} = 0.49$ ;  $p = 0.48$ ) or an interaction effect ( $F_{(2,58)} = 0.08$ ;  $p = 0.17$ ) (**Fig S1C**).

### **Dorsal Raphe Nucleus predicts the confidence rating during social and non-social learning**

A mixed-effects linear regression was computed to predict the confidence of victories for each of the three opponents (social) or slot machines (non-social) during the tasks. To do this, we selected potential explanatory variables and ran a stepwise regression comparison. This stepwise regression procedure allows step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. The explanatory variables were: the trial number of the confidence rating (Trial), the task condition (social or non-social learning), the reward probability categories for the opponent or slot machine types, log(RTs), block of social competition (to control for an effect of block) and the BP<sub>ND,DRN</sub>. We compared the three regression models that were generated by this stepwise procedure and selected the one that explained the most variance. A significant effect was found ( $F_{(3,1000)} = 57,16$ ;  $p < 0,001$ ; Durbin-Watson = 1,404), with an  $R^2$  of 0.147. The trial number of the confidence rating, the reward probability categories, the previous victory/reward and the BP<sub>ND</sub> of the DRN were all significant predictors of the level of confidence (**Fig S3B, table S5**, regression model comparison). Participant's predicted confidence rating is equal to:

$$[\text{Confidence rating}] = \varepsilon + \beta_1 * \text{Trial} + \beta_2 * \text{Reward Probability} + \beta_3 * \text{BP}_{\text{ND}} \quad \text{Eq. 6}$$

Where  $\varepsilon$  represents the error term ( $\varepsilon = 56.69 \pm 3.62$ ), and  $\beta_i$  are the estimated parameters of the variables explaining the confidence rating. The results show that larger BP<sub>ND</sub> in the DRN lead to a higher confidence ( $\beta_3 = 2.98 \pm 1.44$ ,  $p < 0.001$ ). Subjects also tend to be more confident over the duration of task ( $\beta_1 = 0.22 \pm 0.02$ ,  $p < 0.001$ ) but as expected this confidence decreased as the probability of winning against the opponent or the slot machine decreased ( $\beta_2 = -6.94 \pm 0.57$ ,  $p < 0.001$ ). The block number, the task condition and the Log(RT) for the opponent selection did not explain the confidence rating ( $p = 0.733$ ,  $p = 0.325$  and  $p = 0.406$  for the block, task condition and RT for the rating respectively). We also plotted confidence rating and winning probabilities to illustrate time variation in confidence ratings for social and non-social task (**Fig S6C**).

### **Score of the reward responsiveness scale correlates the SERT availability in the DRN**

The SERT availability in the DRN and individuals' score on the BIS/BAS personality questionnaire revealed a positive correlation with the reward responsiveness subscale ( $p = 0.019$ ,  $r = 0.427$ ). Higher SERT availability in the DRN was associated with higher reward responsiveness. No correlation between SERT availability and other subscales of the BIS/BAS were observed, and no other relationships were observed with the other personality scales (**Fig 8**).

## **Supplementary discussion**

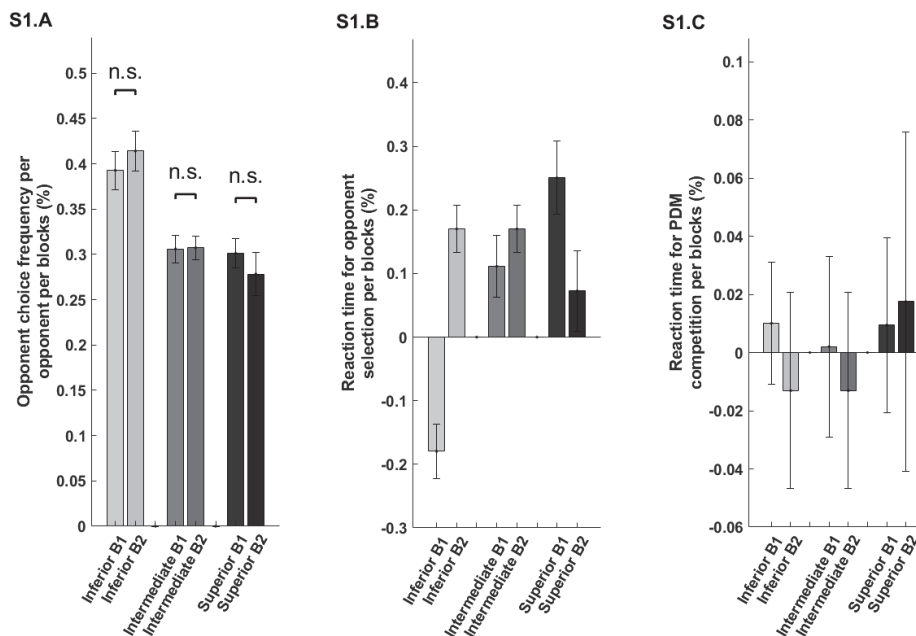
### **Relationship between 5-HT and choice confidence**

Bayesian decision theory proposes that confidence is defined as the belief associated with the proposition that the observer has chosen or intends to choose. More precisely, confidence can be defined as the observer's belief that the chosen action maximizes utility (De Martino et al., 2013). Because organisms can make better decisions if they have a representation of the uncertainty and confidence associated with task-relevant variables (Ma & Jazayeri, 2014), we observed that the confidence rating was modulated by SERT availability, both in the social and non-social learning tasks. Individuals showing high SERT availability were overconfident with respect to their probability of winning, especially concerning the worst option. This result reveals an effect of the serotonergic system on a personal trait. The modulation of the confidence rating

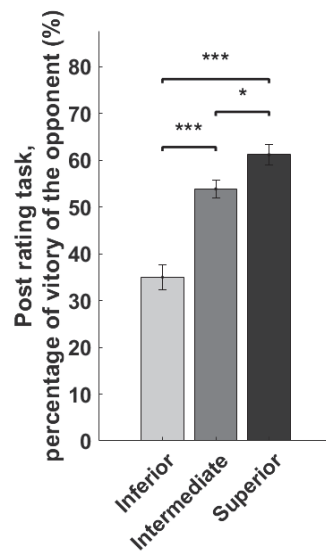
could be accounted for by the fact that Individuals showing high SERT availability selected the strongest opponent less often, and therefore were less informed of the relative strength of the strongest opponents.

Individuals with high SERT availability were also more confident in their choices (regardless of social or non-social context) (**Fig S3B and S3C**) and exhibited higher reward seeking traits (as assessed from BAS, **Fig S6**). Thus, high SERT individuals, who are more attracted to win in general, may seek to select inferior opponents (or the most rewarding slot machines), in agreement with the fact that they show a lower competitive index (**Fig 2D**). These findings indicate that lower 5-HT levels in high SERT individuals could facilitate confidence responses according to existing predispositions. These neurobehavioral findings could also explain why trait anxiety, presumably modulated by 5-HT, leads to differences in competitive confidence under stress (Goette et al., 2015).

### Supplementary figure and table

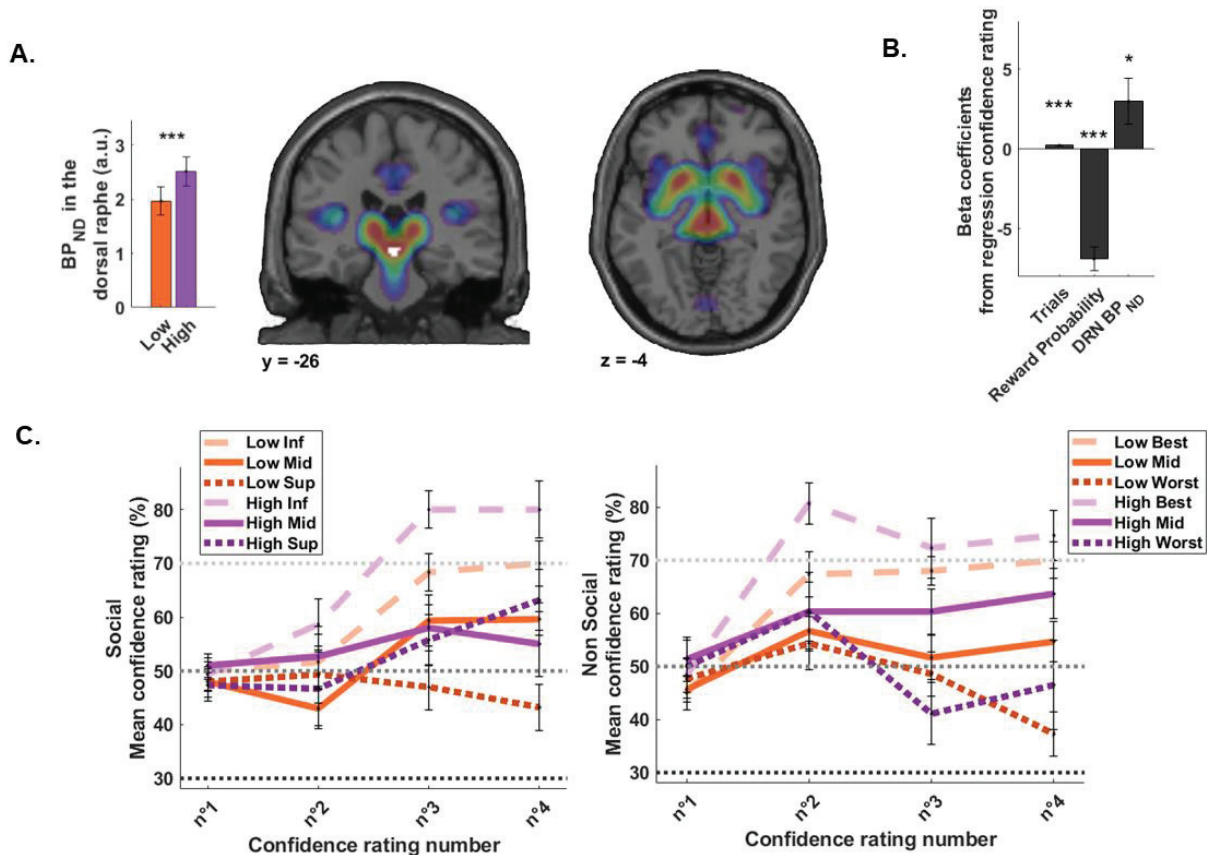


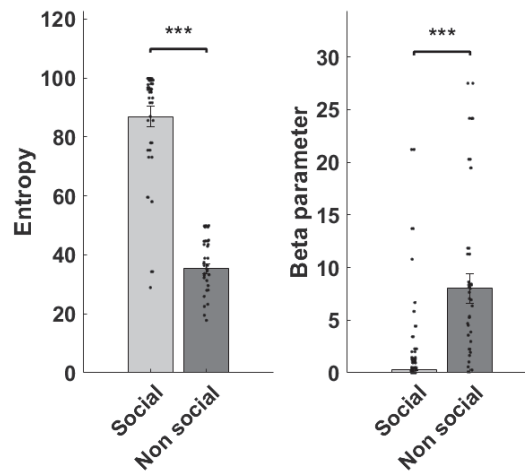
**Figure S1: control for block effect in the social competition task.** Results revealed significant interaction effect between the opponent and the block for the reaction time of the opponent selection ( $F_{(2,58)} = 5.32$ ;  $p = 0.010$ ) (**Fig S1.B**). No differences were observed for the frequency of choice nor for the reaction time at the PDM stage.



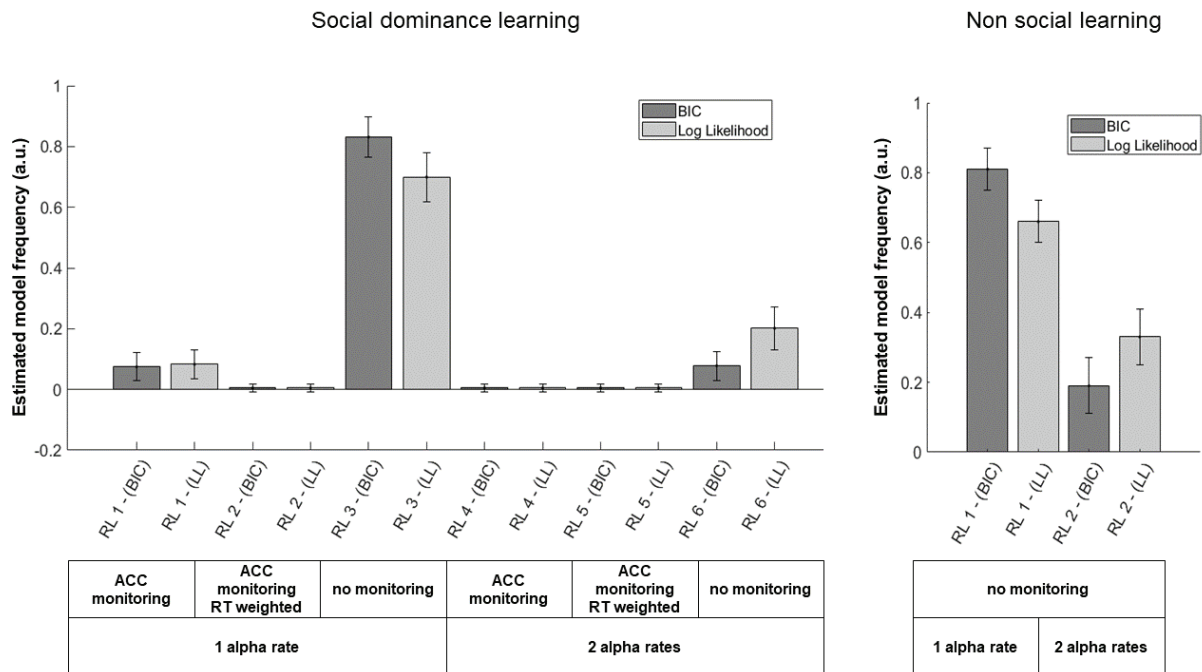
**Figure S2: Post-ratings of the percentage of victories against different types of opponents** (averaged across the two blocks of social competition). Results of the repeated measures ANOVA revealed a significant difference in the evaluations ( $F_{(2,58)} = 4.842$ ,  $p = 0.012$ ). Participants rated the inferior opponents as having less victories ( $M=35.17$   $SEM=2.63$ ) compared to the intermediate opponent ( $M=53.87$   $SEM=1.94$ ) ( $t_{(28)} = -5.88$ ,  $p < 0.001$ ) and the superior opponent ( $M=61.21$   $SEM=2.14$ ) ( $t_{(28)} = -7.09$ ,  $p < 0.001$ ). They also rated the intermediate opponents as having less victories compared to the superior opponent ( $t_{(28)} = -4.14$ ,  $p = 0.048$ ).





