# Approches computationnelles d'analyse des relations entre cerveau et comportement

Jules Brochard

SCIENCES SORBONNE UNIVERSITÉ

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

École doctorale n°158 ED3C

Cerveau, Cognition & Comportement

Institut du Cerveau — Équipe Motivation, cerveau & comportement

# Approches computationelles d'analyse des relations entre cerveau et comportement.

Thèse de doctorat en Neurosciences


Présentée par

Jules Brochard

Dirigée par

Jean Daunizeau


Présentée et soutenue publiquement le 15 janvier 2021

Devant un jury composé de :

| | | |
|---|---|---|
| Jean Daunizeau | Chargé de recherche, Sorbonne Université | Directeur de thèse |
| Andréa Brovelli | Chargé de recherche, - | Rapporteur |
| Alexandre Gramfort | Directeur de recherche, - | Rapporteur |
| Benoit Girard | Directeur de recherche, - | Examinateur |

Colophon This document was typeset in LaTeX, using the beautiful `tufte-latex` class[1]. I am forever thanksfull to Marie-Morgane Paumard for her assistance and the hand-made overlay inspired from her own thesis, *Solving jigsaw puzzles with deep learning*. Firmin Didot's `GFS Didot` acts as the typeface. The bibliography is typeset using `biblatex`.



*First printing, November 2020*

1. `tufte-latex` is based on on the work of the famous statistician Edward Tufte.

# *Remerciements*

Je remercie tout d'abord mon jury de thèse : Alexandre Gramfort, Andréa Brovelli et Benoît Girard, ainsi que les nombreuses personnes que nous avons sollicité pour ce jury, mais qui ont été refusées par l'école doctorale. Vos avis scientifiques, vos commentaires sur mon manuscrit et vos conseils m'ont permis de faire aboutir cette thèse, après plus de quatre longues années.

Il me faut également remercier les institutions, et l'État français, qui ont financé ma thèse. En premier lieu, l'École Normale Supérieure de Rennes qui m'a permis d'effectuer les trois premières années de ma thèse, à l'abri des demandes de bourses et autres anxiétés administratives. Puis le labex Biopsy qui a accordé sa confiance aux projets de recherche de Jean Daunizeau et Mathias Pessiglione auxquels j'ai été intégré.

Je n'aurais pas pu faire cette thèse sans les conseils des différents chercheurs de l'équipe « Motivational Brain Behavior » de l'Institut du cerveau à Paris. Côtoyer Mathias Pessiglione m'a permis de mieux appréhender la modélisation cognitive et son lien avec la neuroimagerie. Mes innombrables discussions avec Sébastien Bouret m'ont bien souvent permis de clarifier mes pensées, notamment vis-à-vis de l'électrophysiologie, d'étendre ma réflexion et de raviver mon enthousiaste sur mes thèmes de recherche. Raphaël Le Bouc, lui, a également toujours été une oreille attentive, m'offrant des conseils tantôt sur le bon usage d'une analyse IRMf, tantôt sur des questions médicales. Fabien Vinckier a aussi toujours été à mon écoute, bienveillant, compréhensif et rassurant, et je lui dois tout ce que je sais sur la psychiatrie. Enfin, Jean Daunizeau m'a accompagné tout au long de cette thèse, m'expliquant les bases de nombreuses approches statistiques, détaillant ses opinions scientifiques à chaque étape et me laissant libre d'explorer des méthodes très variées, couvrant un large spectre des neurosciences.

Et bien sûr, mes remerciements vont en très grande partie aux autres membres du laboratoire. À Lindsay Rondot pour sa bonne humeur perpétuelle particulièrement contagieuse et ses histoires sans fins. À Alizée Lopez-Persem que j'ai eu le plaisir de connaître à la fois comme mentor et comme amie. À Antonius Wiehler pour son écoute, sa gentillesse et son calme. À Emmanuelle Bioud pour ses conseils, sa malice et nos nombreuses soirées. Et bien sûr, à Nicolas Clairis pour toutes nos discussions, sérieuses, idiotes, politiques, scientifiques. Je suis parvenu au bout de cette thèse et devenu un neuroscientifique en très grande partie grâce à lui.

Enfin, et c'est de loin le plus important, je tiens à remercier ma fiancée Marie-Morgane Paumard, qui, en plus de m'apporter son amour et sa bienveillance, m'a soutenu à la fois émotionnellement durant cette longue thèse, mais également techniquement (je lui dois notamment le magnifique format de ce manuscrit) et scientifiquement. Discuter avec elle de ma recherche, de mon équipe, de mes objectifs et de mes difficultés m'a permis de prendre le recul dont j'avais besoin, de me concentrer sur l'essentiel et d'avancer plus sereinement.

# *Abstract*

Modern neuroimaging can measure the activity of our brain while we perform various cognitive tasks in response to a given context. Classic analysis methods can determine and quantify which areas of the brain are activated, when they are activated, and what information can be decoded from them. However, they rarely establish the complete scenario of neural computations that takes place, from initial perception of the environment to the execution of a motor action. This thesis focuses directly on this problem. I am developing and studying two analysis methods : the first of which implements a massively univariate mediation analysis of fMRI data and the second compares artificial neural networks constrained by physiological mechanisms. In both cases, they are evaluated i) mathematically, ii) through numerical simulations, and iii) through the analysis of risky decision-making experiments.

My first study focuses on mediation analysis. I investigate its statistical properties in the context of behavior production. By comparing five seminal statistical tests of mediation, I show that a conjunctive test is systematically valid and has a higher sensitivity than other tests. This difference grows stronger as the tests' statistical threshold is increased, which is a classic procedure in massively univariate fMRI analyses. I then identify a counter-intuitive but critical property of mediation analysis : its bilateral dependence on noise levels. When noise is non-existent in the data or too intense, no mediation can be detected. It follows that a mediation analysis will never be able to identify a complete information processing chain as the first steps are necessarily undetectable. Nevertheless, such an analysis can still identify some neural determinants of behavior. After showing how to adapt a massively univariate mediation analysis to avoid the inversion of the time series of each voxel, I apply this method to real data. I then identify six brain regions involved in decision making and show that the mediating strength of one of them, the left posterior dorsolateral prefrontal cortex, also predicts subjects' loss aversion.

My second study focuses on the physiological constraints imposed on neural decision-making processes. I investigate the dynamics of neural codes from either a range adaptation mechanism—predicted by the efficient coding theory—or Hebbian plasticity. This type of mechanism does allow a neural code to adapt to the statistical properties of its stimuli but also induces a form of instability, altering neural behavioral processes. By injecting these constraints into artificial neural networks, I show that they rely on unique internal representations specific to each constraint, when they are fitted to a series of choices between two options. I then use a model-based approach to fit these networks to real data—specifically the choices of several subjects in a decision-making experiment—and compare their internal representations to the multivariate fMRI activity of various brain regions. In doing so, I show that Hebbian plasticity, but not range adaptation, allows us to describe bilateral variations of the striatum and the amygdala of the participants in the experiment. Moreover, the Hebbian-ness of both the left striatum and the right amygdala predicts the degree of inconsistency in the subjects' choices.

Finally, I compare these two methods and the type of biological mechanisms that they detect and I discuss their extensions to brain network analysis.

Keywords FMRI, Decision-Making, Mediation, Inter-Individual Differences, Artificial Neural Networks, Physiological Constraints

# *Résumé*

La neuro-imagerie moderne permet de mesurer l'activité de notre propre cerveau pendant que nous employons divers processus cognitifs pour réagir à notre environnement. Les méthodes d'analyses classiques permettent de déterminer et de quantifier quelles aires cérébrales s'activent, à quel moment, et quelles informations peuvent y être lues. Cependant, elles permettent rarement de déterminer un scénario complet des calculs neuronaux, de la perception de notre environnement à l'exécution d'une action motrice. Cette thèse s'intéresse directement à ce problème. J'y développe et étudie une méthode d'analyse de médiation massivement univariée de données IRMf, et une méthode de comparaison de réseaux de neurones artificiels contraints par des mécanismes physiologiques. Dans les deux cas ces méthodes sont évaluées sur le plan mathématique, via des simulations numériques et par l'analyse d'expériences de prises de décisions risquées.

Ma première étude se concentre sur l'analyse de médiation. J'étudie tout d'abord ses propriétés statistiques, dans le contexte de la production du comportement. En comparant les principaux tests statistiques permettant une analyse de médiation, je montre qu'un test conjonctif est d'une part systématiquement valide, et d'autre part possède une plus grande sensibilité que les autres tests. Cette différence est d'autant plus marquée lorsque chaque test utilise un seuil statistique fort, habituel en analyse IRMf massivement univariée. J'identifie ensuite une propriété contre-intuitive mais essentielle de l'analyse de médiation : sa dépendance bilatérale aux niveaux de bruit dans les données. Lorsque le bruit est inexistant ou trop intense, aucune médiation ne peut être détectée. Il en découle qu'une analyse de médiation ne pourra jamais identifier une chaîne complète de traitement de l'information, les premières étapes étant nécessairement indétectables. Néanmoins, une telle analyse permet d'identifier certains déterminants neuronaux du comportement. Après avoir montré comment utiliser une analyse de médiation massivement univariée sans avoir à inverser la série temporelle de chaque voxel, j'applique cette méthode à des données réelles. J'identifie alors six régions cérébrales impliquées dans la prise de décision, et montre que la force de médiation de l'une d'entre elle, le cortex préfrontal dorsolatéral postérieur gauche, permet également de prédire l'aversion à la perte des sujets.

Ma seconde étude se concentre sur les contraintes physiologiques s'imposant aux processus de décisions neuronaux. Je m'intéresse notamment à la dynamique des codes neuronaux prenant la forme, soit d'une sensibilité adaptative, prédite par la théorie du codage efficace, soit d'une plasticité hebbienne. Si ce type de mécanisme permet au code neural de s'adapter aux propriétés statistiques des stimuli, il induit également une forme d'instabilité altérant les processus comportementaux neuronaux. En injectant ces contraintes dans des réseaux de neurones artificiels, je montre qu'en les ajustant à des séries de choix entre deux options, chacun d'entre eux développe une représentation interne unique de la valeur des options. J'emploie ensuite une approche model-based pour ajuster ces réseaux à des données réelles, les choix de plusieurs sujets lors d'une expérience de prise de décision, et je compare leurs représentations internes à l'activité IRMf multivariée de diverses régions cérébrales. Ce faisant, je montre que la plasticité hebbienne, mais pas la sensibilité adaptative, permet de décrire les variations bilatérales du striatum et de l'amygdale des participants à l'expérience. De plus, l'hebbiannité du striatum gauche et de l'amygdale droite permet également de prédire le degré d'incohérence dans leur choix. Enfin, j'effectue une comparaison de ces deux méthodes et du type de mécanismes qu'elles permettent d'étudier, puis je discute de leurs extensions à l'analyse de réseaux cérébraux.

Mots-clés IRMf, Prise de décision, Médiation, Différences inter-individuelles, Réseaux de neurones, Contraintes physiologiques

# *Préface*

Qu'est-ce qui détermine nos choix ? Qu'est-ce qui nous pousse à prendre des risques ? Pourquoi sommes-nous parfois irrationnels ? Ces questions agitent depuis longtemps les philosophes et les psychologues. Aujourd'hui, elles intéressent également les économistes, les mathématiciens et les biologistes. Chacun apporte sa grille de lecture, propose des règles et de nouvelles raisons pour comprendre le comportement humain. Au croisement de toutes ces approches se trouvent les neurosciences cognitives. Elles s'articulent autour d'une idée simple : celle que le cerveau est le support de l'esprit, et donc qu'étudier l'un revient à étudier l'autre.

Si cette idée est facile à énoncer, étudier la relation cerveau-esprit est une entreprise particulièrement délicate. Tout d'abord, il n'existe pas une théorie cognitive unique caractérisant nettement chacune de nos actions. Il en existe en fait une multitude, chacune décrivant nos mécanismes mentaux à partir d'opérations élémentaires différentes. Pour les uns, nos comportements dépendront avant tout de notre environnement social, pour les autres, d'associations inconscientes que l'on effectue quotidiennement, ou encore de complexes calculs utilitaristes maximisant notre satisfaction. Toutes ces théories peuvent cependant être comparées sur la base des mécanismes élémentaires qu'elles proposent, et notamment sur leurs implémentations neurales. Cela permet, d'une part, d'offrir un vaste champ de tests empiriques à ces approches, et d'autre part de relier ces modèles à notre organisation neurale. On peut ainsi étudier comment les comportements changent selon les organisations cérébrales de chacun, selon la sensibilité des systèmes neuraux impliqués, leur plasticité, leur degré de communication, etc. Analyser ces liens entre (des modèles du) comportement et activités cérébrales est l'objet central de cette thèse.

Toutefois, l'organisation du cerveau est encore mal comprise et l'étude des mécanismes neuraux qui déterminent le comportement reste encore balbutiante à bien des égards. On peut associer de manière relativement spécifique certaines régions cérébrales à des opérations cognitives, comme la détection de visage, l'analyse sémantique ou la préparation d'un mouvement, mais on comprend à peine les mécanismes biologiques sous-jacents. Comment une opération mentale est-elle physiquement réalisée par des neurones ? Est-elle réalisée par une région cérébrale précise, ou émerge-t-elle de l'interaction de plusieurs régions ? Comment cette organisation biologique se répercute-t-elle sur notre comportement ? De nombreux laboratoires s'attellent aujourd'hui à répondre à ces questions et développent de nouvelles techniques pour décrire et comprendre la production neurale du comportement.

Au cours du dernier siècle, les progrès des neurosciences ont crû exponentiellement : de nouveaux dispositifs permettent des mesures de plus en plus fines de l'activité neuronale, de nouvelles techniques mathématiques permettent de modéliser et d'interpréter cette explosion de données, et les théories cognitives unifient progressivement ces découvertes. Chaque combinaison d'un dispositif, d'une analyse et d'une expérience cognitive permet de tester des hypothèses de plus en plus fines. Cependant, chaque méthode est limitée, tant sur le plan technique que sur le plan conceptuel.

Dans cette thèse je m'intéresse à ce que l'imagerie par résonance magnétique fonctionnelle (IRMf) peut nous apprendre du comportement humain, et dans quelle mesure elle permet d'en quantifier les déterminants neurocognitifs. Plus précisément, j'étudie ici dans quelle mesure l'IRMf permet d'évaluer l'impact de l'organisation neuronale sur nos mécanismes décisionnels.

Dans le premier chapitre, j'introduis tout d'abord le contexte de mes travaux : la neuro-imagerie, ses outils, ses méthodes et son utilité pour les neurosciences cognitives. Le second chapitre est centré sur l'analyse quantitative des déterminants neurocognitifs du comportement. Comment se positionne ce problème au sein des neurosciences ? Quelles sont les approches existantes ? Cet état de l'art nous permettra d'identifier le contour neuroscientifique de mon sujet de thèse, et d'en déterminer les enjeux. Dans le troisième chapitre, j'étudie une méthode de détection des étapes intermédiaires des processus cérébraux de

traitement de l'information déterminant le comportement : l'analyse de médiation cerveau-comportement. Le quatrième chapitre présente une technique permettant d'étudier les contraintes biologiques qui s'imposent sur ces mêmes processus. Enfin, le cinquième chapitre compare les intérêts et limites de ces deux méthodes, et en propose quelques extensions.

# *Table des matières*

# *Liste des figures*

# *Liste des tableaux*

# Première partie

# INTRODUCTION

# *Synopsis*

Comment mesurer l'activité cérébrale ? Que faut-il en extraire ? Est-ce interprétable en termes psychologiques ? Au cours du dernier siècle, de nombreux travaux ont proposé de nouvelles méthodes, de nouvelles analyses, et de nouvelles théories pour expliquer le cerveau et son rapport à l'esprit. Dans cette partie, je présente successivement les modalités de neuro-imagerie, les méthodes d'analyses du cerveau et les interprétations cognitives des données IRMf. Plutôt que de donner des descriptions exhaustives et techniques, cette partie a pour principale ambition de vulgariser [2] la neuro-imagerie et la manière dont elle permet d'étudier l'esprit humain. J'y propose une sélection subjective des avancées scientifiques historiques afin d'offrir des clés de compréhension [3] permettant d'aborder les travaux, plus techniques, des parties suivantes.

En commençant par les dispositifs de mesures, je décris la fameuse balance circulatoire d'Angelo Mosso, ancêtre des principales techniques modernes de la neuro-imagerie : l'électroencéphalogramme (EEG), le magnéto-encéphalogramme (MEG), la tomographie par émission de positrons (PET en anglais) et enfin l'imagerie par résonance magnétique fonctionnelle (IRMf). J'aborde ensuite les grandes méthodes d'analyses ayant permis de comprendre le fonctionnement du cerveau et sa relation avec notre comportement : dissections, corrélations, modèles mathématiques et apprentissages automatiques. Puis, je m'intéresse aux apports plus spécifiques de l'IRMf aux théories cognitives : corrélats neuronaux, comparaison d'hypothèses et décodage d'informations. Cette partie se conclut enfin sur une revue des approches déjà mises en œuvre pour étudier les relations entre cerveau et comportement.

2. J'utilise cependant de nombreuses notes afin de compléter le texte de précision, d'éléments techniques, de détails d'expériences et d'anecdotes.

3. Ou plus modestement, celles que j'aurai aimé posséder en commençant ma thèse

# 1

# *Appareils de mesure : l'avènement de l'IRMf*

Chapitre 2 ➤

Synopsis Très populaire, l'IRMf n'est pas la seule manière d'étudier l'activité cérébrale. Son usage a émergé et évolué au milieu d'une multitude de dispositifs de mesures. Cette partie expose l'évolution et les caractéristiques des principales techniques de neuro-imagerie. J'y présente d'abord la première expérience de neuro-imagerie(§1.1), puis comment l'EEG (§1.2) et la MEG (§1.3) permirent d'abord de mesurer l'activité électromagnétique des neurones au niveau de notre scalp, ensuite comment les tomographies assistées par ordinateur (CT-scan en anglais) et les IRMs nous offrent une vue détaillée de l'anatomie cérébrale(§1.4), et enfin comment la PET (§1.5) et l'IRMf nous permettent aujourd'hui de mesurer l'activité de chaque parcelle de notre cerveau.

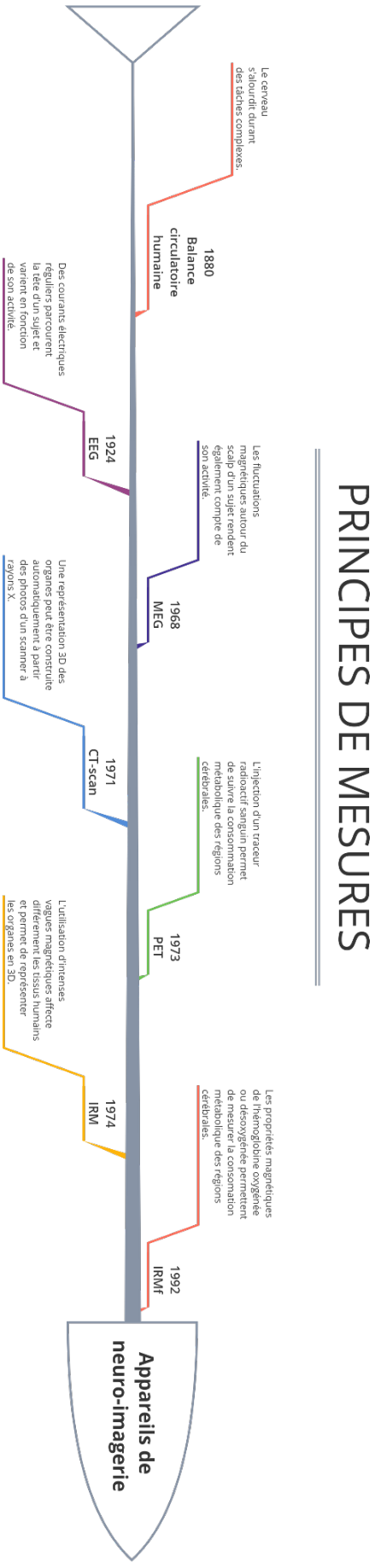



FIGURE *1.1* – *Résumé et chronologie des appareils de neuro-imagerie présentés.*

## 1.1 LA BALANCE CIRCULATOIRE HUMAINE

Une des premières études de neuro-imagerie remonte au moins aux années 1880 et aux travaux d'Angelo Mosso [1]. Fasciné par le système circulatoire sanguin, cet italien créa la « balance circulatoire humaine », une délicate bascule sur laquelle il allongeait ses sujets, puis leur demandait d'effectuer divers calculs mentaux. En comparant la difficulté des calculs aux oscillations de son dispositif, il pouvait ainsi mesurer le déplacement de masse induit par un effort mental. Il découvrit ainsi que plus un calcul était complexe plus le cerveau du sujet recevait de sang et s'alourdissait. On sait à présent que cet afflux de sang est dû au coût métabolique de l'activité neuronale, ou en d'autres termes, leur réapprovisionnement. Après avoir émis une série d'impulsions électriques, les neurones extraient plus d'oxygène du sang et l'organisme compense en augmentant le débit sanguin.

1. Sandrone, S., Bacigaluppi, M., Galloni, M., Cappa, S., Moro, A., Catani, M., and Martino, G. (2014). Weighing brain activity with the balance : Angelo mosso's original manuscripts come to light. *Brain*, 137(2):621–633

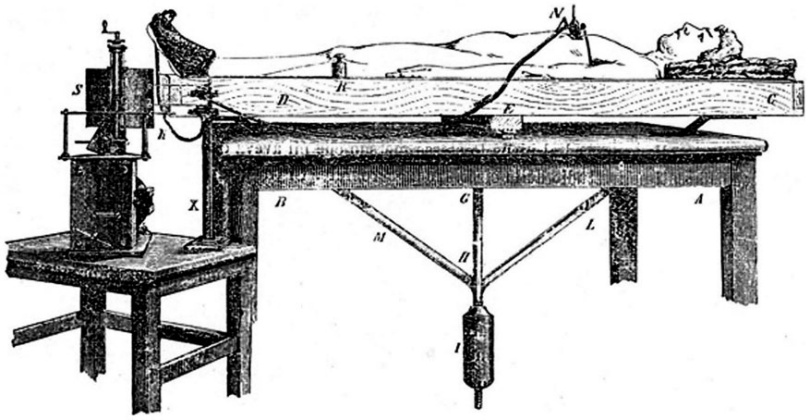


FIGURE *1.2 – Balance circulatoire humaine d'Angelo Mosso. Figure extraite de (Sandrone et al., 2014).*

Cette expérience d'Angelo Mosso fut la première étude de neuro-imagerie non invasive de l'histoire. Sans aucune procédure chirurgicale, il pouvait mesurer l'activité cérébrale de n'importe quelle personne. Sa mesure était peu fiable : le cerveau tout entier était réduit à une seule variable, et le décalage temporel entre l'oscillation de la balance et l'activité cérébrale n'était pas connue. Cependant, on pourrait considérer qu'Angelo Mosso avait, largement avant son temps, proposé le principe de la neuro-imagerie, et démontré sa faisabilité.

## 1.2 EEG

Une des premières avancées a eu lieu 40 ans plus tard, en 1924. On découvrit qu'en plaçant des électrodes directement sur le scalp d'un sujet, on pouvait y détecter d'infimes variations électriques. La première utilisation d'une électro-encéphalographie (ou EEG) sur des humains est attribuée à Hans Berger [2]. En s'appuyant sur les travaux animaliers (lapins et singes) de Richard Caton [3], il découvrit que ces variations électriques correspondaient à l'activité mentale

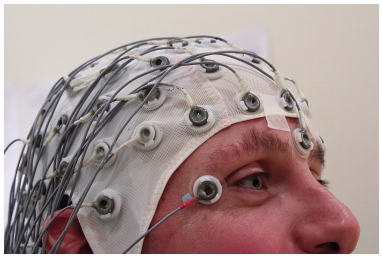


FIGURE *1.3 – Casque d'enregistrement EEG. Source : Wikimédia.*

2. Haas, L. and Caton, R. (2003). Hans berger (1873-1941. *Journal of Neurology, Neurosurgery, and Psychiatry*, 74(1):9

3. Caton, R. (1875). The electric currents of the brain. *British Medical Journal*, 2(765)

de ses sujets. Dans une série de vingt-trois publications, il étudia leurs niveaux d'attention, d'endormissement et d'effort mental en les comparants systématiquement à ses mesures EEG. Bien qu'il ne découvrit pas les phénomènes parapsychologiques qu'il cherchait initialement, son étude permit de mesurer l'activité cérébrale d'un sujet, en temps réel, sans procédure chirurgicale.

Avec sa précision à la milliseconde près, l'EEG permet ainsi de faire correspondre les fluctuations électriques d'un cerveau aux variations des états mentaux du sujet testé. Cette correspondance est critique pour la neuro-imagerie cognitive et ouvre la porte à une foule de questions subsidiaires. En manipulant l'environnement d'un sujet, en le confrontant à divers scénarios, on pouvait désormais étudier le traitement cérébral (conscient ou inconscient) de l'information.

## 1.3 MEG

Plus de 40 ans après, David Cohen 1968 [4] proposa une nouvelle approche s'appuyant sur le lien entre ondes électriques et magnétiques, la magnétoencéphalographie (ou MEG). Très proche de l'EEG, cette technique utilise non pas des électrodes mais des bobines d'induction placées à proximité du sujet. Les deux méthodes mesurent la même activité neuronale, mais la MEG détecte les variations non pas électriques, mais magnétiques de cette activité [5]. Cela permet à une MEG d'être plus sensible à l'orientation des courants qu'un EEG, et de déterminer plus précisément leur source [6] à l'intérieur du crâne du sujet [7].

Si l'EEG et la MEG permettent de suivre en temps réel l'activité cérébrale moyenne d'un sujet, ces techniques n'offrent qu'une résolution spatiale limitée. Elles permettent de mesurer si le cerveau répond à un stimulus, mais n'offrent qu'une information réduite sur les régions ayant contribué à cette réponse.

## 1.4 CT-SCAN ET IRM

En 1971, Alan Cormack et Godfrey Hounsfield développèrent indépendamment la tomographie à rayon X (ou CT-scan : *computer-assisted-tomography-scans* en anglais), ou, en d'autres mots, une méthode pour combiner plusieurs images en une représentation en trois dimensions des organes et des os [8]. Trois ans après, Raymond Damadian déposa un brevet pour une machine à résonance magnétique nucléaire (appelée aujourd'hui IRM : imagerie par résonance magnétique). Plutôt qu'utiliser l'absorption de rayons X pour différencier les tissus, une IRM les différencie selon leur réaction à de puissants champs magnétiques [9].

Chaque technique répond à des contraintes pratiques différentes, mais toutes deux ont permis aux neurologues de visualiser en détail l'organisation cérébrale de leurs patients. Cela permet, entre autres, d'étudier les différences anatomiques et de les comparer à

FIGURE *1.4 – Machine d'imagerie MEG. Source : Wikimédia.*

4. Cohen, D. (1966). Magnetoencephalography : Evidence of mapetic fields produced by alpha-rhylhm currents. *Science*, 161

5. Les deux mesures étant liées. Lorsque le champ électrique autour du scalp varie, il génère un champ magnétique (quatrième équation de Maxwell). Ce champ, à son tour, va générer un nouveau champ électrique dans les bobines d'inductions (troisième équation de Maxwell), qui pourra alors être mesuré.

6. Silva, F. (2013). Eeg and meg : Relevance to neuroscience. *Neuron*, 80(5):1112–1128

7. Pour être détectable, les activités de 10 000 à 50 000 neurones doivent être synchronisées.

8. Une analyse IRM utilise les différences d'absorption des rayons X par les différents tissus du corps. Comme un même tissu va absorber les radiations de la même manière quelques soit l'angle du rayon, on peut l'identifier à travers plusieurs photos. En variant le point focal des rayons on peut ainsi obtenir une bonne résolution d'une zone précise, et ainsi observer un organe sans autre procédure chirurgicale. Cette technique value d'ailleurs le prix Nobel à Cormack et Hounsfield le prix Nobel en 1979.

des différences comportementales. Eleanor Maguire et ses collègues montrèrent ainsi en 2000 que des conducteurs de taxi avaient un plus large hippocampe qu'un groupe témoin, mais également que la taille de cette région était associée à leurs années d'expérience [10]. L'hippocampe étant associé à la capacité à se repérer dans l'espace, cette étude démontra surtout l'adaptabilité de l'anatomie cérébrale. Elle suggère que plus l'importance et le besoin d'exécuter une tâche cognitive sont conséquents, plus la taille des structures impliquées augmente.

Certes, ces techniques de tomographie n'apportent pas de preuves définitives de telles associations, mais néanmoins elles permettent d'étudier un autre aspect de la relation entre le cerveau et le comportement. Non pas en termes d'activité cérébrale, mais en termes d'organisation et d'altérations structurelles.

9. Plus précisément, une machine IRM plonge le sujet dans un premier champ magnétique intense (un million de fois le champ magnétique terrestre) qui oriente tous les atomes d'hydrogènes du sujet dans une même direction. Ensuite l'IRM émet un second et bref champ perturbant cette orientation. Les atomes d'hydrogènes, toujours plongés dans le premier champ, vont alors progressivement revenir à leur orientation initiale. C'est la vitesse de ce retour qui est mesuré dans une IRM mesure. Puisque cette vitesse dépend de la densité du milieu dans lequel les atomes se trouvent, l'IRM permet de distinguer les os et différents types de tissues.
10. Eleanor Maguire reçu d'ailleurs un prix « Ignobel » pour ces travaux.

[]  []

Figure 1.5 – *À gauche : Machine d'imagerie CT-scan, à droite : Machine d'imagerie IRM. Source : Wikimédia.*

## 1.5 PET

En 1975, Michel Ter-Pogossian, Michael Phelps, Edward Hoffman et Nizar Mullani [11] publièrent leur propre technique de tomographie par émission de positrons (abrégée PET en anglais). En suivant un traceur radioactif injecté dans le sang d'un sujet, cette technique permet de reconstruire le réseau veineux d'un patient. La particularité majeure de cette troisième tomographie est qu'elle dépend de la quantité de sang se trouvant dans une zone. Plus une zone reçoit de sang, plus on y détectera le traceur. En comparant l'intensité du traceur entre deux images, on peut déduire les variations de débit sanguin d'une région : son hémodynamique. Tout comme la balance d'Angelo Mosso, les variations identifiées par la PET peuvent être associées à diverses tâches cognitives. Il fallut cependant attendre 25 ans avant que Michael Posner, Steven Peterson, Peter Fox et Marcus Raichle utilisent cette méthode pour étudier une fonction cognitive : la lecture [12] [13]. Grâce à la PET, ils mesurèrent l'hémodynamique de chaque région cérébrale d'une dizaine de sujets tandis qu'ils lisaient

11. Ter Pogossian, M., Phelps, M., Hoffman, E., and Mullani, N. (1975). A positron emission transaxial tomograph for nuclear imaging (pett. *Radiology*, 114(1):89–98

12. Fait amusant, l'institut ayant permis à Raichle de mener son étude fut fondé par James McDonnell dont l'intention première était la même qu'Hans Berger, l'étude des phénomènes parapsychologiques.
13. Posner, M., Petersen, S., Fox, P., and Raichle, M. (1988). Localization of cognitive operations in the human brain. *Science*, 240(4859):1627–1631

ou écoutaient des mots. Dans certains cas, ils devaient juste les répéter, et dans d'autres, en proposer un usage. En comparant les images obtenues dans chaque condition, ils déterminèrent qu'une zone, le cortex préfrontal inférieur gauche, recevait plus de sang lorsque le sujet devait activement comprendre le sens des mots.

Bien que la résolution temporelle de la PET soit nettement inférieure à celle offerte par un EEG ou un MEG (minute vs. milliseconde), elle possède la même précision spatiale que les autres méthodes de tomographie. Elle permet ainsi d'étudier les opérations cognitives sous un nouvel angle : celui de la localisation de leur coût métabolique.

FIGURE *1.6 – Machine d'imagerie PET. Source : Wikimédia.*

## 1.6 IRMF

En 1992, soit 15 ans après la première machine IRM, mais juste quatre ans après les travaux de Raichle, trois groupes proposèrent d'utiliser les IRMs pour localiser anatomiquement les opérations cognitives cérébrales alors même qu'elles étaient effectuées par un sujet [14] [15] [16]. Cette nouvelle approche, baptisée IRM fonctionnelle étend fortement les perspectives de la PET, puisqu'elle se passe de traceur radioactif. L'IRMf s'appuie sur un résultat établi par Linus Pauling en 1936 : les propriétés magnétiques du sang diffèrent lorsque celui-ci perd son oxygène [17] [18]. C'est-à-dire que le sang réagira différemment aux champs magnétiques après avoir alimenté une région cérébrale. Ainsi, une machine IRM peut mesurer la consommation métabolique d'une région cérébrale, tout comme la PET. En 1992, les trois groupes montrèrent que cette nouvelle mesure, nommée imagerie dépendant du niveau d'oxygène dans le sang (abrégée BOLD pour *blood oxygen level dependent imaging*), varie effectivement selon l'activité mentale d'un sujet. Par exemple, lorsqu'un sujet observe des motifs géométriques, son cortex visuel s'allume : son activité BOLD augmente par rapport à son niveau normal. Lorsqu'il bouge les doigts, c'est son cortex moteur qui s'allume. L'IRMf promet ainsi de pouvoir faire aussi bien que la PET, mais bénéficie de trois avantages cruciaux : sa résolution temporelle est meilleure (secondes vs. minutes), elle ne nécessite pas d'injection de traceur radioactif dans l'organisme, et les machines IRM, bien qu'onéreuses, sont déjà répandues dans les hôpitaux.

Malgré cela, les études d'IRMf furent longtemps remises en cause. Contrairement à l'EEG ou la MEG, le signal BOLD ne mesure pas directement l'activité neuronale. Il mesure son coût métabolique qui, lui, dépend de nombreux autres facteurs : par exemple, le rythme cardiaque du sujet, sa consommation de caféine, ou ses mouvements lorsqu'il est allongé dans le scanner. Ce n'est qu'en 2001, soit onze ans après les premières études d'IRMf, que le lien entre BOLD et activité neuronale fut établi expérimentalement par Nikos Logothethis [19]. Il travaillait sur des singes anesthésiés dont il mesurait à la fois les impulsions neuronales par des électrodes internes, le champ électrique entourant ces neurones et le signal BOLD de la région. En présentant aux singes des séries de motifs géométriques, il provoquait

14. Bandettini, P., Wong, E., Hinks, R., Tikofsky, R., and Hyde, J. (1992). Time course epi during task activation. *Magnetic Resonance in Medicine*, 25:390–397

15. KWONG, K., BELLIVEAU, J., CHESLER, D., GOLDBERG, I., WEISSKOFF, R., PONCELET, B., and ROSEN, B. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679

16. Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation : Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5951–5955

17. Le sang oxygéné n'est pas magnétique, tout comme le sang désoxygéné. En revanche ce dernier est paramagnétique, il acquiert des propriétés magnétiques quand il est soumis à un champ intense.

18. Pauling, L. and Coryell, C. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, 22(4):210–216

19. Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157

une décharge des neurones, qu'il compara ensuite aux deux autres mesures. Ses résultats furent sans équivoque. Le champ électrique local suivait fidèlement l'activité neuronale, tout comme le signal BOLD qui, lui, mettait plusieurs secondes à réagir. Cette étude et la série d'articles complémentaires qui suivirent levèrent le doute sur les origines neurales du signal BOLD.