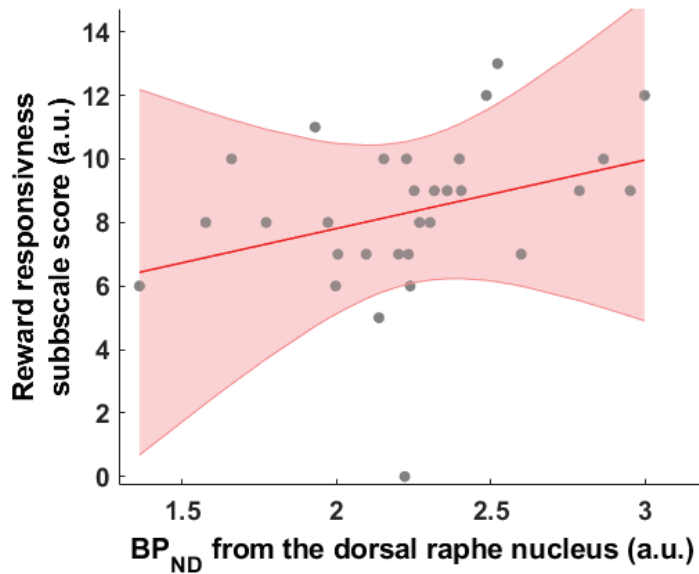


**Figure S4:** Choice entropy and beta parameters comparison between social and non-social task. On the left side the entropy of the choice is plotted. On the right side, the beta parameters estimated from the models. Both bar graph represents the mean value of the metric and dots represent the distribution of the participants. Entropy of the choice was calculate by subtracting Shannon entropy of the chosen option from the Shannon entropy of the unchosen option using the following formula:  $p \cdot \log(p) - (1-p) \cdot \log(1-p)$ .



**Figure S5. Group Bayesian model selection.** On the left side, the estimated model frequency for the model set for the choices made during the social hierarchy learning task. The model comparison indicated that the model with one alpha and no monitoring was the best describing

the data. Light and dark bars represent the estimated frequency of each model in the population using both Bayesian Information Criterion (BIC) and Log likelihood (LL) as comparison metrics, respectively. Note that the model with the highest exceedance probability using the BIC criterion was used. On the right side the comparison for the model set for the non-social learning task is displayed. (see also table S6 and supplemental experimental procedures).



**Figure S6:** Positive correlation between the reward responsiveness subscale of the BAS scale and the SERT availability in the dorsal raphe nucleus ( $p = 0.019$ ,  $r = 0.427$ ). Results are consistent if the participant who had a score of 0 at this subscale is removed ( $p = 0.019$ ,  $r = 0.433$ ).

**Table S1. Table S1. Regions parametrically activated by the Q(t). GLM1**

| Region                                    | MNI coordinates |     |     | k     | Z score |
|---|-----------------|-----|-----|-------|---------|
|   | x               | y   | z   |       |         |
| Ventral striatum R                        | 14              | 14  | -8  | 4450  | 6.17    |
| Ventral striatum L                        | -18             | 14  | -9  |       | 5.5     |
| vmPFC L                                   | -4              | 57  | -4  | 693   | 4.3     |
| dIPFC R                                   | 50              | 32  | 18  | 416   | 4.3     |
| Anterior cingulate cortex                 | -15             | -46 | 38  | 863   | 4.49    |
| Posterior cingulate cortex /<br>precuneus | 6               | -46 | 48  | 424   | 3.96    |
| Inferior temporal gyrus                   | -57             | -56 | -6  | 810   | 4.36    |
| Angular gyrus R                           | 40              | -72 | 46  | 1287  | 4.31    |
| Cerebellum                                | 36              | -74 | -38 | 13068 | 5.74    |

**Table S1. Regions parametrically varying with Q(t). GLM1.** All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected. Labeling of all regions was done using the peak activity and the AAL3 atlas (Rolls et al., 2020).

|                 | Region                    | MNI coordinates |     |       | k    | Z score |
|-----------------|---------------------------|-----------------|-----|-------|------|---------|
|                 |                           | x               | y   | z     |      |         |
| Q(t-1)          | Putamen L                 | -16             | 14  | -6    | 777  | 5.27    |
|                 | Putamen R                 | 16              | 14  | -2    | 1552 | 4.7     |
|                 | vmPFC                     | -4              | 52  | -3    | 689  | 4.3     |
|                 | Temporal middle gyrus R   | 57              | -58 | 2     | 8656 | 5.19    |
| PE(t)           | Caudate R                 | 10              | 18  | 8     | 2768 | 5.93    |
|                 | Caudate L                 | -9              | 14  | 2     | 2279 | 5.85    |
|                 | vmPFC                     | 4               | 62  | 3     | 1638 | 4.66    |
|                 | Superior frontal gyrus R  | 27              | 18  | 51    | 752  | 4.57    |
|                 | Superior frontal gyrus L  | -21             | 21  | 54    | 2508 | 5.74    |
|                 | Middle cingulate gyrus L  | -3              | -39 | 33    | 4442 | 5.73    |
|                 | Angular gyrus R           | 40              | -60 | 28    | 4367 | 5.04    |
|                 | Angular gyrus L           | -44             | -66 | 45    | 5980 | 6.09    |
|                 | Inferior temporal gyrus L | -46             | -51 | -18   | 609  | 4.71    |
| Lingual gyrus R | 12                        | -81             | -9  | 12308 | 5.92 |         |

**Table S2. Regions parametrically varying with Q(t-1) and PE(t). GLM2.** All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected. Labeling of all regions was done using the peak activity and the AAL3 atlas (Rolls et al., 2020).

| Region                  | MNI coordinates |     |    | k    | Z score |
|-------------------------|-----------------|-----|----|------|---------|
|                         | x               | y   | z  |      |         |
| Caudate R               | 15              | 16  | 8  | 1164 | 5.05    |
| Putamen L               | -18             | 12  | -6 | 682  | 4.9     |
| Temporal middle gyrus R | 58              | -58 | 3  | 1288 | 5.13    |
| Temporal Middle gyrus L | -56             | -56 | 2  | 670  | 4.16    |
| Occipital L             | -15             | -94 | 14 | 867  | 4.73    |
| Occipital L (2)         | -15             | -88 | -8 | 878  | 4.41    |

**Table S3. Regions activated by the contrast [Victory>Defeat]. GLM3.** All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected. Labeling of all regions was done using the peak activity and the AAL3 atlas (Rolls et al., 2020).

|                    |                                    | MNI coordinates |     |     |      |         |
|--------------------|------------------------------------|-----------------|-----|-----|------|---------|
| Region, hemisphere |                                    | x               | y   | z   | k    | Z score |
| Q(t) (GLM4)        | vmPFC L                            | -6              | 50  | -10 | 876  | 4.17    |
|                    | Caudate L †                        | -12             | 9   | -2  | 72   | 4.18    |
| PE(t) (GLM5)       | Caudate R                          | 12              | 22  | -4  | 1023 | 4.24    |
|                    | Medial PFC gyrus L                 | -18             | 28  | -16 | 338  | 4.65    |
|                    | Medial PFC gyrus L (2)             | -9              | 51  | -14 | 389  | 3.91    |
|                    | Superior frontal gyrus R           | 9               | 66  | 0   | 395  | 4.24    |
|                    | Medial posterior cingulate gyrus L | -8              | -33 | 39  | 818  | 4.68    |
|                    | Medial OFC gyrus L                 | -18             | 28  | -16 | 338  | 4.65    |
|                    | Superior temporal gyrus R          | 58              | -16 | 2   | 447  | 4.41    |
| Q(t-1) (GLM5)      | vmPFC L *                          | -2              | 39  | -2  | 1053 | 4.71    |

**Table S4. fMRI activity in the non-social task.** GLM4 and GLM5. All statistical analyses were performed at a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.001$  uncorrected otherwise noted. \* denotes a  $p < 0.05$  cluster level corrected for Family Wise Error at the whole brain level, with an initial cluster forming threshold of  $p < 0.005$  uncorrected. † denotes a small volume correction. Labeling of all regions was done using the peak activity and the AAL3 atlas (Rolls et al., 2020).

| Model   | R Square Change | F Change | Unstandardized Coefficients |            | Standardized Coefficients | t      |      | Sig.      | Collinearity Statistics |
|---|-----------------|----------|-----------------------------|------------|---------------------------|--------|------|-----------|-------------------------|
|   |                 |          | B                           | Std. Error | Beta                      |        |      | Tolerance | VIF                     |
| 1 (Constant)<br>Reward probability                                  | 0.073           | 78.598   | 69.821                      | 1.570      |                           | 44.468 | .000 |           |                         |
|   |                 |          | -6.879                      | .776       | -.270                     | -8.866 | .000 | 1.000     | 1.000                   |
| 2 (Constant)<br>Reward probability<br>Trial                         | 0.07            | 81.715   | 63.318                      | 1.673      |                           | 37.851 | .000 |           |                         |
|   |                 |          | -6.912                      | .746       | -.271                     | -9.261 | .000 | 1.000     | 1.000                   |
|   |                 |          | .217                        | .024       | .265                      | 9.040  | .000 | 1.000     | 1.000                   |
| 3 (Constant)<br>Reward probability<br>Trial<br>BP <sub>ND</sub> DRN | 0.004           | 4.262    | 56.687                      | 3.620      |                           | 15.658 | .000 |           |                         |
|   |                 |          | -6.935                      | .745       | -.272                     | -9.306 | .000 | 1.000     | 1.000                   |
|   |                 |          | .217                        | .024       | .265                      | 9.062  | .000 | 1.000     | 1.000                   |
|   |                 |          | 2.981                       | 1.444      | .060                      | 2.064  | .039 | 1.000     | 1.000                   |

**Table S5: Stepwise mixed effect regression summary of the confidence rating.**

The R square change represents the significant proportion of variance explained by the addition of the explanatory variable. The F Change represents the statistic associated to the R square. Unstandardized and Standardized beta coefficient for each variable included in the linear model and their p-values associated. The tolerance metric allows verification that the multicollinearity assumption is not violated. If tolerance is around 0.1, it means that at least two explanatory variables are too collinear. We compared three different models using the stepwise procedure. The selected model was the model number 3.

|            | Description | BIC - sum                            | LL - sum | n param | Alpha |                 | Beta            | $\omega$        | Exceedance probability |      |  |
|------------|-------------|--------------------------------------|----------|---------|-------|-----------------|-----------------|-----------------|------------------------|------|--|
|            |             |                                      |          |         | win   | loss            |                 |                 | BIC                    | LL   |  |
| Social     | RL 1        | 1 $\alpha$ - error update            | -3259    | -2958   | 2     | 0.37 $\pm$ 0.21 | 3.15 $\pm$ 5.42 |                 |                        |      |  |
|            | RL 2        | 1 $\alpha$ - 1 $\omega$ error update | -3345    | -2996   | 3     | 0.37 $\pm$ 0.19 | 2.60 $\pm$ 4.64 | 0.6 $\pm$ 0.19  |                        |      |  |
|            | RL 3        | 1 $\alpha$ - no error update         | -3232    | -2931   | 1     | 0.39 $\pm$ 0.22 | 3.17 $\pm$ 5.03 |                 | 0.99                   | 0.99 |  |
|            | RL 4        | 2 $\alpha$ - error update            | -3300    | -2951   | 3     | 0.39 $\pm$ 0.21 | 0.35 $\pm$ 0.24 | 3.29 $\pm$ 5.48 |                        |      |  |
|            | RL 5        | 2 $\alpha$ - 1 $\omega$ error update | -3393    | -2996   | 4     | 0.39 $\pm$ 0.19 | 0.36 $\pm$ 0.22 | 2.85 $\pm$ 5.02 | 0.65 $\pm$ 0.18        |      |  |
|            | RL 6        | 2 $\alpha$ - no error update         | -3284    | -2936   | 2     | 0.38 $\pm$ 0.21 | 0.41 $\pm$ 0.25 | 3.68 $\pm$ 5.5  |                        |      |  |
| Non social | RL 1        | 1 $\alpha$                           | -1482    | -1236   | 2     | 0.27 $\pm$ 0.23 | 8.92 $\pm$ 9.7  |                 | 0.85                   | 0.7  |  |
|            | RL 2        | 2 $\alpha$                           | -1511    | -1228   | 3     | 0.31 $\pm$ 0.2  | 0.30 $\pm$ 0.24 | 6.7 $\pm$ 5.05  | 0.15                   | 0.3  |  |

**Table S6. Model selection and parameters explaining the choices for the social and non-social learning tasks.** Bayesian Information criterion (BIC). Log Likelihood (LL).  $\alpha$  = learning rate.  $\beta$  = inverse temperature.  $\omega$  = sensitivity to response reaction time. (see also the modeling section of the supplementary methods for the detail of the algorithm used for each model).

|                              | Mean  | STD   | SEM  |
|------------------------------|-------|-------|------|
| Age                          | 23.5  | 2.8   | 0.5  |
| BMI                          | 22.07 | 2.64  | 1.33 |
| BECK                         | 7.33  | 7.67  | 1.33 |
| BAS (total)                  | 48.27 | 11.32 | 1.97 |
| drive                        | 10.00 | 2.44  | 0.42 |
| fun seeking                  | 6.97  | 2.43  | 0.42 |
| reward *                     | 8.33  | 2.48  | 0.43 |
| BIS (total)                  | 15.03 | 4.83  | 0.84 |
| fillers                      | 7.93  | 2.15  | 0.37 |
| STAI (total)                 | 94.57 | 5.98  | 1.04 |
| trait                        | 49.43 | 4.39  | 0.76 |
| state                        | 45.13 | 4.01  | 0.70 |
| Rathus                       | 98.57 | 8.48  | 1.48 |
| Social dominance orientation | 63.40 | 12.77 | 2.22 |

**Table S7. Demographic summary and results of the correlation between personality traits and the binding potential of [<sup>11</sup>C]-DASB in the dorsal raphe nucleus.** Results revealed a significant positive correlation between the reward responsiveness subscale and the level of SERT available in the dorsal raphe nucleus ( $p = 0.019$ ,  $r = 0.427$ ), denoted by the Asterix. No other significant correlation was observed.



# Conclusion générale

Nous avons dans l'introduction posé le cadre d'étude de la prise de décision en milieu social. Nous avons notamment revu la littérature concernant les régions cérébrales engagées dans les processus de décision qui sont l'évaluation des options, leur composante émotionnelle ainsi que les régions sensibles à l'aspect social de la prise de décision. Nous avons aussi revu les différentes composantes des interactions sociales en terme d'implication de soi et de l'autre dans l'interaction, mais aussi en terme d'intention. Nous avons étudié comment peut émerger une hiérarchie dans un groupe comme la sommes de ces interactions sociales.

Dans les chapitres 2 et 3 nous étudions comment les individus prennent en compte l'intention de coopérer des autres pour prendre leurs décisions. Ainsi nous considérons dans le chapitre 2 un environnement d'information minimal ne laissant observable que le choix d'un autre et la récompense associée pour le participant. Dans le chapitre 3 nous étudions un environnement impliquant de multiples personnes et où chacun exprime anonymement mais explicitement sa coopération ou sa défection. Dans ces deux situations écologiquement plausibles nous avons vu que les humains traquent la coopérativité de l'autre et adaptent leur stratégie de manière dynamique en fonction de cette coopérativité.

Dans l'expérience une, la manière dont nous avons modélisé le comportement fait implicitement l'hypothèse que le participant adapte sa coopérativité à la coopérativité de l'autre. Nous pouvons faire cette hypothèse étant donné que le contexte d'information minimale ne donne pas un grand sentiment d'agentivité ou de contrôle



de la situation aux participant. C'est en tout cas ce qu'il en est ressortit des entretiens post-expérimentation. Ainsi les participants étaient dans une situation d'adaptation. Il est cependant important de préciser que dans la nature, plutôt qu'être en réaction et en adaptation face à la coopération ou à la compétition de l'autre, les humains peuvent influencer sur celle-ci. C'est notamment ce que l'on observe dans l'expérience du jeu de bien public avec un modèle qui permet d'anticiper les effets de sa propre action sur le maintien de la coopérativité des autres et ainsi pour maximiser les gains dans un plus long terme par exemple.

Le premier travail qu'est le chapitre 2 prouve l'existence au moment de la décision d'une pondération du choix par l'intention de l'autre de coopéré inférée. Il prouve aussi qu'en fonction de l'intention qui est attribuée les récompenses ne provoquent pas la même réaction cérébrale. Le deuxième travail dans le chapitre 3 prouve que dans un groupe anonyme les humains utilisent des inférences bayésiennes, prenant donc en considération l'incertitude, pour modéliser le groupe comme ayant une coopérativité propre dont découle le comportement des individus. De plus, si l'on fait l'hypothèse que les autres individus du groupe font le même raisonnement au sujet de la coopération global du groupe, il est possible d'utiliser notre décision pour influencer le groupe. Ainsi ceci démontre que la théorie de l'esprit (de profondeur deux, c'est à dire lire les intentions des autres mais aussi prendre en compte le fait que nos actions influencent les pensées et actions des autres) prend aussi place dans un large groupe de personne anonyme.

Ces deux travaux sont complémentaires en ce sens que le premier prouve qu'il y a une recherche dynamique de l'intention cachée derrière les actions de l'autre afin d'adapter son comportement à celle-ci. Le deuxième prouve que l'estimation de cette intention peut être étendue à un groupe et qu'elle est utilisée afin de faire évoluer la coopérativité moyenne du groupe et de maximiser (ou minimiser) ses propres profits (ou ses propres pertes).

Connaissant les intentions des autres ou tout du moins ayant une appréciation de celles-ci, il est aussi nécessaire pour bien interagir socialement de connaître les capacités

d'actions de ces individus. Cette capacité d'interagir avec son environnement est aussi appelée "*agentivité*". Un bon prototype de l'*agentivité* est ce que l'on appelle la hiérarchie. C'est pourquoi nous avons étudié dans les chapitres 4 et 5 l'apprentissage de la hiérarchie pour indirectement analyser les mécanismes d'attribution d'*agentivité* aux autres.

Ainsi l'étude dans le chapitre 4 a permis de mettre en évidence le rôle d'une hormone clé, la sérotonine, dans la variabilité interindividuel lors de l'apprentissage d'une hiérarchie par interaction directe. Nous avons dans ce travail prouvé que la sérotonine permet l'encodage de la valeur d'un opposant comme l'intégration des victoires et défaites passées. Ainsi le taux de sérotonine module la vitesse d'apprentissage de la hiérarchie, ou tout du moins, elle module la réaction des individus face à des personnes de différents niveaux hiérarchiques. Dans le chapitre 5 nous montrons qu'il est possible de moduler l'apprentissage par observation d'une hiérarchie sociale avec une stimulation transcranienne du cortex préfrontal médian. Ainsi nous avons mis en évidence la différence entre l'apprentissage d'une hiérarchie sociale d'une hiérarchie non social. De plus, la modélisation nous permet d'avancer l'hypothèse que le cortex préfrontal médian module le maintien de l'information concernant la hiérarchie sociale, mais qu'il n'est pas impliqué dans l'utilisation transitive des connaissances acquises.

Nous remarquons finalement que nous avons appliqué l'approche par les modèles dans tout ces travaux, mais pas toujours dans le même objectif. Dans les chapitres 2 et 4 nous les avons principalement utilisé pour générer des variables qui selon la théorie sont sous-jacentes à la prise de décision et nous avons recherché comment elles étaient utilisées par le cerveau. Dans les chapitres 3 et 5, l'objectif principal de l'utilisation des modèles était de comprendre quelle stratégie cognitive était appliqué dans des comportements observés. Chaque fois donc, bien qu'introduisant un biais, l'utilisation des modèles permet de tirer des inférences plus puissantes des mêmes observations que le permettent les analyses statistiques classiques.



# Appendix 1 - Perturbation of Right Dorsolateral Prefrontal Cortex (rDLPFC) Makes Power-Holders Less Resistant to Tempting Bribe

Yang Hu<sup>1,2</sup>, Rémi Phillippe<sup>2,†</sup>, Valentin Guigon<sup>2,†</sup>, Sasa Zhao<sup>2,†</sup>, Edmund Derrington<sup>2,3</sup>,  
Brice Corgnet<sup>4</sup>, Xiaolin Zhou<sup>1</sup>, Jean-Claude Dreher<sup>2,3</sup>

<sup>1</sup>School of Psychological and Cognitive Sciences, Peking University, Beijing, China

<sup>2</sup>Neuroeconomics, Reward and Decision Making Laboratory, Institut des Sciences Cognitives Marc Jeannerod, CNRS, France <sup>3</sup>Université Claude Bernard Lyon 1, Lyon, France

<sup>4</sup>EmLyon, Ecully, France

† These authors equally contributed to this study.

---

<sup>1</sup>Mon travail a essentiellement consisté en la mise en place du protocole, l'acquisition des données ainsi qu'à aider à l'analyse des données et à la rédaction.

## **Perturbation of Right Dorsolateral Prefrontal Cortex (rDLPFC) Makes Power-Holders Less Resistant to Tempting Bribes**

### **Abstract**

Bribery is a common form of corruption that takes place when a briber suborns a power-holder to achieve an advantageous outcome at a cost of moral transgression. While bribery has been extensively investigated in behavioral science, its underlying neurobiological basis remains poorly understood. Here we employed transcranial direct current stimulation (tDCS) in combination with a novel paradigm to investigate whether and how disruption of right dorsolateral prefrontal cortex (rDLPFC) causally changed bribe-taking decisions of a power-holder. Perturbing rDLPFC via tDCS specifically made participants more willing to take bribes when the offer proportion ramped up. This tDCS-induced effect via rDLPFC on corrupt behaviors could not be explained by changes in other affective and cognitive measures. Computational modelling analyses further unveiled a causal link between the disruption of the rDLPFC and a decrease in advantageous inequity aversion specific to bribes. These findings reveal a causal role of rDLPFC in modulating corrupt behavior.

### **Statement of Relevance**

Corruption is one of the most pervasive and complex social problems, As one of the common forms, bribery often occurs in interpersonal contexts when a briber suborns a power-holder who can exert an impact on the briber's interest. While confronted with a bribe, a power-holder needs to weigh personal profits and moral costs in determining whether to take the bribe or not. Combining transcranial direct current stimulation (tDCS) with a novel task, we pinpointed the causal role of the right dorsolateral prefrontal cortex (rDLPFC) in modulating the bribe-taking behaviors of a power-holder and the underlying computational process. In particular, disrupting rDLPFC via tDCS specifically made power-holder more willing to accept bribes with increasing proportion, putatively through reduced advantageous inequity aversion to bribes. These findings provide insights for the neurobiological roots of corruption and indicate interventions to modify corrupt behaviors using non-invasive brain stimulation techniques.

## Introduction

As one of the most common forms of corruption, bribery pervasively exists in governments, enterprises and other organizations all over the world (Dreher, Kotsogiannis, & McCorriston, 2007). Aside from purely ethical concerns, it is costly with respect to economics (Mauro, 1995; Shleifer & Vishny, 1993), and brings severe societal consequences such as aggravating income inequality and poverty (Gupta, Davoodi, & Alonso-Terme). In real life, bribes usually occur in interpersonal contexts where there is an asymmetry in power between the parties involved, such as an official with entrusted power (hereafter referred to as the ‘power-holder’) who can exert an impact on the briber’s interest (Köbis, van Prooijen, Righetti, & Van Lange, 2016). Bribes often result in mutual benefits (or advantageous outcomes) via collaboration between two parties involved, but transgress moral principles (e.g. justice and honesty), and even break legal rules. For instance, a company might evade taxes by bribing a tax officer who is in charge of the financial audit. Despite the fact that bribery and its determinants have been widely investigated in the social sciences in past decades using survey-based methods (Martin, Cullen, Johnson, & Parboteeah, 2007; Treisman, 2000) and incentivized economic experiments (Abbink, 2006; Serra & Wantchekon, 2012), the neurobiological roots of bribery and their underlying computation remain largely elusive.

Here, we tested whether transcranial direct current stimulation (tDCS) over the right dorsolateral prefrontal cortex (rDLPFC) can have a causal influence in determining whether a power-holder would accept a bribe or not. According to the framework of value-based decision-making (Rangel et al., 2008) and the account of social preference (Fehr & Krajbich, 2014), a power-holder is supposed to pit material self-interest against the briber’s profit (e.g., fairness concern), along with moral principles so that to reach a final decision. We focused on rDLPFC because several studies implementing non-invasive stimulation techniques have uncovered a crucial role of rDLPFC in evaluating the trade-off between personal profits and other’s welfare (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Ruff, Ugazio, & Fehr, 2013; Speitel, Traut-Mattausch, & Jonas, 2019; Strang et al., 2014) and moral values (Maréchal, Cohn, Ugazio, & Ruff, 2017; Zhu et al., 2014), which is critically involved in the computational processes underlying corrupt decision-making.

To examine this core hypothesis, we performed a tDCS study in which a total of 120 healthy participants were randomly assigned to one of three tDCS groups to causally modulate (i.e., anodal or cathodal tDCS), or maintain (i.e., sham tDCS) the neural excitability of rDLPFC (see **Figure 1 and S1**). Corrupt behaviors of a power-holder were measured using a novel experimental paradigm. That is, participants played the role of a power-holder, who decides whether a (fictitious) proposer would earn a given amount of money or not in a game of chance in which one of two different payoffs was randomly determined by the computer. Depending on the payoff indicated by the computer, the proposer could obtain a larger profit by either

telling a lie (i.e., the bribe condition) or reporting the truth (i.e., the control condition). To achieve this, the proposer offered an amount of money from this larger profit to bribe the power-holder (i.e. participant), with the offer proportion ranging from 10% to 90%. The task for the participants, as power-holders, was to decide whether to accept or reject offers made by the proposer. If accepted, both the proposer and the power-holder would profit from the offer, whereas neither would earn any money if the power-holder rejected the offer (see **Figure 2**). Crucially, bribery was operationally defined as accepting an offer from a proposer who had cheated by reporting a more advantageous offer than the one they should have reported (as indicated by a computer). Thus, the proposer in the bribe condition can be regarded as a briber. Since there was no other financial cost of taking the bribe, the only motivation for power-holders to reject offers, in addition to their notion of fairness, was their ethical concern of colluding with a briber. Notably, the moral cost of taking the bribe critically distinguishes from the psychological cost of dishonesty which have been well documented in the literature (Fischbacher & Föllmi-Heusi, 2013; Gneezy, Kajackaite, & Sobel, 2018; Shalvi, Dana, Handgraaf, & De Dreu, 2011). Importantly, in these studies, participants decided whether to lie for personal profits, whereas, here, they determined whether to collude with another dishonest individual by accepting a bribe to jointly benefit.

Based on previous studies that revealed a critical role of moral cost on ethical decision-making (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Gneezy et al., 2018; Shalvi et al., 2011), we hypothesized that participants would be generally less willing to accept the offers in the bribe (vs. control) condition. More importantly, according to the neural literature mentioned above (Knoch, Pascual-Leone, et al., 2006; Maréchal et al., 2017; Speitel et al., 2019), we expected that disrupting the rDLPFC would alter the acceptance of offers in the bribe (vs. control) condition and that such modulation would critically depend upon offer proportions between the briber and the power-holder (i.e. participant). Specifically, we hypothesized that compared with the sham group, participants receiving cathodal tDCS over the rDLPFC would be more likely to accept offers in the bribe condition, especially when larger offers were proposed, whereas anodal tDCS over rDLPFC would render participants more resistant to such tempting bribes. To further quantify the computational process underlying such decisions, we adopted the Fehr-Schmidt model (Fehr & Schmidt, 1999), a widely-used model to formally characterize other-regarding preferences, to estimate the degree of inequity aversion for both advantageous (i.e., offers that exceed 50% of the large profits) and disadvantageous (i.e., offers that are smaller than 50% of the large profits) domains. Finally, we also explored whether tDCS-induced modulation of bribery behaviors and its relevant computation was susceptible to individual differences in related personality traits, such as empathic concern and immoral preferences.

## Methods

### Participants

One-hundred and twenty French-speaking students from University of Lyon I and local residents who lived nearby (54 females; mean age:  $22.4 \pm 4.4$  years) were recruited via online advertisements. The sample size was adopted based on previous tDCS studies in similar topics (Maréchal et al., 2017; Ruff et al., 2013). All participants were psychiatrically and neurologically healthy and were not taking any medication, as confirmed by a standardized clinical screening. The tDCS study was performed at the Institute of Cognitive Science Marc Jeannerod and was approved by the local ethics committees. All experimental protocols and procedures were conducted in accordance with the IRB guidelines for experimental testing and were in compliance with the latest revision of the Declaration of Helsinki (BMJ 1991; 302: 1194).

### Task and Design

Participants were randomly assigned to one of the three tDCS treatment conditions with 40 persons in each: (i) anodal stimulation (18 females; mean age:  $22.6 \pm 5.5$  years), (ii) cathodal stimulation over the right DLPFC (17 females; mean age:  $21.9 \pm 2.6$  years), or (iii) sham stimulation (19 females; mean age:  $22.6 \pm 4.8$  years), which were unbeknownst to them (For tDCS protocol, see **Supplementary Information**; for visualization of the simulated tDCS effect over rDLPFC, see **Figure 1**).

The main experiment included a computerized incentive task and a follow-up paper-and-pencil rating task, which lasted around 30 min in total (For procedure details, see **Supplementary Information**). In the computerized task, participants were assigned the role of the power-holder who decides to accept or reject financial offers (see **Figure 2A**). In a cover story, they were informed that they would be presented with a series of choices from an independent group, whose data were previously collected by the experimenter. Specifically, participants were led to believe that this independent group of online attendants (for the sake of clarity and convenience, we call them proposers hereafter) played a “game of chance”. This independent group did not actually exist and the choices made by this group were pre-determined by the task software (see below for details). Each proposer was presented with two options that would earn them different payoffs. The larger payoff ranged from 60 to 130 (in €) and the smaller payoff was fixed at 5 (see details of the payoff matrix below). One of the two payoffs was randomly indicated by the computer as the one to be received. According to



the rules of the game, the proposer should report the option indicated, which determined his final payoff. However, the response of the proposer was never checked by the experimenters. This allowed the proposer to lie sometimes by reporting the alternative option that had not been indicated when this would earn them more profit. Importantly, participants were told that each proposer had been informed that whether or not they obtained the payoff of the reported option crucially depended on the decisions of a power-holder (i.e., the participants themselves). Therefore, to obtain the profits in the reported option, the proposer could “share” a portion of the money from their potential gain (i.e., the payoff in the reported option) to influence the power-holder. The task for the power-holder was to decide whether to accept or reject the offer given the information above (i.e. the two potential payoffs, the option indicated by the computer, the reported option, and the offered bribe). If the power-holder accepted the offer, both of them would benefit from the payoff. If the power-holder rejected the offer, neither of them earned anything. Participants were informed that one of their decisions would be randomly selected for payment in that trial at the end of the experiment.