Malgré cela l'IRMf reste régulièrement critiquée dans le milieu des neurosciences cognitives, notamment pour la manière dont le signal est analysé. Par exemple, en 2016, un article défrayait la chronique en prétendant que 70% des études IRMf étaient fausses à cause d'une erreur statistique [20]. Des articles similaires sont publiés à peu près tous les ans [21].

S'il ne faut pas ignorer les problèmes qu'elles identifient, la portée de ces critiques est souvent surestimée. Ces travaux étudient généralement une technique mathématique précise, en décrivent les limitations et pointent les mauvaises utilisations qui peuvent en être faites. Cependant les études IRMf ne se résument pas à l'utilisation d'une technique mathématique. Elles élaborent un argumentaire logique, s'appuient sur des études similaires, se positionnent par rapport à d'autres travaux et théories. Elles sont le fruit d'une réflexion s'étendant au-delà de l'article lui-même et ne reposent que rarement sur une seule analyse mathématique. En bref, prétendre que 70% des études sont fausses ignore entièrement l'écosystème d'un projet scientifique.

20. Eklund, A., Nichols, T., and Knutsson, H. (2016). Cluster failure : Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413

21. Au moment d'écrire ces lignes deux articles viennent d'ailleurs d'être publiés en ce sens (Helmera et al., 2020 ; Marek et al., 2020).

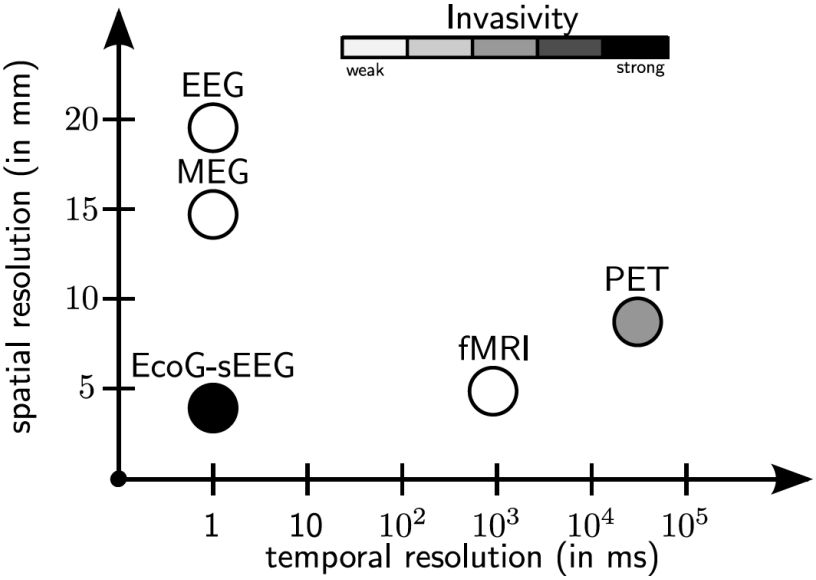



FIGURE *1.7 – Précision temporelle et spatiale des techniques de neuro-imagerie. EcoH-sEEG correspond à l'enregistrement des champs électriques locaux autour des neurones à l'aide d'électrodes implantées dans le cerveau. Plus un cercle est sombre, plus la technique est invasive. Figure extraite de (Olivi, 2011).*

## 1.7 CONCLUSION SUR LES APPAREILS DE MESURES

Depuis une centaine d'années, les méthodes de neuro-imagerie ont permis de mesurer de plus en plus finement notre activité cérébrale. On peut à présent cartographier le cerveau de n'importe qui à l'échelle

d'un grain de sel, mesurer ses réactions à la milliseconde près et étudier la propagation d'un signal à travers tout le cerveau. Le cerveau n'est plus une « boîte noire » insondable.

Cependant, l'interprétation de ces signaux s'avère aussi complexe que leurs captures. Mesurer n'est pas comprendre. La richesse de ces dispositifs se mesure autant à leur degré de précision qu'aux analyses et aux raffinements théoriques qu'ils permettent.

# 2

# *Principes d'analyses : Du scalpel à l'apprentissage automatique*

◂ Chapitre 1 Chapitre 3 ▸

Synopsis Il va sans dire que la méthode scientifique a fortement évolué au cours des siècles, et l'étude du cerveau et de l'esprit ne fait pas exception. En partant de la dissection anatomique, les méthodes de déduction ont progressivement évolués pour inclure des modèles cognitifs et comportementaux, et l'utilisation des statistiques inférentielles. Elles s'appuient désormais sur la personnalisation de modèles mathématiques. Chaque méthode éclaire ainsi le fonctionnement du cerveau et du comportement d'une manière différente. Ce chapitre présente leur évolution et leur contribution à la crédibilité des méthodes modernes de neuro-imagerie [1]. J'y aborde les premiers travaux identifiant le cerveau comme siège de nos perceptions, puis la naissance des statistiques comme outils de déduction scientifique, la modularité fonctionnelle du cerveau, les approches comportementales, la comparaison systématique des cerveaux, et enfin les modèles cognitifs du traitement cérébral de l'information.

1. Je n'aborderai cependant pas les méthodes de microbiologie car elles sont moins directement liées à mes travaux.

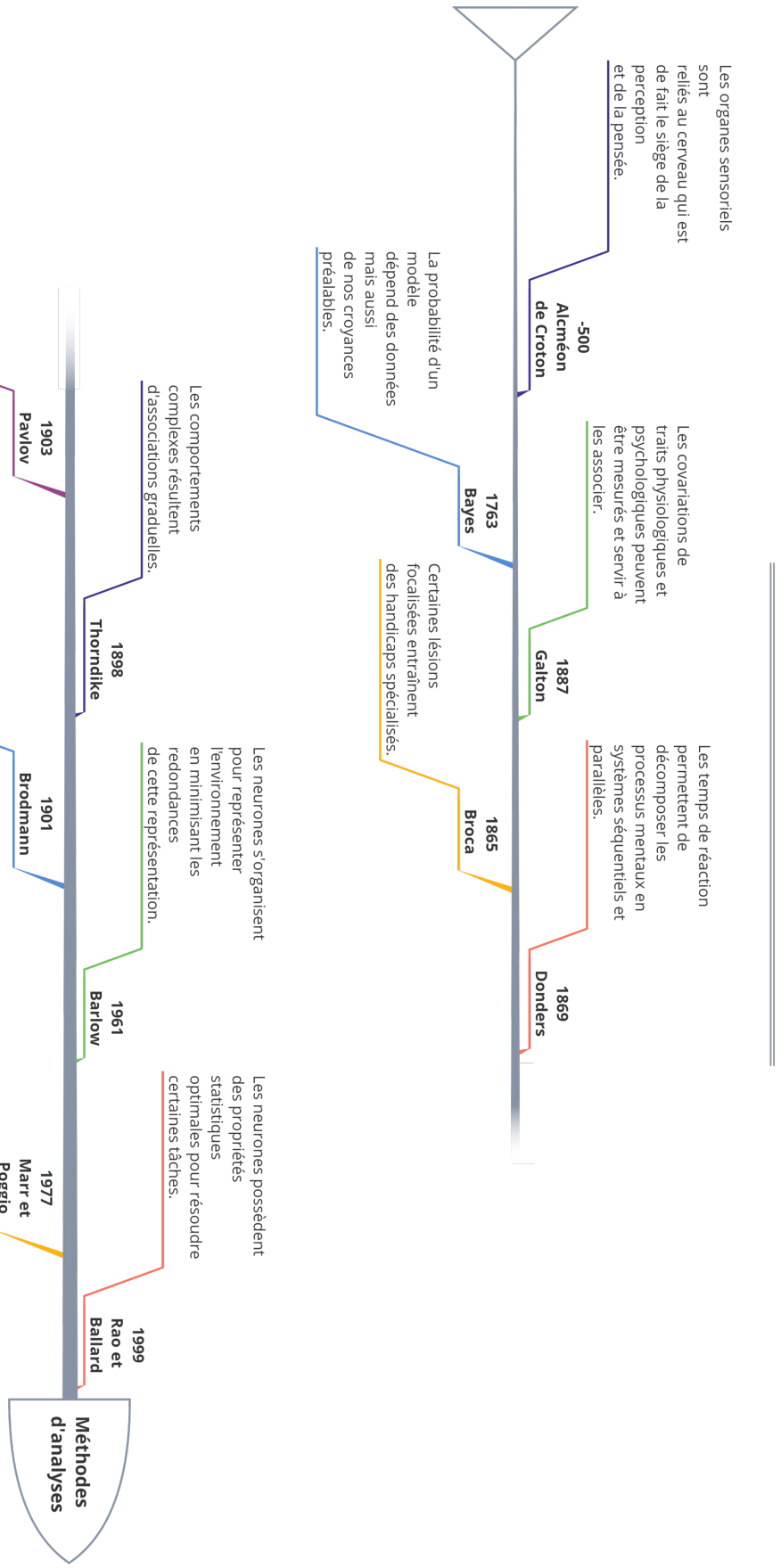



FIGURE 2.1 – *Résumé et chronologie des méthodes d'analyses abordées.*

## 2.1 Alcméon de Crotone et les premières dissections

Les premières méthodes d'analyse du cerveau remontent au 5e siècle avant notre ère, dans la ville portuaire de Crotone en Italie. À l'époque, elle accueillait une fameuse école de médecine, où travaillait Alcméon de Crotone [2]. Pionnier des méthodes de dissection, il travaillait sur la perception humaine et notamment le nerf optique.

En réalisant des cartes du système nerveux humain, il proposa une idée controversée : le cerveau serait le centre névralgique de la pensée et des sens [3]. À l'époque, ces qualités étaient plus volontiers attribuées au cœur. Alcméon est l'un des premiers, dans le monde occidental, à faire le lien entre le cerveau et la pensée.

Figure 2.2 – *Médaille de bronze à l'effigie d'Alcméon. Source : Wikimédia.*

2. Médecin, physiologiste, astronome, philosophe et contemporain de Pythagore c'est notamment à lui que l'on doit la popularisation de la théorie des humeurs

3. Huffman, C. (2017). *Alcmaeon*. The Stanford Encyclopedia of Philosophy

## 2.2 Bayes et la mise à jour des croyances

Bien plus tard, en 1763, les travaux du révérend Thomas Bayes furent publiés à titre posthume [4]. Dans ces essais, il proposa une formule mathématique qui attise depuis controverse et fascination [5]. Avant même d'être utilisée comme clé d'interprétation du cerveau [6], cette formule offre un outil scientifique de premier ordre : elle permet de mesurer la plausibilité d'un modèle, en fonction, d'une part, des données expérimentales et, d'autre part, de nos croyances préalables en ce modèle.

Depuis lors, ce dernier point fait débat. Certains reprochent à la formule d'introduire de la subjectivité dans l'exercice scientifique, là où d'autres défendent l'utilisation des connaissances préalablement établies.

Les travaux de Bayes fondèrent néanmoins l'un des plus grands courants des statistiques modernes : les statistiques bayésiennes. Ce courant étudie le degré de confiance à donner à chaque modèle scientifique, en dérivant la formule de Bayes pour chacun d'entre eux. Aujourd'hui, ces techniques permettent de comparer les modèles scientifiques entre eux, et de déterminer le plus probable.

Figure 2.3 – *Révérant Thomas Bayes. Source : Wikimédia.*

4. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4)

5. Hoang, L. (2018). La formule du savoir

6. Le cerveau bayésien est une théorie considérant que notre système nerveux représente l'information par des distributions de probabilité et s'organise pour constamment les mettre à jour par une approximation de la formule de Bayes.

## 2.3 Galton et la corrélation

En 1887, Francis Galton popularisa le concept de corrélation [7]. Père fondateur de la psychométrie [8], il s'intéressait à l'intelligence humaine et aux mécanismes de l'hérédité. Sa technique de « co-relation » permet de comparer systématiquement deux traits psychologiques [9] et de déterminer leur degré d'association. Son intérêt majeur est qu'elle prend explicitement en compte les différences entre les individus mesurés. Deux traits n'ont pas besoin d'être parfaitement liés pour qu'on puisse établir l'existence d'une relation. Chacun peut être influencé par des facteurs inconnus, tant que leur influence n'est pas trop forte. Une corrélation mesure justement le ratio [10] des variations communes aux deux traits, par rapport aux variations inexpliquées. Ainsi, si les fluctuations d'un trait permettent de suffisamment bien

décrire celle du second, on les considérera associés. Pour les analyses IRMf, si les variations d'un stimulus corrèlent avec les fluctuations du signal BOLD d'une région, on considérera que la région est associée au traitement du stimulus [11].

Les travaux de Galton fondèrent [12] le second grand courant des statistiques modernes : les statistiques « fréquentistes ». Ce courant étend le principe de la corrélation à d'autres formes d'associations et en mesure la fiabilité. Omniprésent dans les articles scientifiques, c'est la méthode de référence pour évaluer la solidité des résultats expérimentaux. Les statistiques fréquentistes doivent leur succès au fait de formaliser élégamment le caractère falsifiable d'une théorie scientifique (voir Karl Popper et la théorie de la falsifiabilité [13]). Si le hasard peut expliquer facilement le résultat, alors il n'est pas possible de considérer que les données valident la théorie. C'est la raison d'être de la fameuse « p-value », grâce à laquelle les statistiques fréquentistes mesurent la facilité avec laquelle le hasard (i.e.l'hypothèse nulle) explique les résultats.

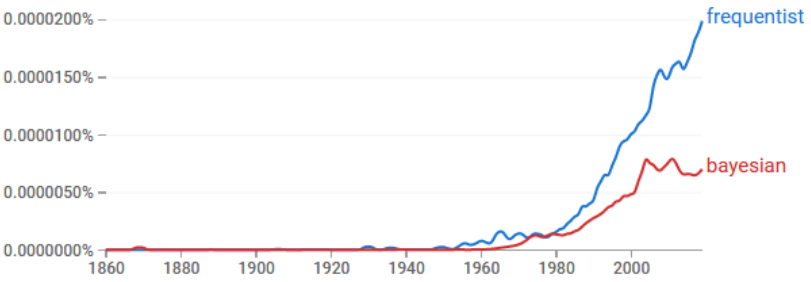



FIGURE 2.6 – *Fréquence d'apparition des mots "fréquentiste" et "bayésien" dans les livres anglais depuis 1860. Source : googleNgram.*

## 2.4 BROCA, WERNICKE, GAGE ET LA LOCALISATION DES FONCTIONS

Renouant avec les méthodes de dissections d'Alcméon, Pierre Broca démontra qu'une partie du cerveau pouvait être spécifiquement associée à un mécanisme mental. En 1865, il publia une étude exposant la dissection post mortem de douze patients, tous présentant une lésion de la même région du lobe frontal gauche, et ayant tous eu une aphasie avant leur mort : ils avaient perdu la capacité à parler, tout en conservant la compréhension du langage [14]. En 1874, Carl Wernicke montra le phénomène inverse : ses patients ne comprenaient plus le langage parlé, mais pouvaient articuler normalement [15]. Tous présentaient une lésion du lobe temporal gauche. La génération et la compréhension du langage apparaissent donc comme deux fonctions dissociées dans le cerveau humain.

Cependant, l'exemple historique le plus spectaculaire est probablement celui de Phineas Gage : après avoir eu le lobe frontal gauche traversé par une barre de fer, ce contremaître des chemins de fer changea fortement de personnalité. Son docteur, John Harlow décrivit en 1868 que Gage était passé d'un employé modèle, assidu et apprécié de ses pairs, à un homme aux comportements impulsifs,

$$P(\text{modèle}|\text{données}) = \frac{P(\text{données}|\text{modèle})}{P(\text{données})} P(\text{modèle}) \quad (2.1)$$

FIGURE 2.4 – *Formule de Bayes. Elle permet d'exprimer la probabilité d'un modèle aux vues de nouvelles données en fonction de 3 termes : la probabilité des données en supposant qu'elles proviennent directement du modèle, la probabilité des données indépendamment d'un modèle et enfin la probabilité préalable que l'on accorde au modèle.*

FIGURE 2.5 – *Francis Galton. Source : Wikimédia.*

7. Galton, F. (1888). Co-relations and their measurement , chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45:135–145
8. C'est à dire la mesure des capacités psychologiques humaines.
9. ou physiologique.
10. Techniquement : la racine carrée du ratio.
11. Cependant cette association ne dit rien des détails de ce lien. Plusieurs scénarios sont possibles : 1) le stimulus cause l'activité neuronale et est effectivement traité par la région 2) le stimulus cause l'activation d'un autre système cérébral, qui a son tour entraîne l'activité neuronale 3) l'activité neuronale et le stimulus sont effectivement concomitant, mais ils sont tous deux liés à une autre variable (le temps depuis le début de l'expérience par exemple) et n'ont pas d'autre type de lien etc. Distinguer ces scénarios est bien souvent difficile, et cette difficulté a donné lieu au célèbre adage : "corrélation ne fait pas causation".
12. Notamment par le biais de son protégé Karl Pearson.
13. Popper, K. R. (1989). *Logik der Forschung*, volume 9. JCB Mohr Tübingen

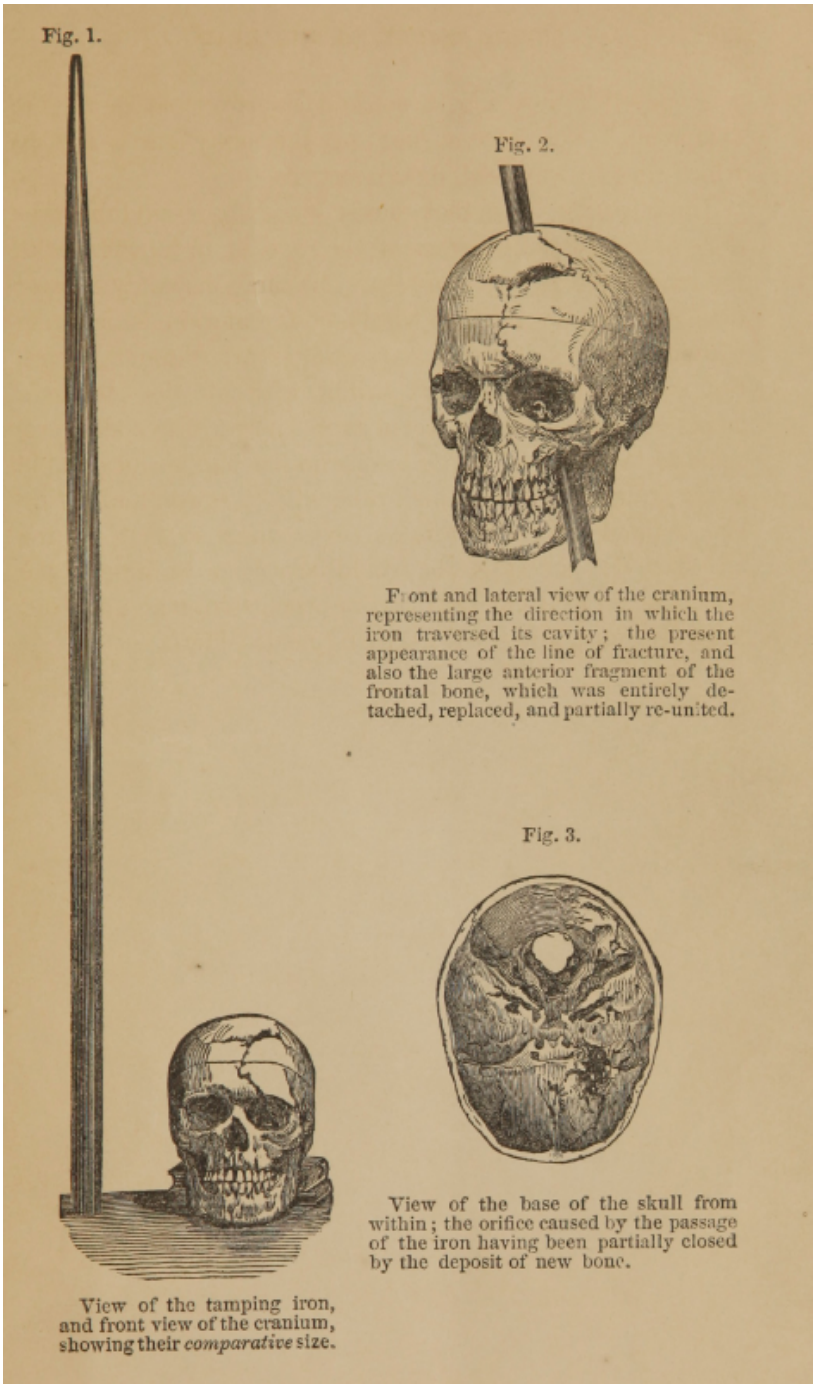


Fig. 1.

View of the tamping iron, and front view of the cranium, showing their *comparative* size.

Fig. 2.

Front and lateral view of the cranium, representing the direction in which the iron traversed its cavity; the present appearance of the line of fracture, and also the large anterior fragment of the frontal bone, which was entirely detached, replaced, and partially re-united.

Fig. 3.

View of the base of the skull from within; the orifice caused by the passage of the iron having been partially closed by the deposit of new bone.

FIGURE 2.7 – *Extrait du livre de John Harlow décrivant l'accident de Phineas Gage.*

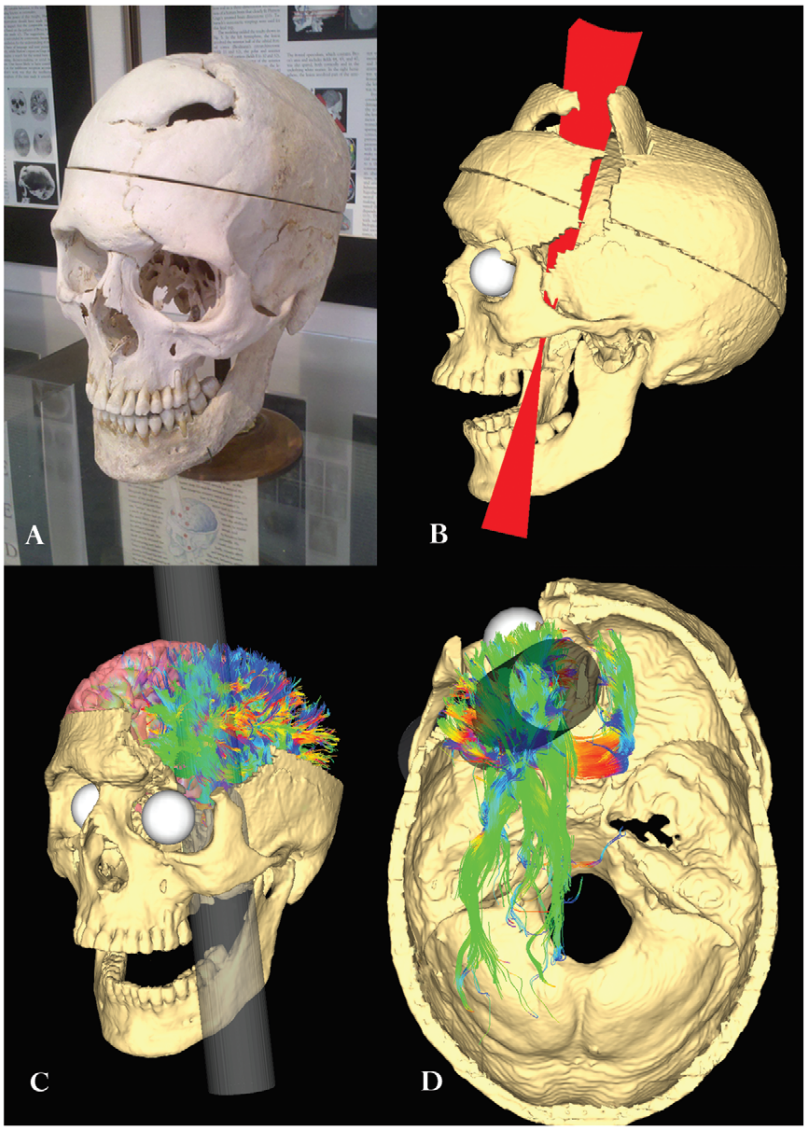



FIGURE 2.8 – *Reconstruction moderne des lésions subies par Phineas Gage. Figure extraite de (Horn et al., 2012).*

grossiers et capricieux[16]. Ses autres capacités mentales semblaient relativement peu affectées[17]. Gage démontra ainsi que notre personnalité, aussi complexe soit-elle, est la résultante du fonctionnement, et donc de l'intégrité, de certaines régions cérébrales[18]. La spécialisation fonctionnelle du cerveau ne concerne donc pas uniquement nos capacités mentales élémentaires, telles que la perception et le langage. La neuropsychologie était née.

De nos jours, l'étude des déficits comportementaux induits par des lésions cérébrales reste une des méthodes les plus convaincantes pour étudier le rôle d'une région cérébrale[19]. Cependant, face à la rareté des patients présentant une lésion anatomique focale, et face à la multitude des comportements à évaluer, ce type d'étude a été progressivement remplacée par la neuro-imagerie sur des volontaires « sains ». Cela dit, la nature de la relation cerveau-comportement que l'on met en évidence est différente. Plutôt que de rechercher l'implication causale d'une région cérébrale dans un comportement ou un processus cognitif, on se limitera à évaluer la force du lien statistique existant entre ces processus et l'activité cérébrale. Nous discuterons de l'importance de ces différences au prochain chapitre.

14. Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins de La Société d'anthropologie de Paris*, 6(1):377–393
15. Wernicke, C. (1970). The aphasic symptom-complex. a psychological study on an anatomical basis. *Archives of Neurology*, 22(3):280–282
16. Harlow, J. (1868). *Passage of an iron bar through the head*. Massachusetts Medical Society
17. Par la suite Gage récupéra l'essentiel de ses facultés de contrôles.
18. en particulier : le lobe frontal
19. Vaidya, A., Pujara, M., Petrides, M., Murray, E., and Fellows, L. (2019). Lesion studies in contemporary neuroscience. In *Trends in Cognitive Sciences*, volume 23

## 2.5 DONDERS ET LA CHRONOMÉTRIE MENTALE

En 1869, Franciscus Donders proposa une méthode à la fois simple et versatile pour étudier le cerveau, bien avant l'avènement de la neuro-imagerie[20]. Donders compara les temps de réaction de ses sujets lorsqu'ils effectuaient deux tâches similaires. Dans chacune d'entre elles, ils réalisaient un mouvement simple, comme appuyer sur un bouton. Dans la seconde, ils devaient en plus effectuer une opération mentale *avant* de bouger : un calcul mathématique par exemple. En soustrayant les deux temps de réaction, Donders proposa qu'on puisse mesurer la vitesse de réflexion des sujets : le temps pris par l'opération mentale. Certes, la méthode « soustractive » ne tient pas compte de l'existence d'interactions entre les différentes composantes cognitives et/ou motrices des tâches, mais elle séduit par sa simplicité. Judicieusement adaptée, elle permet notamment d'étudier le degré « d'automatisation » des traitements cérébraux de l'information[21]. Ce type d'approche a notamment permis de mettre en évidence l'existence de deux systèmes distincts (l'un automatique/inconscient/rapide et l'autre, contrôlé/conscient/lent) entrant en compétition pour le contrôle du comportement[22].

De nos jours, le principe de soustraction popularisé par Donders est omniprésent. En neuro-imagerie, par exemple, le signal BOLD est rarement analysé directement. Il est comparé à lui-même entre deux conditions expérimentales, qui diffèrent par la présence ou l'absence du processus cognitif d'intérêt. On interprétera alors les différences du signal BOLD comme résultant de la présence ou de l'absence du processus d'intérêt. La spécificité du résultat dépend donc de la correspondance des deux conditions. S'il existe des différences entre les conditions qui ne relèvent pas du processus d'intérêt, alors ces

FIGURE *2.9 – Franciscus Donders. Source : Wikimédia.*

20. Donders, F. (1969). On the speed of mental processes. *Acta Psychologica*, 30(C):412–431
21. Par définition, un processus est automatique s'il n'est pas altéré par l'exécution d'une autre tâche effectuée en parallèle. On peut donc évaluer le degré d'automatisation d'un processus en mesurant la différence de temps de réaction entre une condition où le processus est isolé et une condition « duale », où le processus est sollicité pour l'exécution de deux tâches effectuées en parallèle.
22. Ses travaux sur les deux « vitesses de la pensée » (Kahneman, 2011) ont valu à Daniel Kahneman le prix Nobel d'économie

différences peuvent expliquer les différences du signal BOLD. On parle de « facteurs confondants ». Tout l'enjeu du design expérimental est donc de minimiser ces facteurs confondants.

## 2.6 Pavlov et l'établissement des réflexes

Ivan Petrovich Pavlov est probablement l'un des physiologistes les plus célèbres du siècle dernier, notamment pour ses travaux sur le conditionnement classique en 1903 [23]. Il s'intéressait aux biomarqueurs du comportement : des mesures biologiques indiquant le comportement dans lequel un animal s'engage. Ses travaux montrèrent qu'un réflexe physiologique, c'est-à-dire une réponse automatique à une stimulation naturelle (chez le chien : la salivation à l'approche de la nourriture), pouvait être déclenché par une autre stimulation si celle-ci était associée de manière systématique à la première (par exemple : le son d'une cloche prédisant l'apparition de la nourriture). En d'autres termes, il existe des mécanismes cérébraux visant à établir des associations stimulus-réponse automatiques. Il s'avéra ensuite que de nombreux réflexes physiologiques et/ou psychologiques relevaient de telles associations.

Dans le contexte des sciences cognitives, on parle de « comportement pavlovien » lorsque la réponse comportementale est automatiquement déclenchée par un stimulus. Ce type de comportement s'oppose au comportement « intentionnel » ou « instrumental », qui a la particularité d'être « dirigé vers un but » (par exemple, une action menée pour obtenir une récompense). De même, le « conditionnement pavlovien » est le mécanisme par lequel l'association stimulus→comportement se construit. Cette expression est aujourd'hui passée dans le langage courant, popularisant ainsi la vieille idée selon laquelle notre comportement est régi par des automatismes inconscients, construits sur des associations fortuites et purement contextuelles [24].

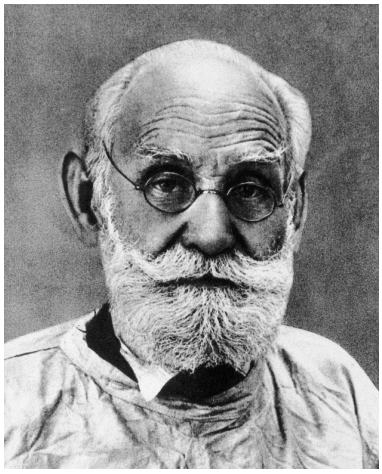


Figure *2.10 – Ivan Petrovich Pavlov. Source : Wikimédia.*

23. Pavlov, I. (1903). The experimental psychology and psychopathology of animals. In *The 14th International Medical Congress*, page 23–30

24. Jarius, S. and Wildemann, B. (2017). Pavlov's reflex before pavlov : Early accounts from the english, french and german classic literature. *European Neurology*, 77(5–6):322–326

## 2.7 Thorndike et le connexionnisme

Avec Pavlov, Edward Thorndike fut l'un des pionniers du « behaviorisme », un courant de pensée qui domina la psychologie jusqu'à la « révolution cognitive ». Les behavioristes promeuvent le principe du « canon de Morgan » [25], qui propose qu'aucun comportement animal ne devrait être expliqué par un processus psychologique avancé, s'il peut être expliqué plus simplement par des processus évolutifs et/ou développementaux. En son temps, Thorndike étudia le « conditionnement instrumental » chez les animaux. C'est-à-dire l'automatisation progressive d'une action induite par son association à une récompense [26]. En observant des chats apprendre des comportements complexes [27], il proposa une série de lois générales guidant l'apprentissage d'un comportement. La plus fameuse étant la loi de l'effet : une action récompensée tendra à se répéter. Cette loi est intéressante parce qu'elle permet d'expliquer l'émergence de

Figure *2.11 – Edward Thorndike. Source : Wikimédia.*

25. Madsen, K. (1988). A history of psychology in metascientific perspective. In *Advances in Psychology*, volume 53, page 193–257. Retrieved from

26. Thorndike, E. (1898). *Animal intelligence : an experimental study of the associative processes in animals*. Macmillan, New York

comportements intentionnels complexes sans invoquer la nécessité d'une planification sophistiquée. Les chats de Thorndike apprenaient graduellement, par « essai-erreur », à pousser une série de leviers et de boutons dans un ordre très précis pour s'échapper de leur cage. D'après Thorndike, leur comportement s'expliquait, non pas par une capacité d'introspection, de réflexion et/ou de planification, mais simplement par le renforcement progressif des séquences d'actions associées à la récompense. Le principe de « l'apprentissage par renforcement » était né. Le célèbre modèle de Robert Rescorla et Allan Wagner, développé en 1972 [28], en capture l'essence de manière remarquablement simple et élégante.

Plus tard, le courant du « connexionnisme » proposa des réseaux de neurones artificiels capables de s'ajuster à une tâche, c'est-à-dire à émettre la réponse qui maximise le taux de récompense. La particularité de ce type de modèles est de produire des comportements complexes à partir d'opérations très simples effectuées par les neurones. Plus récemment, ces travaux ont abouti à la révolution de « l'apprentissage profond » en informatique, une branche de l'intelligence artificielle qui propose de résoudre des problèmes complexes simplement grâce à la flexibilité des réponses permises par les réseaux de neurones artificiels à plusieurs couches.

27. Thorndike construisit une cage à chat, dont les félins ne pouvaient s'enfuir qu'un effectuant une série de mouvement très précis, par exemple appuyer sur un barreau, puis pousser un levier et enclencher un interrupteur. Thorndike testa s'ils pouvaient apprendre la bonne séquence juste par l'observation. Il compara la vitesse d'apprentissage des chats confrontés à la cage pour la première fois, à celle des chats ayant d'abord observé leurs congénères s'échapper. Il ne trouva aucune différence, avec des chats ou d'autres animaux. A partir de ces résultats il s'opposa aux thèses vitalistes de son époque. L'apprentissage des animaux ne semblaient pas s'appuyer sur une quelconque forme d'introspection et de « déclic » leur permettant de sortir immédiatement de la cage. Au contraire, leur apprentissage était graduel et suivait un loi simple, la loi de l'effet : si une action est récompensée, elle tendra à se répéter. Thorndike proposa par la suite une série de lois de cet ordre pour expliquer le comportement animal et humain.

28. Wagner, R. and R, A. (1972). A theory of classical conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, 21(6):64–99

## 2.8 BRODMANN ET LA CARTOGRAPHIE DU CERVEAU

En 1909, le neurologiste Korbinian Brodmann publie ses travaux sur l'architecture du cortex cérébral (Brodmann, 1909). Il y délimite le cerveau en 52 zones distinctes, chacune identifiée par les propriétés de ses cellules, de ses tissus et de leurs organisations. Brodmann pensait que ces différences histologiques entraînaient des différences de fonctions, et permettraient d'identifier le rôle de chaque zone (Guimaraes, Santos, Freire, 2016).