Several aspects of this task merit additional note. First, previous studies have shown that the wording of instructions can produce framing effects (Abbink & Hennig-Schmidt, 2006; Banerjee, 2016) so we adopted neutral wording (e.g., “persuade” instead of “bribe” and “corrupt”) in the instruction. Second, participants were informed that each decision was independent and was matched with different proposers to avoid possible learning effects or strategic responses. Third, each participant was always paid €30 at the end, as required by the ethics approval board. We disclosed this fact to participants during the post-study debriefing when we paid them. Finally, we designed the task such that the proposer always reported the option with a larger payoff, and his/her personal profits after “sharing” with the power-holder were always more than the €5 option. This feature ensured that selfish motivation was the only source that drove the proposer to cheat for a higher payoff, and ruled out other motivations perceived by participants that might influence their subsequent behaviors.

We implemented a  $3 \times 2$  mixed design by manipulating the *tDCS treatment* (i.e., a between-subject factor) and the *task condition* (i.e., a within-subject factor). Crucially, we operationally defined corrupt behaviors as the acceptance of offers proposed by the proposer only when the proposer lied (i.e., the bribe condition). Compared with the other condition where the proposer earned more via telling the truth (i.e., the control condition), accepting offers in the bribe condition incurred the moral cost of colluding with the proposer’s dishonesty. Importantly, we manipulated the *offer proportion*, which was defined as the proportion of the amount the proposer decided to share with the power-holder from the payoff the proposer

would have earned in the reported option, ranging from 10% to 90% (in steps of 10%; 9 levels). This allowed us to investigate whether and how the degree of temptation of a bribe modulated corrupt behavior. To further increase the variance of offers in order to facilitate the model-based analyses, we orthogonalized offer proportions and potential gains that could be earned by the proposer (i.e., the larger payoff, which ranged from 60 to 130 in steps of 10; 8 levels). As a result, this yielded 72 trials, each involving a unique offer, which appeared once in each condition.

Each trial began with a screen displaying two payoff options in the “game of chance”, the computer’s choice (indicated by a computer icon), the proposer’s report (indicated by a blue arrow) together with the identity of the proposer (indicated by initials of the name), and the proposer’s offer. Participants were asked to decide whether to accept or reject the offer by pressing relevant buttons with either left or right index finger at their own pace. A yellow bar appeared below the corresponding option for 0.5 s once the decision was made. Each trial ended up with an inter-trial interval of random duration (i.e., 1 ~ 2 s; see **Figure 2B**) showing a fixation cross. The positions of payoff options were randomized within participants and those of the decision options (i.e., accept or reject) were counterbalanced across participants. All stimuli were presented using Presentation v14 (Neurobehavioral Systems Inc., Albany, CA, USA).

The follow-up rating task aimed to measure the overall subjective feelings of participants about the task and evaluations of behaviors of either proposers or themselves by means of a Likert scale (0 indicated none, 100 indicated very much). In particular, they indicated the degree of 1) moral inappropriateness of the proposers’ behaviors and their decisions (had they accepted offers), 2) moral conflict during the decision period, 3) the guilt they felt (had they accepted offers) in each condition. They also reported the degree to which they had a power advantage over proposers and they perceived offers from proposers as bribes.

## **Data Analyses**

One participant in the cathodal group was excluded for having incomplete data recording due to technical issues, thus leaving a total of 119 participants whose data were further analyzed (overall: 54 females; mean age  $\pm$  SD = 22.4  $\pm$  4.5 years; anodal group: 18 females; mean age  $\pm$  SD = 22.6  $\pm$  5.5 years; cathodal group: 17 females; mean age  $\pm$  SD = 22.0  $\pm$  2.5 years; sham group: 19 females; mean age  $\pm$  SD = 22.6  $\pm$  4.8 years). Overall, participants did

not report any uncomfortable feeling after the experiment and were not able to correctly identify the treatment they were assigned ( $\chi^2(1) = 1.89, p = 0.169$ ). Since no difference in age ( $F(2, 116) = 0.26, p = 0.775$ ) and gender ( $\chi^2(2) = 0.13, p = 0.939$ ) was observed between tDCS groups, we did not include these variables as covariates for later analyses.

Behavioral analyses were conducted using R (<http://www.r-project.org/>) and relevant packages (R Core Team, 2014). Model-based analyses were performed using the “hBayesDM” package (Ahn, Haines, & Zhang, 2017). All reported  $p$  values are two-tailed and  $p < 0.05$  was considered statistically significant (see **Supplementary Information** for details).

## Results

### tDCS over rDLPFC increased the probability of accepting bribes with higher offer proportions

We first tested our main hypothesis regarding choice behavior. Using mixed-effect logistic regression, we observed that participants were less likely to accept an offer in the bribe (vs. control) condition (a main effect of *task condition*:  $\chi^2(1) = 126.94$ ,  $p < 0.001$ ) and more likely to do so when the offer proportion increased (a main effect of offer proportion:  $\chi^2(1) = 96.34$ ,  $p < 0.001$ ). We also detected a significant two-way interaction between *task condition* and *offer proportion* ( $\chi^2(1) = 33.05$ ,  $p < 0.001$ ). *Post-hoc* analyses indicated that compared with the control condition, participants were more likely to accept offers when the offer proportion increased in the bribe condition ( $z = 5.41$ ,  $p < 0.001$ ).

More importantly, we found a significant three-way interaction between *tDCS group*, *task condition*, and *offer proportion* with respect to whether the offer was accepted ( $\chi^2(2) = 8.04$ ,  $p = 0.018$ ; see **Figure 3**). To follow up the three-way interaction, we performed *post-hoc* analyses on choice for each tDCS group that incorporated *task condition*, *offer proportion*, and their interaction as fixed-effect predictors. As a result, compared with the control condition, participants receiving either type of tDCS stimulation were more likely to accept offers when the offer proportion increased in the bribe condition (anodal:  $z = 4.67$ ,  $p < 0.001$ ; cathodal:  $z = 4.34$ ,  $p < 0.001$ ), which was not the case in the sham group ( $z = 0.67$ ,  $p = 0.501$ ; see **Table S1** for details of the regression output).

Notably, we did not observe any tDCS main effect or related interaction on a series of other behavioral measures, including decision time (DT), task-related subjective ratings, and task-irrelevant measures (see **Supplementary Information** for details).

### tDCS over rDLPFC reduced advantageous inequity aversion when taking bribes

Having shown that tDCS modulated bribe acceptance, we next performed a model-based analysis on choice behavior to understand how tDCS over the right DLPFC altered the computations underlying corrupt behaviors. To this end, we adopted the Fehr-Schmidt model to fit choice behaviors because it was designed to delineate how an individual weighs payoff inequality between oneself and the other person depending on task conditions, defined as follows:

$$SV(p_P, p_{PH}) = p_{PH} - \alpha \max(p_P - p_{PH}, 0) - \beta \max(p_{PH} - p_P, 0)$$
$$\alpha, \beta = \begin{cases} \alpha_{control}, \beta_{control} & \text{if the control condition} \\ \alpha_{bribe}, \beta_{bribe} & \text{if the bribe condition} \end{cases}$$

where, in a given trial, SV denotes the subjective value,  $p_P$  and  $p_{PH}$  represents the payoff (i.e., monetary gain) for the proposer and power-holder given different choices (i.e., accept or reject the offer),  $\alpha$  and  $\beta$  measure the degree of aversion to payoff inequality in disadvantageous and advantageous situations respectively. In other words, these parameters capture how much the participant (i.e., power-holder) dislikes the offer when they earn less, (measured by  $\alpha$ ), or more (measured by  $\beta$ ), than the proposer in the bribe and the control condition respectively.

Parameters were estimated using the hierarchical Bayesian approach (HBA) via the “hBayesDM” package. R-hat values of all estimated parameters were smaller than 1.02, indicating adequate convergence of the MCMC chains (Gelman & Rubin, 1992). The posterior predictive check revealed that the proportion of acceptance predicted by this model could capture the proportion of observed acceptance across individuals (both conditions in all groups:  $r_s > 0.88$ ,  $p_s < 0.001$ ; see **Figure S2**), which further justified the validity of our model.

To examine how tDCS treatment modulated the bribe-specific effect on each of these parameters, we calculated the between-condition difference scores of  $\alpha$  and  $\beta$  within each participant (i.e.,  $\Delta\alpha = \alpha_{\text{bribe}} - \alpha_{\text{control}}$ ;  $\Delta\beta = \beta_{\text{bribe}} - \beta_{\text{control}}$ ). Regression analyses showed a significant cathodal tDCS effect on the  $\Delta\beta$  ( $b = -1.96$ , 95% CI: -3.90 to -0.03,  $b_z = -0.21$ ,  $SE = 0.98$ ,  $t(116) = -2.01$ ,  $p = 0.047$ ), with a similar trend observed in the anodal group ( $b = -1.77$ , 95% CI: -3.69 to 0.15,  $b_z = -0.19$ ,  $SE = 0.97$ ,  $t(116) = -1.82$ ,  $p = 0.071$ ). No similar effect was observed in  $\Delta\alpha$  (anodal:  $b = -0.44$ , 95% CI: -3.43 to 2.55,  $b_z = -0.03$ ,  $SE = 1.51$ ,  $t(116) = -0.29$ ,  $p = 0.77$ ; cathodal:  $b = 1.05$ , 95% CI: -1.96 to 4.06,  $b_z = 0.07$ ,  $SE = 1.52$ ,  $t(116) = -0.69$ ,  $p = 0.49$ ), indicating an interaction between tDCS group and task condition on advantageous, but not on disadvantageous inequality aversion. *Post-hoc* analyses revealed that the cathodal tDCS significantly reduced  $\beta$  in the bribe condition (i.e., Cathodal tDCS vs. sham:  $b = -2.05$ , 95% CI: -4.04 to -0.06,  $b_z = -0.21$ ,  $SE = 1.00$ ,  $t(116) = -2.04$ ,  $p = 0.044$ ). This pattern appeared similar in the anodal group, however the effect was not statistically significant (Anodal tDCS vs. sham:  $b = -1.59$ , 95% CI: -3.57 to 0.39,  $b_z = -0.17$ ,  $SE = 1.00$ ,  $t(116) = -1.59$ ,  $p = 0.114$ ). No between-group difference was observed in  $\beta$  in the control condition ( $p_s > 0.35$ ; see **Figure 4A**; also see **Table S2** for the descriptive summary of both parameters), indicating that the tDCS-induced effect is selective for corrupt decision-making.

Explorative analyses further revealed that such modulation elicited by anodal tDCS (vs. sham) on  $\Delta\beta$  depended on individual variations in other-oriented empathy, as measured by the Prosocial Personality Battery (PSB) (Penner, Fritzsche, Craiger, & Freifeld, 1995) ( $b = -2.09$ , 95% CI: -4.03 to -0.14,  $b_z = -0.29$ ,  $SE = 0.98$ ,  $t(112) = -2.13$ ,  $p = 0.036$ ; see **Figure 4B**). This effect was driven by a differential relationship between  $\beta$  and empathy scores modulated by anodal tDCS in the bribe condition ( $b = -1.75$ , 95% CI: -3.75 to 0.25,  $b_z = -0.24$ ,  $SE = 1.01$ ,  $t(112) = -1.73$ ,  $p = 0.086$ ) and the control condition ( $b = 0.34$ , 95% CI: -0.05 to 0.73,  $b_z = 0.25$ ,  $SE = 0.20$ ,  $t(112) = 1.74$ ,  $p = 0.085$ ; see **Figure S3**). A similar trend was observed for pattern of the relationship between  $\Delta\beta$  and empathy score was also found in the cathodal group, but

the effect was not statistically significant ( $b = -1.79$ , 95% CI: -3.76 to 0.19,  $b_z = -0.24$ ,  $SE = 1.00$ ,  $t(112) = -1.79$ ,  $p = 0.077$ ). All these results still held after we additionally controlled for the effect of helpfulness, as measured by PSB, and immoral preference, as measured by the Machiavellianism scale (Mach-IV) (Christie & Geis, 1970). No interaction effect on  $\Delta\beta$  was found between tDCS groups and other personality traits ( $ps > 0.05$ ). No tDCS  $\times$  personality interaction effects were observed on  $\Delta\alpha$  ( $ps > 0.12$ ; see Methods for details about personality measures; see **Table S3** and **S4** for details of the regression output).

## Discussion

In the present study, we combined a novel task that captures the essence of real-life bribery with tDCS to examine whether and how rDLPFC causally influences the corrupt behaviors of a power-holder. Compared with the sham condition, the disruption of rDLPFC (i.e., both anodal and cathodal groups) made participants, as power-holders, more likely to accept offers in the bribe (vs. control) condition as the size of the prospective payoff increased. Our results in the cathodal group are consistent with several previous studies investigating cost-benefit decisions in social contexts. Specifically, the inhibition of rDLPFC via either repetitive TMS (Knoch, Gianotti, et al., 2006) or cathodal tDCS (Speitel et al., 2019) has been found to cause receivers in the ultimatum game to be less resistant to offers that are disadvantageously unfair to them but bring themselves extra financial profits. Furthermore, patients with lesions of DLPFC are more likely to break moral principles by lying more frequently for higher personal payoffs when they balance moral rules against material self-interests (Zhu et al., 2014). In line with these findings, our results also indicate a crucial role of DLPFC (especially the right part) in overriding selfish motivation when it is in conflict with other-regarding or moral concerns (Carlson & Crockett, 2018).

Surprisingly, the excitation of rDLPFC via anodal tDCS shows a similar effect as cathodal tDCS in promoting bribe-taking behaviors. This seems to run against our prediction, based on a recent study reporting that anodal tDCS over rDLPFC increases honest behaviors (Maréchal et al., 2017). One explanation for such a difference is that the specific (im)moral behaviors induced and measured in their study and ours are quite different. The study by Maréchal and colleagues investigated how an individual decides whether to commit misconduct (i.e., misreporting an outcome) for higher self-profits, via a simple die-rolling task. However, in the present study, participants did not lie, but rather decided whether to collaborate with another's misconduct depending on his or her "gift", which required complex tradeoff calculations. This difference possibly results in distinct cognitive processes that underlie each type of decision. In the former task, participants balance self-interest solely against honesty concerns, whereas here they are confronted with a more complex trade-off involving the relationship between self-other source allocation, and the moral cost of colluding with another's immoral conduct. Thus, it is possible that the modulation of rDLPFC via anodal tDCS involves a complicated mechanism that interacts with behaviors in a more complex social context (Miller & Cohen, 2001; Tanji & Hoshi, 2008). Another explanation relates to the mechanisms underlying the relationship between tDCS and behavioral changes. It is known that anodal and cathodal tDCS increase and decrease neural excitability respectively on a microscopic scale (Bikson & Rahman, 2013), and in motor cortex on a macroscopic scale (Nitsche & Paulus, 2000).

However, the way in which the micro-scale effects of tDCS on neural circuits translate into macro-scale changes in behavior is often unclear, especially for complex social behaviors of the type studied here (Bestmann, de Berker, & Bonaiuto, 2015).

Moreover, we explored possible mechanisms underlying the offer proportion-dependent increase in bribe-taking behaviors induced by both anodal and cathodal tDCS. Our task allowed us to further investigate the exact computational component that was altered by stimulation. Paralleling the behavioral findings noted above, we observed that compared with the sham condition, both tDCS conditions, but especially cathodal tDCS, selectively dampened the level of advantageous inequity aversion in the bribe condition, without affecting the same parameter in the control condition. No such effect was observed in the level of disadvantageous inequity aversion in either condition. Given the psychological meaning of the parameter, the advantageous inequity aversion measures how an individual feels guilty about earning more than the other person. Thus, our results may reflect that power-holders feel less guilty to take bribes (rather than normal offers) with higher proportions when the rDLPFC was temporarily perturbed. These findings reveal a dedicated causal role of rDLPFC in regulating the advantageous inequity aversion that elicits implicit guilt in guiding morally-compliant decisions. Importantly, none of the additional measures regarding subjective feelings related to the task (i.e., moral conflict, moral inappropriateness, guilt, sense of power, sense of being bribed), the general emotional state, and cognitive reflection ability were modulated by tDCS or its interaction with task conditions. These results further corroborate our inference on a specific role for rDLPFC in gating moral behaviors via modulation of advantageous inequity aversion.

Further exploratory analysis revealed that tDCS-induced attenuation of advantageous inequity aversion specific to bribery is dependent on the prosocial personality (especially the disposition of other-oriented empathy) of individuals as power-holders. More specifically, the more empathic concern the individual has, the more the advantageous inequity aversion parameter is attenuated by anodal tDCS (vs. sham), but only in the bribe condition. This suggests that the regulation of corrupt behaviors by rDLPFC was contingent on the level of empathic concern, both working together to sustain a higher level of intrinsic guilt in a power-holder to prevent them from being tempted by a bribe.

Several issues merit further consideration. First, while the present study focuses on the role of rDLPFC, it remains unknown that how rDLPFC interacts with other regions, such as the mentalizing or valuation network (Ruff & Fehr, 2014; Suzuki & O'Doherty, 2020) during corrupt decision-making. Second, our study lays the groundwork for additional research questions, such as how rDLPFC is involved in weighing additional factors (e.g., the risk of

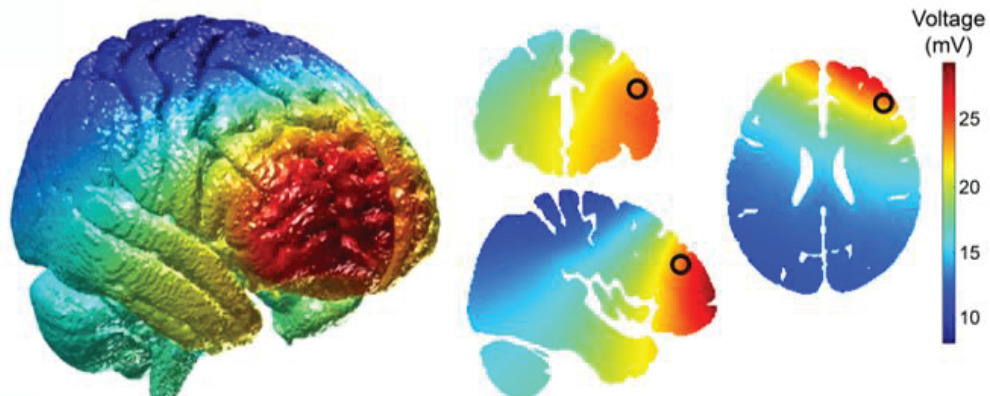


being caught) during corrupt decision-making.

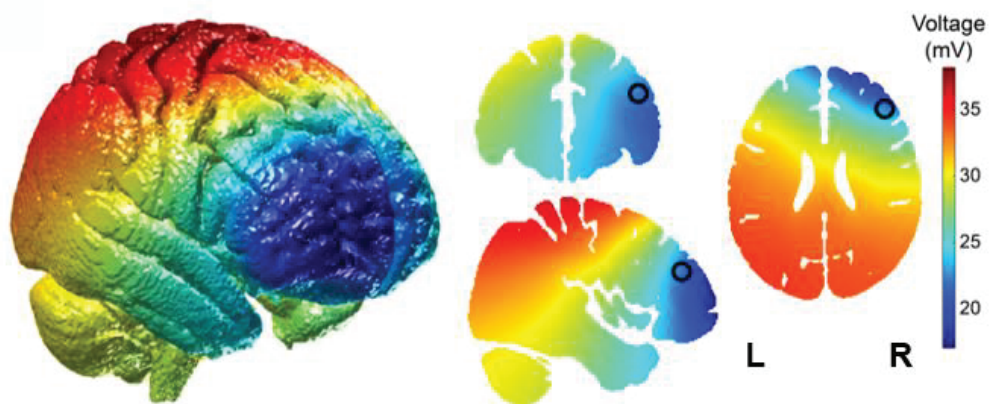
In sum, the present study provides empirical evidence that perturbing rDLPFC via tDCS causally influences decisions by a power-holder of whether or not to accept a tempting bribe. This influence is likely exerted via suppression of advantageous inequity aversion, which also depends on the empathic concern disposition of power-holders. These findings shed light on the neurobiological substrates of corrupt behavior and elaborate our understanding of the function of rDLPFC in complex social behaviors. Our study offers a new window to investigate corrupt behaviors using a multi-disciplinary research approach and provides practical insights to identify people with variant tendency to succumb to corruption after tDCS stimulation.

## Figures

**A**

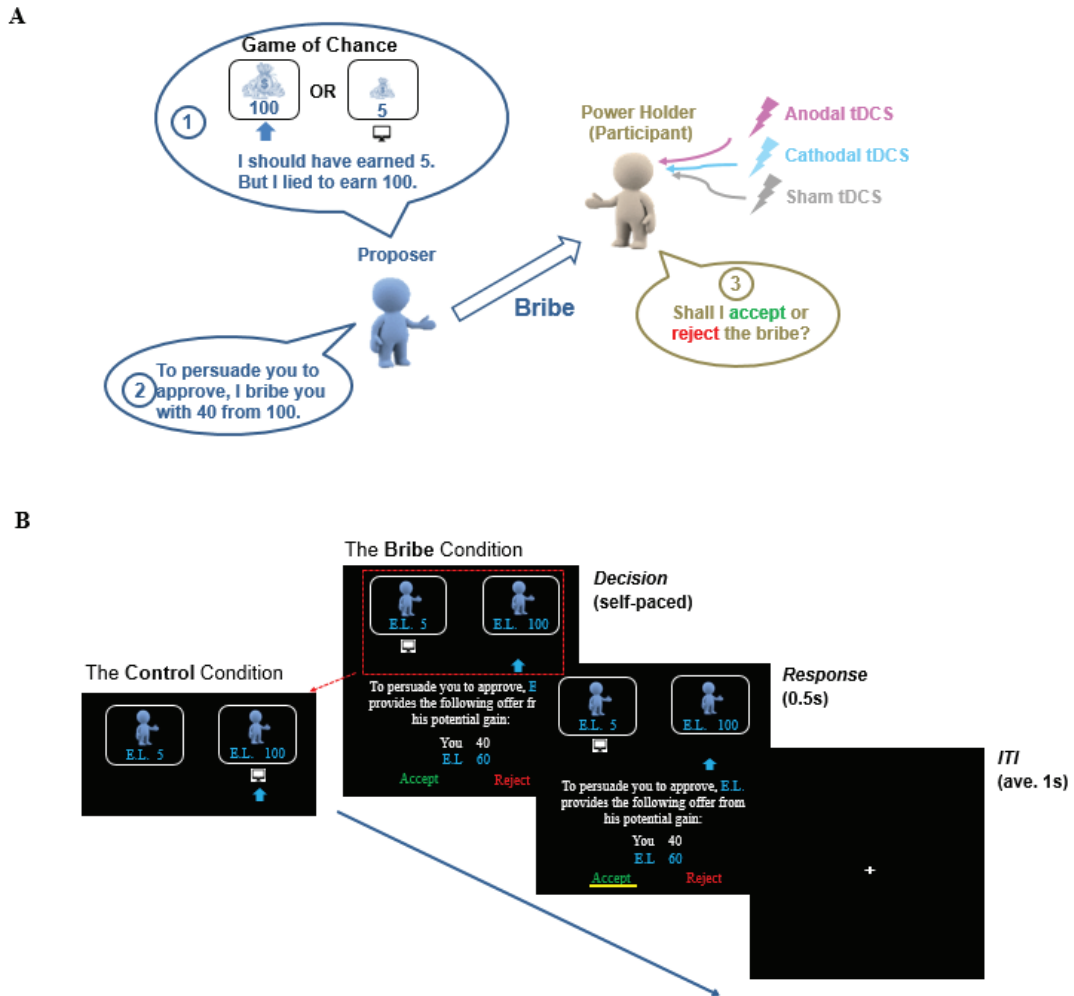


**B**



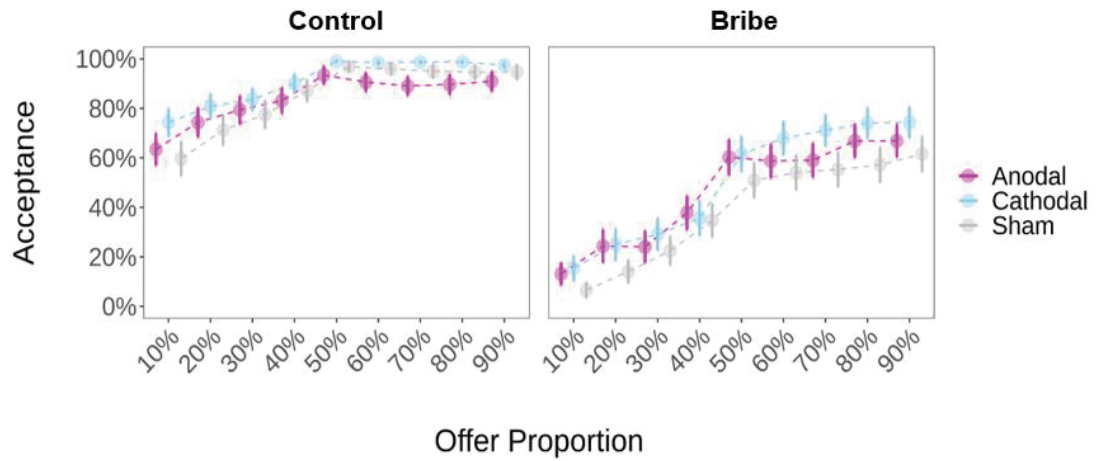
**Figure 1. Electric field simulation for (A) anodal and (B) cathodal tDCS stimulation.**

Based on previous literature closely relevant to the current study (Knoch et al., 2006; Strang et al., 2014), we chose the position centering around the Talarach coordinate of 39/37/22 as our target site. This location approximately corresponds to the electrode position of AF4 in the 10-20 system of EEG cap (the right panel; marked with a black circle). The vertex was chosen as the reference electrode based on the study by Marechal et al (2017), which corresponds to the electrode position of Cz. Electrodes were simulated as pads, with a 100x100x3mm pad located over Cz and a 70x50x3mm pad located over AF4, using standard 10-10 system locations. Tissue conductivities were set as white matter=0.11 S/m, gray matter=0.21 S/m, CSF=0.53 S/m, bone=0.02 S/m, and skin=0.90 S/m. For the anodal simulation, 1.5mA was set as inward flowing current from the AF4 pad, and -1.5mA outward flowing current from the Cz pad, and vice versa for the cathodal simulation. The simulation was done via ROAST (Huang, Datta, Bikson, & Parra, 2019; <https://github.com/andypotatohy/roast>). Abbreviations: L: left; R: right.



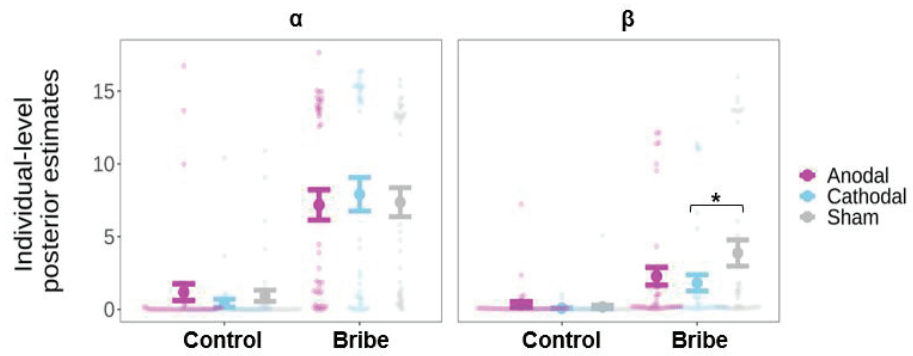
**Figure 2 Task design. (A) Schematic illustration of the tDCS manipulation and the behavioral paradigm.** All participants were randomly assigned to one of the three tDCS groups (i.e., anodal, cathodal or sham). The task consists of two roles, a proposer (i.e., fictitious participants of a previous online study where they played a “Game of Chance”) and a power-holder (i.e., the real participants of the current study). In the control condition, the proposer truthfully reports the reward amount selected by the computer. In the bribe condition (as shown here), the proposer lies about the selected reward amount. In both conditions the proposer offers a certain amount of money to the power-holder, whose task was to decide whether to accept or reject the offer. **(B) Trial procedure.** In this example trial in the bribe condition, a proposer (E.L.) lied by reporting the non-selected option with a larger payoff (as indicated by the misalignment of the blue arrow and the icon of a computer), and bribed the power-holder with a certain amount of money from his/her potential gain (i.e., 40 out of 100 Euros). The participant needed to decide whether to accept or reject the offer. Once the

decision was made (i.e., accepting the bribe here), a yellow bar appeared on the corresponding option to highlight the choice for 0.5 s, which was followed by a fixation (i.e., 0.6~1.4 s with a mean of 1s). Trials in the control condition followed the same procedure except that the proposer truthfully reported the selected option with a larger payoff (as indicated by the alignment of the blue arrow and the icon of a computer).

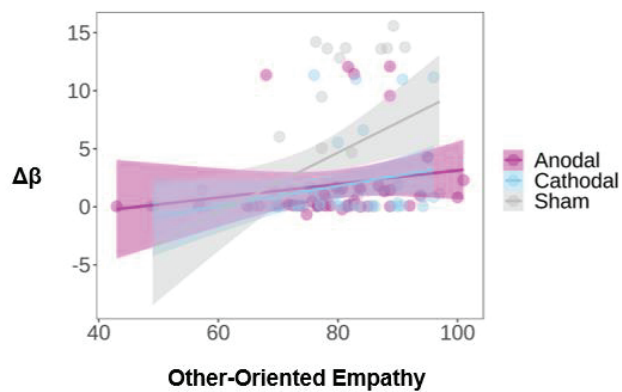


**Figure 3. Results of acceptance rate (%).** Mean acceptance rate plotted as a function of *tDCS* group (anodal/cathodal/sham), *task condition* (control / bribe), and *offer proportion* (10% to 90% in steps of 10%). Error bars represent SEM.