Aujourd'hui, bien que l'on sache que cette association histologie-fonction est imparfaite, les aires de Brodmann restent largement utilisées par la communauté scientifique. Elles fournissent un référentiel anatomique permettant de positionner une aire cérébrale et d'étudier systématiquement les processus cognitifs l'impliquant. Cependant, d'autres méthodes de positionnement anatomique existent aujourd'hui, qui permettent de dépasser les limites de l'approche de Broadman. En effet, les aires de Brodmann ont été établies à partir d'un seul sujet uniquement, et n'offrent qu'un degré de précision limité. Par exemple, en 1988, Jean Talairach et Pierrre Tournoux proposèrent un système de coordonnées spatiales s'adaptant à la taille et la forme du cerveau de chaque individu [29]. Six ans plus tard, Alan Evans et ses collègues proposèrent un autre système de coordonnées, les coordonnées MNI [30] (pour Montreal Neurological Institute), basé sur la standardisation de 250 cerveaux [31]. De nos jours, de nouveaux atlas cérébraux sont régulièrement proposés pour répondre aux besoins spécifiques de la neuro-imagerie fonctionnelle. On trouve ainsi des atlas définissant des régions selon la configuration

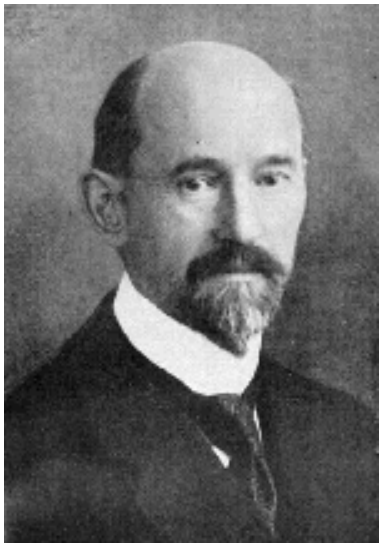


FIGURE *2.12 – Korbinian Brodmann. Source : Wikimédia.*

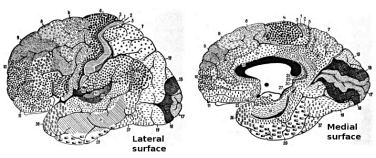



FIGURE *2.13 – Aires de Brodmann. Source : Wikimédia.*

des circonvolutions cérébrales : l'atlas AAL [32] et MarsAtlas [33] [34]. On trouve également des atlas de connectivités cérébrales, basés sur la propagation des impulsions électriques [35], ou identifiant des groupes de régions par la similarité de leurs activités [36]. Certains atlas proposent également une partition cérébrale basée sur des propriétés à la fois anatomiques et fonctionnelles [37]. Comme tous sont convertibles entre eux, le choix d'un atlas cérébral ne relève plus de la nécessité d'adopter un langage commun, mais des propriétés pertinentes à une étude. De nos jours il est également possible de définir des régions d'intérêt simplement à l'aide de mots clés, à la manière d'un moteur de recherche internet. Des outils de méta-analyses automatiques comme https://neurosynth.org/ [38] ou https://neuroquery.org [39] agrègent des milliers d'études de neuro-imagerie pour fournir des cartes d'associations entre les mots clés et les aires cérébrales.

## 2.9 BARLOW, LAUGHLIN ET LE CODAGE EFFICACE

En 1961, Horace Barlow appliqua les principes de la théorie de l'information, proposée 13 ans plus tôt par Claude Shannon [40], au « code neural » [41], c'est-à-dire la manière dont les neurones représentent et/ou traitent l'information. En comparant le cerveau à un système de traitement de l'information soumis à des contraintes biologiques, il proposa une forme d'optimalité biologique dans l'organisation du code neural. Barlow raisonna de la manière suivante : si le but des neurones est de transmettre une information sur l'environnement (comme c'est le cas pour les neurones des systèmes sensoriels), alors il doit être possible d'extraire cette information à partir de leur activité. On parle de « représentation » neurale de l'information [42]. Or la représentation neurale de l'information induit un coût biologique qui augmente avec la sollicitation des ressources neuronales. Barlow suggéra que le code neural s'organise de manière efficace, c'est-à-dire en effectuant un compromis optimal entre la qualité de la représentation neurale de l'information et le coût biologique qu'elle engendre. Le codage « efficace » suppose simplement que les neurones maximisent la quantité d'informations transmises par chaque impulsion électrique (potentiel d'action), en minimisant la « redondance » du code neural. Par exemple, si certaines propriétés de l'environnement sont systématiquement contingentes, il est inefficace de les représenter de manière indépendante. Il en découle alors que les neurones doivent s'adapter aux statistiques naturelles de l'environnement. Il faut noter ici que l'argument de Barlow couvre plus que l'environnement sensoriel d'un individu : il couvre l'environnement de chaque neurone : toutes les impulsions qu'il reçoit. D'un point de vue macroscopique, la spécialisation fonctionnelle d'un neurone, ou d'une région, n'est donc pas paradoxale, elle relève simplement d'une sélection progressive des propriétés à conserver pour représenter un environnement.

Vingt ans plus tard, Simon Laughlin apporta une démonstration expérimentale de l'une des variantes du principe de codage efficace [43]. Laughlin nota que la marge d'activation d'un neurone impose une

29. Talairach, J. and Tournoux, P. (1988). *Co–Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System : An Approach to Cerebral Imaging*. Thieme Medical Publishers
30. Cette méthode propose de récupérer une modèle 3D d'un cerveau avant de le distordre pour faire correspondre à un modèle moyen, que l'on décrit alors en coordonnées cartésiennes.
31. Evans, A., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., and Bub, D. (1992). Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage*, 1(1):43–53
32. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289
33. Contrairement à AAL, MARSAtlas s'appuie sur une détermination automatique des régions, et non sur des annotations manuelles.
34. Auzias, G., Coulon, O., and Brovelli, A. (2016). Marsatlas : A cortical parcellation atlas for functional mapping. *Human Brain Mapping*, 37(4):1573–1592
35. Trebaul, L., Deman, P., Tuyisenge, V., Jedynak, M., Hugues, E., Rudrauf, D., and David, O. (2018). Probabilistic functional tractography of the human cortex revisited. *NeuroImage*, 181(January):414–429
36. Power, J., Cohen, A., Nelson, S., Wig, G., AnneBarnes, K., Church, J., and Petersen, S. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678
37. Glasser, M., Coalson, T., Robinson, E., Hacker, C., Harwell, J., Yacoub, E., and Van Essen, D. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, page 1–11
38. Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., and Wager, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670
39. Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., and Varoquaux, G. (2020). Neuroquery : comprehensive meta-analysis of human brain mapping. Retrieved from

limite fondamentale sur la transmission d'informations. Si la sensibilité d'un neurone à ses entrées est trop forte, alors les entrées saturent la réponse, ce qui induit une perte d'information. Si elle est trop basse, alors une large part de la variation exploitable de la réponse du neurone est perdue. Pour coder ses entrées de manière efficace, la sensibilité d'un neurone doit donc correspondre à la gamme de variation naturelle de ses entrées. En étudiant la réponse des neurones de la rétine de la mouche, Laughlin découvrit que la première couche neurale du système visuel était adaptée aux variations naturelles de l'intensité lumineuse dans l'environnement.

D'un point de vue méthodologique ces travaux revêtent également un grand intérêt. En effet, ils permettent de traduire une hypothèse sur le rôle fonctionnel d'une aire cérébrale, représenter l'intensité lumineuse par exemple, en une hypothèse sur son activité : le codage neuronal sera adapté aux statistiques naturelles de l'intensité lumineuse.

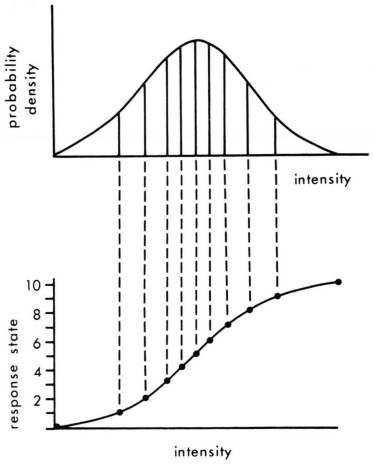



FIGURE 2.14 – *Stratégie d'encodage étudié par Simon Laughlin. En haut : la distribution de probabilité de l'intensité d'un stimuli visuel. En bas : la réponse d'une cellule permettant de transmettre autant d'information que possible avec uniquement 10 niveaux de réponses.*

## 2.10 MARR ET LES TROIS NIVEAUX D'ANALYSES

David Marr était un neuroscientifique anglais qui étudiait le système visuel. Fortement influencés par la recherche en intelligence artificielle, Tomaso Poggio et lui définirent en 1977 trois niveaux d'analyse [44] pour comprendre un système neural [45]. Ils proposèrent de considérer 1) le but du système : « quels problèmes résout-il ? », 2) ses méthodes : « sur quelles représentations s'appuie-t-il, comment les manipule-t-il ? », et 3) son support, « quelle est l'implémentation neurobiologique de ses calculs élémentaires ? ». En proposant ces niveaux d'analyses, Marr et Poggio insistèrent d'une part, sur leur indépendance relative, et d'autre part, sur la nécessité d'y positionner adéquatement les résultats neuroscientifiques. Ils critiquaient notamment l'interprétation des propriétés neuronales en termes psychologiques. Selon eux, les propriétés des neurones ne permettent pas de tirer des conclusions sur le but ou les méthodes du système auquel ils appartiennent. Elles ne doivent être interprétées qu'en termes d'opérations entrées/sorties et ne peuvent pas être directement reliées à des notions supérieures. Pour Marr, un neurone d'œil de mouche ne détecte pas la lumière, il répond juste à un signal électrique. La fonction de détecteur de lumière et son utilité pour le système visuel de la mouche n'ont pas sens au niveau du neurone. C'est au niveau de l'activité globale du réseau (le système visuel) que les représentations sont manipulées, et que la fonction émerge (but du système).

40. Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656
41. Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, page 216–234
42. Plus précisément, on dit qu'un neurone ou un ensemble de neurone « représente » une propriété de l'environnement lorsque son activité varie de manière spécifique et systématique avec cette propriété.
43. Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capactiy
44. Il existait un $4^e$ niveau, intermédiaire entre le 2 et le $3^e$, mais il n'a pas été retenu. Il proposait d'étudier le lien entre les méthodes employées par un système et le support les implémentant.
45. Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*

| Computational theory | Representation and Algorithm | Hardware Implementation |
|---|---|---|
| What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out ? | How can this computational theory be implemented ? In particular, what is the representation for the input and output, and what is the algorithm for the transformation ? | How can the representation and algorithm be realized physically ? |

TABLE 2.1 – *Description originale (donc en anglais) des 3 niveaux de Marr.*

Sans cesse débattu depuis lors, ce cadre théorique a pour vertu d'organiser les travaux de recherche de manière systématique, et d'identifier les problèmes d'interprétation que peuvent soulever les résultats neuroscientifiques. Par exemple, comme la plupart des études IRMf ne mesurent les variations hémodynamiques qu'au cours d'une seule tâche, elles ne nous informent, au mieux, que sur les représentations induites par un contexte cognitif. En principe, elles ne permettent pas de déduire le but d'une région ou ses opérations élémentaires. Pour identifier directement le but d'une région (niveau 1), il faudrait par exemple, comprendre comment l'information représentée par cette région est utilisée par le reste du cerveau, en étudiant son activité dans une vaste gamme de contextes cognitifs.

FIGURE 2.15 – *Rajesh Rao : Source : Wikimédia.*

## 2.11 RAO, BALLARD ET LE CODAGE PRÉDICTIF

En 1999, Maximilian Riesenhuber et Tomaso Poggio ont construit un modèle mathématique s'inspirant de l'architecture du cortex visuel : un réseau de neurones artificiels alternant des cellules à simple et double contraste [46] [47]. Ils montrèrent que, malgré sa simplicité, leur modèle pouvait être ajusté pour reconnaître automatiquement des objets présents dans le champ visuel, quelles que que soient leur taille et leur orientation. De plus, cet ajustement faisait correspondre la sensibilité de leurs neurones artificiels à celle de véritables neurones. Ils démontrèrent ainsi qu'un système artificiel pouvait exhiber des propriétés similaires à celle du cerveau, dès lors qu'il en partageait le but et les méthodes, c'est-à-dire les niveaux un et deux de Marr [48].

La même année, Rajesh Rao et Dana Ballard ont obtenu un résultat similaire, mais en relâchant les hypothèses de la modélisation [49]. Ils proposèrent un réseau de neurones artificiels qui n'était pas à priori contraint de respecter l'organisation de véritables neurones du système visuel, mais qui devait simplement vérifier l'hypothèse de Barlow sur la redondance minimale du code neural. Plus précisément, leur modèle est simplement construit de manière à extraire une représentation condensée de ses entrées, grâce à un mécanisme lui permettant de minimiser les erreurs de cette représentation [50]. Rao et Ballard démontrèrent alors qu'en ajustant le réseau pour qu'il représente des images de scènes naturelles, le réseau montrait des propriétés correspondantes à celles du cortex visuel. D'une part, les neurones des deux couches du réseau répondaient aux variations de contraste lumineux de manière remarquablement similaire à celle de

46. Les cellules à simple contraste répondent à une couleur et à son absence, celle à double contrastes répondent à la différence de deux couleurs, ou à son absence.
47. Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025
48. Le niveau 1 correspond à la détection d'objet, le but du système, et le niveau 2 à l'organisation des couches de neurones artificiels et la forme de leur sensibilité : cellules à simple ou double contraste.
49. Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex : A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87
50. Leur modèle utilise un codage « prédictif » en 2 étapes. Une première couche correspond à une représentation condensée de l'image, et une seconde couche à une représentation condensée de celle-ci. Les deux représentations sont ajustées afin de minimiser l'erreur de reconstruction de l'image initiale, et de sa première représentation, à partir respectivement de la première et de la seconde représentation.

véritables neurones du cortex visuel primaire et secondaire, d'autre part, les neurones d'erreurs de leur modèle fournissaient une explication originale aux cellules hypercomplexes[51] observées dans le cortex visuel primaire. Ces cellules ont la particularité d'être inhibées par certaines caractéristiques des stimuli visuels situées hors de leur champ récepteur, ce qui rendait les chercheurs perplexes puisqu'elles ne devraient pas y être sensible. Équipés uniquement d'un principe algorithmique général (le codage « prédictif »[52]) et d'un but tout aussi général (la représentation du champ visuel), Rao et Ballard pouvaient rendre compte de propriétés surprenantes de l'activité neurale et en proposer de nouvelles interprétations.

51. Découverte par Hubel et Wiesel, ces cellules sont sensibles à l'orientation de barres dans un environnement visuel, quel que soit leur emplacement. La particularité des cellules hyper-complexes par rapport aux cellules simples ou complexes, est que leur réponse croit jusqu'à ce que la barre atteigne une longueur donnée, puis est rapidement inhibée.

52. Le principe du codage prédictif consiste à prédire l'activité des premières étapes d'une chaîne de traitement de l'information à partir des étapes suivantes.

## 2.12 Conclusion sur les principes d'analyses

Au-delà des dispositifs de mesures, on bénéficie aujourd'hui d'un vaste éventail de méthodes et principes théoriques nous permettant d'étudier l'organisation fonctionnelle du cerveau. Dans ce chapitre, j'ai pris le parti de présenter leurs intérêts et leurs ambitions plutôt que leurs limites. Il faut cependant noter qu'aucune de ces approches ne peut, isolément, aboutir à une compréhension exhaustive du système. Les dissections anatomiques sont aveugles aux signaux échangés entre régions. Les analyses statistiques des signaux de neuro-imagerie sont toujours biaisés par leurs hypothèses. Les lésions nous informent sur l'importance d'une région dans un processus, mais pas sur le but ou les moyens (au sens de Marr) du système sous-jacent. Les modèles cognitifs du comportement étudient des phénomènes isolés, privés du contexte naturel de notre vie quotidienne. Les atlas cérébraux peinent à décrire les phénomènes distribués. Les niveaux de Marr ne rendent pas compte de leurs interactions. Enfin, les principes théoriques ne peuvent pas générer des prédictions valides si le critère d'optimalité retenu n'est pas adéquat pour le système étudié.

De fait, rares sont les théories neuroscientifiques qui permettent de lier ces éléments de preuves entre eux et de dépasser leurs limitations individuelles.

# 3

# *Comment l'analyse IRMf permet-elle d'affiner les théories cognitives ?*

◂ Chapitre 2 Chapitre 4 ▸

Synopsis La neuro-imagerie s'appuie fortement sur les théories issues de la psychologie cognitive pour étudier le rôle des différents systèmes cérébraux. Ce chapitre décrit la manière dont les analyses IRMf sont utilisées pour soutenir une hypothèse psychologique, la décomposer, comparer des hypothèses concurrentes et/ou extraire une information sur les états mentaux (croyances, préférences, émotions, etc.) d'un individu.

Du point de vue psychologique, identifier la localisation cérébrale d'une fonction cognitive ne revêt paradoxalement qu'un intérêt secondaire. L'intérêt premier de la neuro-imagerie est de *confirmer* l'existence d'un concept psychologique. Par exemple, savoir qu'une région s'active spécifiquement lorsqu'un sujet regarde un visage, révèle l'importance de la détection des visages pour le cerveau. Plus exactement, cela révèle que la détection de visages est si importante que nos cerveaux sont pourvus d'une région spécialisée pour réaliser cette fonction. En revanche, dans ce type de raisonnement, la mécanique interne et la localisation de cette région, la Fusiform Face Area (FFA), ne sont que secondaires.

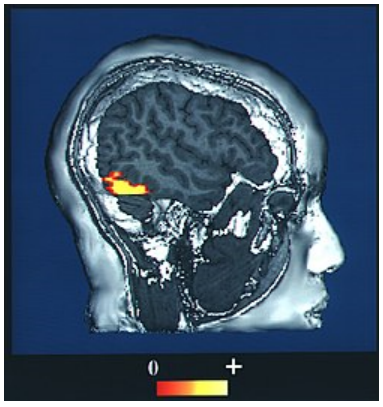



FIGURE *3.1 – Fusiform Face Area (FFA) identifiée par la corrélation entre l'activité BOLD des voxels d'un scan IRMf et la présentation de visages au sujet. Source : Wikimédia.*

Il existe de nombreuses méthodes statistiques pour tester la validité neuroscientifique d'un concept psychologique. Cependant, toutes reposent sur l'idée simple que si le concept est représenté par le cerveau, alors la manipulation expérimentale de ce concept devrait induire des variations systématiques d'activité cérébrale. Par exemple, sous l'hypothèse que la reconnaissance d'un visage repose sur la présence et l'organisation spatiale de ses éléments constitutifs (yeux, nez, bouche), l'activité d'une région dont la fonction est d'identifier un visage devrait être sensible à la variation de ces propriétés [1].

1. Liu, J., Harris, A., and Kanwisher, N. (2010). *Perception of Face Parts and Face Configurations : An fMRI Study*, volume 22. Liu, J., Harris, A., Kanwisher, N

Typiquement, on cherchera donc à vérifier si l'activité cérébrale répond aux variations contrôlées des facteurs expérimentaux d'intérêt (ici : présence et organisation spatiale des éléments du visage). Bien entendu, la forme de la réponse est inconnue, mais il est toujours possible d'en proposer une approximation. En particulier, une expansion de Taylor au premier ordre permet de décomposer la réponse cérébrale comme la somme pondérée des facteurs expérimentaux :

$$BOLD(t) = facteur_1(t) * poids_1 + \cdots + facteur_n(t) * poids_n + bruit(t) \tag{3.1}$$

FIGURE *3.2 – L'équation fondamentale d'un GLM. Pour chaque étape $t$ de l'expérience on explique le signal IRMf, $BOLD(t)$, comme une combinaison de la valeur de chaque facteur expérimental, $facteur_1(t)$. $bruit_t$ regroupe toutes les variations du signal échappant à cette combinaison.*

On parle de « modèle linéaire général » ou GLM (General Linear Model) [2]. L'approche statistique du GLM consiste alors à ajuster les poids du modèle pour décrire le signal BOLD au mieux, puis à étudier leur amplitude, leur signe et leur significativité [3]. Chaque poids correspond à la corrélation [4] entre le facteur et le signal BOLD, après avoir retiré l'influence des autres facteurs. Ainsi, lorsqu'un poids diffère de zéro, on conclura que le signal BOLD est associé au facteur correspondant. Si un poids est plus grand qu'un autre, on conclura que l'influence de son facteur prédomine.

De manière générale, l'approche du GLM s'avère extrêmement pratique et versatile pour étudier le fonctionnement cerveau. Les psy-

2. En pratique on ne l'utilise pas directement dans cette forme et on tient compte de divers facteurs confondants, les mouvements du sujet dans l'IRMf par exemple, ou le décalage entre une activation neurale et la manifestation de son coût métabolique (voir plus bas le paragraphe sur la HRF).

3. C'est-à-dire si leurs valeurs pourraient-être dues au hasard uniquement (voir la « p-value » au chapitre précédent).

4. Si on utilise une régression des moindres carrées pour ajuster les poids, alors ils correspondent à la corrélation partielle entre le signal et le facteur, pondérée par le rapport des déviations standard.

chologues peuvent ainsi scanner des cerveaux entiers, répéter l'analyse GLM à chaque voxel [5], trouver les régions corrélant avec leurs manipulations expérimentales, et comparer l'influence de chaque facteur. Plus ces corrélations sont fortes et systématiques, plus le concept sous-jacent au facteur expérimental peut être considéré comme fondamental.

5. On parlera d'analyse massivement univariée

## 3.2 DÉCOMPOSITION NEURONALE

Il est possible qu'un facteur expérimental corrèle avec plusieurs aires cérébrales. Cela suggère que le concept sous-jacent n'est pas traité par une seule aire, mais par un réseau de régions cérébrales. Dans ce cas, on peut s'intéresser plus précisément à l'organisation fonctionnelle du réseau, c'est-à-dire comprendre les interactions et les contributions respectives de chaque région au phénomène étudié.

Une extension directe du GLM consiste à inclure le signal BOLD des autres régions à la place des facteurs expérimentaux. Cela permet de mesurer et de comparer l'influence des régions les unes sur les autres. C'est l'essence des approches SEM [6] et DCM [7] qui proposent d'évaluer les relations statistiques entre les activités de différentes régions cérébrales. Par exemple, en utilisant plusieurs GLMs :

6. Pour Structural Equation Modeling
7. Pour Dynamical Causal Modeling

$$
\begin{aligned}
BOLD_1 &= facteur_1 * poids_{0\to1} + bruit_1 \\
BOLD_2 &= BOLD_1 * poids_{1\to2} + bruit_2 \\
BOLD_3 &= BOLD_1 * poids_{1\to3} + BOLD_2 * poids_{2\to3} + bruit_3
\end{aligned}
$$

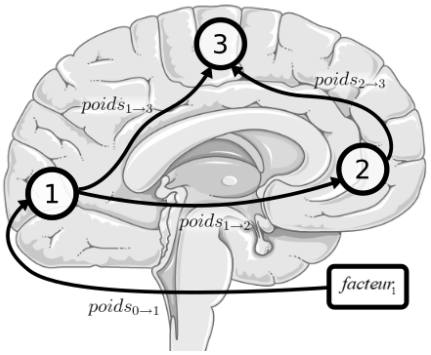



FIGURE *3.3 – Exemple de modèle SEM. Ici on a omis les indices temporels par simplicité. Chacune des trois régions est représentée par la notation* $BOLD_{i\in\{1,2,3\}}$*. La structure de chaque équation représente directement les hypothèses de connectivité du modèle graphique.*

En analysant les interactions d'un tel réseau, on va chercher à comprendre précisément le rôle de chaque région. Par exemple, on peut s'interroger sur le rôle de la FFA. Son rôle est-il d'extraire l'identité ou l'émotion d'un visage, ou se contente-t-elle d'intégrer les éléments faciaux [8], eux-mêmes analysés par d'autres régions [9] ? Les analyses SEM ou DCM permettent d'étudier et de comparer ces scénarios pour déterminer lequel est le plus probable. Idéalement, elles permettent de tester et/ou comparer des hypothèses sur les composantes d'un processus cognitif en identifiant l'architecture

8. La position des yeux, du nez, de la bouche etc.
9. He, Y. and Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Curr. Opin. Neurol*, 23:341–350

cérébrale qui le sous-tend.

En poussant cette approche plus loin, on peut aller jusqu'à comparer des ontologies de concepts cognitifs, leur organisation et leur dépendance mutuelle, à l'organisation de réseaux neuronaux [10]. Ce type d'analyse est certes plus complexe, mais l'argument reste le même. Plus l'ontologie sera similaire à l'organisation de l'activité cérébrale, plus on la considérera comme un scénario plausible du fonctionnement cérébral.

10. Poldrack, R. and Yarkoni, T. (2016). From brain maps to cognitive ontologies : Informatics and the search for mental structure. *Annual Review of Psychology*, 67(1):587–612

## 3.3 Comparaison d'hypothèses concurrentes

La psychologie expérimentale cherche à identifier les mécanismes mentaux qui expliquent la nature de la réponse comportementale à certaines informations ou caractéristiques de l'environnement. Il s'agit là d'un problème inverse : déduire des causes à partir de leur effet. C'est de plus un problème mal posé [11]. En effet, il existe généralement plusieurs explications à un comportement donné. La manière la plus simple de déterminer expérimentalement l'explication la plus probable consiste à toutes les mettre à l'épreuve. On cherchera alors à mettre en place une expérience où chaque hypothèse prédira un résultat différent, et à mesurer laquelle décrit le mieux le comportement des sujets. Cependant, il n'est pas toujours possible de construire un cadre expérimental dans lequel les scénarios candidats prédisent des comportements qualitativement différents. En effet, plus un scénario est flexible, plus il pourra justifier une large gamme de comportements. Deux hypothèses concurrentes pourront même faire exactement les mêmes prédictions, mais pour des raisons différentes. Dans ce contexte, la neuro-imagerie peut permettre d'arbitrer entre les scénarios candidats, si les mécanismes qu'ils invoquent peuvent se distinguer sur le plan neural.

11. Au sens mathématique du terme.

Par exemple, les neurosciences de la décision s'intéressent aujourd'hui à la manière dont on attribue une valeur à un objet ou une action. Attribue-t-on une valeur à un produit alimentaire de la même manière qu'à un ticket de loterie, ou à un accessoire de mode ? On peut formuler deux hypothèses : soit il existe plusieurs modules périphériques évaluant chaque catégorie indépendamment, soit un module central traite tous les objets, quelle que soit leur catégorie (hypothèse de la « monnaie commune »). Sans suppositions supplémentaires, il sera difficile de les distinguer. Il suffit d'imaginer que le module central possède les mêmes propriétés que les modules périphériques, ou inversement, pour que les deux hypothèses fassent des prédictions similaires sur les choix d'un individu.

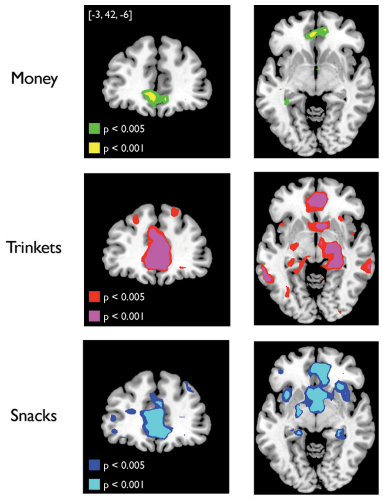



Figure 3.4 – *Corrélat neuronaux de la valeur d'options monétaires, d'accessoires et de produit alimentaires. Le code couleur indique la p-value de l'association entre la valeur des objets présentés aux sujets et les variations de leur signal BOLD. Figure extraite de (Chib et al., 2009).*

En 2009, Vikram Chib et ses collègues ont ainsi utilisé l'IMRf pour montrer que le cortex préfrontal ventromédian (vmPFC) corrélait systématiquement avec la valeur des objets auxquels ils étaient exposés, quelle que soit la catégorie des objets [12] [13]. Il aurait été très difficile de valider l'hypothèse de la « monnaie commune » uniquement par des observations comportementales.

Dès lors que deux hypothèses proposent deux organisations dif-

12. Cette expérience ne discrédite cependant pas entièrement l'hypothèse des modules alternatifs. Chib et ses collègues ont également identifié des régions corrélant spécifiquement chaque catégorie.

13. Chib, V., Rangel, A., Shimojo, S., and O'Doherty, J. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, 29(39):12315–12320

férentes des processus mentaux, on peut tenter de les différencier grâce à la neuro-imagerie. Cela permet de travailler à la frontière des niveaux de Marr : comprendre le rôle fonctionnel des régions cérébrales d'une part, et d'autre part raffiner les théories cognitives en les confrontant à leurs implications sur le plan de l'architecture cérébrale [14].

14. Mather, M., Cacioppo, J., and Kanwisher, N. (2008). How fmri can inform cognitive theories. *Bone*, 23(1):1–7

## 3.4 INFÉRENCE INVERSE ET DÉCODAGE

En octobre 2018, Russell Poldrack publia un livre intitulé « Les nouveaux télépathes : ce que la neuro-imagerie peut et ne peut pas révéler de nos pensées », où il décrit longuement l'analyse IRMf [15]. Il y critique notamment une utilisation controversée de l'IRMf : « l'inférence inverse ». Cette approche consiste à exploiter une relation connue entre l'activité d'une région et un concept psychologique (par exemple : effrayer le sujet entraîne une hausse de l'activité de l'amygdale), puis à interpréter n'importe quelle activation de cette région comme une manifestation de ce concept (son amygdale s'active, donc [16] il a peur). Le problème est que cette inférence n'est absolument pas spécifique : l'amygdale peut s'activer pour de nombreuses raisons, la peur n'étant que l'une d'entre elles. Cela révèle une limite des approches précédentes : l'activation spécifique d'une région par un stimulus ne signifie pas que la région traite spécifiquement ce type de stimulus.

15. Poldrack, R. (2018). The new mind readers : What neuroimaging can and cannot reveal about our thoughts

16. C'est ici que la logique de l'inférence est « inversé »

Cependant, ce type d'approche peut s'avérer très informatif, lorsqu'il est employé plus rigoureusement. Le problème que soulève la « lecture » des états mentaux d'un sujet à partir de son activité cérébrale est que l'on ne connaît pas tous les contextes dans lesquels une région est activée. Pour s'assurer que la lecture d'une information reste valide hors du contexte ou le modèle de lecture est ajusté, il faut donc tester une gamme de concepts la plus large possible. L'inférence inverse consiste alors à déterminer, étant donné le profil d'activité observé, quels concepts ont vraisemblablement été activés (et avec quelle probabilité). Encore balbutiant, ce type d'approche se base aujourd'hui sur l'agrégation de milliers de résultats expérimentaux [17] [18].

17. Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., and Varoquaux, G. (2020). Neuroquery : comprehensive meta-analysis of human brain mapping. Retrieved from

18. Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., and Wager, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670

On peut également construire des modèles « prédictifs » plus modestes, dont la validité est restreinte au contexte de l'expérience en cours. Dans leur forme la plus simple, ils s'appuient également sur des GLMs, où le rôle des facteurs expérimentaux et du signal BOLD sont inversés :

$$Facteur = BOLD_1 * poids_1 + \cdots + BOLD_n * poids_n + bruit$$

Les poids vont alors être ajustés pour décrire au mieux les variations du facteur. Cependant le modèle ne va pas être évalué par rapport à ses poids, mais par rapport à la qualité de ses prédictions [19].

$$Prdiction = BOLD_1 * poids_1 + \cdots + BOLD_n * poids_n$$

19. Pour ce faire, la plupart des modèles vont employer une même méthode : la validation croisée, c'est-à-dire ajuster les paramètres du modèle sur des données, puis évaluer la qualité de ses prédictions sur d'autres données.

Il ne s'agit plus de comprendre ce qui compose un signal neuronal, mais de mesurer la qualité de l'information qu'on peut en extraire.

On peut ainsi « décoder » des informations très diverses [20] par exemple : une sensation de douleur [21], son intensité subjective [22], le visage que l'on regarde [23 24 25], les lettres que l'on imagine [26], la décision que l'on s'apprête à prendre [27] etc.

Plus généralement, les approches de décodage permettent d'étudier si une information est présente dans le cerveau d'un sujet, sous quelle forme, dans quelle région, et à quel moment [28]. Cependant, il faut noter la différence de nature entre cette méthode et les précédentes. Ici, on ne s'intéresse plus à la manière dont le cerveau s'organise et réagit à des stimuli, mais plutôt aux informations qui peuvent y être lues par un observateur extérieur.

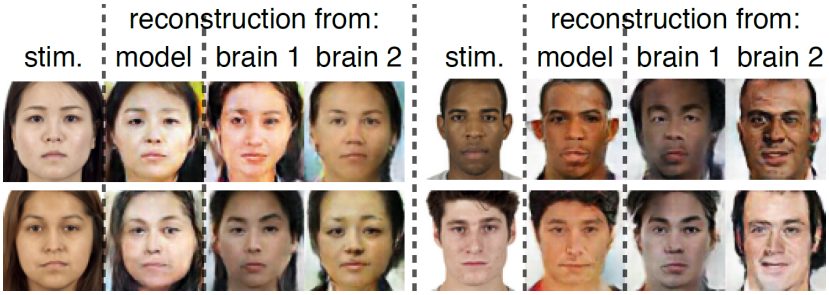



FIGURE 3.5 – *Reconstruction de visages (stim) à partir de la représentation interne d'un modèle de réseaux de neurones artificiels (model) et de la prédiction des états internes de ce modèle par les scans IRMf de deux sujets (brain 1 et brain 2). Images extraites de (Güçlütürk et al., 2017).*

## 3.5 CONCLUSION SUR L'INTERPRÉTATION DES ANALYSES IRMF

Les analyses IRMf permettent d'aborder l'étude des mécanismes psychologiques d'un individu de plusieurs manières. Comme précédemment, j'ai pris le parti de mettre l'accent sur leurs avantages plutôt que leurs limitations. Il ne faut pourtant pas oublier que ces analyses utilisent des mesures moyennes et indirectes de l'activité cérébrale, et non directement les impulsions électriques neuronales. D'autre part, elles s'appuient souvent sur des modèles mathématiques assez éloignés des mécanismes biologiques réels [29 30]. Par ailleurs, même si ces approximations restent valides dans un contexte donné, les interprétations des résultats de ces études sont généralement peu généralisables à d'autres contextes cognitifs [31 32 33].

Finalement, si ces analyses permettent d'évaluer la validité neuroscientifique de concepts psychologiques, cela ne garantit pas, formellement, que ces concepts soient utilisés par le cerveau (pour déterminer le comportement, par exemple).

Cependant, ces limitations peuvent être dépassées. En particulier, il est possible de raffiner la modélisation des mécanismes cognitifs et neuraux. Bien entendu, la calibration de ce type de modèle est plus délicate, et nécessite l'utilisation de données supplémentaires. Ici réside tout l'intérêt de la mesure concomitante des réponses cérébrales et comportementales aux facteurs expérimentaux. Comme

20. Avec des modèles linéaires ou des variantes plus complexes.
21. Wager, T., Atlas, L., Lindquist, M., Roy, M., Woo, C., and Kross, E. (2013). An fmri-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397
22. Tu, Y., Tan, A., Bai, Y., Sam Hung, Y., and Zhang, Z. (2016). Decoding subjective intensity of nociceptive pain from pre-stimulus and post-stimulus brain activities. *Frontiers in Computational Neuroscience*, 10(APR):1–11
23. Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., Van Lier, R., and Van Gerven, M. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, page 4247–4258
24. (Güçlütürk et al., 2017) utilisent ici (voir figure 3.5) un réseau de neurones entrainé à extraire les principales dimensions d'un visage. Un second réseau utilise ensuite ces représentations pour tenter de reconstruire le visage initial. Le décodage consiste alors à utiliser l'activité BOLD des sujets pour prédire les représentations du premier réseau et reconstruire un visage à partir de ces dernières par le biais du second réseau.
25. Vanrullen, R. and Reddy, L. (XXXX). Reconstructing faces from fmri patterns using deep generative neural networks, 31052
26. Senden, M., Emmerling, T., Hoof, R., Frost, M., and Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, 224(3):1167–1183
27. Hampton, A. and O'Doherty, J. (2007). Decoding the neural substrates of reward-related decision making with functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4):1377–1382
28. King, J. and Dehaene, S. (2014). Characterizing the dynamics of mental representations : The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210
29. de Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain ? *Psychonomic Bulletin Review*, page 1–14
30. Ritchie, J., Kaplan, D., and Klein, C. (2019). Decoding the brain : Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607

on le verra, l'analyse conjointe de ces données permet de rendre l'inférence plus spécifique et/ou plus informative sur la mécanique neuronale sous-jacente.

31. Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., and Hariri, A. (2020). What is the test-retest reliability of common task-functional mri measures ? new empirical evidence and a meta-analysis. *Psychological Science*, 31(7):792–806

32. Ghuman, A. and Martin, A. (2019). Dynamic neural representations : An inferential challenge for fmri. *Trends in Cognitive Sciences*, 23(7):534–536

33. Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., and Liu, T. (2014). Survey of encoding and decoding of visual stimulus via fmri : An image analysis perspective. *Brain Imaging and Behavior*, 8(1):7–23

# 4

# *Études des déterminants biologiques du comportement*

◂ Chapitre 3 Chapitre 5 ▸

Synopsis Durant ma thèse, j'ai développé de nouvelles méthodes d'analyses IRMf permettant d'identifier les mécanismes biologiques qui déterminent notre comportement. À terme, l'objectif est double : 1) caractériser la contribution d'une région cérébrale d'intérêt dans la production du comportement, et 2) évaluer l'impact de mécanismes et contraintes biologiques sur les processus de traitement neuronal de l'information participant au contrôle du comportement.

Dans ce chapitre, je présente les trois approches classiques utilisées pour étudier les déterminants biologiques du comportement : les interventions causales, les études statistiques IRMf et les modèles formels de traitement neuronal de l'information. Cela permettra de positionner les contributions de mon travail de thèse dans le contexte de la littérature neuroscientifique et d'en dégager l'intérêt et les limites.

## 4.1 APPROCHES CAUSALES PAR PERTURBATION

L'approche causale par perturbation est l'approche classique de l'étude des systèmes biologiques : il s'agit de perturber un système et d'observer les conséquences de cette perturbation.

Lorsque nous réagissons à une stimulation environnementale, nos actions sont issues d'un processus de traitement neuronal de l'information, c'est-à-dire une série de réactions électriques et chimiques, transformant progressivement un signal sensoriel en une action motrice. Les approches causales cherchent à caractériser ce processus en perturbant une à une ces réactions afin de déterminer leur rôle. Chez l'homme, ce type d'approche est typiquement utilisé pour valider des hypothèses de nature anatomique, par exemple la contribution d'une région cérébrale d'intérêt, ou des hypothèses neuro-chimiques, par exemple la contribution d'un neurotransmetteur d'intérêt.

Historiquement, c'est tout d'abord l'étude du comportement des patients cérébro-lésés qui a permis la validation d'hypothèses de nature anatomique (voir les travaux de Broka et Wernicke au chapitre précédent). Cependant, la spécificité anatomique des lésions est faible et les processus de récupération fonctionnelle limitent la sensibilité de ces approches. Par ailleurs, les différences comportementales entre patients cérébro-lésés et groupes « contrôle » peuvent être expliqués par d'autres facteurs confondants difficiles à éliminer [1]. Les techniques modernes permettent d'affiner considérablement ces analyses.

1. Par exemple la récente expérience hospitalière des patients.

En 2016, par exemple, Michael Frank et ses collègues démontrent l'implication du noyau sous-thalamique dans la prise de décision, en le stimulant directement à l'aide d'électrodes implantées (pour des raisons thérapeutiques) dans le cerveau de patients parkinsoniens [2]. Au cours de l'expérience, leurs sujets choisissaient entre deux stimuli visuels et apprenaient progressivement à discriminer quels stimuli étaient fréquemment récompensés. Les participants étaient organisés en trois groupes, un groupe de patients parkinsoniens traités par une stimulation profonde du noyau subthalamique, un autre groupe de patients traités par un médicament dopaminergique, et un groupe contrôle de sujets sains du même groupe d'âge. Les auteurs montrèrent alors que le premier groupe prenait moins de temps à comparer deux stimuli positifs, c'est-à dire récompensés dans plus de 50% des cas. En d'autres termes, la stimulation profonde du noyau subthalamique rend les sujets plus impulsifs face à une décision conflictuelle. Ce type de résultat peut également être obtenu de manière moins invasive. En 2017, Derek Nee et Mark D'Esposito iden-

2. Frank, M. J., Samanta, J., Moustafa, A. A., and Sherman, S. J. (2007). Hold your horses : impulsivity, deep brain stimulation, and medication in parkinsonism. *science*, 318(5854):1309–1312

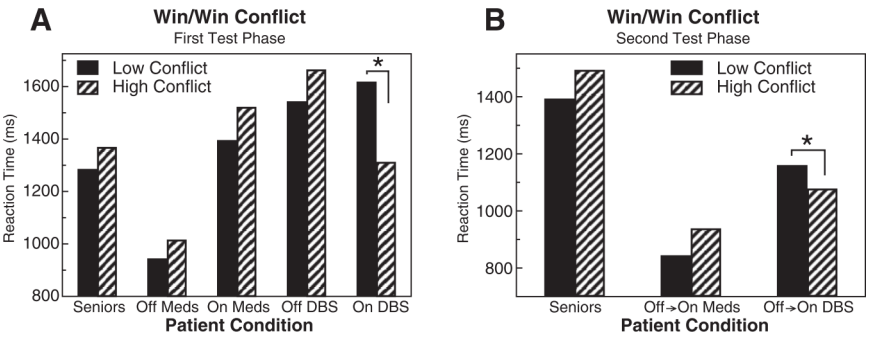



FIGURE 4.1 – *Temps de réaction de chaque groupe (groupe contrôle, groupe médicalisé avec et sans traitement, groupe stimulé avec et sans stimulation) face à un conflit faible ou fort. Dans la seconde phase (à droite) les patients ont appris les associations avant leur traitement. Figure extraite de (Frank et al., 2007)*

tifiaient ainsi l'implication du cortex préfrontal latéral dans le contrôle cognitif en le perturbant par de brèves d'impulsions magnétiques [3] [4]. Au cours de leur expérience, on présentait à un groupe de sujets sains une série de lettres positionnées au centre de formes géométriques (carrées, cercles, croix, etc.). Leur tâche comportait deux modalités. Une première modalité, spatiale, où les participants devaient indiquer si la forme géométrique présentée était au même endroit qu'à sa dernière apparition, et une seconde modalité, verbale, où ils devaient indiquer si la lettre présentée suivait alphabétiquement la précédente lettre apparue dans la même forme. Le but de cette expérience est de manipuler deux types de contrôle cognitif : un contrôle contextuel, et un contrôle temporel. Face à des stimuli identiques les participants doivent, d'une part, appliquer une règle différente selon la règle qu'on leur donne (verbal ou spatiale), et d'autre part, retenir une information apparue plus au moins récemment, deux à cinq essais avant.

Les auteurs ont ensuite mesuré les pertes de performance induites par la stimulation magnétique du cortex préfrontal. Ils mirent en évidence une augmentation du taux d'erreurs des sujets dans la tâche spatiale par rapport à la tâche verbale ainsi qu'une plus grande difficulté à retenir la position et le contenu des formes géométriques.

3. Par Stimulation Magnétique Transcrânienne (TMS).
4. Nee, D. and D'Esposito, M. (2017). Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *ELife*, 6:28040

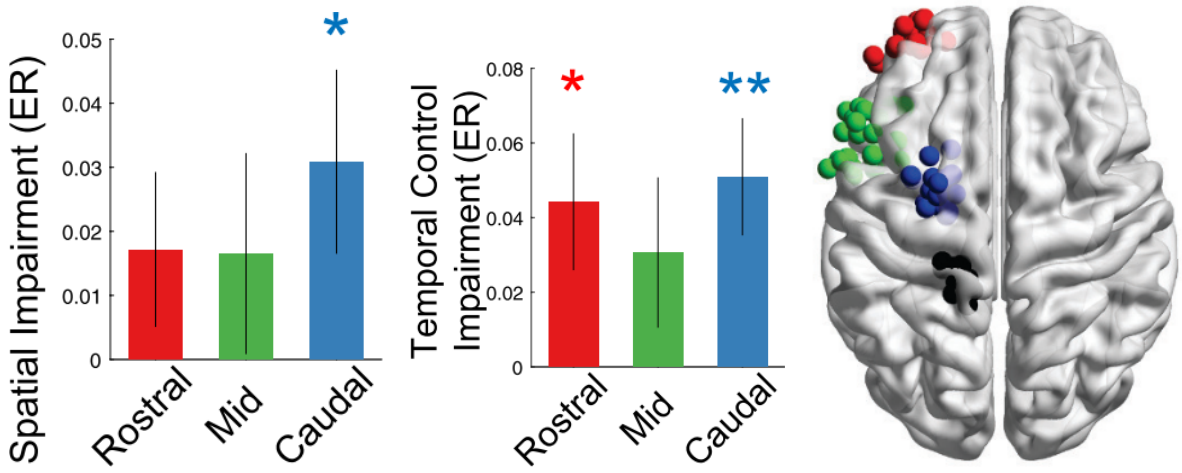



FIGURE 4.2 – *À gauche : Différence des taux d'erreurs entre la tache spatiale et la tâche verbale, au cours de la TMS. Au milieu : Taux d'erreur expliqué par la dimension temporelle des tâches, au cours de la TMS. À droite : Sites d'application de la TMS. Rouge : lPFC rostral, Vert : mid lPFC, Bleu : lPFC caudal, Noire : Cortex primaire moteur (région témoin). Figure extraite de (Nee and D'Esposito, 2017)*

La stimulation électrique ou magnétique d'une région cérébrale d'intérêt permet d'établir solidement l'implication causale de la région dans la production du comportement. Cependant ces approches ne permettent pas d'identifier les mécanismes physiologiques qui sous-tendent le traitement neuronal de l'information.

Chez l'humain, ce type de question peut cependant être étudié par le biais d'études pharmacologiques. La question ne sera plus de savoir si une région participe à la génération du comportement, mais plutôt de savoir si tel ou tel neurotransmetteur y joue un rôle. En 2016, par exemple, Fabien Vinkier et ses collègues montrèrent qu'une injection de kétamine [5] altérait spécifiquement la capacité d'apprentissage de sujets sains [6].

Au cours de leur expérience, des participants effectuaient des choix entre deux symboles visuels abstraits et apprenaient progressivement quels symboles étaient associés à des récompenses positives ou négatives. Ils montrèrent alors que les sujets drogués apprenaient moins bien les associations action-récompense. Plus précisément, par rapport au groupe contrôle, leur vitesse d'apprentissage dépendait moins de leurs succès précédents. Les auteurs démontrèrent ainsi un rôle causal des récepteurs NMDA dans l'intégration de nos expériences passées à l'évaluation de nos décisions.

5. La kétamine est un antagoniste du récepteur NMDA, auquel se lie le glutamate, un neurotransmetteur excitateur du système nerveux central. Elle inhibe donc l'action du récepteur NMDA.

6. Vinckier, F., Rigoux, L., Oudiette, D., and Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature communications*, 9(1):1–12

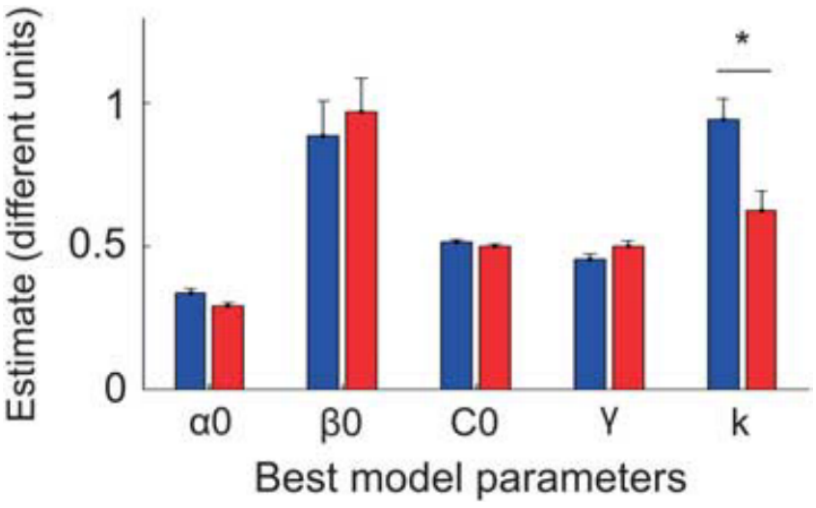



FIGURE 4.3 – *Paramètres moyens du modèle d'analyse. $\alpha_0$ : vitesse d'apprentissage initial, $\beta_0$ : température initiale du choix, C0 : confiance initiale, $\gamma$ : vitesse d'apprentissage de la confiance (en l'optimalité des choix passés), k : poids de cette confiance dans la vitesse d'apprentissage.*

Ce type d'approche revêt un rôle tout particulier dans la recherche clinique en psychiatrie, où l'on cherche précisément à traiter de telles altérations comportementales. Cependant, elle ne permet pas d'établir un modèle neuronal détaillé des processus cognitifs déterminant le comportement. En termes de niveau de Marr, elles ne nous informent que sur une partie de l'implémentation physique du système, et non sur ses buts ou ses méthodes.

De plus, ces méthodes sont quasi exclusivement des méthodes confirmatoires. Étant donné leurs difficultés pratiques, elles se prêtent mal à la découverte de nouveaux déterminants biologiques et sont généralement réservées à l'évaluation d'hypothèses déjà soutenues par d'autres éléments de preuve.

## 4.2 ANALYSES STATISTIQUES IRMF

La neuro-imagerie, et tout particulièrement l'IRMf, permet, elle aussi d'étudier les déterminants biologiques du comportement. Il existe deux méthodes majeures permettant cette étude. Les méthodes dites d'encodage, où l'on cherche à détecter une réponse cérébrale induite par une manipulation expérimentale, et les méthodes de décodages, où l'on cherche à lire une information dans l'activité cérébrale (voir chapitre précédent).

Les analyses d'encodages permettent d'étudier comment un stimulus est perçu et décomposé par le cerveau. Par exemple, en 2007, Sabrina Tom et ses collègues montrèrent que le phénomène d'aversion à la perte était directement lié à la réponse du striatum aux options présentées lors d'un choix[7]. Dans son expérience, une vingtaine de participants étaient allongés dans une machine IRM et devaient soit accepter soit refuser des paris. Pour chaque pari, deux sommes étaient présentées, une perte potentielle et un gain potentiel, le résultat du pari étant déterminé par un lancer de pièce[8]. Les auteurs montrèrent non seulement que le striatum ventral répondait aux gains et aux pertes en jeu, mais également que l'amplitude relative de ces réponses déterminait l'aversion à la perte de leurs sujets. Ils établirent ainsi un parallèle entre un mécanisme cognitif du comportement, l'aversion à la perte, et un déterminant biologique, la sensibilité neuronale relative des pertes par rapport aux gains. Dans ce type d'approche, le lien entre activité cérébrale et comportement est indirect. Dans un premier temps, l'influence de la manipulation expérimentale sur le comportement est établie. Dans un deuxième temps, l'influence de la manipulation expérimentale sur l'activité cérébrale est établie. Dans un troisième temps, on étudie le lien statistique entre les deux types d'influences.

7. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518

8. La probabilité de gagner ou de perdre le pari était de 50%

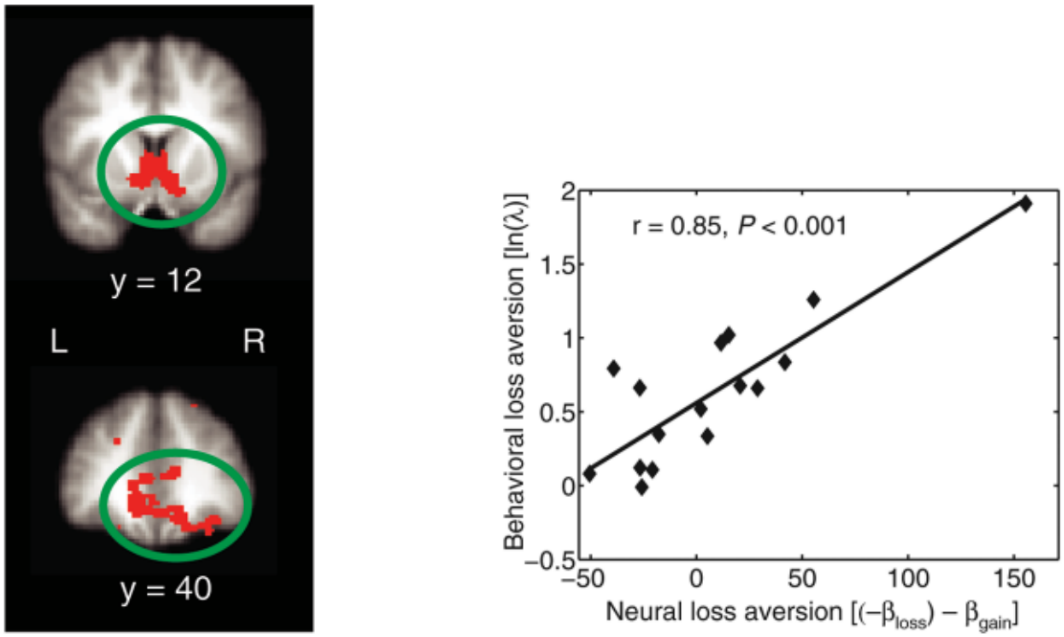



FIGURE 4.4 – *Zones cérébrales où la sensibilité aux gains et la sensibilité aux pertes sont significatives. Relation entre la sensibilité relative neurale et comportementale aux gains et aux pertes.. Figures extraites de (Tom et al., 2007)*

Les analyses de décodage, elles, peuvent être utilisées pour prédire le comportement essai par essai, directement à partir de l'activité neuronale. Par exemple, en 2008, Soon et ses collègues montrèrent qu'un choix pouvait être décodé de l'activité BOLD dix secondes avant que les sujets prennent conscience de leur décision [9]. Pendant l'expérience, on présentait aux sujets une succession rapide de lettres à l'écran. La tâche des participants consistait à déclencher un mouvement volontaire (appuyer sur un bouton) à leur convenance, autant de fois qu'ils le désiraient. Par ailleurs, ils devaient mémoriser la lettre qui apparaissait à l'écran au moment où ils décidaient d'initier leur mouvement et par la suite la communiquer aux chercheurs. Les auteurs purent ainsi évaluer le délai entre le moment où les sujets rapportent la prise de conscience de leur décision et celui à partir duquel le choix des sujets est lisible dans leur activité cérébrale. De manière remarquable, les choix des sujets pouvaient être décodés dix secondes avant qu'ils en aient conscience.

9. Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543–545

Bien entendu, les approches de décodage peuvent aussi être utilisées pour évaluer la quantité d'information qu'il est possible d'extraire, non pas sur un comportement, mais sur un stimulus ou une instruction manipulée par l'expérimentateur. Dans ce cas, l'analyse renverse la relation causale « naturelle » de l'expérience (c'est-à-dire le fait que la manipulation expérimentale induise une réponse cérébrale), dans le but d'exploiter l'information distribuée sur les profils d'activité cérébrale.

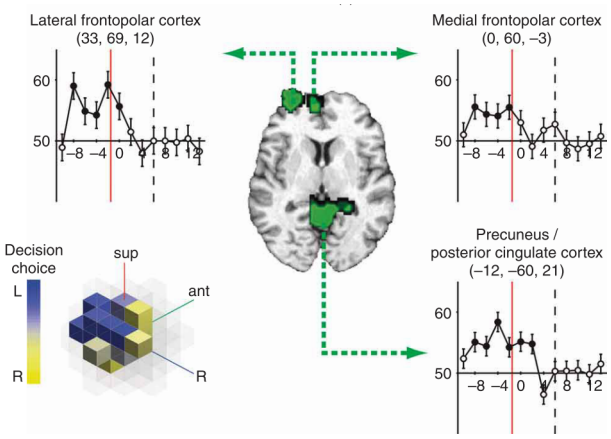



FIGURE 4.5 – *Régions permettant la prédiction du choix des sujets. Les graphes décrivent le pourcentage de décodages réussis par rapport au temps les séparant de la prise de conscience des sujets. Figure extraite de (Soon et al., 2008).*

Ces méthodes d'encodage et de décodage permettent ainsi d'étudier le lien entre une activité physiologique et, d'une part le traitement d'un stimulus ou d'une instruction, et d'autre part la prédiction du comportement. Elles permettent également de comparer, de manière exploratoire, toutes les régions cérébrales et ainsi de localiser les territoires anatomiques impliqués. Certes, elles n'offrent pas de preuves causales, puisqu'elles ne s'appuient que sur des corrélations, mais elles permettent de formuler des théories sur l'organisation cérébrale qui sous-tend nos actions, celles-ci pouvant ensuite être validées par une approche causale.

Malheureusement, ces deux approches n'offrent qu'un point de

vue partiel de la production du comportement. Dans les études d'encodage, l'analyse ne tient généralement pas compte du comportement [10]. Or, si un stimulus évoque une augmentation d'activité dans une région, rien ne garantit que cette sensibilité régionale détermine le comportement. La réponse cérébrale peut par exemple être liée à une perception passive de l'environnement, ou à un processus mnésique. Dans les études de décodage du comportement, c'est le contexte dans lequel les sujets agissent qui est ignoré, or celui-ci détermine, au moins partiellement, le comportement. Par ailleurs, le fait de pouvoir prédire un comportement à partir d'une activité neuronale ne garantit pas que le cerveau utilise cette activité pour générer le comportement. Même lorsque la prédiction précède l'action, on ne peut exclure les activités mémorielles ou métacognitives observant simplement d'autres processus cérébraux sans y prendre part. En résumé, les études d'encodage ne disent rien de l'utilisation d'une information et les études de décodage ne disent rien de son traitement.

10. Il ne sert que de prétexte pour inciter les participants à prêter attention aux stimuli.

Une alternative à ces deux méthodes consiste à modéliser conjointement la génération des données neuronales et comportementales [11 12 13]. Plus précisément, il s'agit de considérer la chaîne de traitement cérébral de l'information, depuis les stimuli manipulés expérimentalement jusqu'à la réponse comportementale, en passant par les étapes de traitement opérées par le réseau impliqué. À ce jour, au moins deux types d'approches existent, qui permettent d'aborder l'identification des déterminants biologiques du comportement de cette manière : le modèle bDCM et l'analyse de médiation.

11. Palestro, J., Bahg, G., Sederberg, P., Lu, Z.-L., Steyvers, M., and Turner, B. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *J. Math. Psychol*, 84:20–48

12. Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206

13. Turner, B. M., Palestro, J. J., Miletić, S., and Forstmann, B. U. (2019b). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews*, 102:327–336

En 2015, Lionel Rigoux et Jean Daunizeau proposent le modèle bDCM, qui permet d'étudier la propagation de l'information utile à la détermination du comportement, à travers un réseau de régions cérébrales d'intérêt [14].

14. Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221

Ce faisant, ils étendent la méthode DCM pour évaluer la contribution d'une connexion à la génération du comportement. La méthode DCM classique consiste à estimer la structure de connectivité entre régions cérébrales, ainsi que leurs sensibilités aux stimuli, pour expliquer les variations temporelles du signal BOLD observé. Lionel Rigoux et Jean Daunizeau ajoutent alors une contrainte supplémentaire à cette inférence : la structure de connectivité doit également prédire les variations de comportement, et donc *in fine* les prédire à partir de la dynamique neuronale induite par les manipulations expérimentales. Ce faisant, le modèle bDCM ne cherche plus à capturer n'importe quelles connexions, mais uniquement celles permettant de prédire le comportement. Réciproquement, il ne capture pas non plus n'importe quelles fluctuations comportementales, mais uniquement celles expliquées par les interactions des régions étudiées.

Ce type d'approche permet une modélisation réellement neurobiologique du comportement. En principe, à l'issue d'une analyse bDCM, le chercheur dispose d'un modèle quantitatif capable de prédire la réponse comportementale d'un individu à une séquence de stimuli donnée. L'approche bDCM reste néanmoins restreinte à la

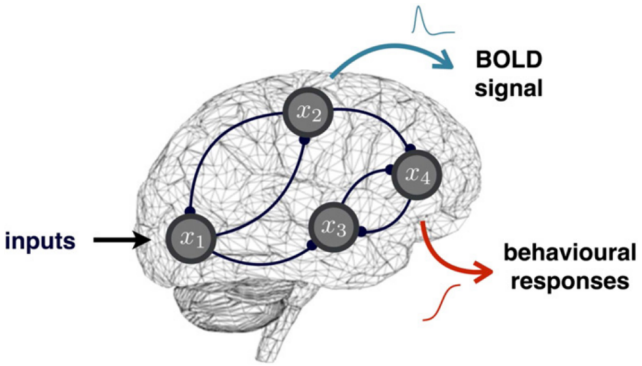



FIGURE 4.6 – *Schéma illustratif du modèle bDCM. Figure extraite de (Rigoux and Daunizeau, 2015)*

confirmation d'hypothèses. En particulier, les régions incluses dans le réseau étudié doivent être préétablies. Par ailleurs, la pertinence de l'inférence dépend de la validité des hypothèses de modélisation sur la nature des transformations de l'information opérées par le réseau (à partir des stimuli expérimentaux).

Une autre approche de modélisation conjointe des données IRMf et du comportement permet d'étudier le cerveau entier et non juste un ensemble de régions préétablies : l'analyse de médiation neurale proposée par Tor Wager en 2008 [15]. Dans une série d'études sur le traitement neuronal de la douleur, la peur et la régulation émotionnelle [16] [17] [18] [19], il proposa d'allier le principe de l'analyse d'encodage à celui du décodage. Il identifia des régions cérébrales qui, d'une part, répondaient aux stimuli expérimentaux, et d'autre part, permettait de prédire le comportement de chaque sujet (au-delà des prédictions issues de la manipulation expérimentale).

L'analyse de médiation considère deux critères que doit vérifier un processus de traitement neural de l'information déterminant le comportement : 1) être sensible aux entrées de la chaîne de traitement : les stimuli expérimentaux, et 2) contenir une information présente en aval dans la chaîne de traitement, c'est-à-dire présente dans le comportement du sujet. Cette approche permet d'explorer la totalité du cerveau pour identifier les régions qui répondent à ces deux critères.

Bien entendu, et comme pour les autres approches statistiques du même genre (y compris bDCM), il ne s'agit toujours pas ici d'une approche à proprement parler « causale » [20]. De plus, si l'analyse de médiation permet d'identifier des régions susceptibles de faire partie intégrante du processus cérébral de détermination du comportement, elle ne permet pas d'identifier les mécanismes biologiques qui traitent l'information utile à l'élaboration du comportement. En d'autres termes, l'analyse de médiation cherche à détecter une transformation de l'information opérée localement (par une région cérébrale) de manière agnostique. Par comparaison avec bDCM, l'inférence ne dépend donc pas de la validité d'hypothèses de modélisation sur la nature de cette transformation. En revanche, il n'est pas possible d'utiliser les résultats de l'analyse de médiation pour effectuer une prédiction quantitative du comportement sans disposer de données

15. David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., and Depaulis, A. (2008). Identifying neural drivers with functional mri : An electrophysiological validation. *PLoS Biol*, 6:315
16. Atlas, L., Bolger, N., Lindquist, M., and Wager, T. (2010). Brain mediators of predictive cue effects on perceived pain. *J. Neurosci. Off. J. Soc. Neurosci*, 30:12964–12977
17. Wager, T., Waugh, C., Lindquist, M., Noll, D., Fredrickson, B., and Taylor, S. (2009b). Brain mediators of cardiovascular responses to social threat : part i : Reciprocal dorsal and ventral subregions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47:821–835
18. Wager, T., Ast, V., Hughes, B., Davidson, M., Lindquist, M., and Ochsner, K. (2009a). Brain mediators of cardiovascular responses to social threat, part ii : Prefrontal-subcortical pathways and relationship with anxiety. *NeuroImage*, 47:836–851
19. Wager, T. (2008). canlab/ m3 mediationtoolbox (cognitive and affective neuroscience laboratory
20. On reviendra sur ce point plus en détail au chapitre suivant.

IRMf au moment où le comportement est produit.

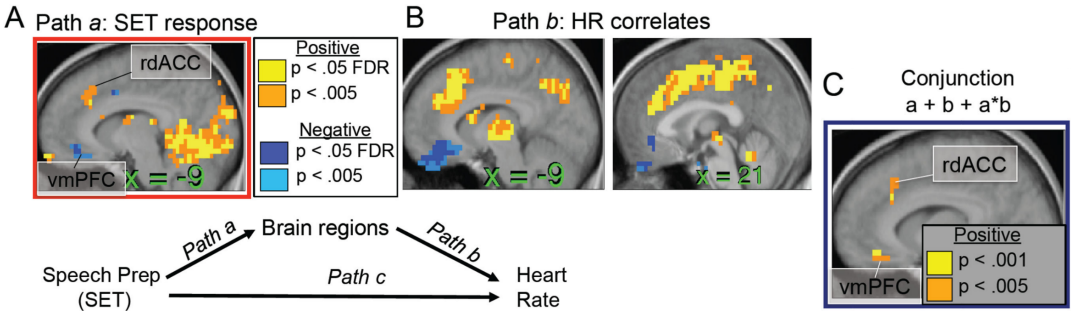



FIGURE 4.7 – *Cartes de médiation de l'impact d'un stress social sur les battements cardiaques. A : carte de la sensibilité au stress social, B : carte des zones prédicatrices des battements cardiaques, C : carte des médiateurs. Figures extraites de (Wager et al., 2009a).*

## 4.3 MODÈLES FORMELS DU TRAITEMENT DE L'INFORMATION

La production du comportement peut également être étudiée via l'utilisation de modèles mathématiques du traitement neural de l'information. Typiquement, ces modèles « computationnels » décrivent la manière dont l'information entrante (stimuli et/ou instructions) est transformée par le cerveau pour déterminer le comportement, dans un contexte cognitif donné.

Dans leur variante la plus simple, les études de neurosciences computationnelles du comportement cherchent à valider, grâce aux données neurales, l'existence d'une étape de traitement de l'information prédite par le modèle théorique. En termes de niveaux de Marr, il s'agit de faire le lien entre les deux premiers niveaux et le troisième, entre les méthodes nécessaires à un but et leur implémentation physique. Une fois formalisé mathématiquement, un modèle computationnel capable de prédire le comportement dans une tâche donnée prédit aussi la manière dont certaines quantités nécessaires au traitement de l'information varient au cours de la tâche. Ces quantités ne sont pas des mesures directes de l'expérience, ce sont des constructions mathématiques, des calculs intermédiaires permettant de comprendre et de prédire le comportement d'un sujet. Leurs corrélats neuronaux ne correspondent donc pas à un élément tangible de l'expérience. Ils indiquent une étape intermédiaire du traitement de l'information qui est utilisée pour déterminer le comportement.

De par ses caractéristiques techniques, l'IRMf peut paraître inappropriée pour étudier l'utilisation d'une information neurale. Il est notamment difficile d'établir des liens causaux par une analyse IRMf [21] [22]. Cependant, ces limitations peuvent être partiellement contournées grâce aux analyses dites model-based [23] que l'on vient de décrire. En résumé, l'approche model-based consiste à comparer les prédictions des variables internes d'un modèle computationnel directement aux variations du signal BOLD du sujet afin de valider le modèle. Si une région cérébrale montre une activité qui corrèle avec la variable computationnelle décrivant l'étape de traitement, on suppose alors que cette région est impliquée dans cette étape de

21. Seth, A., Chorley, P., and Barnett, L. (2013). Granger causality analysis of fmri bold signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–555

22. Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., and Woolrich, M. (2011). Network modelling methods for fmri. *NeuroImage*, 54:875–891

23. Gläscher, J. and O'Doherty, J. (2010). Model-based approaches to neuroimaging : Combining reinforcement learning theory with fmri data. *Wiley Interdisciplinary Reviews : Cognitive Science*, 1(4):501–510

traitement. C'est-à-dire qu'elle opère des calculs similaires à ceux du modèle afin de produire la réponse comportementale.

En 2003, John O'Doherty et ses collègues montrèrent ainsi que l'activité du striatum corrélait aux variations de l'erreur de prédiction prescrites par un modèle d'apprentissage associatif [24]. Au cours de leur expérience, leurs sujets, légèrement assoiffés, observaient passivement des formes géométriques abstraites, des fractales. Chaque forme était associée, trois secondes plus tard, à la délivrance d'une petite quantité de liquide au goût plaisant ou neutre [25], ou à l'absence de liquide. Les auteurs utilisèrent alors un modèle dont le but est d'apprendre progressivement l'association entre les indices (les fractales) et la récompense (positive, neutre ou négative). Dans de telles conditions, le modèle d'apprentissage prédit les variations d'un signal dit « d'erreur de prédiction [26] » au cours des essais successifs et pour chaque association indice→récompense. En particulier, il prédit une hausse d'activité au moment de la délivrance du liquide en début d'expérience, puis un décalage progressif de ce signal vers l'instant où la fractale était présentée. Il prédit également une hausse du signal d'erreur de prédiction lorsqu'une fractale généralement associée à l'absence de liquide est suivie par la délivrance du liquide plaisant. Inversement, il prédit une diminution du signal d'erreur de prédiction lorsque le liquide n'est pas délivré après la présentation d'une fractale généralement associée au liquide plaisant. Les auteurs montrèrent alors que le signal BOLD du striatum suivait ces prédictions et donc validait leur modèle d'apprentissage associatif. Par la suite, cette étude a été étendue à un contexte d'apprentissage instrumental (voir Thorndike au chapitre 2), et le signal d'erreur de prédiction correspondant corrélait avec l'activité du cortex préfrontal ventromédian. [27] [28].

Le principe des approches model-based est résumé dans les figures 4.9 et 4.10.

24. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337

25. Bien que les auteurs n'emploient pas ce terme, on peut faire l'amalgame entre ce liquide et le jus de fruit utilisé dans les recherches sur le singe.

26. Le modèle d'apprentissage « associatif » s'appuie sur l'erreur de prédiction pour actualiser ses prédictions, c'est-à-dire qu'il apprend à partir de la différence entre l'amplitude de la récompense reçue et la valeur prédite par le stimulus présenté (selon le modèle).