A



B



**Figure 4. Model-based results. (A) Posterior mean of individual-level estimates of disadvantageous ( $\alpha$ ) and advantageous inequity aversion ( $\beta$ ).** Each small dot represents the data of a single participant; filled dots represent the group-level mean of parameters. Error bars represent SEM; Significance: \* $p < 0.05$ . **(B) Relationship between  $\Delta\beta$  (i.e.,  $\beta_{\text{bribe}} - \beta_{\text{control}}$ ) and other-oriented empathy across individuals.** The trait score of other-oriented empathy was measured by the prosocial personality scale (PSB). Lines represent the linear fits; shaded areas represent the confidence intervals.

## Reference

- Abbink, K. (2006). Laboratory experiments on corruption. In S. Rose-Ackerman (Ed.), *International handbook on the economics of corruption* (pp. 418-437).
- Abbink, K., & Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9(2), 103-121.
- Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, 1, 24-57.
- Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: moral frames and social norms. *Experimental Economics*, 19(1), 240-267.
- Bestmann, S., de Berker, A. O., & Bonaiuto, J. (2015). Understanding the behavioural consequences of noninvasive brain stimulation. *Trends in Cognitive Sciences*, 19(1), 13-20.
- Bikson, M., & Rahman, A. (2013). Origins of specificity during tDCS: anatomical, activity-selective, and input-bias mechanisms. *Frontiers in Human Neuroscience*, 7.
- Carlson, R. W., & Crockett, M. J. (2018). The lateral prefrontal cortex and moral goal pursuit. *Current Opinion in Psychology*.
- Christie, R., & Geis, F. L. (1970). *Studies in machiavellianism*. NY: Academic Press.
- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, 111(48), 17320-17325.
- Dreher, A., Kotsogiannis, C., & McCorrison, S. (2007). Corruption around the world: Evidence from a structural model. *Journal of comparative economics*, 35(3), 443-466.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, 817-868.
- Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525-547.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.
- Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2), 419-453.
- Gupta, S., Davoodi, H., & Alonso-Terme, R. Does corruption affect income inequality and poverty? *Economics of Governance*, 3(1), 23-45.
- Huang, Y., Datta, A., Bikson, M., & Parra, L. C. (2019). Realistic vOlumetric-Approach to Simulate Transcranial Electric Stimulation -- ROAST -- a fully automated open-source pipeline. *Journal of Neural Engineering*, 16(5).
- Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., & Brugger, P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *The Journal of Neuroscience*, 26(24), 6469-6472.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829-832.
- Köbis, N. C., van Prooijen, J.-W., Righetti, F., & Van Lange, P. A. (2016). Prospection in individual and interpersonal corruption dilemmas. *Review of General Psychology*, 20(1), 71.
- Maréchal, M. A., Cohn, A., Ugazio, G., & Ruff, C. C. (2017). Increasing honesty in humans with

- noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, 114(17), 4360-4364.
- Martin, K. D., Cullen, J. B., Johnson, J. L., & Parboteeah, K. P. (2007). Deciding to bribe: A cross-level analysis of firm and home country influences on bribery activity. *Academy of Management Journal*, 50(6), 1401-1422.
- Mauro, P. (1995). Corruption and growth. *The quarterly journal of economics*, 110(3), 681-712.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1), 167-202.
- Nitsche, M., & Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *The Journal of physiology*, 527(3), 633-639.
- Penner, L. A., Fritzsche, B. A., Craiger, J. P., & Freifeld, T. S. (1995). Measuring the prosocial personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment*, Vol. 10. (pp. 147-163). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- R Core Team. (2014). R: A language and environment for statistical computing.
- Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8), 549-562.
- Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, 342(6157), 482-484.
- Serra, D., & Wantchekon, L. (2012). *New advances in experimental research on corruption* (Vol. 15): Emerald Group Publishing.
- Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2), 181-190.
- Shleifer, A., & Vishny, R. W. (1993). Corruption. *The quarterly journal of economics*, 108(3), 599-617.
- Speitel, C., Traut-Mattausch, E., & Jonas, E. (2019). Functions of the right DLPFC and right TPJ in proposers and responders in the ultimatum game. *Soc Cogn Affect Neurosci*, 14(3), 263-270. doi:10.1093/scan/nsz005
- Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. (2014). Be nice if you have to—The neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, nsu114.
- Suzuki, S., & O'Doherty, J. (2020). Breaking human social decision making into multiple components and then putting them together again. *Cortex*.
- Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological reviews*, 88(1), 37-57.
- Treisman, D. (2000). The Causes of Corruption: A Cross-National Study. *Journal of Public Economics*, 76(3), 399-457.
- Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., . . . Hsu, M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature neuroscience*, 17(10), 1319-1321.



**Supplementary Information for**

**Perturbation of Right Dorsolateral Prefrontal Cortex (rDLPFC) Makes  
Power-Holders Less Resistant to Tempting Bribes**

**This PDF file includes:**

Supplementary Methods

Supplementary Results

Figures S1 to S4

Tables S1 to S7

## Supplementary Methods

### tDCS Protocol

tDCS was administered using a multichannel stimulator (NeuroConn, Munich) and pairs of standard electrodes covered with conductive paste. Sites of stimulation were fixed through a 10-20 EEG system cap and noted with a marker on the participant's scalp. According to the fairness-related activation foci reported by previous studies (i.e., Talarach x/y/z: 39/37/22; Knoch, Pascual-Leone, et al., 2006; Strang et al., 2014), we placed one of these electrodes (5 cm × 7 cm) over AF4 on the 10-20 EEG system for stimulation of the right DLPFC. The other electrode (10 cm × 10 cm) was placed over Cz (i.e., vertex), based on previous tDCS studies on social decision-making (Maréchal et al., 2017). Following well-established technical guidelines for tDCS studies (Woods et al., 2016), we applied stimulation at an intensity of 1.5 mA for up to 30 min in the anodal and cathodal groups during the experiment. To verify that the chosen electrode montage targeted the right DLPFC, we performed current flow simulations were performed using ROAST (Huang, Datta, Bikson, & Parra, 2019) with the MNI152 template brain (see **Figure 1**). For the sham group, stimulation with the same intensity was set to emit for 1s per minute to simulate the tingling sensations. To minimize the sensations at stimulation onset, the current was linearly ramped up (at the start) and down (at the end) over periods of 20 s.

### Procedure

Participants were invited to group sessions with up to 4 in each. Prior to the experiment, participants signed a written informed consent form according to the Declaration of Helsinki. Next, they underwent a clinical screen performed by an experienced neurological doctor in the hospital affiliated with the university, and answered questions from standard health screening questionnaires. Having been confirmed to meet the inclusion criteria for the experiment, they were led to the tDCS room and were randomly placed at seats (desktops), which were separated from each other by shelves. They were then provided with the general instructions and completed the Multidimensional Mood Questionnaire (MDMQ) to report their baseline emotion state. Then, they were given the task instructions, and answered a series of comprehension questions to ensure that they fully understood the task. Meanwhile, two experimenters fitted the participants with the tDCS electrodes. Before the main

experiment, participants also practiced a few example trials to get familiar with the paradigm and the response button.

The main experiment included a computerized incentive task and a follow-up paper-and-pencil rating task (see Task and Design for details), which lasted around 30 min in total. Once all participants in the session were prepared, the experimenter started the tDCS stimulation for 45s and then commenced the incentive task. To further protect their privacy, curtains behind the participants' seats were drawn during the whole experiment.

The tDCS was maintained until participants in the session finished the main experiment. After that, they took a short break and then filled out a battery of questionnaires for control measures. In particular, they indicated whether they felt comfortable after the stimulation, declared their belief about treatment (stimulation, placebo, or unknown), reported their emotional state again by filling out the Multidimensional Mood Questionnaire (Steyer, 2014), and finished a Cognitive Reflection Test as a measure of their cognitive reflection ability (Frederick, 2005). We also measured individual differences in machiavellian personality using Mach-IV (Christie & Geis, 1970), and altruistic preferences using Prosocial Personality Questionnaire (Penner, Fritzsche, Craiger, & Freifeld, 1995) for explorative analyses. Finally, participants were debriefed on all task-relevant information, and informed about their final payoffs.

## Data Analyses

All analyses and visualization were conducted using R (<http://www.r-project.org/>) and relevant packages (R Core Team, 2014). All reported p values are two-tailed and  $p < 0.05$  was considered statistically significant. For choice data, we performed repeated measures mixed-effect logistic regression on the decision of choosing the “accept” option, using the *glmer* function in the “lme4” package (Bates, Maechler, & Bolker, 2013), with group (dummy variable; reference level: Sham), task condition (dummy variable; reference level: control), offer proportion (continuous variable), and their interactions as fixed-effects of interest. The effect of the larger payoff the proposer would earn in the reported option (continuous variable; z-scored) was also incorporated as a fixed-effect covariate. The random-effects were established using a “maximal” principle such that we allowed intercepts and slopes (i.e., task condition, offer proportion and their interaction) to vary across participants (Barr, Levy,

Scheepers, & Tily, 2013). For statistical inference on each fixed effect, we performed a Type II Wald chi-square test on the model fits by using the *Anova* function in the “car” package (Fox et al., 2016). For decision time (DT), we first log-transformed the data due to its non-normal distribution (i.e., Anderson-Darling normality test:  $A = 1411.1$ ,  $p < 0.001$ ) and then performed a mixed-effect linear regression on the log-transformed DT using the *lmer* function in the “lme4” package. Random-effect predictors were specified in the same way as above. When a model failed to converge, we dropped one or more of the random slopes until the estimation converged. We followed the procedure recommended by Luke (2017) to obtain the statistics of each predictor by applying the Satterthwaite approximations on the restricted maximum likelihood model (REML) fit via the “lmerTest” package (Luke, 2017). We performed post-hoc analyses of interaction effects using *emtrends* function of the “emmeans” package. For subjective rating, we used mixed analysis of variance (ANOVA) or simple linear regression analyses depending on specific items (see Results for details). Furthermore, we reported the odds ratio as an index of effect size of each predictor on choice. For decision time and other continuous dependent measures (e.g., rating, parameter estimates), we computed the standardized coefficient ( $b_z$ ) as an index of effect size using the “lm.beta” package (<https://cran.r-project.org/web/packages/lm.beta/>). We also used *partial*  $\eta^2$  via the “sjstats” package (<https://cran.r-project.org/web/packages/sjstats/>) to indicate the effect size of main effects or interactions in ANOVA analyses when applicable.

As mentioned earlier, we implemented a model-based analysis by adopting a modified version of Fehr-Schmidt model to fit choice behavior. The probability of accepting the offer was determined by the softmax function:

$$prob(SV_{accept}) = \frac{e^{\tau SV_{accept}}}{e^{\tau SV_{accept}} + e^{\tau SV_{reject}}}$$

where  $SV$  denotes the subjective value (of accepting or rejecting the offer), calculated by the model mentioned earlier.  $\tau$  is the inverse softmax temperature parameter ( $0 \leq \tau \leq 10$ ) denoting the sensitivity of an individual’s decision to the difference in  $SV$  between the choice of accepting versus rejecting the offer.

The above model was fit using a hierarchical Bayesian approach (HBA) via the “hBayesDM” package (Ahn, Haines, & Zhang, 2017), which adopts a Markov Chain Monte Carlo (MCMC) sampling scheme to perform full Bayesian inference. Convergence of the MCMC chains was assessed through the Gelman-Rubin R-hat

Statistics (Gelman & Rubin, 1992). Here, R-hat values of all estimated parameters of each tDCS group were smaller than 1.02, indicating adequate convergence of the MCMC chains. We also performed the posterior predictive check (Zhang, Langersdorff, Mikus, Glaescher, & Lamm, 2020) following the procedure of a similar study (Qu, Hu, Tang, Derrington, & Dreher, 2020) to examine whether the prediction of the model could capture the features of real behaviors of participants.

For each individual, we obtained the posterior mean of individual-level parameter estimates ( $\alpha$  and  $\beta$ ) in each condition and calculated  $\Delta\alpha$  and  $\Delta\beta$  to characterize the bribe-specific effect on each of these parameters. Then, we performed simple linear regression on  $\Delta\alpha$  and  $\Delta\beta$  with tDCS group as the predictor. In addition, we also explored whether related personality traits could modulate the effect of tDCS on these parameters. To this end, we ran linear regressions with tDCS group, personality scores (e.g., other-oriented empathic concern and helpfulness as measured by the Prosocial Personality Battery [PSB]; immorality preference as measured by the Machiavellianism scale [Mach-IV]), and their interactions on  $\Delta\alpha$  and  $\Delta\beta$ , respectively.

## Supplementary Results

### No tDCS effect was observed in other behavioral measures

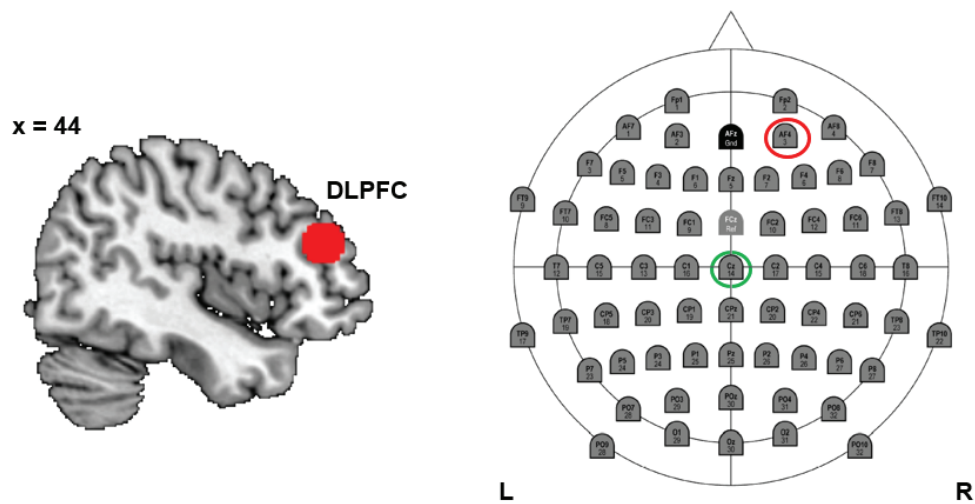
We investigated whether a similar effect of tDCS existed in other behavioral measures. Analyses on log-transformed DT revealed that participants responded slightly slower in the bribe condition (vs. control; a main effect of task condition:  $F(1,325) = 5.97, p < 0.001$ ) and when the offer proportion increased (a main effect of offer proportion:  $F(1,17012) = 67.03, p < 0.001$ ). In addition, we observed a two-way interaction between *task condition* and *offer proportion* ( $F(1,16937) = 16.59, p < 0.001$ ; see **Figure S4**). *Post-hoc* analyses indicated that participants responded faster when the offer proportion increased in both conditions ( $z_s < -3.15, p_s < 0.002$ ) but the slope was less steep in the bribe condition (vs. control;  $z = 4.07, p < 0.001$ ; see **Table S5** for details of the regression output).