27. Gläscher, J., Hampton, A., and O'Doherty, J. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19(2):483–495

28. Kim, H., Shimojo, S., and O'Doherty, J. (2006). Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. *PLoS Biology*, 4(8):1453–1461

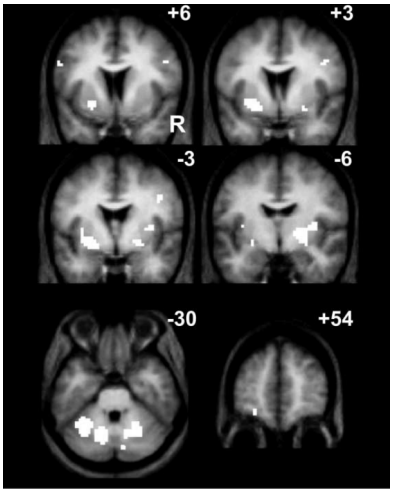



FIGURE 4.8 – *Cartes des régions corrélant significativement avec l'erreur de prédiction au moment de présentation des fractales. Figure extraite de (O'Doherty et al., 2003).*

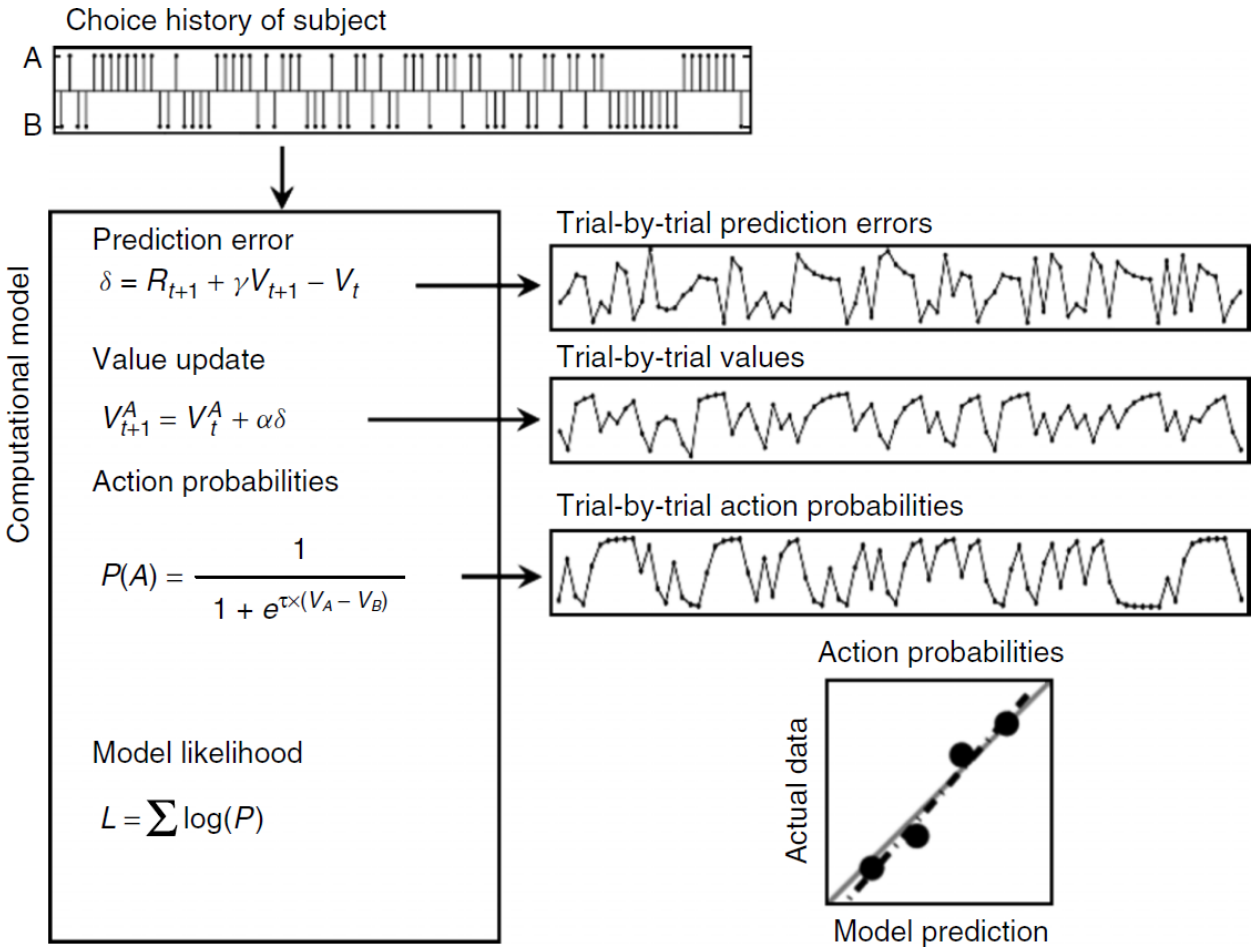



FIGURE 4.9 – *Exemple de la partie concernant le modèle comportementale d'une analyse IRMf model-based. Tout d'abord on ajuste les paramètres d'un modèle permettant de décrire les choix effectués par un sujet. Ici un modèle d'apprentissage par renforcement est utilisé pour décrire les choix binaires d'un sujet. Il utilise 5 variables internes : une erreur de prédiction, les valeurs des deux options présentées aux sujets et les probabilités que le sujet choisisse l'une ou l'autre. Chaque variable du modèle fluctuera au cours de l'expérience en fonction des mécanismes du modèle et des choix du sujet. Figure extraite de (Gläscher and O'Doherty, 2010).*

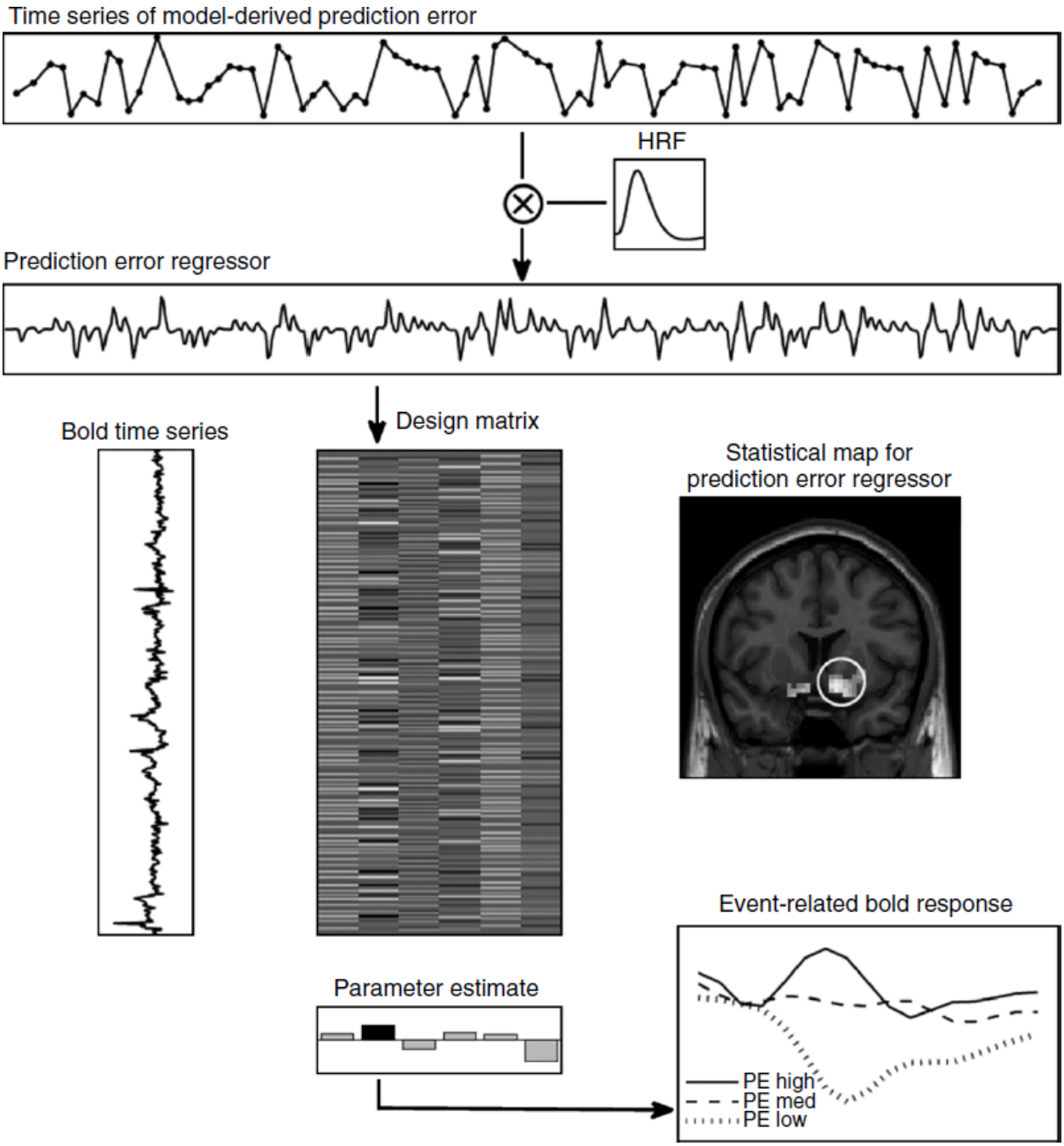



FIGURE 4.10 – *Exemple de la partie concernant les corrélats neuronaux d'une analyse IRMf model-based. Les variables internes du modèle comportementale sont extraites et converties dans le format du signal BOLD (voir texte). On les réunit ensuite dans une matrice (Design matrix) afin de les comparer au signal BOLD grâce à un GLM. Enfin on scanne le cerveau pour identifier les zones pour lesquels le GLM fournit des poids statistiquement non nuls, et où les variables du modèle comportemental permettent de décrire les niveaux d'activations du signal. Figure extraite de (Gläscher and O'Doherty, 2010).*

Cette approche permet ainsi de comparer l'activité physiologique d'une région cérébrale à un mécanisme de traitement de l'information directement interprétable. L'un des intérêts de ce type d'approche est que chaque variable du modèle computationnel a un rôle et une interprétation fonctionnelle précise [29], ce qui permet d'associer la région cérébrale qui l'opère à un but et un mécanisme, au sens de Marr [30]. D'autre part, l'analyse model-based identifie des signaux utilisables par le cerveau, à défaut de vérifier qu'ils soient effective-

29. Notamment quand le modèle s'appuie sur la littérature psychologique,

30. Krakauer, J., Ghazanfar, A., Gomez-Marin, A., MacIver, M., and Poeppel, D. (2017). Neuroscience needs behavior : Correcting a reductionist bias. *Neuron*

ment utilisés (comme pourrait le faire une analyse par perturbation). En d'autres termes, s'il ne s'agit pas ici d'une approche causale, l'approche model-based permet néanmoins d'interpréter, sous les hypothèses du modèle, le rôle computationnel d'une région cérébrale dans la détermination du comportement.

Cependant, les approches *model-based* de ce type ignorent souvent les contraintes biologiques qui peuvent moduler le traitement neural de l'information. Ces modèles sont formulés dans un cadre psychologique (niveau deux de Marr), et les données neuronales sont typiquement utilisées pour identifier la région cérébrale qui implémente (au sens de Marr) l'étape de traitement prédite par le modèle. Or l'interprétation du lien statistique entre l'activité d'une région et la variable computationnelle d'intérêt est en fait moins triviale qu'il n'y paraît. En résumé, pour qu'un tel signal soit détectable avec l'IRMf, des centaines de milliers de neurones doivent effectuer la même opération, afin qu'elle laisse une trace dans le coût métabolique global de la région. Pourtant les modèles utilisés jusqu'ici tendent à présenter des mécanismes relativement simples. Alors pourquoi faudrait-il des centaines de milliers de neurones pour les implémenter ? Pour expliquer cela, certains auteurs avancent que le traitement neuronal de l'information est en fait extrêmement redondant [31]. Chaque neurone effectue des calculs si similaires à ceux de ses voisins que leur trace est préservée dans l'activité moyenne de la population. Par ailleurs, les approches model-based de ce type ignorent généralement le problème du format de l'information accessible aux neurones. Si le striatum calcule l'erreur de prédiction, il doit le faire à partir d'une représentation neurale des informations pertinentes (les options présentées au participant et leurs valeurs, la récompense, etc.). Or ces représentations peuvent ne pas être similaires à celles manipulées par l'expérimentateur [32] [33]. Par exemple, on sait que la perception des quantités est logarithmique et non linéaire [34]. En principe, ce type de transformation devrait être pris en compte dans les modèles décrivant l'utilisation de ces informations par des systèmes situés en aval des systèmes perceptifs et impliqués dans le contrôle du comportement [35].

Une seconde approche s'intéresse directement aux contraintes biologiques qui pourraient s'imposer sur le code neural. Il s'agit ici de formaliser ces contraintes, d'en déduire les limitations qu'elles imposent sur le traitement de l'information, et/ou d'étudier les organisations neurales permettant de compenser ces limites.

Par exemple, le codage efficace proposé par Horace Barlow (voir chapitre 2) prend explicitement en compte les contraintes dues aux nombres finis de neurones d'une région cérébrale et à leurs activations bornées. Le modèle de codage efficace prédit que les signaux les moins fréquemment rencontrés dans l'environnement seront filtrés par les traitements neuronaux, afin de maximiser l'information totale retenue. Depuis lors, cette prédiction a été vérifiée sur les neurones de la rétine de la mouche par Simon Laughlin en 1981 (voir chapitre précédent). Ces contraintes, cependant, ont d'autres conséquences.

31. Guest, O. and Love, B. (2016). What the success of brain imaging implies about the neural code. *BioRxiv*, 1:071076

32. Brette, R. (2019). Is coding a relevant metaphor for the brain ? behavioral and brain sciences

33. de Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain ? *Psychonomic Bulletin Review*, page 1–14

34. Nieder, A. and Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, 32:185–208

35. Wood, G., Nuerk, H., Sturm, D., and Willmes, K. (2008). Using parametric regressors to disentangle properties of multi-feature processes. *Behavioral and Brain Functions*, 4(i):1–12

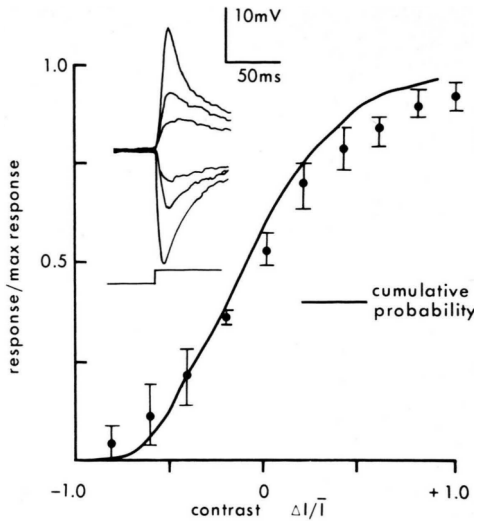



FIGURE 4.11 – *Comparaison de la sensibilité des neurones de la rétine d'une mouche à l'intensité lumineuse (points noirs), et de la probabilité cumulée de ces fréquences (trait continu). Figure extraite de (Laughlin, 1981)*

Dans une étude publiée en 2010, Shunsuke Kobayashi et ses collègues ont mis en évidence le mécanisme d'adaptation neuronale prédit par la théorie du codage efficace [36]. Dans leur expérience, deux macaques rhésus effectuent une tâche d'observation forcée pour obtenir une récompense alimentaire : une certaine quantité de jus de fruits. Ils doivent tout d'abord stabiliser leur regard sur le centre de l'écran, où apparaît un carré ou un cercle, puis une fractale leur indique la quantité de récompenses en jeu. Enfin, le jus est délivré s'ils maintiennent leur regard sur la fractale durant 1.5 seconde. Les auteurs utilisent six fractales représentants deux distributions de quantité de jus différentes, une associée au carré, l'autre au cercle [37]. Elles ont toutes deux la même moyenne, mais des déviations standard différentes. La théorie du codage efficace prédit que les neurones qui encodent la valeur du jus doivent adapter leur sensibilité à la quantité de jus de telle manière qu'ils étalent leurs réponses sur la plage de variations associée à chaque fractale.

36. Kobayashi, S., De Carvalho, O., and Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience*, 30(2):534–544

37. Dans leur expérience, chaque distribution est utilisée séparément de l'autre au cours de petits blocs de 4 à 13 essais, ou de longs blocs de 14 à 93 essais.

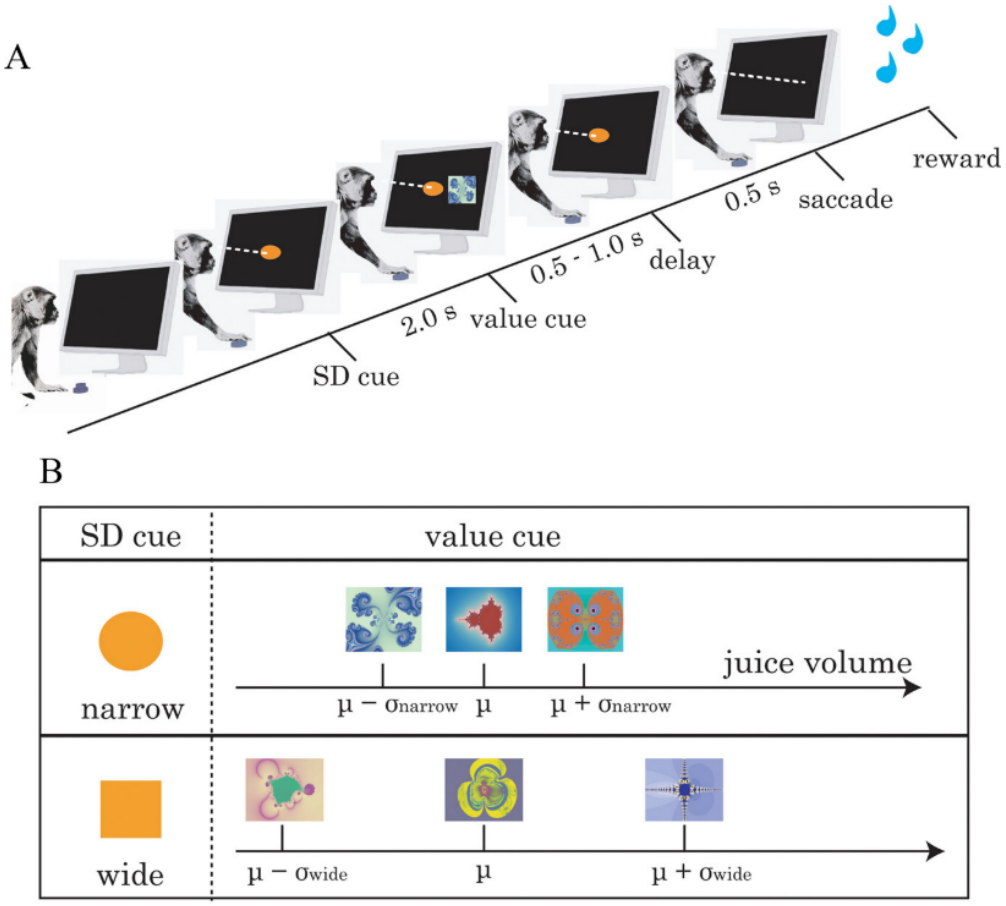



FIGURE 4.12 – *Présentation de l'expérience de Kobayashi. En haut : la succession des étapes présentées au singe. En bas : les deux distributions de jus de fruits utilisées et les fractales y étant associées.*

Shunsuke Kobayashi et ses collègues montrent alors que lorsque la variance de la distribution de récompense augmente, les neurones du cortex orbitofrontal (OFC) sont moins sensibles à un incrément d'un millilitre de jus.

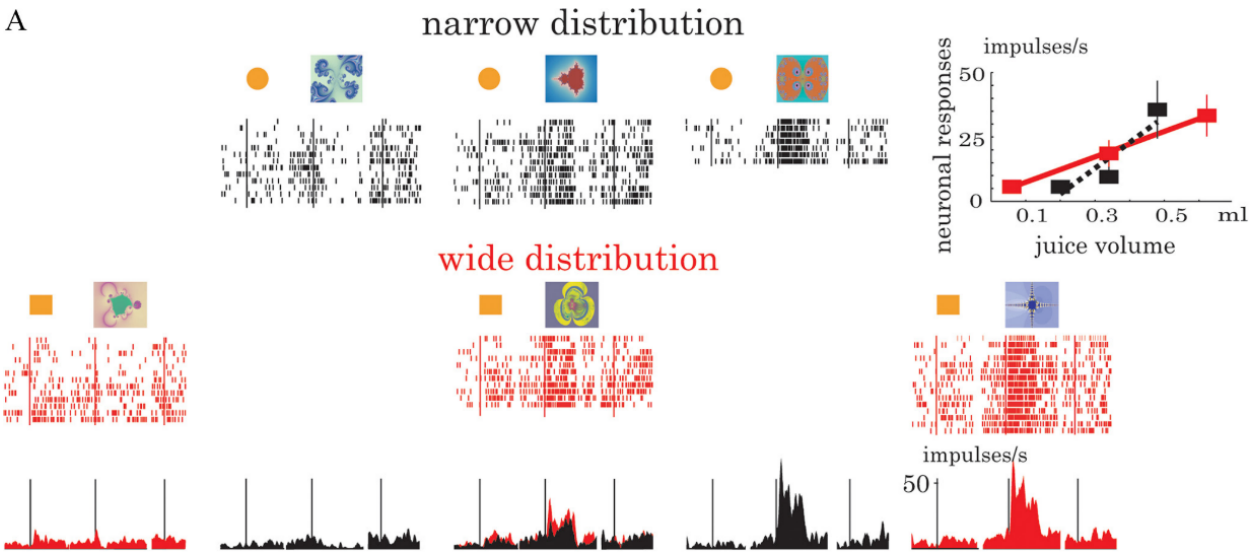



FIGURE 4.13 – *Réponses neuronales des singes par rapport à chaque fractale. Les traits correspondent à l'instant de présentation de la distribution (cercle ou carré) et de la fractale, chaque ligne correspond à l'enregistrement d'un neurone et chaque point à l'émission d'un potentiel d'action. En haut en noir, les réponses à la distribution fine, au milieu en rouge, les réponses à la distribution large. En bas, le nombre total de potentiels d'action à chaque instant. En haut à droite la sensibilité globale des neurones à la récompense (quantité de jus) pour chaque distribution. Figure extraite de (Kobayashi et al., 2010).*

Dans une étude publiée en 2018, Polania et ses collègues ont également démontré deux prédictions comportementales issues de la théorie du codage efficace [38]. Dans leur expérience, les participants (humains) commençaient d'abord par attribuer une valeur à différents objets communs, puis indiquaient leur préférence lorsque deux de ces mêmes objets leur étaient présentés (tâche de choix « binaire »). Si la valeur des objets est encodée efficacement au sens de Barlow, les valeurs les plus fréquentes devraient être les mieux encodées. C'est-à-dire que la perte d'information sur la valeur induite par le bruit neural devrait être minimale sur les valeurs les plus fréquentes. En conséquence, les choix effectués entre des objets dont la valeur est fréquente devraient être plus cohérents que les choix effectués entre des objets dont la valeur est rare. C'est effectivement ce que Polania et ses collègues observèrent dans les choix de leurs participants.

Les auteurs s'intéressèrent également à la précision de l'encodage de la valeur. Sous l'hypothèse que la valeur est décodée par un système bayésien qui utilise une distribution a priori de la valeur des objets, la valeur rapportée par les participants devrait être biaisée vers la valeur centrale de cette distribution. De plus, si l'encodage de la valeur est efficace, alors le bruit neural d'encodage devrait moduler ce biais de manière non triviale. En effet, le bruit neural fait varier la précision de l'information décodable par le système en fonction de la fréquence des valeurs dans l'environnement. Polania et ses collègues exposèrent leurs sujets plus ou moins longtemps aux objets durant la phase d'évaluation, postulant que cela réduirait le bruit neural

38. Polania, R., Woodford, M., and Ruff, C. (2018). Efficient coding of subjective value. *BioRxiv*, 358317

d'encodage. Ils montrèrent ainsi que leurs participants, tout comme un modèle formel mêlant codage efficace et décodage bayésien, ne rapportaient pas la même valeur après une exposition longue ou courte.

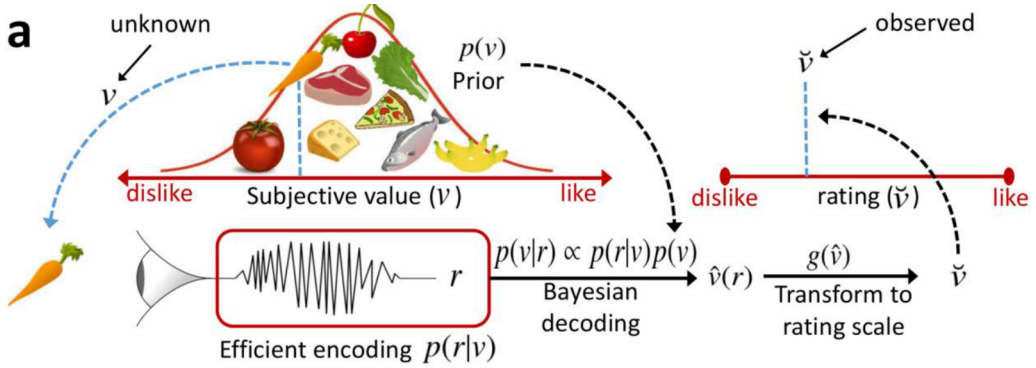



FIGURE 4.14 – *Modèle de Polania. La valeur de chaque objet peut être plus ou moins fréquente et est modélisée par une gaussienne représentant toutes les valeurs possibles. La valeur de chaque objet est ensuite encodée selon un codage efficace privilégiant les valeurs les plus fréquentes. Enfin cet encodage est décodé par une méthode bayésienne avant d'être transformé pour être énoncé par le sujet dans l'échelle de l'expérience. Figure extraite de (Polania et al., 2018).*

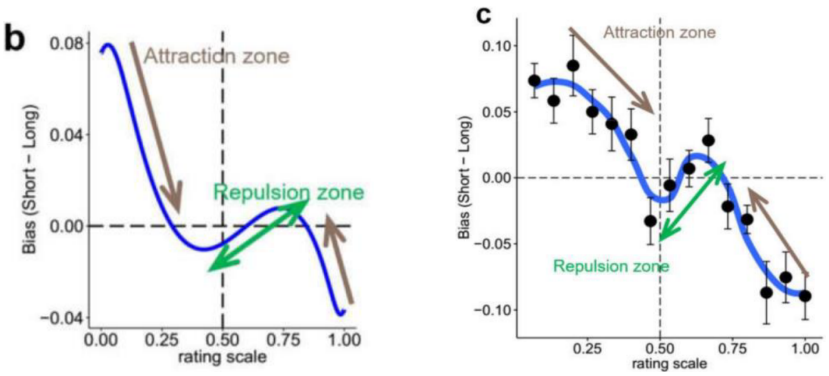



FIGURE 4.15 – *Bais d'évaluation selon le modèle (gauche) et mesuré chez les participants (droite). Figure extraite de (Polania et al., 2018).*

Plus précisément, lorsque la valeur était peu fréquente (basses et hautes valeurs dans leur expérience), le biais vers le centre de la distribution (0.55 dans la figure ci-dessous) était augmenté par un bruit neural fort. Inversement, lorsque la valeur était plus fréquente, le biais vers le centre de la distribution était diminué, et même inversé !

Utilisé pour interpréter le comportement, le modèle de codage efficace joue donc le rôle d'un « microscope mathématique » capable de révéler les limitations biologiques du traitement neural de l'information.

D'autres types de contraintes biologiques peuvent également être formalisées et utilisées pour étudier le code neural de la production du comportement. On peut mentionner par exemple le codage populationnel [39 40 41], et le codage prédictif [42 43]. Tout comme pour les approches model-based en IRMf, la validation de ces modèles par des données comportementales et/ou neurales est confirmatoire et ne fournit pas un niveau de preuve équivalent aux approches causales. Néanmoins ces théories proposent des scénarios quantitatifs modélisant explicitement l'impact de certains déterminants biologiques sur le comportement. En principe, ces modèles sont donc compatibles

39. Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438

40. Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132

41. Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural computation*, 10(2):403–430

42. Aitchison, L. and Lengyel, M. (2017). With or without you : predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227

43. Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711

avec une approche expérimentale causale, qui perturberait directement les mécanismes biologiques sous-jacents et vérifierait que les conséquences observées valident les prédictions des modèles. Par ailleurs, il est possible d'utiliser les données neurales pour calibrer les prédictions comportementales du modèle, puisque celles-ci sont, en théorie, informatives vis-à-vis des mécanismes biologiques étudiés.

## 4.4 CONTRIBUTIONS ATTENDUES

Le tableau ci-dessous récapitule l'intérêt des approches précédentes.

| Méthodes | Question comportementale typique | Localisation anatomique | Implication d'un mécanisme physiologique | Modèle causal, génératif ou descriptif | Exploratoire / Confirmatoire |
|---|---|---|---|---|---|
| Intervention causale régionale | La région X est-elle impliquée dans le comportement ? | Oui | Mitigé | Causal | Confirmatoire |
| Intervention causale globale | Le neurotransmetteur X est-il impliqué dans le comportement ? | Non | Fort | Causal | Confirmatoire |
| Décodage | L'information X est-elle neuralement prédictible ? | Oui | Faible | Descriptif | Exploratoire |
| Modèles neuro-comportementaux | Les réponses neurales influencent-elles le comportement ? | Oui | Fort | Génératif | Majoritairement confirmatoire |
| Model-based | La transformation f(X) des stimuli X est-elle effectuée par le cerveau ? | Oui | Faible | Génératif | Exploratoire (substrat), confirmatoire (mécanisme) |
| Code neural | Comment un réseau neuronal s'adapte-t-il à la contrainte X ? | Non | Fort | Génératif | Confirmatoire |

TABLE 4.1 – *Récapitulatif des méthodes d'analyse des déterminants biologiques du comportement, et de leurs caractéristiques*

Dans la suite de cette thèse, je cherche à exploiter les avantages des approches non causales et propose deux méthodes d'analyses des données de neuro-imagerie pour étudier les déterminants biologiques du comportement.

La première s'intéresse à l'analyse de médiation des données IRMf. L'originalité de cette approche consiste en l'utilisation jointe de méthode d'encodage et de décodage pour étudier la génération du comportement (voir chapitre 5,6 et 7). L'ambition de ce premier travail est de déterminer les limites statistiques et conceptuelles de cette méthode, ainsi que proposer des solutions pour améliorer ses propriétés statistiques vis-à-vis de l'identification des régions cérébrales impliquées dans la détermination du comportement.

La seconde méthode proposée permet d'approfondir l'étude des

contraintes biologiques imposées sur le code neural. Plus précisément, les chapitres 8,9 et 10 s'intéressent à la trace de mécanismes physiologiques façonnant les processus décisionnels. La méthode que je propose s'appuie sur des réseaux de neurones artificiels peu profonds contraints à utiliser des fonctions d'activations bornées, une plasticité hebbienne ou encore une adaptation dynamique de leurs sensibilités [44]. Ils sont d'abord ajustés directement au comportement des sujets, avant d'être comparés aux profils multivariés des signaux BOLD d'une région d'intérêt. En particulier, cette seconde méthode permet d'évaluer et de comparer plusieurs types de contraintes biologiques imposées sur le code neural.

Comme nous le verrons, ces deux approches permettent d'expliquer les différences interindividuelles comportementales par des différences de nature biologique (sur un plan anatomique et/ou sur un plan physiologique).

44. voir chapitre 9 pour l'explicitation de ces termes.

# Deuxième partie

# ANALYSE DE MÉDIATION MASSIVEMENT UNIVARIÉES DE DONNÉES IRMF

# 5

# *Introduction sur l'analyse de médiation*

◂ Chapitre 4 Chapitre 6 ▸

Les analyses IRMf peuvent généralement se classer en deux grandes catégories : les analyses d'encodages et de décodages (voir chapitres précédents). Elles permettent d'étudier d'une part comment le cerveau répond à une manipulation expérimentale, et d'autre part quelles informations peuvent y être lues. Ces approches sont simples, rapides [1] et versatiles, mais ne permettent d'étudier qu'une partie de la chaîne de traitement neural de l'information.

Dans ce chapitre j'étudie les propriétés d'une méthode d'analyse combinant ces deux approches pour identifier les mécanismes neuronaux de production du comportement : l'analyse de médiation. Son principe est simple : scanner entièrement le cerveau d'un sujet et sélectionner les voxels correspondant à des médiateurs d'un processus comportemental. C'est-à-dire des voxels à la fois sensibles aux facteurs expérimentaux et prédisant le comportement de chaque sujet.

Le principe de l'analyse de médiation existe depuis 1982 et de nombreux tests de médiation ont déjà été proposés et discutés [2] [3] [4]. Plus récemment, de nouveaux cadres théoriques permettent de décomposer les médiateurs en plusieurs dimensions [5], d'étudier les variations d'une médiation selon le contexte [6] [7] ou de se passer d'hypothèses sur la forme mathématique de la médiation [8]. Depuis quelques années, l'analyse de médiation est même appliquée directement aux données IRMf pour étudier par exemple, la régulation émotionnelle [9], la peur sociale [10] ou encore la perception de douleur [11] [12].

Plutôt que de proposer une nouvelle méthode, ou une nouvelle application de l'analyse de médiation, ce chapitre propose d'étudier les propriétés statistiques de l'analyse de médiation, dans le contexte de son application à l'identification des médiateurs neuraux du comportement (grâce à l'IRMf). J'y compare les principaux tests statistiques proposés depuis 1982 en évaluant leur sensibilité et leur spécificité à l'aide de simulations numériques. En analysant les formes mathématiques de ces tests, j'identifie une propriété contre-intuitive, mais essentielle de l'analyse de médiation : sa dépendance bilatérale aux niveaux de bruit dans les données. Cette propriété me permet alors de caractériser les médiateurs pouvant ou non être identifiés, quel que soit le test. Enfin, j'étudie plus spécifiquement les spécificités liées aux données IRMf. Je compare deux manières de préparer les données : l'une issue des analyses d'encodages, l'autre issue de méthodes de décodages. Je souligne ensuite une ambivalence

1. De nos jours, le cerveau d'un sujet effectuant 1h d'expérience peut être scanné en quelques heures.
2. Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol*, 13:290–312
3. Kisbu-Sakarya, Y., MacKinnon, D. P., and Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate behavioral research*, 49(3):261–268
4. Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies : new procedures and recommendations. *Psychological methods*, 7(4):422
5. Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136
6. Fairchild, A. J. and MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, 10(2):87–99
7. Lindquist, M. A. (2012b). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309
8. Pearl, J. (2012). The mediation formula : A guide to the assessment of causal pathways in nonlinear models. In *Causality*, pages 151–179. John Wiley Sons, Ltd
9. Wager, T. (2008). canlab/ m3 mediationtoolbox (cognitive and affective neuroscience laboratory
10. Wager, T., Waugh, C., Lindquist, M., Noll, D., Fredrickson, B., and Taylor, S. (2009b). Brain mediators of cardiovascular responses to social threat : part i : Reciprocal dorsal and ventral subregions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47:821–835
11. Atlas, L., Bolger, N., Lindquist, M., and Wager, T. (2010). Brain mediators of predictive cue effects on perceived pain. *J. Neurosci. Off. J. Soc. Neurosci*, 30:12964–12977

fondamentale de ce type d'analyse, ayant trait à l'interprétation causale des résultats. J'introduis alors une métrique issue de la théorie de l'information et permettant de lever partiellement cette ambivalence. Je conclus ensuite sur une analyse de données IRMf collectées lors d'une expérience de prise de décision risquée [13]. J'y illustre la sensibilité des tests de médiations en conditions réelles, ainsi que l'intérêt qu'ils représentent pour profiler différents mécanismes neuraux déterminant le comportement.

12. Atlas, L., Lindquist, M., Bolger, N., and Wager, T. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, 155:1632–1648

13. Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., and Liu, T. (2014). Survey of encoding and decoding of visual stimulus via fmri : An image analysis perspective. *Brain Imaging and Behavior*, 8(1):7–23

# 6

# *Meet me in the middle : brain-behavior mediation analysis for fMRI experiments*

◂ Chapitre 5 Chapitre 7 ▸

Ce chapitre présente l'article "*Meet me in the middle : brain-behavior mediation analysis for fMRI experiments*" que j'ai publié en préimpression sur le site bioRxiv. Il est en anglais et est présenté sans modifications.

Adresse web :
https://www.biorxiv.org/content/10.1101/2020.10.17.343798v1.