In addition, we also examined whether tDCS over rDLPFC affected subjective ratings, in order to rule out alternative accounts that might explain the effect of tDCS on bribe-taking behaviors. First, compared with the control condition, participants in the bribe condition felt a higher level of moral conflict during the decision period ( $F(1,116) = 103.50, p < 0.001, \text{partial-}\eta^2 = 0.157$ ). They thought that the proposer's offering act ( $F(1,116) = 21.65, p < 0.001, \text{partial-}\eta^2 = 0.472$ ) and their hypothetical acceptance were more morally inappropriate ( $F(1,115) = 157.73, p < 0.001, \text{partial-}\eta^2 = 0.578$ ). They also felt more guilty for their hypothetical acceptances of offers provided by the proposer ( $F(1,115) = 101.64, p < 0.001, \text{partial-}\eta^2 = 0.469$ ). However, none of these measures were modulated by tDCS ( $F_s < 1.01, p_s > 0.36, \text{partial-}\eta^2_s < 0.02$ ) nor its interaction with task conditions ( $F_s < 1.34, p_s > 0.26, \text{partial-}\eta^2_s < 0.03$ ). Second, participants from the three tDCS groups reported similar levels of the sense of power over the proposer ( $F(2,116) = 0.52, p = 0.597, \text{partial-}\eta^2 = 0.009$ ) and the sense of being bribed ( $F(2,116) = 1.04, p = 0.357, \text{partial-}\eta^2 = 0.018$ ).

Regarding task-irrelevant measures, no difference between the three tDCS groups was found in emotional state, as measured by the Multidimensional Mood Questionnaire (MDMQ) (Steyer, 2014), reported before the main task (the awake-tired [AT] subscale:  $F(2,115) = 0.85, p = 0.429, \text{partial-}\eta^2 = 0.015$ ; the calm-nervous [CN] subscale:  $F(2,114) = 0.22, p = 0.804, \text{partial-}\eta^2 = 0.004$ ; the good-bad [GB] subscale:  $F(2,115) = 0.44, p = 0.645, \text{partial-}\eta^2 = 0.008$ ) or after (AT:  $F(2,116) = 0.39, p = 0.677, \text{partial-}\eta^2 = 0.007$ ; CN:  $F(2,116) = 1.18, p = 0.312, \text{partial-}\eta^2 = 0.020$ ; GB:  $F(2,116) = 0.95, p = 0.389, \text{partial-}\eta^2 = 0.016$ ). Cognitive reflection ability, as measured by the Cognitive Reflection Test (Frederick, 2005), was unaffected by the tDCS manipulation

( $\chi^2(4) = 5.28$ ,  $p = 0.260$ ; see **Table S6** and **S7** for a descriptive summary of these measures).

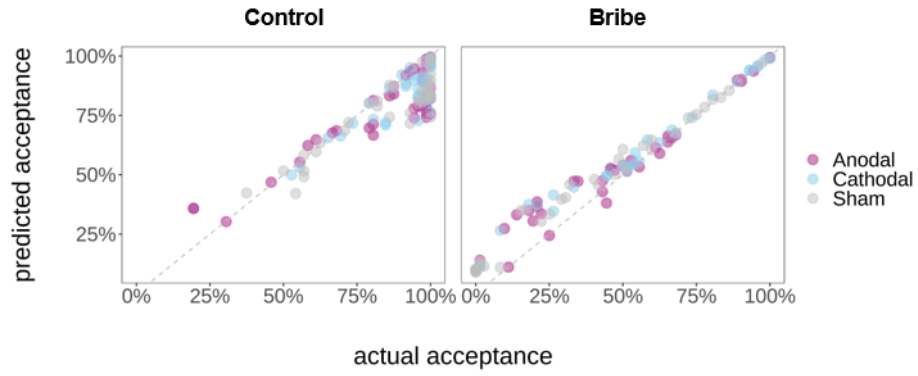
## Supplementary Figures



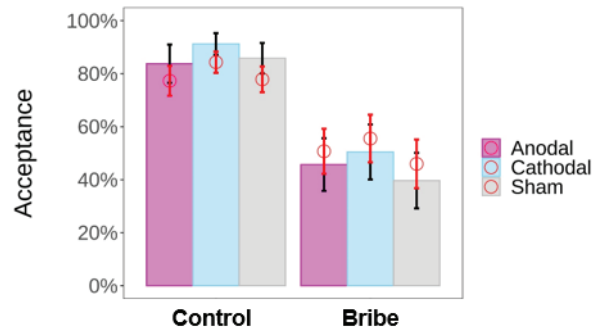
**Figure S1. Display of the tDCS electrode localization.** Based on previous literature closely relevant to the current study (Knoch *et al.*, 2006; Strang *et al.*, 2014), we chose the position centering around the MNI coordinate of 39/37/22 as our target site (the left panel; a sphere of a 10mm radius was used for visualization). This location approximately corresponds to the electrode position of AF4 in the 10-20 system of 64-channel EEG cap (the right panel; marked with a red circle). The vertex was chosen as the reference electrode based on the study by Marechal *et al* (2017), which corresponds to the electrode position of Cz (the right panel; marked with a green circle).



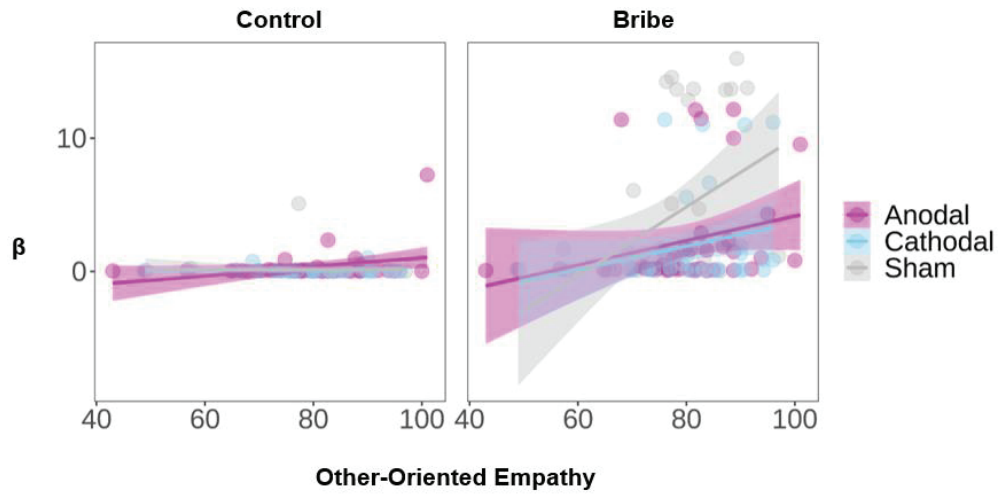
**A**



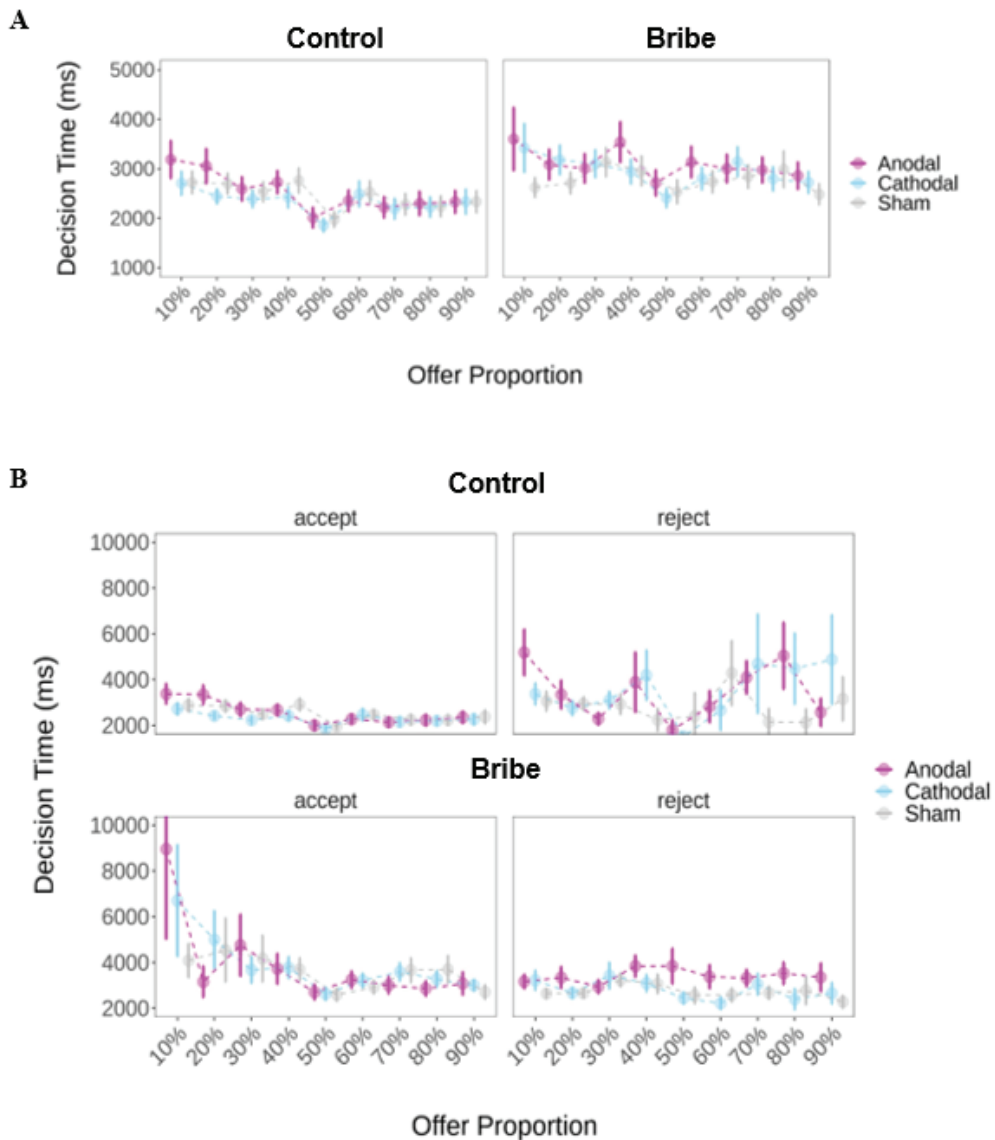
**B**



**Figure S2. Posterior predictive check. (A) Relationship between predicted acceptance rates and actual behaviors across individuals.** Each dot represent individual data. **(B) Mean predicted acceptance rates (red circle) and actual behaviors (filled bar) plotted as a function of tDCS group and task condition.** Error bars represent 95% CI.



**Figure S3. Relationship between  $\beta$  and other-oriented empathy in each task condition across individuals.** The trait score of other-oriented empathy was measured by the prosocial personality scale (PSB). Lines represent the linear fits; shaded areas represent the confidence intervals.



**Figure S4. Results of decision time (DT; ms). (A) Mean DT are plotted as a function of tDCS group (anodal/cathodal/sham), task condition (control/bribe), and offer proportion (10% to 90% in a step of 10%). (B) Mean DT are plotted as a function of these independent variables for acceptance trials and rejections trials respectively. Error bars represent SEM.**

## Supplementary Tables

Table S1 Results of mixed-effect logistic regressions predicting acceptance

|  | all                | control            | bribe              |
|--|--------------------|--------------------|--------------------|
|  | <i>b</i>           | <i>b</i>           | <i>b</i>           |
|  | (SE)               | (SE)               | (SE)               |
| Intercept  | 0.25<br>(0.80)     | 0.23<br>(0.88)     | -6.58***<br>(0.83) |
| Group (Anodal)   | 0.72<br>(1.12)     | 0.67<br>(1.20)     | 0.44<br>(1.17)     |
| Group (Cathodal)   | 1.49<br>(1.14)     | 1.64<br>(1.23)     | 0.14<br>(1.18)     |
| Condition  | -6.79***<br>(1.03) |                    |                    |
| Offer Proportion   | 10.47***<br>(1.58) | 10.26***<br>(1.78) | 11.51***<br>(1.87) |
| Group (Anodal) × Condition                                     | -0.23<br>(1.43)    |                    |                    |
| Group (Cathodal) × Condition                                   | -1.29<br>(1.45)    |                    |                    |
| Group (Anodal) × Offer Proportion                              | -3.22<br>(2.17)    | -3.19<br>(2.25)    | 1.90<br>(2.65)     |
| Group (Cathodal) × Offer Proportion                            | -2.86<br>(2.22)    | -3.11<br>(2.30)    | 2.37<br>(2.66)     |
| Condition × Offer Proportion                                   | 1.06<br>(1.57)     |                    |                    |
| Group (Anodal) × Condition × Offer Proportion                  | 5.33*<br>(2.08)    |                    |                    |
| Group (Cathodal) × Condition × Offer Proportion                | 5.20*<br>(2.13)    |                    |                    |
| Larger payoff for proposer in the reported option <sup>a</sup> | 0.29***<br>(0.03)  | 0.18***<br>(0.05)  | 0.37***<br>(0.04)  |
| AIC  | 7400.6             | 3211.6             | 4243.8             |
| BIC  | 7578.8             | 3282.2             | 4314.4             |
| N (Observation)  | 17136              | 8568               | 8568               |
| N (Participant)  | 119                | 119                | 119                |

Note: <sup>a</sup> This variable was standardized before the analyses. Reference levels in dummy variables were set as follows: Group = sham, Condition = control. Table also shows goodness-of-fit statistics: AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. Significance: \**p* < 0.05, \*\**p* < 0.01, \*\*\**p* < 0.001.