Abstract Functional outcomes (e.g., subjective percepts, emotions, memory retrievals, decisions, etc...) are partly determined by external stimuli and/or cues. But they may also be strongly influenced by (trial-by-trial) uncontrolled variations in brain responses to incoming information. In turn, this variability provides information regarding how stimuli and/or cues are processed by the brain to shape behavioral responses. This can be exploited by brain-behavior mediation analysis to make specific claims regarding the contribution of brain regions to functionally-relevant input-output transformations. In this work, we address four challenges of this type of approach, when applied in the context of mass-univariate fMRI data analysis : (i) we quantify the specificity and sensitivity profiles of different variants of mediation statistical tests, (ii) we evaluate their robustness to hemodynamic and other confounds, (iii) we identify the sorts of brain mediators that one can expect to detect, and (iv) we disclose possible interpretational issues and address them using complementary information-theoretic approaches. En passant, we propose a computationally efficient algorithmic implementation of the approach that is amenable to whole-brain exploratory analysis. We also demonstrate the strengths and weaknesses of brain-behavior mediation analysis in the context of an fMRI study of decision under risk. Finally, we discuss the limitations and possible extensions of the approach.

## 6.1 INTRODUCTION

Functional outcomes (e.g., subjective percepts, emotions, memory retrievals, decisions, etc...) are partly determined by external stimuli and/or contextual cues. But they may also be strongly influenced by irreducible variability in brain responses to incoming information [1 2]. In particular, neural noise may be a critical determinant of illusory percepts, aberrant emotions, erroneous memory retrievals, biased decisions, etc... [3 4 5]. For most existing statistical data analyses of neurophysiological data, neural noise is typically treated as a statistical nuisance, since it compromises the identification of relationships between measured brain activity and experimental variables [6 7]. This perspective is unfortunate however, since neural noise can provide complementary information regarding how incoming information is processed and/or distorted by the brain to yield functional outcomes [8 9 10].

The critical point is that a brain system may encode functionally-relevant information that is not used by the brain when producing a functional outcome. This has been repeatedly demonstrated in neurological patients who do not exhibit significant behavioral impairments despite being lesioned in brain regions that are known to encode behaviorally-relevant information [11 12]. But what if one can show that neural noise contributes to -otherwise unexplained- behavioral variability ? This is the essence of brain-behavior mediation analysis, which aims at detecting neural systems that both respond to behaviorally-relevant cues or stimuli and eventually impact overt behavior [13].

Recall that any cognitive function can be seen as some form of -potentially complex, context-dependent, redundant, partially unconscious, etc- neural transformation of relevant stimuli into adaptive behavioural outcomes [14]. By adaptive, we simply mean that cognitive functions serve a specific purpose, which can be abstracted and put to a (behavioural) test. At the limit, one could argue that understanding cognitive functions reduces to assessing input-output relationships, where inputs are experimentally controlled stimuli and/or task instructions, and outputs are overt behavioural outcomes. In this view, neuroimaging in healthy subjects should serve to identify how brain networks contribute to the input-output transformation [15 16 17]. A reasonable strategy here is to identify intermediary neural states that mediate the impact of incoming information onto overt behavior and/or subjective reports.

In its simplest form, brain-behavior mediation analysis reduces to a twofold regression analysis that aims at detecting uncontrolled variability in brain responses that significantly improves behavioral predictability. The ensuing statistical tests typically reason as follows : if region M responds to experimental factor X, and explains behaviour Y above and beyond the effect of X, then M mediates the effect of X onto Y. For example, brain-behavior mediation analysis was used to identify the prefrontal and/or subcortical systems that


1. Ferster, D. (1996). Is neural noise just a nuisance ? *Science*, 273:1812–1812
2. Shadlen, M. and Newsome, W. (1994). Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol*, 4:569–579
3. Bays, P. (2014). Noise in neural populations accounts for errors in working memory. *J. Neurosci*, 34:3632–3645
4. Faisal, A., Selen, L., and Wolpert, D. (2008). Noise in the nervous system. *Nat. Rev. Neurosci*, 9:292–303
5. Hong, S. and Rebec, G. (2012). A new perspective on behavioral inconsistency and neural noise in aging : compensatory speeding of neural communication. *Front. Aging Neurosci*, 4
6. Doi, E. and Lewicki, M. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Comput*, 23:2498–2510
7. Naselaris, T., Kay, K., Nishimoto, S., and Gallant, J. (2011). Encoding and decoding in fmri. *NeuroImage*, 56:400–410
8. Dinstein, I., Heeger, D., and Behrmann, M. (2015). Neural variability : friend or foe ? trends cogn. *Sci*, 19:322–328
9. McDonnell, M. and Ward, L. (2011). The benefits of noise in neural systems : bridging theory and experiment. *Nat. Rev. Neurosci*, 12:415–426
10. Stein, R., Gossen, E., and Jones, K. (2005). Neuronal variability : noise or part of the signal ? *Nat. Rev. Neurosci*, 6:389–397
11. Aerts, H., Fias, W., Caeyenberghs, K., and Marinazzo, D. (2016). Brain networks under attack : robustness properties and the impact of lesions. *Brain J. Neurol*
12. Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Comput. Biol*, 5:1000408
13. MacKinnon, D., Fairchild, A., and Fritz, M. (2007). Mediation analysis. *Annu. Rev. Psychol*, 58:593
14. Robbins, T. (2011). Cognition : The ultimate brain function. *Neuropsychopharmacology*, 36:1–2
15. Palestro, J., Bahg, G., Sederberg, P., Lu, Z.-L., Steyvers, M., and Turner, B. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *J. Math. Psychol*, 84:20–48
16. Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221
17. Turner, B., Palestro, J., MiletiÄ‡, S., and Forstmann, B. (2019a). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neurosci. Biobehav. Rev*, 102:327–336

mediate successful emotional regulation [18], threat response [19 20] or risk avoidance [21]. More recently, the anterior cingulate cortex, the anterior insula, the thalamus and some brain stem nuclei were shown to mediate various aspects of pain perception [22 23 24 25 26 27]. Most of these studies were performed using the multilevel mediation/moderation or M3 toolbox [28], which was first derived for probing effective connectivity from fMRI signals. Since then, a few multivariate extensions of brain-behavior mediation analysis were proposed, aiming at improving either spatial or temporal resolution [29 30 31]. But these approaches neither lay out nor address the specific methodological and interpretational challenges posed by brain-behavior analysis, when applied to typical fMRI experiments. In our view, progress in brain-behavior mediation analysis requires answering at least four important (and related) questions :

1. (Q1) Which test statistics should be used ? Not only should the test statistics be valid (i.e. yield controlled false positive rate), but they also should be maximally powerful. The latter is a pressing issue because fMRI induces a massive multiple comparison problem, which can only be solved by using more stringent significance thresholds [32 33]. We will summarize and compare the statistical properties of the most established test statistics of mediation analysis.
2. (Q2) How robust is brain-behavior mediation analysis to assumptions regarding the hemodynamic response function (HRF) and other confounds ? Recall that virtually all forms of fMRI time-series analyses rely on HRF models to assess effects of interest [34 35 36 37]. Although brain-behavior mediation analysis involves similar assumptions, different modelling strategies may be employed that yield distinct bias-variance tradeoffs. We will compare the statistical properties of these candidate approaches in the presence of deviations to modelling assumptions.
3. (Q3) What sort of brain mediators can we expect to detect ? Consider the bottom-up chain of neural information processing stages that eventually yield behavioral outcomes (from low-level sensory processing to high-level cognitive treatment of stimuli and/or cues). It turns out that these stages do not have the same chance of being detected. As we will see, this is a corollary consequence of the nontrivial (and yet undisclosed) impact of neural noise onto the statistical properties of mediation analysis.
4. (Q4) Does mediation analysis induce potential interpretational issues ? As we will see, some interpretational issues are specific to the chosen statistical testing approach, but others are generic to any brain-behavior mediation analysis. In particular, significant mediated effects are compatible with two distinct scenarios regarding the causal relationship between brain activity and behavioral responses. We discuss the importance of this and related issues and identify ways to address them.

In this work, we address these four questions from a user-oriented


18. Wager, T., Davidson, M., Hughes, B., Lindquist, M., and Ochsner, K. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, 59:1037–1050
19. Wager, T., Waugh, C., Lindquist, M., Noll, D., Fredrickson, B., and Taylor, S. (2009b). Brain mediators of cardiovascular responses to social threat : part i : Reciprocal dorsal and ventral subregions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47:821–835
20. Wager, T., Ast, V., Hughes, B., Davidson, M., Lindquist, M., and Ochsner, K. (2009a). Brain mediators of cardiovascular responses to social threat, part ii : Prefrontal-subcortical pathways and relationship with anxiety. *NeuroImage*, 47:836–851
21. Yamamoto, D., Woo, C.-W., Wager, T., Regner, M., and Tanabe, J. (2015). Influence of dorsolateral prefrontal cortex and ventral striatum on risk avoidance in addiction : a mediation analysis. *Drug Alcohol Depend*, 149:10–17
22. Atlas, L., Bolger, N., Lindquist, M., and Wager, T. (2010). Brain mediators of predictive cue effects on perceived pain. *J. Neurosci. Off. J. Soc. Neurosci*, 30:12964–12977
23. Atlas, L., Lindquist, M., Bolger, N., and Wager, T. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, 155:1632–1648
24. Geuter, S., Losin, E., Roy, M., Atlas, L., Schmidt, L., Krishnan, A., Koban, L., Wager, T., and Lindquist, M. (2018). Multiple brain networks mediating stimulus-pain relationships in humans. *BioRxiv*, 298927
25. Koban, L., Kross, E., Woo, C.-W., Ruzic, L., and Wager, T. (2017). Frontal-brainstem pathways mediating placebo effects on social rejection. *J. Neurosci*, 37:3621–3631
26. Koban, L., Jepma, M., LÃ³pez-SolÃ , M., and Wager, T. (2019). Different brain networks mediate the effects of social and conditioned expectations on pain. *Nat. Commun*, 10:4096
27. Woo, C.-W., Roy, M., Buhle, J., and Wager, T. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLOS Biol*, 13:1002036
28. Wager, T. (2008). canlab/ m3 mediationtoolbox (cognitive and affective neuroscience laboratory
29. Chen, O., Crainiceanu, C., Ogburn, E., Caffo, B., Wager, T., and Lindquist, M. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostat. Oxf. Engl*, 19:121–136
30. Lindquist, M. (2012a). Functional causal mediation analysis with an application to brain connectivity. *J. Am. Stat. Assoc*, 107:1297–1309

statistical perspective. Our aim here is to set a methodological standard for brain-behavior mediation analysis. The Methods section serves as the statistical and conceptual basis for addressing the four questions (Q1-Q4) above. It starts with a description of the brain-behavior mediation model and its associated null-hypothesis testing alternatives. Specific issues that arise in the context of typical fMRI experiments (factorial designs and condition contrasts, group-level random effects analysis, etc) are shortly discussed. We then consider the critical role of neural noise in brain-behavior mediation analyses, and present alternative solutions to the issue of HRF deconvolution. We close this section with a note on causality and its accompanying interpretational issue. We address the latter using a complementary information-theoretic approach (so-called I/O test). En passant, we show how to exploit the underlying mathematical degeneracy to drastically reduce the computational cost of whole-brain mediation analysis. In the Results section, we use numerical Monte-Carlo simulations to answer questions Q1-Q4. We compare the specificity and sensitivity of candidate mediation tests, as a function of neural noise, and in the presence of hemodynamic confounds. We also evaluate the utility and robustness of our I/O test. We then strengthen our in-sillico conclusions with an application to an experimental fMRI dataset acquired when people make decisions under risk. We exemplify the use of brain-behavior mediation analysis to ask questions regarding intra- and between-subjects variations in behavioral responses and attitudes. Finally, we discuss our results in the light of the existing literature and highlight potential weaknesses and perspectives (Discussion section).

## 6.2 METHODS

In what follows, we will consider behavioral paradigms akin to decision tasks, whereby subjects need to process some (experimentally-controlled) information $X$ to provide a (measured) behavioral response $Y$. Brain-behavior mediation analysis then aims at identifying whether some (anatomically-specific) feature of their observed brain activity $M$ mediates the effect of $X$ onto $Y$ . In our example fMRI application (see Results section), we will focus on a value-based decision making task, whereby participants have to accept or reject (response$Y$ ) a risky gamble composed of a 50% chance of winning a gain G and a 50% chance of loosing L (input information $X$ ). But more generally, $X$ is an experimental manipulation of some sort, $M$ is a measure of neural activity at the time of processing the stimulus, and $Y$ is some overt expression of the stimulus-induced covert mental state of interest.


31. Zhao, Y. and Luo, X. (2017). Granger mediation analysis of multiple time series with an application to fmri
32. Lindquist, M. and Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosom. Med*, 77:114
33. Worsley, K. and Friston, K. (1995). Analysis of fmri time-series revisited–again. *NeuroImage*, 2:173–181
34. Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on granger causality analysis of fmri. *NeuroImage*, 52:884–896
35. Gitelman, D., Penny, W., Ashburner, J., and Friston, K. (2003). Modeling regional and psychophysiologic interactions in fmri : the importance of hemodynamic deconvolution. *NeuroImage*, 19:200–207
36. Liao, C., Worsley, K., Poline, J.-B., Aston, J., Duncan, G., and Evans, A. (2002). Estimating the delay of the fmri response. *NeuroImage*, 16:593–606
37. Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., and Gramfort, A. (2015). Data-driven hrf estimation for encoding and decoding models. *NeuroImage*, 104

### 6.2.1 *The brain-behavior mediation model*

Let be the number of trials in a typical experimental session. Let $X$, $M$ and $Y$ be $n \times 1$ column vectors encoding the trial-by-trial experimental manipulation, the brain's response to the experimental manipulation (e.g., the magnitude of the fMRI BOLD response to the stimulus at each trial, in some voxel or region of interest) and the behavioral response to the experimental manipulation, respectively. For the sake of mathematical simplicity, and without loss of generality, we will assume that $X$, $M$ and $Y$ have all been z-scored.

From the perspective of identifying the determinants of behavior, one may first ask whether $X$ has an effect on $Y$ or not. In its simplest mathematical form, this question reduces to considering the following simple linear regression model :

$$Y = Xc + \epsilon_Y^0 \tag{6.1}$$

where $s$ is an unknown regression coefficient that measures the strength of the statistical relationship between the independent ($X$) and dependent ($Y$) variables, and $\epsilon_Y^0$ are model residuals. On would then simply test for the statistical significance of $c$ , under some assumptions regarding the distribution of model residuals $\epsilon_Y^0$.

Now, one may also ask whether $M$ mediates the effect of $X$ onto $Y$. In its simplest mathematical form, this question relies on the following pair of linear regression models :

$$\begin{aligned} M &= Xa + \epsilon_M^0 \\ Y &= Mb + Xc' + \epsilon_Y \end{aligned} \tag{6.2}$$

where the first equation expresses the fact that $M$ responds to $X$ (with some unknown susceptibility $a$), and the second equation states that $Y$ depends upon both $M$ (with some unknown susceptibility $b$) and $X$ (with some unknown susceptibility $c'$). On may think of residuals $\epsilon_M^0$ in terms of some form of *neural noise*, because they capture trial-by-trial variations in $M$ that are independent of $X$. As we will see, they play a pivotal role in brain-behavior mediation analysis.

Although simple, Equation 6.2 does not explicitly quantify the size of a mediated effect. But this can be done by noting that Equation 6.2 can be rewritten as follows :

$$\begin{aligned} Y &= Xab + \epsilon_M^0 b + Xc' + \epsilon_Y \\ &= X(ab + c') + \epsilon_M^0 b + \epsilon_1 \end{aligned} \tag{6.3}$$

where $M$ has simply been replaced by its expression from Equation 6.2. Equation 6.3 is helpful in realizing that the *total effect* of $X$ onto $Y$ is partitioned into a *direct* effect (whose size is $c'$) and an *indirect effect* (whose size is $ab$). This distinction is important because the latter is the effect of $X$ onto $Y$ that is mediated by $M$. This is why established mediation tests rely on assessing the statistical significance of the indirect effect [38]. Note that so-called *full mediation* occurs when

38. MacKinnon, D., Fairchild, A., and Fritz, M. (2007). Mediation analysis. *Annu. Rev. Psychol*, 58:593

$c' = 0$ (no direct path), and one speaks of *partial mediation* whenever $c' \neq 0$.

Importantly, when we perform mass-univariate mediation analysis, we effectively consider each voxel or region of interest in isolation, and ask whether the local indirect effect is statistically significant. If mediation tests are repeated over voxels, then they form a statistical mediation map, which can localize which brain structure(s) mediate(s) the effect of $X$ onto $Y$. In this context, Equations 1-3 have two interesting implications, which we will highlight now.

To begin with, recall that the incoming information $X$ is processed by a distributed brain system, whose elements (sampled across large voxel sets) concurrently contribute to the behavioural response $Y$. The structure of this distributed brain system is likely to involve multiple processing pathways that work both in series and in parallel, as in Figure 6.1 below.

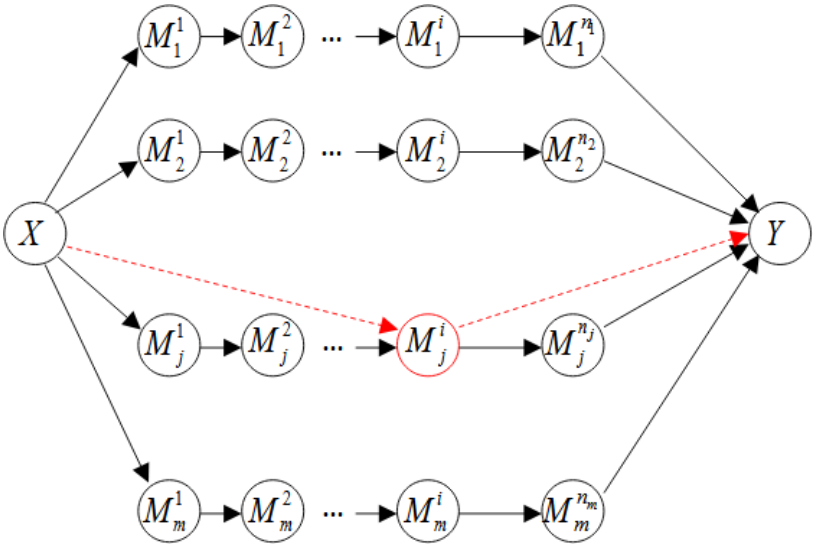



FIGURE 6.1 – *Example structure of a processing hierarchy in the brain. Here, $X$ and $Y$ encode the experimental manipulation and the ensuing behavioral response, respectively. Variables $M_j^i$ are activity within brain regions that act as intermediary processing steps. In this oriented graphical model, arrows represent causal relationships. Although processing pathways operate both in series and in parallel, mass-univariate brain-behavior mediation analysis ignore this and treat each region/voxel independently of each other (red dotted arrows).*

In simple bottom-up hierarchical architectures such as this one, lower levels would correspond to e.g., occipital low-level visual processes, whereas higher levels would map to, e.g., prefrontal decision making processes. Clearly, Figure 6.1 is already an oversimplification because it ignores reciprocal connections, branching processes and/or context-dependent gating mechanisms [39] [40] [41]. But when we perform mass-univariate brain-behavior mediation analysis, we reduce the complexity even further by considering each voxel or region of interest in isolation, effectively ignoring any hierarchical structure of this sort.

First, given the likely parallel nature of processing pathways, one would not expect that any isolated voxel or region of interest may ever fully mediate the impact of $X$ onto $Y$. The implicit assumption of mass-univariate brain-behavior mediation analysis is that, in each voxel, the direct path $c'$ effectively captures, in a non specific manner, mediated effects that go through other (parallel) pathways. This, however, places a very heavy load on the statistical sensitivity of

39. Friston, K. (2011). Functional and effective connectivity : A review. *Brain Connect*, 1:13–36

40. He, Y. and Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Curr. Opin. Neurol*, 23:341–350

41. Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity : Uses and interpretations. *NeuroImage*, 52:1059–1069

mediation tests, which need to be able to detect potentially small indirect effect sizes, even when correcting for multiple comparisons (e.g., across voxels).

Second, nothing prevents different processing pathways to have strong but opposing impacts on the behavioral response. An example here would be opponent brain systems that yield strong but ambivalent (e.g., appetitive-aversive) cognitive states, whose idiosyncratic balance may explain one's specific ability to suppress e.g., impulsive behavioral responses [42] [43]. In particular, if the impact of different pathways balance out, then the total effect of $X$ onto $Y$ may become undetectable ($c \approx 0$). It follows that brain-behavior mediation analyses may be required for faithfully identifying the determinants of behavior. Alternatively, the relative contribution of different pathways may vary across individuals, which may drive inter-individual behavioral differences. We will see an example of this in the Results section below.

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

43. Zhang, S., Mano, H., Lee, M., Yoshida, W., Robbins, T., Kawato, M., and Seymour, B. (2017). The control of tonic pain by active relief learning. *BioRxiv*, 222653

### 6.2.2 *Statistical tests of mediation*

In what follows, we recall the most established approaches to null-hypothesis testing of mediated effects. We start with the premise that if $M$ mediates the effect of $X$ onto $Y$, then the corresponding indirect effect has to be different from 0 ($ab \neq 0$). In what follows, we summarize two kinds of statistical testing approaches (namely : the "indirect" and the "conjunctive" approaches) that differ in terms of how they frame the corresponding null hypothesis.

The indirect approach follows from noting that the null hypothesis of mediation analysis can be framed as follows : $H_0^{indirect} : ab = 0$.

Under the simple brain-behavior mediation model in Equations 1-2, the indirect effect equates the difference between total and direct effects, i.e. $ab = c - c'$ . This is why early approaches to mediation testing were assessing the statistical significance of the difference $c - c'$ – [44]. However, theoretical work demonstrated that this equivalence may not always hold [45], which would render the ensuing test invalid. This applies to typical fMRI experiments, because of the effect of confounding variables on path coefficient estimates. Another, more valid, approach is to compare estimates of the indirect effect to their distribution under the null. This is the principle of Sobel's test [46]. Recall that all parameters are identifiable from Equation 6.2, given $X$, $Y$ and $M$. In particular, the ordinary least-squares (OLS) estimates $\hat{a}$ and $\hat{b}$ of unknown path coefficients and are given by (all $X$, $Y$ and $M$ variables are z-scored, see Appendix 6.17 for a mathematical derivation) :

44. Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research : conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol*, 51:1173–1182

45. Pearl, J. (2012). The mediation formula : A guide to the assessment of causal pathways in nonlinear models. In *Causality*, pages 151–179. John Wiley Sons, Ltd

46. Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol*, 13:290–312

$$\begin{aligned} \hat{a} &= X^T M/n \\ \hat{b} &= \frac{1}{\hat{\sigma}^2_{M|X}} \hat{\epsilon}^0_{M|X}{}^T Y/n \end{aligned} \tag{6.4}$$

where the neural noise estimate $\hat{\epsilon}^0_{M|X} = M - X\hat{a}$ is the component of $M$ that cannot be explained by $X$, and $\hat{\sigma}^2_{M|X} = 1 - \hat{a}^2$ is its sample

variance. From6.4, one can see that $\hat{b}$ is simply the sample correlation between behavioral data $Y$ and the neural noise estimate $\hat{\epsilon^0_M}$ . In other words, $\hat{b} \neq 0$ when $M$ has an effect on $Y$ *above and beyond the effect of* $Y$. In addition, the variance of these OLS estimates are given by :

$$\begin{aligned} \hat{\sigma}^2_a &= \frac{\hat{\sigma}^2_{M|X}}{n} \\ \hat{\sigma}^2_b &= \frac{\hat{\sigma}^2_{Y|M,X}}{(n-1)\hat{\sigma}^2_{M|X}} \end{aligned} \tag{6.5}$$

where $\hat{\sigma}^2_{Y|M,X} = 1 - \hat{b}^2 - (X^T Y/n - \hat{a}\hat{b})^2$ is the sample variance of behavioral residuals' estimates $\hat{\epsilon_Y} = Y - M\hat{b} - X\hat{c}'$. Under the assumption that model residuals $\epsilon_Y$ and $\epsilon^0_M$ are i.i.d. normal variables, then both $\hat{a}$ and $\hat{b}$ follow normal distributions : $\hat{a} \sim \mathcal{N}(a, \hat{\sigma}^2_a)$ and $\hat{b} \sim \mathcal{N}(b, \hat{\sigma}^2_b)$. It can then be shown $\hat{a}\hat{b}$ (see Appendix 6.B) that the product approximately follows a normal distribution, i.e. : $\hat{a}\hat{b} \sim \mathcal{N}(ab, \hat{b}^2\hat{\sigma}^2_a + \hat{a}^2\hat{\sigma}^2_b)$ . This implies that, under the null, the following pseudo z-statistics :

$$z^{(Sobel)}_{ab} = \frac{\hat{a}\hat{b}}{\sqrt{\hat{b}^2\hat{\sigma}^2_a + \hat{a}^2\hat{\sigma}^2_b}} \tag{6.6}$$

approximately follows a Student probability density function. This then serves to derive the p-value of Sobel's unsigned (two-tailed) significance test $p^{(Sobel)}_0 = 1 - 2\phi(|z^{(Sobel)}_{ab}|)$, where $\phi$ is Student's cumulative density function with appropriate degrees of freedom. Later improvements over Sobel's test (Hayes and Scharkow, 2013) derived from theoretical statistical works on the distribution of the product of two normal random variables, which essentially include an additional $\pm\hat{\sigma}^2_a\hat{\sigma}^2_a$ term to the denominator of Sobel's pseudo-zscore (Aroian, 1947 ; Goodman, 1960).

$$\begin{aligned} z^{(Aroian)}_{ab} &= \frac{\hat{a}\hat{b}}{\sqrt{\hat{b}^2\hat{\sigma}^2_a + \hat{a}^2\hat{\sigma}^2_b + \hat{\sigma}^2_a\hat{\sigma}^2_a}} \\ z^{(Goodman)}_{ab} &= \frac{\hat{a}\hat{b}}{\sqrt{\hat{b}^2\hat{\sigma}^2_a + \hat{a}^2\hat{\sigma}^2_b - \hat{\sigma}^2_a\hat{\sigma}^2_a}} \end{aligned} \tag{6.7}$$

We refer to these extensions as Aroian's and Goodman's tests, respectively. Alternatively, non-parametric approaches have been proposed to derive the distribution of indirect effect size estimates under the null [47] [48]. Here, we will use the same bias-corrected bootstrap approach as the one proposed in the M3 toolbox [49].

The *conjunctive* approach follows from noticing that the null hypothesis of mediation analysis is composite [50], i.e. : $H^{(conjunction)}_0 : a = 0$ OR $b = 0$ . Of course, both null hypotheses are exactly equivalent, but the composite null highlights the fact there is no mediated effect as long as one path coefficient is null (which breaks the causal cascade). In turn, one may test for the conjunction of both effects, i.e. test for the statistical significance of both and path coefficients. In practice,

47. MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods*, 7:83–104

48. MacKinnon, D., Lockwood, C., and Williams, J. (2004). Confidence limits for the indirect effect : Distribution of the product and resampling methods. *Multivar. Behav. Res*, 39:99

49. Wager, T. (2008). canlab/ m3 mediationtoolbox (cognitive and affective neuroscience laboratory

50. Moran, P. P. (1970). On asymptotically optimal tests of composite hypotheses. *Biometrika*, 57:47–55

conjunctive testing relies on the "maximum p-value" approach (here, two-tailed test) :

$$\begin{aligned} p_0^{(conj)} &= max(2\phi(-|t_a|), 2\phi(-|t_b|)) \\ &= 1 - 2\phi(min(|t_a|, |t_b|)) \end{aligned} \tag{6.8}$$

where $t_a = \hat{a}/\hat{\sigma}_a$ and $t_b = \hat{b}/\hat{\sigma}_b$ are Student's test statistics of $a$ and $b$ path coefficients, respectively. Formally speaking, $p_0^{(conj)}$ rovides an upper bound on the joint probability that, under the null, two independent Student's test statistics take more extreme values than $t_a$ and $t_b$ [51] [52]. This is important, because conjunctive testing cannot be invalid but may have low sensitivity. However, it is trivial to show that Sobel's pseudo z-score is always smaller than the conjunctive test statistics, i.e. : $|z_{ab}^{(Sobel)}| \leq z_{ab}^{(conj)}$, where $z_{ab}^{(conj)} = min(|t_a|, |t_b|)$ is the conjunctive test statistics (see Appendix 6.B). This means that one would expect conjunctive testing to be systematically more efficient than Sobel's approach. At this point, we note that the sensitivity profile of indirect and conjunctive approaches actually depends upon neural noise strength and model misspecifications (see next sections). We will address this and related issues in the Results section, using extensive numerical Monte-Carlo simulations. We refer the reader interested in extending these statistical approaches to experimental designs including multiple conditions (cf., e.g., factorial designs) and/or multiples subjects (cf. group-level random effects analysis) to Appendix 6.C and 6.D, respectively.

51. Friston, K., Penny, W., and Glaser, D. (2005a). Conjunction revisited. *NeuroImage*, 25:661–667

52. Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, 25:653–660

### 6.2.3 *The non-trivial impact of neural noise*

Although indirect and conjunctive null hypotheses are formally equivalent to each other, the latter is helpful to disclose the subtle tension behind mediation testing. In brief, two conditions must be satisfied for detecting a mediated effect : (i) strong evidence for $a \neq 0$ and (ii) strong evidence for $b \neq 0$ . The former means that $X$ partly explains the trial-by-trial variability of $M$ . And the latter means that $M$ partly explains the variability of $Y$ that is unexplained by $X$. The critical point here is to realize that these two conditions are in conflict with each other. This is because they have opposing demands on neural noise $\epsilon_M^0$ . Note that the conjunctive test statistics $z_{ab}^{(conj)}$ is given by :

$$z_{ab}^{(conj)} = min\left(\sqrt{n}\frac{|\hat{a}|}{\hat{\sigma}_{M|X}}, \sqrt{n-1}\frac{|\hat{b}|}{\hat{\sigma}_{Y|M,X}}\hat{\sigma}_{M|X}\right) \tag{6.9}$$

where we simply have inserted Equation 6.5 into the definition of the conjunctive test statistics. One can see that the standard deviation $\hat{\sigma}_{M|X}$ of the neural noise estimate $\hat{\epsilon}_M^0$ will have opposing effects on the conjunctive test statistics. In brief, if $\hat{\sigma}_{M|X} \to 0$, then $z_{ab}^{(conj)} = |t_b|$, which tends towards 0 when $\hat{\sigma}_{M|X} \to 0$. Recall that, by definition, $\hat{\epsilon}_M^0$ is the component of $M$ that cannot be explained by $X$ (cf. Equation 6.2). Thus, in the absence of neural noise, the evidence for $a \neq 0$ is

maximal, but $M$ cannot explain any variability in $Y$ that is unexplained by $X$, i.e. the evidence for $b \neq 0$ is minimal. Reciprocally, if $\hat{\sigma}_{M|X} \to \infty$, then $z_{ab}^{(conj)} = |t_a|$, which tends also towards 0 when $\hat{\sigma}_{M|X} \to \infty$ . In other words, if neural noise strength is very high, then evidence for $a \neq 0$ is weak. Only for *intermediary levels of neural noise* can evidence for both $a \neq 0$ and $b \neq 0$ reach statistical significance. We note that this observation generalizes to any mediation test, irrespective of the mathematical form of the brain-mediation model. We refer the interested reader to Appendix 6.E.

We will quantify the impact of neural noise on the statistical efficiency of candidate mediation testing approaches in the Results section below. But this property of mediation analysis has an important implication, which we now highlight.

Recall the structure of the processing hierarchy in Figure 6.2. Within a given processing pathway, each hierarchical level responds to its (lower-level) parents, eventually changing the information content in an incremental manner, e.g. :

$$\begin{aligned} M_1 &= Xa_0 + \epsilon_M^0 \\ M_2 &= M_1 a_1 + \epsilon_M^1 \\ &\dots \\ M_{i+1} &= M_i a_i + \epsilon_M^i \\ &\dots \\ Y &= M_N b + Xc' + \epsilon_Y \end{aligned} \tag{6.10}$$

where $\{M_1, M_2, \dots, M_i, \dots, M_N\}$ are local neural responses (indexed by their level along the hierarchy), and local neural noise increments $\epsilon_M^i$ effectively capture, in an agnostic manner, the unique contribution of each hierarchical level. Here, one would expect that local neural responses gradually diverge from the initial explanatory variable $X$ . This is simply because the correlation between $X$ and the local neural response $M_i$ degrades as the accumulated neural noise increments $\sum_{j=0}^{i} \epsilon_M^i$ increases. In turn, one would expect that mass-univariate mediation analysis can only detect those neural information processing steps that are positioned at an intermediary hierarchical level, i.e. sufficiently far away from either end of the hierarchy. We will exemplify this in the Results section below.

### 6.2.4 *Dealing with hemodynamic confounds*

Clearly, the brain-behavior mediation model in Equation 6.2 cannot directly be applied to fMRI time series. The reason is twofold. First, behavioral and neural variables are not sampled in the same manner. In brief, the former is collected at each "trial" of the behavioral task, while the latter is typically sampled at a sub-trial temporal resolution. Second, fMRI BOLD dynamics effectively result from the convolution of neural activity with the hemodynamic response function or HRF [53] [54]. This implies that the event-related BOLD response

53. Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157

54. Martin, C., Martindale, J., Berwick, J., and Mayhew, J. (2006). Investigating neural-hemodynamic coupling and the hemodynamic response function in the awake rat. *NeuroImage*, 32:33–48

is delayed in time, when compared to trial onsets. In addition, if the inter-trial interval is smaller than the HRF duration (which is typically the case), BOLD signals measured during a trial may derive from the additive contributions of multiple neural responses (to the current and preceding trials). For the purpose of brain-behavior mediation analysis, there are essentially two ways of dealing with such hemodynamic confounds.

On the one hand, one may deconvolve BOLD signals from the HRF, as follows. Let $\tau_k$ (resp. $\Delta_k$ ) be the onset time (resp., duration) of the $k$-th trial in the experimental design. One first construct "trial" regressors that span the duration of the fMRI session (at the sampling resolution of fMRI ; typically : TR=1-2secs), which are zero everywhere except during the time interval defined as $[\tau_k, \tau_k + \Delta_k]$ . Each of these is then convolved with the canonical HRF and its temporal derivatives, to account for potential mismatches in hemodynamic delays [55]. One then augment the resulting GLM with fMRI confounds (e.g., motion regressors and slow drifts), and fit it to fMRI time series. Fitted regressor weights at each voxel thus provide an estimate $\hat{M}$ of the local neural response to each trial, which is deconvolved from the HRF and corrected for typical fMRI confounds, and can then enter a mediation analysis. We call this the deconvolution approach. On the other hand, one may reframe the brain-behavior mediation model in the HRF-convolved space. One first resample the explanatory and dependent variables at the fMRI temporal resolution by reweighing each "trial'" regressor above with its corresponding $X$ and $Y$ entries and them summing over trials. One then convolves the resulting regressors with the canonical HRF (and its temporal derivatives) and augment the resulting GLM with fMRI confounds prior to entering a mediation analysis. We call this the *convolution* approach.

55. Liao, C., Worsley, K., Poline, J.-B., Aston, J., Duncan, G., and Evans, A. (2002). Estimating the delay of the fmri response. *NeuroImage*, 16:593–606

Both approaches can, in principle, deal with hemodynamic and other fMRI confounds, but they differ in terms of their respective bias-variance tradeoff. The convolution approach effectively yields reliable neural response estimates, under the implicit assumption that the HRF is identical across trials. In contrast, the deconvolution approach allows for trial-by-trial variations in HRF, at the cost of compromising the reliability of neural response estimates. In the Results section below, we evaluate the robustness of these two strategies w.r.t. deviations to canonical HRF models.

### 6.2.5 *A note on causality*

Let us now highlight a possible interpretational issue of mediation analysis. Note that Equation 6.2 implicitly assumes a cascade of causal influences [56], which may be best summarized in terms of the directed acyclic graph depicted on Figure 6.2 below (left panel).

56. MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods*, 7:83–104

One would then be tempted to interpret a statistically significant mediated effect in causal terms, as in : perturbing the independent variable $X$ should result in changes in the mediator variable $M$ that would eventually cascade down to the dependent variable $Y$. In the


FIGURE 6.2 – *The two causal interpretations of mediated effects.*
*Left panel : "native" causal interpretation of brain-behavior mediation analysis (cf. Equation 6.2). Right panel : "swapped" causal interpretation of brain-behavior mediation analysis (cf. Equation 6.11). Corresponding path coefficients are shown in red.*

context of brain-behavior mediation, this causal interpretation aligns with the intuitive notion that behavioral responses to stimuli necessarily has to emerge from an intermediate neural information processing step. This causal reasoning, however, does not hold regarding the relationship between $M$ and $Y$, which are both *observed* data. In Equation 6.2, the strength of this relationship is controlled by the path coefficient $b$. Importantly, statistical inference on the path coefficient $b$ is but a quantitative assessment of the conditional mutual information $I(M, Y|X)$, which is invariant under a reversal of the directionality of the relationship between $M$ and $Y$. In other terms, Equation 6.2 is formally equivalent to the following alternative model :

$$\begin{aligned} M &= Xa + \epsilon_M^0 \\ M &= Yd + Xa' + \epsilon_M \end{aligned} \tag{6.11}$$

where the second line simply derives from swapping (with impunity) the explanatory and response variables in the second line of Equation 6.2. Here, $d$ and $a'$ are "swapped" path coefficients that have a different causal interpretation (cf. Figure 6.1, right panel), and $\epsilon_M$ are model residuals that are not equivalent to the neural noise $\epsilon_M^0$ of Equation 6.2. One can show (see Appendix 6.E) that both native and swapped path coefficients estimates are analytically related as follows :

$$\hat{b} = \hat{d}\frac{\hat{\sigma}^2_{Y|X}}{\hat{\sigma}^2_{M|X}} \tag{6.12}$$

where $\hat{\sigma}^2_{Y|X} = 1 - (X^T Y/n)^2$ is the sample variance of Equation 6.1's residuals estimates $\hat{\epsilon}_Y^0 = (I - XX^T/n)Y$. It should be clear from Equations 11-12 that assessing the conditional mutual information $I(M, Y|Y)$ can be equivalently addressed either by assessing the evidence for $b \neq 0$ (native form of the mediation model, cf. Equation 6.2), or by assessing the evidence for $d \neq 0$ (cf. Equation 6.11). In fact, the ensuing t-statistics and are exactly equal (see Appendix 6.E), i.e. brain-behavior mediation test statistics are invariant under a permutation of $M$ and $Y$ variables.

This has two important consequences.

First, one may rely on Equations 11-12 to improve the computational efficiency of brain-behavior mediation analysis by several orders of magnitude. Recall that in the context of whole-brain fMRI, working with regression models where fMRI signals only enter as dependant variables is computationally very advantageous. This is because many algebraic operations that are required for parameter estimation (e.g.,

here, matrix multiplications and inversions, etc) can be computed once and for all. In brief, the computational gain of performing brain-behavior mediation analysis using Equations 9-10, when compared to Equation 6.2, is of the order of $n_{scan}^2 \times n_{voxel}$, where $n_{scan}$ and $n_{voxel}$ are the number of fMRI time samples and voxels, respectively. This may speed up whole-brain mediation analysis by several orders of magnitude.

Second, a statistically significant mediated effect is compatible with two causal interpretations. In particular, under the "swapped" model of Equation 6.11, variations in behavior $Y$ may cause changes in the neural response $M$ (cf. Figure 6.1, right panel). This alternative causal interpretation ($Y \rightarrow M$) is not as nonsensical as it may first sound. For example, somatosensory cortices will respond to variations in motor actions, eventually enabling proprioceptive sensations. More generally, a given brain system may be collecting and/or processing information regarding overt behavior (which may have been produced elsewhere in the brain) for the purpose of, e.g., learning, memory, metacognition, etc... In any case, this interpretational issue is important, because the implicit intention behind brain-behavior mediation analysis is clearly to provide statistical evidence for the "native" causal scenario ($X \rightarrow M \rightarrow Y$). We will comment on this and related issues in the Discussion section of this manuscript.

One may think that affording evidence for the "native" causal claim of brain-behavior mediation analysis may require non observational studies, e.g., causal perturbations of neural activity (lesion studies, transcranial magnetic stimulation, etc). Nevertheless, we argue that one may perform complementary data analyses that may partially address the interpretational issue above. For example, having assessed the significance of a mediated effect, one may exploit locally multivariate information to provide statistical evidence for or against candidate causal claims. In fact, when considering the set of mediator variables within a significant cluster together, "native" ($M \rightarrow Y$) and "swapped" ($Y \rightarrow M$) causal interpretations induce a many-to-one and a one-to-many M-Y mapping, respectively (see Figure 6.3 below). Because "native" and "swapped" causal scenarios differ in terms of whether $Y$ is viewed as an input or as an output of local neural information processing, we refer to the ensuing test statistics as an *I/O test statistics*.

Let $M_i$ be the trial-by-trial variations of a voxel belonging to a given mediator cluster, where $i \in [1, N]$ and $N$ is the number of voxels in the cluster. We define our I/O test statistics $\bar{\lambda}$ as follows :

$$\bar{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i \tag{6.13}$$

where $\lambda_i$ is the loss of conditional mutual information between $M_i$ and $Y$ when accounting for other neighboring voxels $M_{j \neq i}$ :

$$\lambda_i = I(M_i, Y|X) - I(M_i, Y|X, M_{j \neq i}) \tag{6.14}$$

In Equation 6.14, $I(M_i, Y|X, M_{j \neq i})$ and $I(M_i, Y|X)$ are the condi-

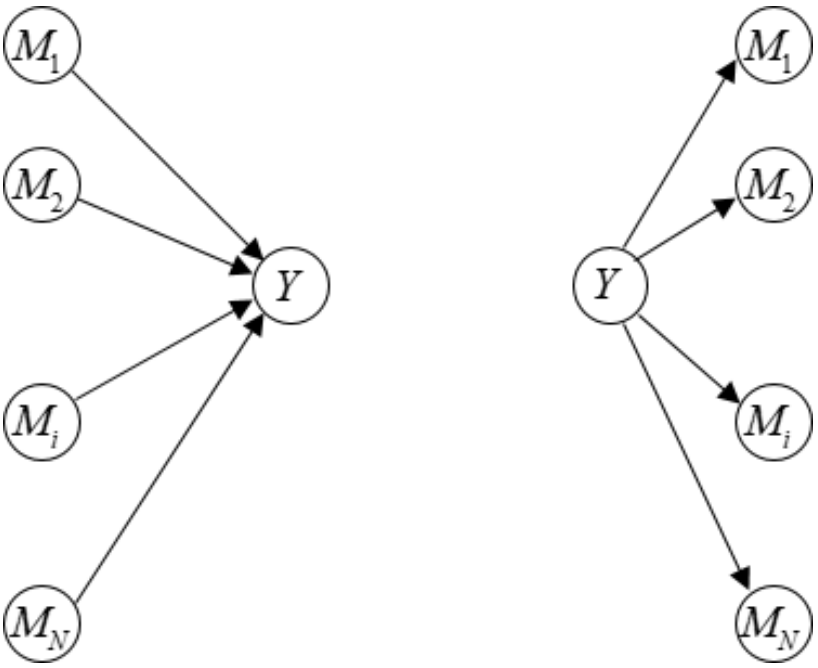



FIGURE 6.3 – *Candidate multivariate input/output M-Y mappings. Left panel : "native" causal interpretation (many-to-one input/output mapping, Y as an output). Right panel : "swapped" causal interpretation (one-to-many input/output mapping, Y as an input).*

tional mutual information between $M_i$ and $Y$, given $X$ and the activity in all other voxels $j \neq i$ or not, respectively. Note that $\lambda_i$ is sometimes coined the *interaction information*[57]. In brief, $\bar{\lambda}$ measures the average improvement or worsening of the mutual information between candidate mediator voxels and behavioral responses, when accounting for variations in neighboring brain activity.

For arbitrary gaussian variables $X$ and $Y$, the mutual information $I(X,Y)$ can be written as $I(X,Y) = -1/2\ log(1-\rho^2_{X,Y})$, where $\rho_{X,Y}$ is the correlation between $X$ and $Y$[58]. In turn, Equation 6.12 can be rewritten as follows :

$$\bar{\lambda} = -\frac{1}{2N}\sum_{i=1}^{N} log\left(\frac{1-\hat{b}_i^2}{1-\tilde{b}_i^2}\right) \tag{6.15}$$

where $\tilde{b}_i$ is the conditional correlation between $M_i$ and $Y$, given $X$ and $M_{j\neq i}$ :

$$\tilde{b_i} = \frac{\tilde{Y_i}^T\tilde{M_i}}{\sqrt{\tilde{Y_i}^T\tilde{Y_i}}\sqrt{\tilde{M_i}^T\tilde{M_i}}} \text{ with } \begin{cases} Z_i = [M_{j\neq i}, X] \\ P_i = I - Z_i(Z_i^TZ_i)^{-1}Z_i^T \\ \tilde{Y} = P_iY \\ \tilde{M_i} = P_iM_i \end{cases} \tag{6.16}$$

It turns out that the sign of $\bar{\lambda}$ provides evidence in favor or against the native causal interpretation of the brain-behavior mediation model. More precisely : if $M \rightarrow Y$, then $E[\bar{\lambda}] \leq 0$ , whereas if $Y \rightarrow M$, then $E[\bar{\lambda}] \geq 0$ . This is because if $Y$ is an output of local brain activity ($M \rightarrow Y$), then any given univariate statistical relationship between $M_i$ and $Y$ (path coefficient $\hat{b}_i$) is obscured by the (partially independent) contributions of all other mediator variables $M_{j\neq i}$. Therefore, when removing all the variability that can be explained with $M_{j\neq i}$, one

57. McGill, W. (1954). Multivariate information transmission. *Psychometrika*, 19:97–116

58. Marrelec, G., Daunizeau, J., Pelegrini-Issac, M., Doyon, J., and Benali, H. (2005). Conditional correlation as a measure of mediated interactivity in fmri and meg/eeg. *IEEE Trans. Signal Process*, 53:3503–3516

reveals the unique contribution of $M_i$ (i.e. $\hat{b}_i < \tilde{b}_i$ ). In contrast, if $Y$ is an input to local brain activity ( $Y \rightarrow M$), then the variability shared by all mediator variables results from the influence of $Y$. Therefore, when removing all the variability that can be explained with $M_{j\neq i}$, one degrades the statistical relationship between $M_i$ and $Y$ (i.e. $\widehat{b}_i > \tilde{b}_i$). We will evaluate the utility and robustness of our I/O test statistics in the Results section below.

## 6.3 RESULTS

In what follows, we will be comparing five testing approaches : Sobel's test, Aorian's test, Goodman's test, the M3 bootstrap test, and the conjunctive approach, in terms of their statistical sensitivity and specificity. Using numerical simulations, we will assess the impact of neural noise and deviations to HRF assumptions. Taken together, these in-silico experiments will serve to address questions Q1 to Q3. Using further numerical simulations, we will demonstrate the utility of our I/O test statistics for addressing the main interpretational issue of brain-behavior mediation analysis (Q4). Finally, we will report the results of a brain-behavior mediation analysis in the context of an fMRI experiment on decision making under risk.

Comparing the statistical specificity and sensitivity of testing approaches First, we ask whether candidate testing approaches yield valid inferences, i.e. whether they allow for a faithful control of false positive rate. To address this question, we simulated data under three different variants of the null hypothesis. More precisely, we simulated 40,000 datasets with Equation 6.2 , using three different settings of the path coefficients, i.e. : (i) $a = 0$ and $b = 1/2$ , (ii) $a = 1/2$ and $b = 0$, or (iii) $a = b = 0$ . In all simulations, we simulated $n = 50$ trials, set the direct effect size to $c' = 1/2$ and used unitary variance for all independent variables in Equation 6.2 (i.e. $X$ ,$\epsilon_M^0$ and $\epsilon_Y$). Across these 40,000 simulations, we then measured the (false positive) detection rate of each candidate testing approach, as one varies the significance threshold $\alpha$. Note that all (indirect or conjunctive) parametric tests were performed with Student's probability distribution functions with $n-2$ degrees of freedom. Finally, we kept the default number of 1000 resamplings in the bias-corrected M3 bootstrap test. Second, we asked how sensitive are candidate testing approaches under moderate mediated effect sizes. Here, we simulated 40,000 datasets with Equation 6.2, using $a = b = 1/2$, and measured the (true positive) detection rate of each candidate testing approach, as one varies the significance threshold $\alpha$.

The results of these analyses are summarized on Figure 6.4 below.

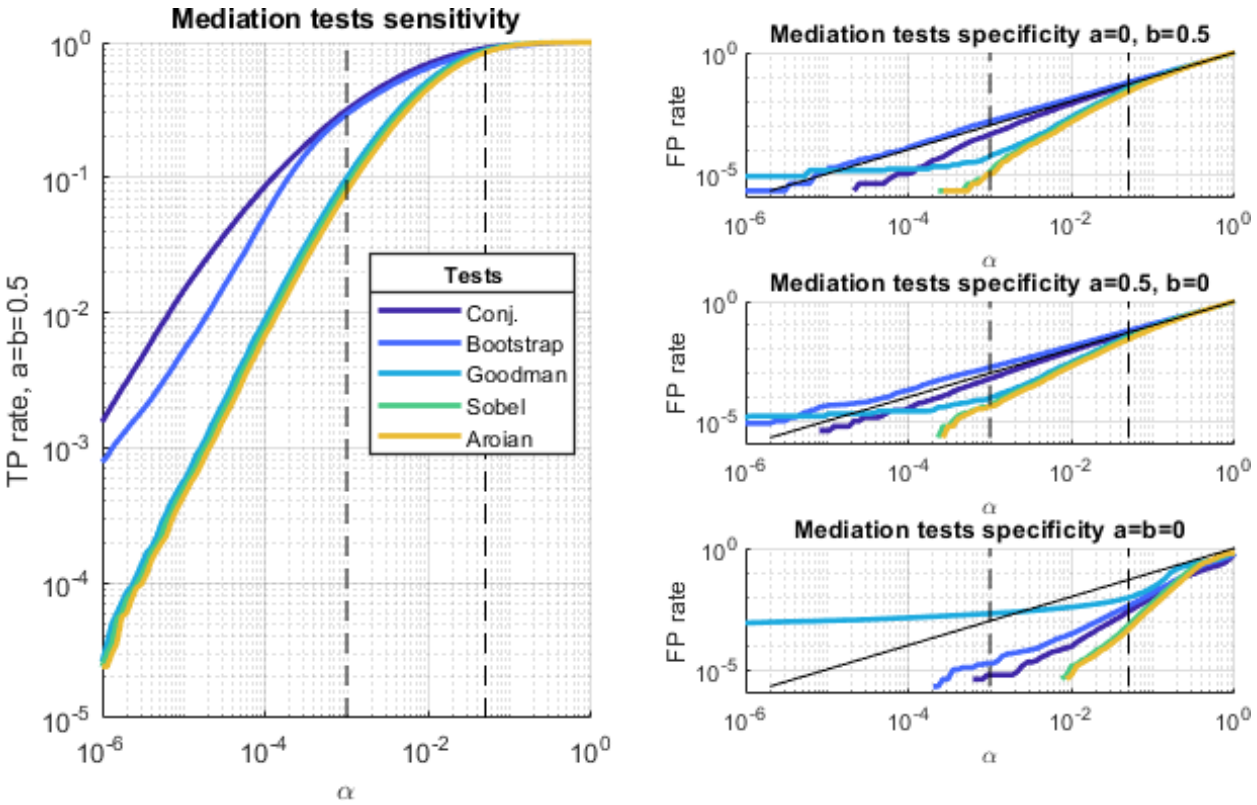



FIGURE 6.4 – *Statistical specificity and sensitivity of variants of mediation significance testing approaches. Left panel : The sensitivity of mediation tests (y-axis) is plotted against the significance threshold $\alpha$ (x-axis), for each candidate testing approach (dark blue : conjunctive testing, blue : M3 bootstrap indirect approach, light blue : Goodman's indirect approach, green : Sobel's indirect approach, yellow : Aorian's indirect approach). Upper right panel : The specificity of mediation tests (y-axis) is plotted against the significance threshold, for $H_0 : a = 0$ AND $b = 1/2$ (same format as left panel). Middle right panel : Same format as upper right panel for $H_0 : a = 1/2$ AND $b = 0$. Lower right panel : Same format as upper right panel for $H_0 : a = b = 0$.*

As expected, the conjunctive test is more sensitive than both Sobel and Aorian tests. Slightly more surprising maybe is the fact that the conjunctive test turns out to also be more sensitive than Goodmans test and the M3 bootstrap test, though the latter reach similar sensitivity levels for significance thresholds higher than 0.001. We will refine our evaluation of statistical sensitivity when assessing the impact of neural noise below.

In addition, all approaches except Goodman and the M3 bootstrap tests are valid, i.e. they yield a false positive rate that is equal or smaller than the significance threshold $\alpha$. Goodman's test always yield invalid inference if the significance threshold is small enough, whereas the M3 bootstrap test only yields invalid inference when $b = 0$. Note that the conjunctive approach is the least conservative of all tests, and this difference grows when the significance threshold decreases.

### 6.3.1 *Assessing the impact of neural noise*

Recall that the magnitude of neural noise is expected to play a critical role for the statistical sensitivity of mediation analysis. To demonstrate this effect, we simulated 10,000 datasets using the same

parameter settings as above, except for neural noise magnitude, which we varied from $Var[\epsilon_M^0] = 10^{-2}$ to $Var[\epsilon_M^0] = 10^2$. For each neural noise magnitude, we then measured the (true positive) detection rate of each candidate testing approach, when setting the significance threshold to $\alpha = 0.05$. The ensuing sensitivity profiles are summarized on Figure 6.5 below (left panel).

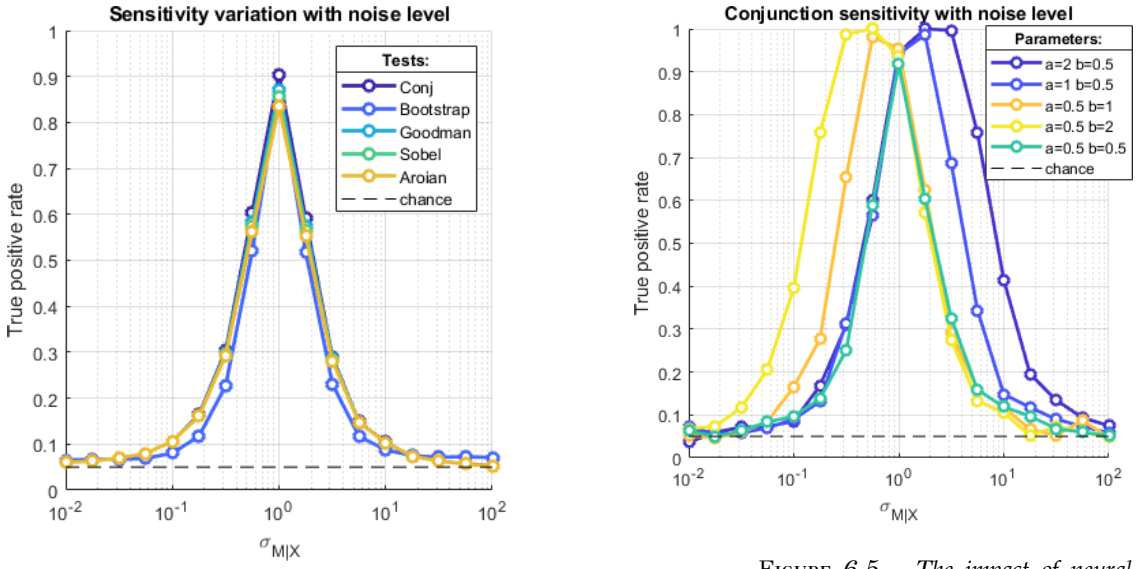



FIGURE 6.5 – *The impact of neural noise on statistical power. Left panel : The sensitivity of mediation tests (y-axis) is plotted against the variance of neural noise (x-axis), for each candidate testing approach (same format as Figure 6.4), when $a = b = 1/2$. Chance level is indicted using a black dotted line. Right panel : The sensitivity of conjunctive mediation tests (y-axis) is plotted against the variance of neural noise (x-axis), when varying path coefficients (dark blue : $a = 2$ and $b = 1/2$, blue : $a = 1$ and $b = 1/2$, cyan : $a = b = 1/2$, orange : $a = 1/2$ and $b = 1$, yellow : $a = 1/2$ and $b = 2$).*

All testing approaches have a similar sensitivity profile, which follows a bell-shaped function of neural noise magnitude, with an apex around $Var[\epsilon_M^0] = 1$. This corresponds to a situation in which about 20% of the trial-by-trial variance in $M$ is explained by $X$. As the amount of explained variance in $M$ departs from this nominal level, the sensitivity of mediation analysis effectively tends towards chance level. Now, everything else being equal, increasing $a$ or the variance of $X$ eventually inflates sensitivity on the right tail of the sensitivity profile, while increasing $b$ rather boosts its left tail. This moves the position of sensitivity apex towards smaller and stronger noise variance, respectively (see Figure 6.5, right panel).

Now, in the Methods section above, we reasoned that the expected sensitivity profile of mediation analysis should eventually favor the detection of neural information processing steps that are positioned away from either end of the processing hierarchy. In what follows, we compare candidate testing approaches w.r.t. their ability to detect levels in a simple feed-forward hierarchy. In brief, we simulated 1,000 datasets under Equation 6.10, using 100 intermediary network nodes. In all simulations, initial and final path coefficients were set to $a_0 = b = 1/2$ and all intermediary path coefficients were set to $a_i, \forall i$. In addition, the variance of all independent variables were set to unity except for the local neural noise increments, whose standard deviation was set to 0.3. Following the principle of mass-univariate mediation analysis, a mediation test was then performed on each node in isolation (significance threshold : $\alpha = 0.05$). For each network node, the ensuing (true positive) detection rate was then measured

across the 1,000 simulations. The result of the ensuing detection profile is shown on Figure 6.6 below.

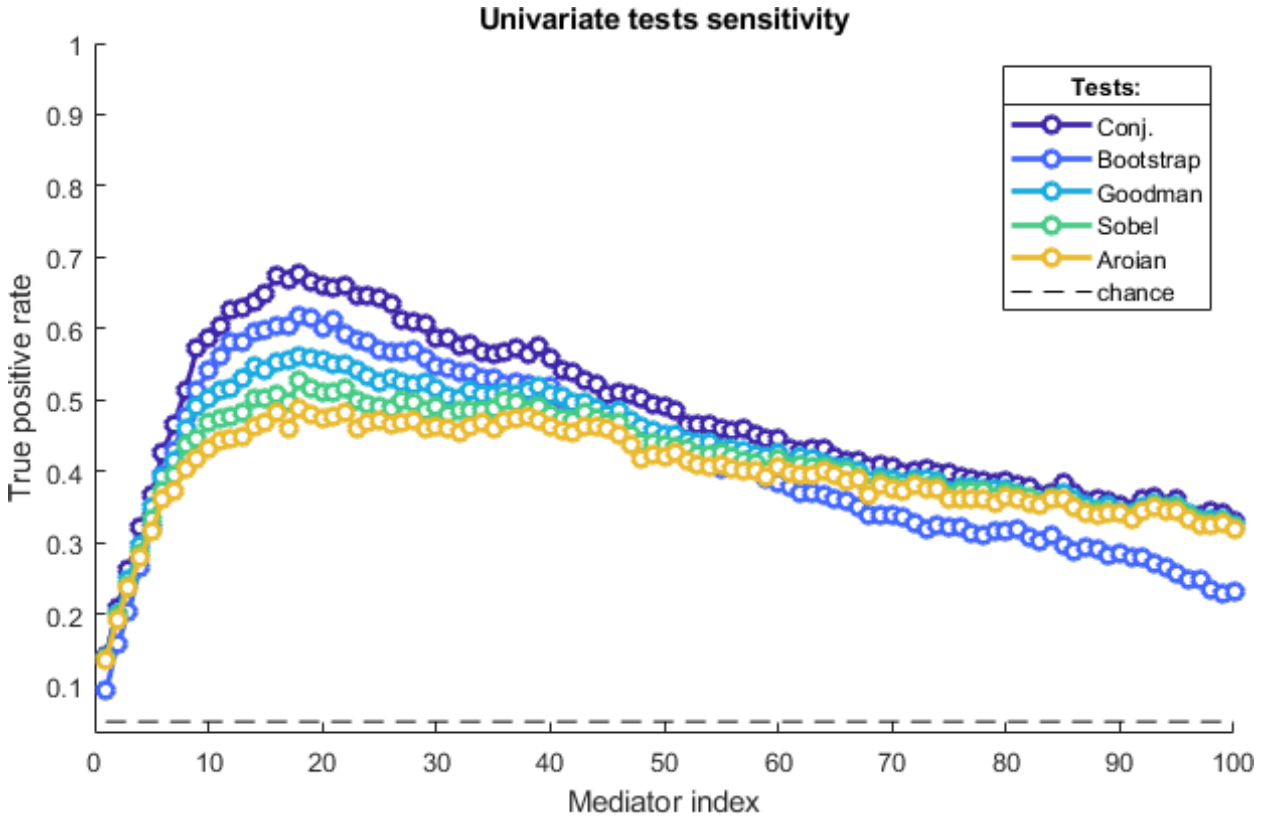



FIGURE 6.6 – *The heterogeneity of statistical sensitivity.*
*The sensitivity of mediation tests (y-axis) is plotted against the hierarchical level of candidate mediators along the serial processing pathway (x-axis), for each candidate testing approach (same format as Figure 6.4).*

As expected, local neural noise increments accumulate along the hierarchy, effectively increasing the neural noise level estimate as the hierarchical level increases. In turn, the detection profile also follows a bell-shaped function of hierarchical level, such that lower and higher hierarchical levels are less easy to detect. Interestingly, one can also see that different testing approaches have different sensitivity profiles. In particular, one can see that the conjunctive approach exhibits a higher sensitivity than all other approaches, irrespective of the hierarchical level of interest. Note that the M3 bootstrap test is better than other indirect approaches for intermediate hierarchical levels, but eventually loses its competitive advantage for higher hierarchical levels.

### 6.3.2 *Assessing the robustness to deviations from hemodynamic assumptions*

Despite the inclusion of HRF derivatives in the mediation model, deviations to the canonical HRF can impair test sensitivity. In this section, we compare the robustness of convolution and deconvolution approaches to unanticipated delays in HRF. We thus simulated 100 datasets using the same parameter settings as above, except that we varied systematically the HRF delay, effectively inducing a shift with the canonical HRF ranging from -5 second to 5 seconds. Each

dataset was then analyzed using all (indirect and conjunctive) testing approaches, under both convolution and deconvolution strategies (with the canonical HRF and its delay derivative). For each HRF delay shift, we then measured the (true positive) detection rate of each candidate mediation analysis strategy, when setting the significance threshold to $\alpha = 0.05$. The ensuing sensitivity profiles are summarized on Figure 6.7 below.

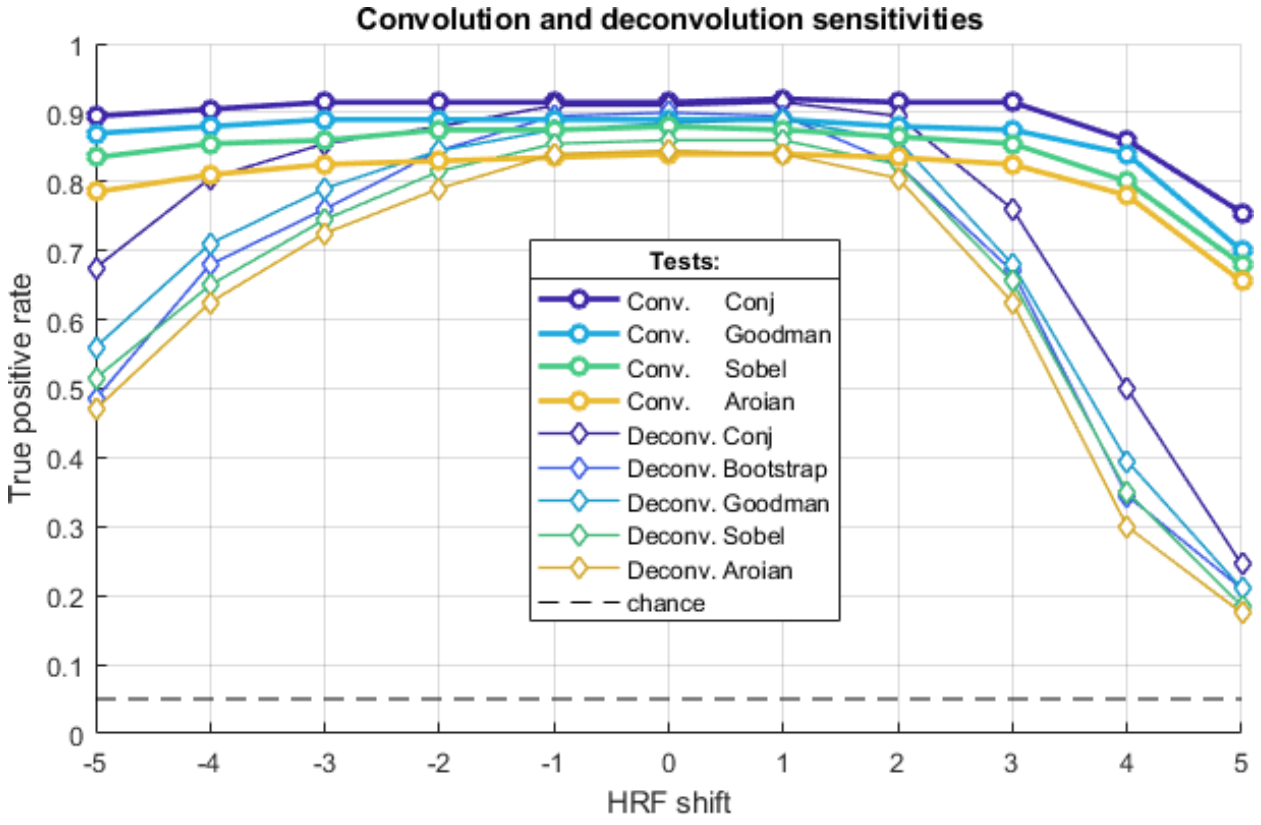



FIGURE 6.7 – *The impact of unmodelled hemodynamic delays.*
*The sensitivity of mediation tests (y-axis) is plotted against the HRF shift (x-axis), for each candidate testing approach (same format as Figure 6.4, thick lines : convolution approach, thin lines : deconvolution approaches).*

All mediation analysis strategies exhibit a bell-shaped sensitivity profile, eventually peaking when there is no deviation to the canonical HRF (i.e. when the HRF delay shift is null). Also, when there is no deviation to the canonical HRF, *deconvolution* and *convolution* strategies yield similar test sensitivity. However, when the deviation to the canonical HRF increases, the loss of statistical sensitivity is much stronger for *deconvolution* than for convolution approaches. For example, with a (realistic) delay shift of 3 seconds, most *deconvolution* approaches lose about 10% to 15% sensitivity on average. In comparison, *convolution* approaches only lose about 2% sensitivity. In addition, the *conjunctive* approach always exhibit higher sensitivity than indirect approaches, irrespective of whether one chooses a convolution or *deconvolution* strategy.

We note that, with a significance threshold of $\alpha = 0.05$, deviations to the canonical HRF has no adverse effect on the validity of statistical tests, i.e. all mediation test approaches yield 5% or less false positive rates under the null.

### 6.3.3 *Addressing the interpretational issue of brain-behavior mediation analysis with the I/O test statistics*

Recall that a significant mediated effect may have two distinct causal interpretations : the behavioral variable may either be an input ($Y \rightarrow M$) or an output ($M \rightarrow Y$) of the brain region where the null has been rejected. To address this issue, we proposed a simple I/O test statistics, whose sign is expected to discriminate between these two scenarios. Here, we evaluate the utility of the I/O test statistics $\bar{\lambda}$, in conditions similar to our fMRI data analysis below, using numerical Monte-Carlo simulations. First, we simulated data under three scenarios :

— H1 (native causal scenario $M \rightarrow Y$) : the independent variable $X$ is sampled under a normal distribution, each multivariate mediator unit $M_i$ is set to a noisy affine transformation of $X$ (with random weights), and the dependent variable $Y$ is set as a noisy mixture of $X$ and all mediator units (with random weights).

— H2 ("swapped" causal scenario $Y \rightarrow M$) : the independent variable $X$ is sampled under a normal distribution, the dependent variable $Y$ is set as a noisy affine transformation of $X$ (with a random weight) and each multivariate mediator unit $M_i$ is set to a noisy mixture of $X$ and $Y$ (with random weights).

— H0 (null scenario) : the independent variable $X$ is sampled under a normal distribution, and all other variables are set to a noisy affine transformation of $X$ (with random weights).

We simulated each scenario 1000 times, with 64 trials and 20 mediating units (all random variables and weights were sampled under a centered normal distribution with unit variance). Note that, in all three scenarios, $M$ and $Y$ variables are correlated with each other (under the null, this is because of the influence of $X$, which acts as a confounding variable). For each simulated dataset, we derive the I/O test statistics $\bar{\lambda}$. The resulting Monte-Carlo distributions are shown on Figure 6.8 below (left panel).

On can see that the three scenarios are very well discriminated. In particular, the distribution of the $\bar{\lambda}$ under the null is centered on zero, and lies in between its distribution under H1 and under H2. Moreover, and as expected, $E[\bar{\lambda}|H_1] < 0$ and $E[\bar{\lambda}|H_2] > 0$.

These simulations however, do not account for the limitations that arise in realistic settings. In particular, the number of neural units or voxels that compose the multivariate set of mediators may largely exceed the number of trials. Here, a pragmatic solution is to perform a PCA decomposition, and keep the K first principal components to summarize the within-region variability. We now ask whether the ensuing I/O test statistic is robust to this dimension reduction. In brief, we performed the same set of simulations as above, this time simulating 100 mediating units and deriving the test statistics from the ensuing K=20 first principal components. The resulting Monte-Carlo distributions are shown on the right panel of Figure 6.8.