Table S2 Descriptive statistics of posterior mean of individual-level parameter estimates ( $\alpha$  and  $\beta$ )

|                          |         | Anodal<br>(N = 40) | Cathodal<br>(N = 39) | Sham<br>(N = 40) |
|--------------------------|---------|--------------------|----------------------|------------------|
| $\alpha$ (mean $\pm$ SD) | control | 1.19 $\pm$ 3.64    | 0.43 $\pm$ 1.75      | 0.94 $\pm$ 2.42  |
|                          | bribe   | 7.18 $\pm$ 6.60    | 7.92 $\pm$ 7.22      | 7.37 $\pm$ 6.28  |
| $\beta$ (mean $\pm$ SD)  | control | 0.35 $\pm$ 1.19    | 0.09 $\pm$ 0.19      | 0.18 $\pm$ 0.80  |
|                          | bribe   | 2.28 $\pm$ 3.88    | 1.83 $\pm$ 3.47      | 3.88 $\pm$ 5.71  |

Table S3 Results of linear regressions predicting parameters

|                                 | $\Delta\alpha$    | $\alpha_{\text{control}}$ | $\alpha_{\text{bribe}}$ | $\Delta\beta$     | $\beta_{\text{control}}$ | $\beta_{\text{bribe}}$ |
|---------------------------------|-------------------|---------------------------|-------------------------|-------------------|--------------------------|------------------------|
|                                 | <i>b</i>          | <i>b</i>                  | <i>b</i>                | <i>b</i>          | <i>b</i>                 | <i>b</i>               |
|                                 | (SE)              | (SE)                      | (SE)                    | (SE)              | (SE)                     | (SE)                   |
| Intercept                       | 6.95***<br>(1.07) | 0.93*<br>(0.45)           | 7.88***<br>(1.07)       | 4.29***<br>(0.67) | 0.18<br>(0.13)           | 4.47***<br>(0.69)      |
| Group (Anodal)                  | -0.99<br>(1.51)   | 0.33<br>(0.64)            | -0.66<br>(1.51)         | -2.41*<br>(0.95)  | 0.13<br>(0.19)           | -2.29*<br>(0.98)       |
| Group (Cathodal)                | 0.38<br>(1.51)    | -0.48<br>(0.63)           | -0.10<br>(1.51)         | -2.62**<br>(0.95) | -0.09<br>(0.19)          | -2.71**<br>(0.97)      |
| PSB: Empathy <sup>a</sup>       | 2.34<br>(1.21)    | -0.04<br>(0.51)           | 2.30<br>(1.21)          | 2.70***<br>(0.76) | 0.01<br>(0.15)           | 2.71***<br>(0.78)      |
| Group (Anodal) × PSB: Empathy   | -1.48<br>(1.56)   | -0.19<br>(0.66)           | -1.67<br>(1.56)         | -2.09*<br>(0.98)  | 0.34<br>(0.20)           | -1.75<br>(1.01)        |
| Group (Cathodal) × PSB: Empathy | 0.02<br>(1.59)    | -0.26<br>(0.67)           | -0.24<br>(1.59)         | -1.79<br>(1.00)   | -0.01<br>(0.20)          | -1.79<br>(1.03)        |
| R <sup>2</sup>                  | 0.09              | 0.02                      | 0.07                    | 0.16              | 0.08                     | 0.16                   |
| Adjusted R <sup>2</sup>         | 0.05              | -0.02                     | 0.03                    | 0.12              | 0.04                     | 0.12                   |
| N (Participant) <sup>b</sup>    | 118               | 118                       | 118                     | 118               | 118                      | 118                    |

Note: <sup>a</sup>This variable was standardized before the analyses. <sup>b</sup> Data of PSB-empathy from one participant in the anodal group was missing. Reference levels in dummy variables were set as follows: Group = sham, Condition = control. Table also shows goodness-of-fit statistics: Significance: \*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001. Abbreviation: PSB: prosocial personality battery.

Table S4 Robustness check of linear regressions predicting parameters

|                                 | $\Delta\alpha$    | $\alpha_{\text{control}}$ | $\alpha_{\text{bribe}}$ | $\Delta\beta$     | $\beta_{\text{control}}$ | $\beta_{\text{bribe}}$ |
|---------------------------------|-------------------|---------------------------|-------------------------|-------------------|--------------------------|------------------------|
|                                 | <i>b</i>          | <i>b</i>                  | <i>b</i>                | <i>b</i>          | <i>b</i>                 | <i>b</i>               |
|                                 | (SE)              | (SE)                      | (SE)                    | (SE)              | (SE)                     | (SE)                   |
| Intercept                       | 7.19***<br>(1.06) | 0.90<br>(0.46)            | 8.09***<br>(1.06)       | 4.41***<br>(0.67) | 0.17<br>(0.14)           | 4.58***<br>(0.69)      |
| Group (Anodal)                  | -0.97<br>(1.48)   | 0.32<br>(0.64)            | -0.65<br>(1.48)         | -2.40*<br>(0.94)  | 0.12<br>(0.19)           | -2.28*<br>(0.97)       |
| Group (Cathodal)                | -0.33<br>(1.51)   | -0.38<br>(0.65)           | -0.71<br>(1.51)         | -2.96**<br>(0.96) | -0.05<br>(0.19)          | -3.01**<br>(0.99)      |
| PSB: Empathy <sup>a</sup>       | 2.66*<br>(1.19)   | -0.07<br>(0.51)           | 2.59*<br>(1.19)         | 2.86***<br>(0.76) | -0.003<br>(0.15)         | 2.86***<br>(0.78)      |
| Group (Anodal) × PSB: Empathy   | -1.92<br>(1.54)   | -0.17<br>(0.66)           | -2.09<br>(1.54)         | -2.31*<br>(0.98)  | 0.35<br>(0.20)           | -1.96<br>(1.00)        |
| Group (Cathodal) × PSB: Empathy | -0.16<br>(1.56)   | -0.28<br>(0.67)           | -0.44<br>(1.56)         | -1.88<br>(0.99)   | -0.01<br>(0.20)          | -1.89<br>(1.02)        |
| PSB: Helpfulness                | -1.21<br>(0.62)   | 0.21<br>(0.27)            | -1.01<br>(0.62)         | -0.58<br>(0.39)   | 0.08<br>(0.08)           | -0.50<br>(0.41)        |
| Machiavellianism                | 1.03<br>(0.60)    | 0.15<br>(0.26)            | 1.18*<br>(0.60)         | 0.54<br>(0.38)    | 0.05<br>(0.08)           | 0.60<br>(0.39)         |
| R <sup>2</sup>                  | 0.14              | 0.03                      | 0.12                    | 0.19              | 0.09                     | 0.19                   |
| Adjusted R <sup>2</sup>         | 0.08              | -0.03                     | 0.06                    | 0.14              | 0.04                     | 0.14                   |
| N (Participant) <sup>b</sup>    | 118               | 118                       | 118                     | 118               | 118                      | 118                    |

Note: <sup>a</sup>This variable was standardized before the analyses. <sup>b</sup> Data of PSB-empathy from one participant in the anodal group was missing. Reference levels in dummy variables were set as follows: Group = sham, Condition = control. Table also shows goodness-of-fit statistics: Significance: \*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001. Abbreviation: PSB: prosocial personality battery.

Table S5 Results of mixed-effect linear regressions predicting decision time (DT)

|  | all                | control <sup>b</sup> | bribe <sup>b</sup> |
|--|--------------------|----------------------|--------------------|
|  | <i>b</i>           | <i>b</i>             | <i>b</i>           |
|  | (SE)               | (SE)                 | (SE)               |
| Intercept  | 7.58***<br>(0.08)  | 7.56***<br>(0.08)    | 7.69***<br>(0.09)  |
| Group (Anodal)   | 0.03<br>(0.12)     | -0.005 (0.11)        | 0.06 (0.12)        |
| Group (Cathodal)   | -0.04<br>(0.12)    | -0.03 (0.11)         | 0.07 (0.12)        |
| Condition  | 0.04<br>(0.06)     |                      |                    |
| Offer Proportion   | -0.22***<br>(0.05) | -0.21***<br>(0.03)   | -0.15***<br>(0.03) |
| Decision   | 0.03†<br>(0.02)    | 0.14***<br>(0.02)    | -0.05*<br>(0.02)   |
| Group (Anoda) × Condition                                      | 0.01<br>(0.08)     |                      |                    |
| Group (Cathodal) × Condition                                   | 0.11<br>(0.08)     |                      |                    |
| Group (Anodal) × Offer Proportion                              | -0.07<br>(0.06)    |                      |                    |
| Group (Cathodal) × Offer Proportion                            | -0.01<br>(0.06)    |                      |                    |
| Condition × Offer Proportion                                   | 0.11† (0.06)       |                      |                    |
| Group (Anodal) × Condition × Offer Proportion                  | 0.11<br>(0.09)     |                      |                    |
| Group (Cathodal) × Condition × Offer Proportion                | 0.01<br>(0.09)     |                      |                    |
| Larger payoff for proposer in the reported option <sup>a</sup> | -0.01**<br>(0.005) | -0.01<br>(0.007)     | -0.02**<br>(0.007) |
| AIC  | 33637.4            | 16653.2              | 17095.3            |
| BIC  | 33776.9            | 16709.6              | 17151.7            |
| N (Observation)  | 17136              | 8568                 | 8568               |
| N (Participant)  | 119                | 119                  | 119                |

Note: <sup>a</sup> This variable was standardized before the analyses.

<sup>b</sup> We did not incorporate interactions between tDCS and offer proportion, as none of these effects was significant in the regression using all trials. Reference levels in dummy variables were set as follows: Group = sham, Condition = control, Decision = acceptance. Table also shows goodness-of-fit statistics: AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. Significance: †p < 0.08, \*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001.



Table S6 Descriptive statistics of task-relevant subjective rating

|   |         | Anodal<br>(N = 40) | Cathodal<br>(N = 39) | Sham<br>(N = 40) |
|---|---------|--------------------|----------------------|------------------|
| Perceived as bribe                            |         | 68.6 ± 31.4        | 67.6 ± 27.4          | 76.1 ± 27.4      |
| Sense of Power                                |         | 71.6 ± 30.9        | 77.9 ± 27.2          | 72.8 ± 29.1      |
| Moral conflict                                | bribe   | 42.2 ± 29.0        | 41.1 ± 31.8          | 36.9 ± 31.3      |
|   | control | 14.5 ± 22.1        | 6.3 ± 13.2           | 13.3 ± 24.0      |
| Guilt <sup>a</sup>                            | bribe   | 44.2 ± 32.8        | 48.0 ± 36.7          | 48.2 ± 37.7      |
|   | control | 14.2 ± 22.8        | 8.7 ± 17.3           | 11.8 ± 22.4      |
| Moral Inappropriateness:<br>Self <sup>a</sup> | bribe   | 56.7 ± 33.8        | 54.7 ± 34.6          | 60.8 ± 33.4      |
|   | control | 11.6 ± 21.0        | 13.9 ± 23.0          | 16.5 ± 25.8      |
| Moral Inappropriateness:<br>Proposer          | bribe   | 56.4 ± 34.0        | 51.3 ± 33.2          | 54.0 ± 33.6      |
|   | control | 25.0 ± 31.9        | 30.6 ± 36.6          | 39.5 ± 33.5      |

Note: <sup>a</sup> Ratings of these items in the bribe condition from one participants in the cathodal group was missing. Thus we dropped this participant for analyses on these two items.

Table S7 Descriptive statistics of other measures

|                  |                   | Anodal<br>(N = 40) | Cathodal<br>(N = 39) | Sham<br>(N = 40) |
|------------------|-------------------|--------------------|----------------------|------------------|
| MDMQ: pre-task   | AT <sup>a</sup>   | 35.2 ± 6.6         | 33.8 ± 6.5           | 35.5 ± 5.7       |
|                  | CN <sup>a,b</sup> | 39.4 ± 6.9         | 39.3 ± 6.7           | 40.2 ± 5.8       |
|                  | GB <sup>a</sup>   | 39.0 ± 5.0         | 40.4 ± 8.9           | 39.8 ± 4.9       |
| MDMQ: post-task  | AT                | 31.9 ± 7.5         | 30.4 ± 6.3           | 31.4 ± 7.8       |
|                  | CN                | 37.3 ± 7.5         | 38.1 ± 6.1           | 39.5 ± 5.9       |
|                  | GB                | 36.4 ± 5.9         | 37.0 ± 5.6           | 38.1 ± 5.7       |
| CRT              |                   | 0.9 ± 0.8          | 1.1 ± 0.9            | 0.8 ± 0.8        |
| Machiavellianism |                   | 58.0 ± 7.6         | 58.2 ± 7.4           | 57.8 ± 7.3       |
| PSB: Empathy     |                   | 80.2 ± 11.2        | 79.3 ± 10.8          | 76.2 ± 9.1       |
| PSB: Helpfulness |                   | 31.0 ± 4.4         | 28.4 ± 4.9           | 30.5 ± 4.1       |

Note: <sup>a</sup>Data of the pre-task MDMQ measures from one participant in the cathodal group was missing

<sup>b</sup>Data of pre-task MDMQ measures (only in CN subscale) from one participant in the sham group was missing.

Abbreviations: MDMQ: multidimensional mood questionnaire; subscales: AT: awake-tired, CN: calm-nervous, GB: good-bad; CRT: cognitive reflection ability; PSB: prosocial personality battery.



## Résumé

Le comportement des autres est souvent plus difficile à prédire que la physique des objets inanimés de notre environnement car il est dirigé par des buts, des croyances et des désirs dont l'accès direct reste souvent caché. La capacité de faire des inférences sur ces états mentaux est connue sous le nom de théorie de l'esprit. Une autre facette de cette théorie est la faculté d'attribuer des aptitudes, des capacités d'actions, ou des expertises à autrui. Ainsi, l'appréhension des aptitudes et l'attribution d'intentions aux autres sont deux éléments clés pour naviguer dans notre environnement social, notamment pour former des alliances et éviter de perdre des ressources lors de conflits. Le but de cette thèse est de mieux comprendre les mécanismes neurocomputationnels des attributions d'intention et des aptitudes hiérarchiques compétitives. L'attribution d'intention compétitive ou coopérative constituera notre premier axe d'étude, la construction mentale d'une hiérarchie constituera notre deuxième axe. Dans ces deux axes nous combinons les outils de la modélisation computationnelle du comportement et des approches de neuroimagerie (IRMf basée sur les modèles, TEP utilisant le traceur du transporteur de la sérotonine) et causale (stimulation magnétique transcrânienne). Dans une première partie, nous montrons comment lors d'interactions sociales simples, les individus attribuent des intentions compétitives ou coopératives à autrui pour ajuster leurs actions et maximiser leurs bénéfices. Nous étudions les mécanismes computationnels mis en jeu pour suivre la coopérativité d'un individu ou d'un groupe, utiles non seulement pour prédire les intentions des autres mais aussi pour simuler l'effet de ses propres actions sur celles des autres. Nous montrons en particulier que le striatum ventral et le cortex préfrontal calculent un signal de coopérativité de l'autre au moment d'un choix social, et que le cortex préfrontal dorsolatéral et la jonction temporopariétale calculent un signal différenciant interaction compétitive et coopérative. Nous montrons aussi comment la connectivité fonctionnelle entre ces deux régions permet la mise à jour dynamique de l'intention de l'autre. Nous mettons aussi en évidence par quels mécanismes neurocomputationnels les individus élargissent leur capacité d'attribuer des intentions à un groupe d'individus et comment ils font leurs choix afin d'influencer sur l'intention des autres. Dans une deuxième partie, nous montrons qu'il est possible d'étudier l'apprentissage d'une hiérarchie sociale par observation d'interaction dyadique en IRMf et de moduler sélectivement cet apprentissage en stimulant le cortex préfrontal médial avec un courant continu (tDCS). Grâce aux méthodes neurocomputationnelles nous montrons de plus que cette modulation provient de la perturbation de la capacité à encoder l'incertitude lors de l'apprentissage. Nous nous intéressons aussi à l'apprentissage d'une hiérarchie par interactions dyadiques et au rôle computationnel de la sérotonine dans la variabilité des comportements interindividuels dans ce type d'apprentissage. Nous montrons que le taux de transporteur disponible dans les noyaux du raphé explique les différences de vitesse d'apprentissage. De plus, nous montrons un lien direct entre disponibilité du transporteur de la sérotonine et l'encodage de la récompense attendu face à un individu lors de l'apprentissage du rang d'autrui dans une interaction sociale compétitive. Finalement, nous discutons de l'intérêt d'utiliser les méthodes neurocomputationnelles pour étudier les mécanismes sous-jacents aux capacités d'attribution d'intention et des aptitudes.

**Mots Clés:** Décision sociale; Hiérarchie; Intention; Compétition; Coopération.