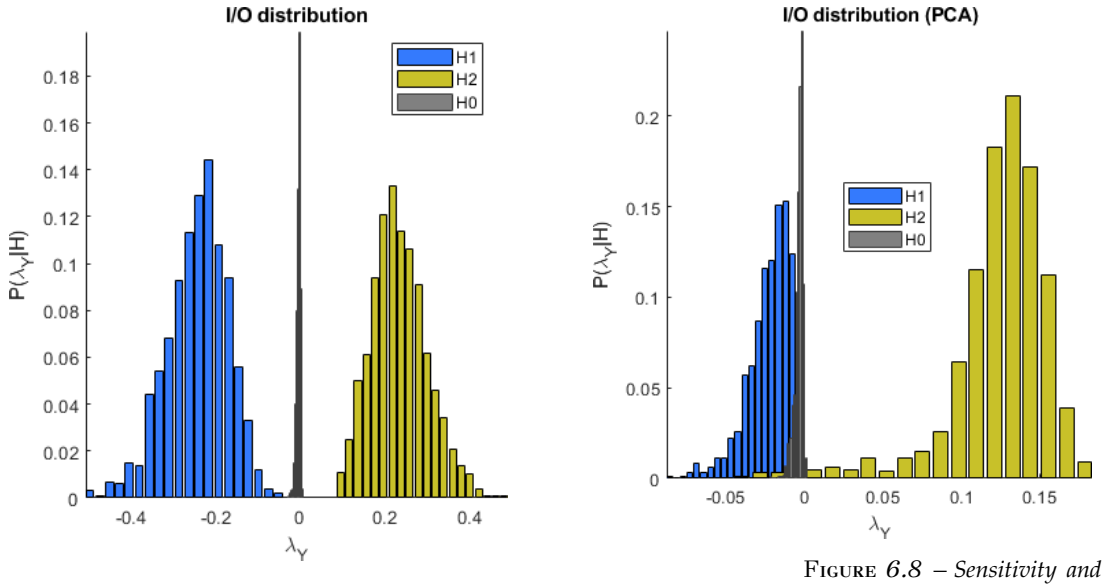



FIGURE 6.8 – *Sensitivity and robustness of the I/O test statistics $\bar{\lambda}$. Left panel : The Monte-Carlo distribution of the I/O test statistics $\bar{\lambda}$ (y-axis) is plotted under alternative scenarios (H1 : blue, H2 : yellow, H0 : grey). Right panel : Same format as left panel, but under data dimension reduction (20 first principal components of a PCA).*

One can see that the dimension reduction strongly reduces the range of variation of the I/O test statistics, when compared to the situation above, where all the relevant variation is available. Furthermore, the magnitude of $\bar{\lambda}$ under scenario H1 and H2 is asymmetrical. More precisely, one can see that $|E[\bar{\lambda}|H_1]| < |E[\bar{\lambda}|H_2]|$. In other terms, when relevant information is lacking, the average evidence in favor of H1 is weaker than the average evidence in favor of H2. Nevertheless, the sign of the I/O test statistics can still be interpreted as evidence for or against the native causal interpretation of the brain-behavior mediation model, i.e. $E[\bar{\lambda}|H_1] < 0$ and $E[\bar{\lambda}|H_2] > 0$.

### 6.3.4 *fMRI study of decision making under risk*

Here, we perform a brain-behavior mediation analysis of previously acquired fMRI data (Chen, 2014) [59], which is openly available as part of the OpenFMRI project [60]. In this study, 60 participants made a series of 64 accept/reject decisions on risky gambles. On each trial, a gamble was presented, entailing a 50/50 chance of gaining an amount G of money or losing an amount L (so-called "baseline" condition). Participants were told that, at the end of the experiment, four trials would be selected at random : for those trials in which they had accepted the corresponding gamble, the outcome would be decided with a coin toss, and for the other ones -if any-, the gamble would not be played. All 64 possible combinations of G/L pairs (10$<G<40$, 5$<L<20$) were presented across trials, which were separated by 7 seconds on average (min 6, max 10). MRI scanning was performed on a 3T Siemens Prisma scanner. High-resolution T1w structural images were acquired using a magnetization prepared rapid gradient echo (MPRAGE) pulse sequence with the following parameters : TR = 2530 ms, TE = 2.99 ms, FA = 7, FOV = 224 × 224 mm, resolution = 1 × 1 × 1 mm. Whole-brain fMRI data were acquired using echo-planar imaging with multi-band acceleration

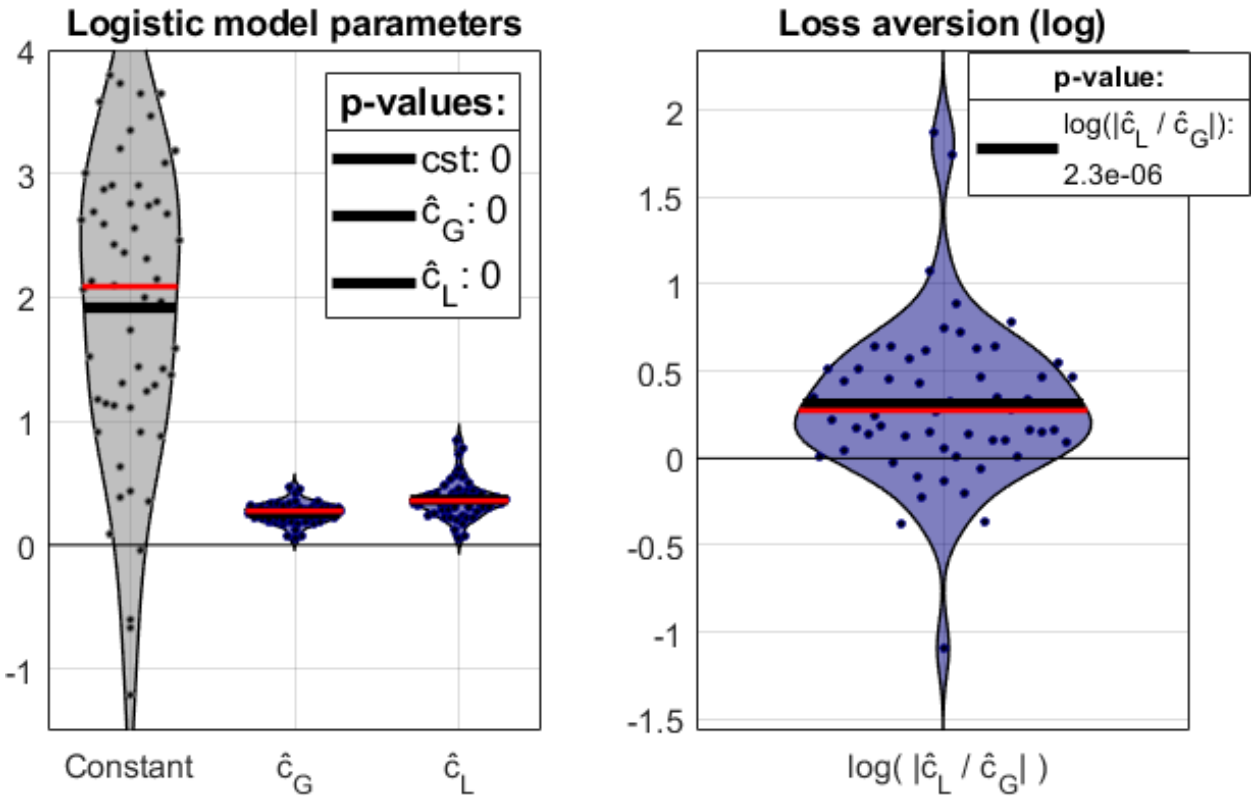
59. Chen, M.-Y. (2014). The development of bias in perceptual and financial decision-making

60. Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., and Milham, M. (2013). Toward open sharing of task-based fmri data : the openfmri project. *Frontiers in Neuroinformatics*, 7(July):12

factor of 4 and parallel imaging factor (iPAT) of 2, TR = 1000 ms, TE = 30 ms, flip angle = 68 degrees, in-plane resolution of 2X2 mm 30 degrees of the anterior commissure-posterior commissure line to reduce the frontal signal dropout, with a slice thickness of 2 mm and a gap of 0.4 mm between slices to cover the entire brain. See https://openneuro.org/datasets/ds000053/versions/00001 for more details.

Data preprocessing included standard realignment and movement correction steps. Note that we excluded 2 participants, either due to missing information or because the misalignment between functional and anatomical scans could not be corrected.

We first regressed, for each participant, the observed choices against gains and losses (Equation 6.1). This yielded estimates of the total effects $\hat{c}_G$ and $\hat{c}_L$ of gains and losses, respectively. This also provided an estimate $\hat{\sigma}_{Y|X}$ of the behavioral residuals' standard deviation. The results of this analysis are shown on Figure 6.9 below.




In brief, both gain and loss factors have a significant effect on decisions under risk (gain factor : $p < 10^{-5}$, loss factor : $p < 10^{-5}$). We note that, together, gain and loss factors explain on average 44.6% (std : 24.2%) of the trial-by-trial variance on participants' decisions (average balanced accuracy : 84.84%, std : 9%).

For each participant, we also derived a loss-aversion index : $\hat{\omega} = log(\hat{c}_L/\hat{c}_G)$, which is positive when losses have a stronger weight on accept/reject decisions than gains. One can see that the average loss-aversion index is significant (p<10-5), i.e. losses have more weight

FIGURE 6.9 – *Summary of behavioral results. Left panel : Between-subject empirical distribution of estimated within-subject parameters (left : constant term in the regression, middle : gain weight $\hat{c}_G$, right : loss weight $\hat{c}_L$). The black and red lines show the group-level mean and median, respectively. Right panel : Between-subject empirical distribution of the loss-version index (same format as left panel).*

on participants' decisions than gains.

Then, we analyzed fMRI time series (at the within-subject level) using both *convolution* and *deconvolution* approaches. The convolution strategy relied upon the following two GLMs :

— Equation 6.10 (first line) : GLM1 included regressors for trial-by-trial gains and losses (temporally aligned with the gamble presentation and convolved with the canonical HRF and its delay derivative), and basic confounding factors (six movement regressors and their squared values, as well as a Fourier basis set for slow drift removal). Fitting GLM1 to each fMRI voxel time series yielded a map of estimates $\hat{a}_G$ and $\hat{a}_L$ that correspond to the local effect of gain and loss on neural activity at the time of gamble presentation, respectively. In addition, we extracted the standard deviation of GLM1's residuals, which form a map of the local neural noise's strength $\hat{\sigma}_{M|X}$.

— Equation 6.11 (second line) : GLM2 is identical to GLM1, but also includes acceptance/rejection choices (convolved with the canonical HRF and its temporal derivatives). Fitting GLM2 to fMRI time series yielded regressor weight estimates that measure the correlation between local neural activity and behavior, above and beyond the effect of gain and losses ($\hat{d}$). The map of local path coefficients $\hat{b}$ was then obtained from $\hat{d}$, $\hat{\sigma}_{Y|X}$ and $\hat{\sigma}_{M|X}$ using Equation 6.12.

The *deconvolution* strategy was implemented as follows. First, we fitted GLM3, which included "trial" regressors (temporally aligned with the gamble presentation) as well as basic fMRI confounds. Regression weight estimates yielded local trial-by-trial neural responses $\hat{M}$. Maps of path coefficients estimates $\hat{a}$ and $\hat{b}$ were obtained using Equation 6.4, given local neural responses $\hat{M}$.

Random-effect group-level inference on the mediation of gain and loss factors was then performed by reporting group averages of path coefficients $\hat{a}$ and $\hat{b}$, after 8mm FWHM smoothing. We applied all indirect and conjunctive approaches except for the M3 bootstrap method (because of its limited statistical gain, when compared to its computational cost). For all approaches, we used unsigned (two-tailed) tests with standard random field theory (RFT) correction for whole-brain multiple comparisons correction.

In brief, no mediation testing approach based upon the *deconvolution* strategy reached statistical significance. This was the case even when using more lenient corrections for multiple comparisons (e.g., FDR). This was however not the case for mediation analyses based upon the *convolution* strategy. Here, indirect approaches yielded group-level significant clusters at low-set inducing thresholds ($p = 0.01$ or $p = 0.05$, uncorrected). In what follows, we discard these results as these thresholds are known to violate RFT assumptions (Flandin and Friston, 2019). Now, under the default set-inducing threshold (p=0.001, uncorrected), the conjunctive approach identified 6 clusters that significantly mediate the effect of gain : the right supra-

marginal gyrus or SMG ($p = 0.011$, RFT-corrected), bilateral posterior dorsomedial PFC or BA8 (left : $p = 0.003$, right : $p = 0.008$, RFT-corrected), the right anterior ventrolateral PFC or BA45 ($p = 0.018$, RFT-corrected) and bilateral posterior dorsolateral PFC or BA8/9 (left : $p = 0.007$, right : $p = 0.009$, RFT-corrected). In addition, there was a trend ($p = 0.06$, RFT-corrected) for 1 cluster mediating the effect of loss, in the left anterior ventrolateral PFC. These clusters are shown on Figure 6.10 below.

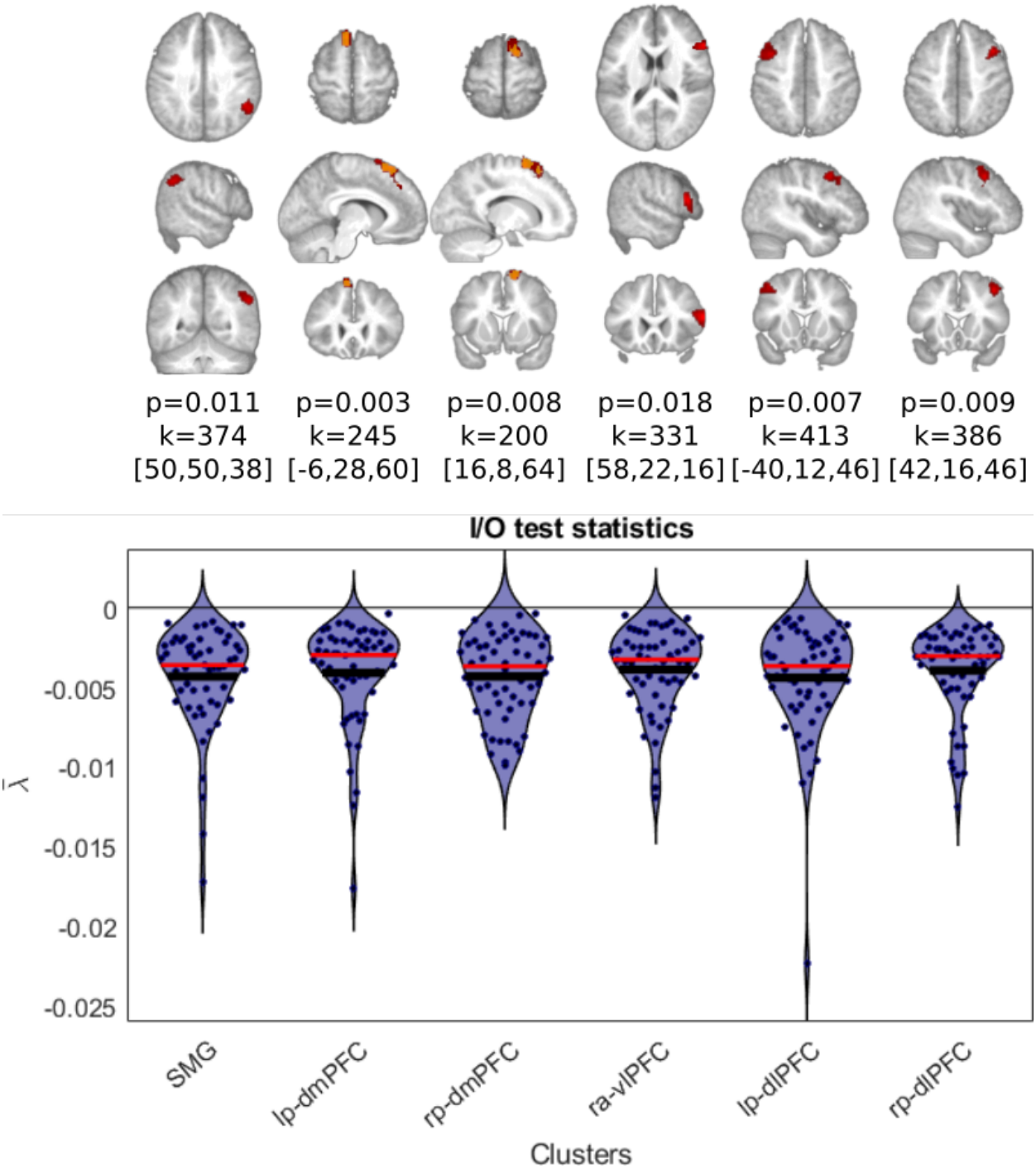



FIGURE *6.10 – Significant mediators of the gain effect on decisions under risk.*
*Upper panel : the six significant clusters of brain-behavior mediation analysis (conjunctive/convolution approach) are shown on axial (up), sagittal (middle) and coronal views (bottom). All maps used a default set-inducing threshold of correction* $p = 0.001$ *uncorrected (red areas) for the RFT correction, except the bilateral dmPFC's map where with* $p = 0.0002$ *uncorrected (yellow areas) in order to separate the two hemispheres. Lower panel : The ensuing between-subject empirical distribution of the I/O test statistics* $\bar{\lambda}$ *(y-axis, group-level mean ±standard deviation) is shown for each significant clusters (x-axis).*

We note that regions contralateral to significant unilateral mediators of the gain effect were all close to statistical significance : c.f. left SMG ($p = 0.094$, RFT-corrected) and left anterior vlPFC or BA45 ($p = 0.177$, RFT-corrected).

At the very least, these analyses demonstrate the superior statistical

efficiency of *conjunctive/convolution* approaches. In brief, no other candidate variant of mediation analysis yields positive results on this dataset.

Now, the significant mediated effects above may have two distinct causal interpretations. To afford evidence in favor or against the "native" causal claim of brain-behavior mediation analysis, we derived, for each participant and each significant cluster, our I/O test statistics. Note that, prior to the analysis, we summarized the trial-by-trial variance in each cluster using the 20 first principal components from the within-cluster PCA decomposition (on average cross clusters and participants, these preserve 89% ±2% of the trial-by-trial variance). The group-level empirical distribution of $\bar{\lambda}$ is shown on the lower panel of Figure 6.10, for each of the 6 significant clusters. Reassuringly, all clusters exhibit strong evidence in favor of the "native" causal interpretation of brain-behavior mediation analysis, i.e. $\bar{\lambda} < 0$ for all subjects and all clusters. We will comment on these results in the Discussion section.

Now, the effect of experimental factors seems to be mediated by a set of anatomically segregated regions in the brain. These regions are likely to be organized into a functional network (cf. Figure 6.1 above), eventually exerting competitive and/or cooperative influences on behavioral responses. The analysis above is agnostic about the functional architecture of this network. However, the extent to which each of these network nodes actually mediates the effect of gains and losses onto choices varies across subjects. Thus, a given individual may have an idiosyncratic structure of brain pathways for processing gain and loss information. In turn, inter-individual differences in the pattern of mediated effect sizes may have behavioral consequences in terms of how strongly gains and/or losses impact decisions under risk.

Recall that the balance between the behavioral effects of gains and losses is measured using the loss aversion index $\hat{\omega} = log(\hat{c}_L/\hat{c}_G)$ (cf. Figure 6.9). One may thus ask whether the pattern of mediated effect sizes predicts loss aversion. We thus extracted, in each voxel of the 6 significant mediating clusters above, the indirect effect size $|\hat{a}_G\hat{b}|$, and average these within each cluster. This resulted in 6 region-specific indirect sizes per participant. We then regressed loss aversion indices against (log-transformed) indirect effect sizes, across participants. The results of this analysis are summarized on Figure 6.11 below.

First, an omnibus F-test shows that the pattern of indirect effect sizes significantly predicts loss aversion ($F = 5.13$, $df = [7, 51]$, $R2 = 12.8\%$, $p = 2 \times 10^{-4}$). This is important, since this means that one can think of loss aversion in terms of a trait that is partly determined by the relative contribution of processing pathways that mediate the effect of gains onto decisions under risk. In addition, one can see that loss aversion increases when the indirect effect size in the left posterior dmPFC increases ($t = 2.66$, $df = 51$, $p = 0.01$). No indirect effect size in any other region has a significant effect on loss aversion (all $p > 0.29$).

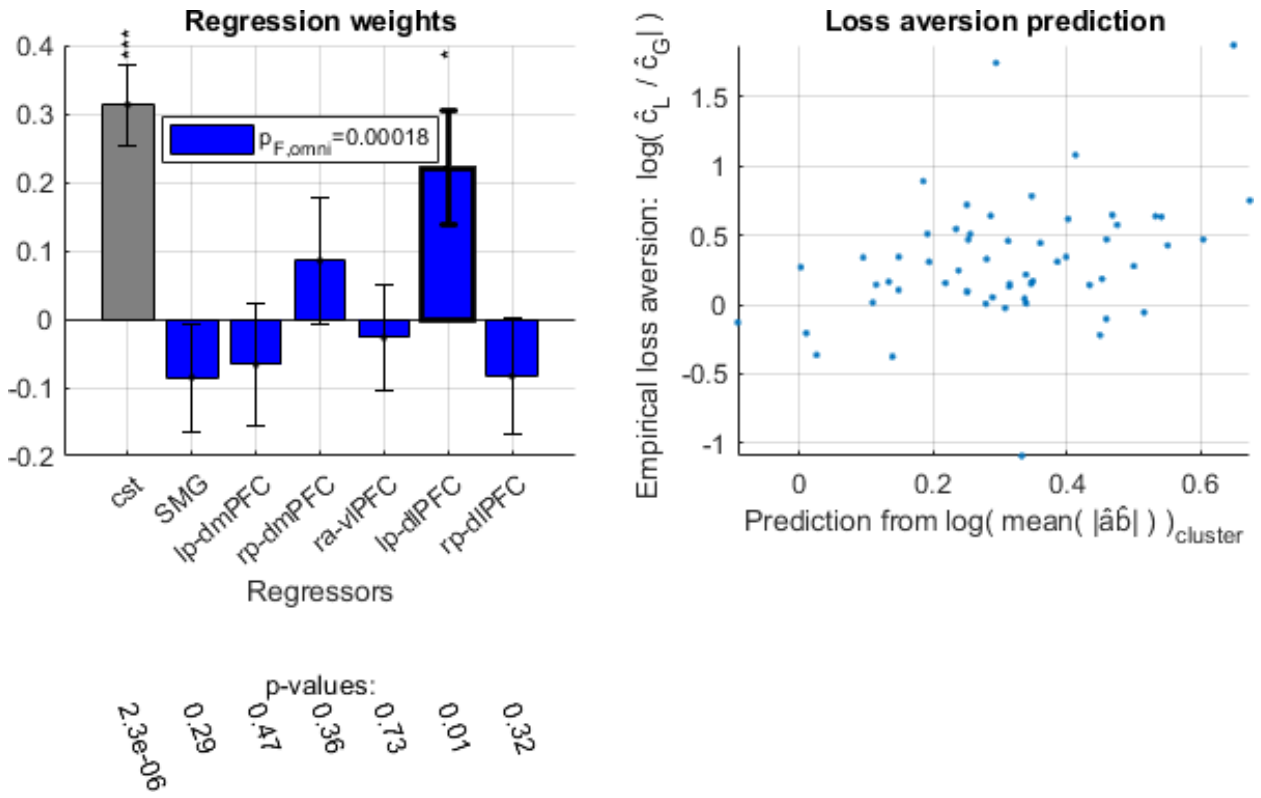



FIGURE 6.11 – *Inter-individual differences in loss-aversion. Left panel : regression coefficients of the analysis of inter-subject differences of loss aversion (grey : constant term, blue : weight of inter-individual differences in cluster-averages of indirect effect size). Error-bars depict standard errors of the mean. Right panel : Observed (y-axis) versus predicted (x-axis) loss aversion. Each dot is a participant.*

## 6.4 DISCUSSION

In this work, we identified the specific challenges of brain-behavior mediation analysis. In particular, we evaluated the specificity and sensitivity of five statistical tests, including so-called indirect and conjunctive approach. In brief, the conjunctive approach systematically shows higher sensitivity, while yielding valid inference. In addition, we disclosed the non-trivial impact of neural noise, and assessed the robustness to deviations from fMRI modelling assumptions. The former implies that brain-behavior mediation analysis cannot detect mediators that are too close from either end of the neural information processing hierarchy. In-silico investigations of the latter eventually favor the convolution approach to HRF modelling. We also disclosed some interpretational issues of mediated effects, in particular : significant mediated effects have two distinct causal interpretations. Importantly, this causal degeneracy may be partially addressed using complementary multivariate I/O test statistics. In addition, it has unexpected favorable computational consequences for whole-brain mediation analysis. Lastly, brain-behavior mediation analysis of fMRI data acquired in the context of decisions under risk further demonstrated the importance of methodological choices regarding brain-behavior mediation analysis. Eventually, the conjunctive/convolution test approach showed that the right SMG, bilateral posterior dmPFC, right anterior vlPFC and bilateral posterior dlPFC mediate the effect of prospective gains on decisions under risk.

Group-level I/O test statistics provided evidence that these regions are contributing to shaping behavioral responses (in a feedforward, causal, manner), rather than collecting and/or processing information about it (cf. interpretational issue). Finally, we showed that inter-individual differences in loss aversion is partly determined by the relative contribution of these six regions to behavioral control.

Taken together, our numerical simulations and analyses of experimental fMRI data demonstrated that conjunctive testing has higher statistical sensitivity than indirect approaches. This is true even for the bias-corrected M3 bootstrap test, despite its huge computational cost. We note that the sensitivity of the M3 bootstrap test may, in principle, be improved by increasing the number of permutations used to approximate the null distribution (here : 1,000). This however, would render whole-brain analysis excessively slow. Note that M3 bootstrap and conjunctive tests had already been compared at the standard 5% significance threshold outside the context of fMRI (Hayes and Scharkow, 2013). Although authors noted that bias-corrected bootstrap tests were slightly invalid (false positive rate greater than 5%), they recommended them because they eventually yielded more reliable confidence interval estimations. We extended these simulations, eventually showing that the invalidity of bias-corrected bootstrap tests increases as one relies on more stringent significance threshold (cf. Figure 6.3), which is required when correcting for multiple comparisons. For all these reasons (test validity, statistical sensitivity and computational cost), we would rather favor conjunctive testing for mass-univariate brain-behavior mediation analysis.

Although computationally expedient, mass-univariate brain-behavior mediation analysis essentially relies upon an incomplete model. Not only is it agnostic about the structure of the distributed brain system that process the incoming information (cf. Figure 6.1), but local, voxel-based, mediation tests simply ignore about 99.999% of the brain. We would argue however, that such incompleteness may be necessary for statistical mediation analysis. Recall that evidence for a mediated effect requires an appropriate amount of neural noise. But neural noise estimates have two entirely distinct sources. On the one hand, it may derive from irreducible variations in neural responses that are inherent to the underlying neurobiological processes. On the other hand, it may arise from imperfections in the way neural responses are modeled. The latter most likely applies to the linear brain-behavior model in Equation 6.2. For example, saturating neural responses to stimuli would, under Equation 6.2, inflate model residuals. However, although the ensuing neural noise estimates $\hat{\epsilon}_M^0$ would be partly artefactual, they would still be very informative to predict behavioral responses $Y$ above and beyond the *linear* effect of $X$. Now, let us assume that a neurocognitive model was available, that would describe how incoming information $X$ would be distorted, transformed and integrated with other (potentially incidental) processes, along the processing hierarchy. For example, such model may derive from recent work in theoretical neuroscience regarding population coding [61 62],

61. Averbeck, B., Latham, P., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366

62. Georgopoulos, A., Schwartz, A., and Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419

predictive coding [63 64 65] or efficient coding [66 67]. Or it could rely on agnostic multivariate and/or nonlinear decompositions that, when properly parameterized, would account for all sorts of complex relationships between $X$ and $M$. In any case, if the model was complete enough, then observed neural activity would not strongly deviate from its predictions. This would preclude the statistical detection of mediated effects. Ironically speaking then, progress in modelling neural information processing may eventually hinder the statistical efficiency of brain-behavior mediation analysis. More practically, this means that statistical brain-mediation analysis may be used in an exploratory manner, to identify brain regions that contribute to behavioral control. Further, complementary, model-based approaches to neural information processing would then help reducing one's epistemic uncertainty regarding neural noise. For example, artificial neural network modeling may be useful to identify either the structure of processing pathways [68] and/or the impact of incidental biological constraints that may distort local neural information processing [69].

This is not to say, however, that statistical brain-behavior mediation analysis cannot be improved.

For example, one may aim at providing more informative inferences regarding the structure of the underlying processing hierarchy. A possibility here is to merge mediation analysis with existing graph analysis techniques that were developed for assessing effective connectivity in the brain [70 71 72]. Another, less exhaustive but simpler, solution is to work iteratively : having identified a brain region that significantly mediates the $X \rightarrow Y$ effect, one may then look for other brain regions that would mediate both $X \rightarrow M$ and $M \rightarrow Y$ relationships, and repeat on subsequent mediators. Note that this would require additional corrections for the natural dependencies between brain regions. We refer the interested reader to (van Kesteren and Oberski, 2019) [73] for an interesting first step in this direction.

We also think that progress can be made regarding the main interpretational issue of brain-mediation analysis. In this context, let us highlight two extensions of linear mass-univariate approaches that sound promising.

First, one may rely on more stringent inferences regarding the causality of the $M \rightarrow Y$ relationship [74]. For example, temporal precedence may be accounted for, and inserted in the brain-mediation model using variants of Granger causality [75]. Note that special care must be taken regarding hemodynamic delays, whose variations across brain regions may confound temporal precedence. In particular, established fMRI applications of Granger causality are known to be prone to such confounds [76 77 78]. Nevertheless, constraining the brain-behavior mediation model with temporal precedence would likely reduce spurious inferences.

Second, one may exploit locally multivariate information to discriminate between many-to-one ($M \rightarrow Y$) and one-to-many ($Y \rightarrow M$) input/output mappings. In this work, we proposed a first step in this direction : namely, the I/O test statistics $\bar{\lambda}$. Numerical simulations


63. Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711

64. Hosoya, T., Baccus, S., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77

65. Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex : A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87

66. Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, page 216–234

67. Doi, E. and Lewicki, M. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Comput*, 23:2498–2510

68. Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221

69. Brochard, J. and Daunizeau, J. (2020). Blaming blunders on the brain : can indifferent choices be driven by range adaptation or synaptic plasticity ? *BioRxiv*

70. Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Comput. Biol*, 5:1000408

71. Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., and Woolrich, M. (2011). Network modelling methods for fmri. *NeuroImage*, 54:875–891

72. Sporns, O. (2013). Making sense of brain network data. *Nat. Methods*, 10:491–493

73. van Kesteren, E.-J. and Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling : A Multidisciplinary Journal*, 26(5):710–723

74. Preacher, K. (2015). Advances in mediation analysis : A survey and synthesis of new developments. *Annu. Rev. Psychol*, 66:825–852

75. Zhao, Y. and Luo, X. (2017). Granger mediation analysis of multiple time series with an application to fmri

76. David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., and Depaulis, A. (2008). Identifying neural drivers with functional mri : An electrophysiological validation. *PLoS Biol*, 6:315

77. Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on granger causality analysis of fmri. *NeuroImage*, 52:884–896

78. Zhao, Y. and Luo, X. (2017). Granger mediation analysis of multiple time series with an application to fmri

demonstrated the utility of this information-theoretic measure, and its robustness to partial information losses that result from data dimensionality reduction. However, this work falls short of an exhaustive analytical treatment of I/O test statistics. For example, neither did we investigate whether and how nonlinearities in causal relationships confound and/or bias $\bar{\lambda}$ estimates, nor did we derive a formal statistical test of the significance of $\bar{\lambda}$ estimates. We note that the difficulty here, is that the null hypothesis may not be the most useful reference point for I/O test statistics. Rather, one aims at comparing two alternative non-nested models. Therefore, an optimal statistical treatment of I/O tests statistics may be best approached using a bayesian approach [79 80]. We will pursue this and related extensions of I/O test statistics in subsequent publications.

Finally, let us discuss the results of our fMRI analysis. Recall that we identified six candidate mediators of the effect of gain onto decisions under risk. Among these, the posterior dmPFC was previously shown to regulate speed-accuracy tradeoffs [81] and its anatomical lesion is known to impair inhibitory control in the presence of response conflict [82]. Also, decades of neuroimaging, stimulation and lesions studies have evidenced the role of posterior dlPFC and vlPFC cortices in cognitive control [83 84 85 86]. In addition, functional and anatomical studies report convergent evidence that the right SMG is crucial for regulating emotional responses [87 88 89]. Now, in the context of decisions under risk, automatic fear responses may induce a default tendency to reject risky gambles, eventually yielding loss aversion [90 91]. This default emotional bias may enter in conflict with the appetitive effect of prospective gains. Whether the appetitive dimension of gambles eventually dominates automatic fear responses may then depend on the potentiation of emotional responses and on the efficiency of downstream cognitive control, which would explain why the SMG, dmPFC, dlPFC and vlPFC cortices mediate the effect of gain on decisions. This is also in line with our analysis of inter-individual differences of loss aversion, which shows that peoples' loss aversion increases when the indirect effect size (of gains on accept decisions) in the left dlPFC pathway increases. This is because a strong involvement of the dlPFC pathway may signal inefficient cognitive control [92 93], which would result in loss aversion worsening.

Although quite self-consistent and elegant, this interpretation really relies on the "native" causal interpretation of brain-behavior mediation analysis. So what if we had not found support for this causal scenario with our I/O test statistics ? In fact, the existing literature may also be queried to find past evidence that may be more compatible with the alternative causal interpretation of brain-behavior mediation analysis. For example, beyond its well-known implication in language processing, the right SMG has been shown to be involved in somatosensory perception [94 95]. Under this perspective, evidence for $b \neq 0$ (or, equivalently, $d \neq 0$) may be interpreted in terms of low-level perceptual representations of (motor ?) action plans. We note that the experimental design is compatible with this interpretation because


79. Kass, R. and Raftery, A. (1995). Bayes factors. *J. Am. Stat. Assoc*, 90:773–795

80. Liu, C. and Aitkin, M. (2008). Bayes factors : Prior sensitivity and model generalizability. *J. Math. Psychol*, 52:362–375

81. Forstmann, B., Dutilh, G., Brown, S., Neumann, J., Cramon, D., Ridderinkhof, K., and Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proc. Natl. Acad. Sci*, 105:17538–17542

82. Nachev, P., Wydell, H., O'Neill, K., Husain, M., and Kennard, C. (2007). The role of the pre-supplementary motor area in the control of action. *Neuroimage*, 36:155– 163

83. Gbadeyan, O., McMahon, K., Steinhauser, M., and Meinzer, M. (2016). Stimulation of dorsolateral prefrontal cortex enhances adaptive cognitive control : A high-definition transcranial direct current stimulation study. *J. Neurosci*, 36:12530–12536

84. Levy, B. and Wagner, A. (2011). Cognitive control and right ventrolateral prefrontal cortex : reflexive reorienting, motor inhibition, and action updating. *Ann. N. Y. Acad. Sci*, pages 40–62

85. Nee, D. and D'Esposito, M. (2017). Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *ELife*, 6:28040

86. Soutschek, A. and Tobler, P. (2020). *Causal role of lateral prefrontal cortex in mental effort and fatigue*. Hum. Brain Mapp

87. Adolphs, R. (2002). Neural systems for recognizing emotion. *Curr. Opin. Neurobiol*, 12:169–177

88. Makovac, E., Meeten, F., Watson, D., Garfinkel, S., Critchley, H., and Ottaviani, C. (2016). Neurostructural abnormalities associated with axes of emotion dysregulation in generalized anxiety. *NeuroImage Clin*, 10:172–181

89. Silani, G., Lamm, C., Ruff, C., and Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *J. Neurosci*, 33:15466–15476

90. De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. (2006). Frames, biases and rational decision-making in the human brain. *Science*, 313(5787):684–687

91. De Martino, B., Camerer, C., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792

92. Braver, T., Cole, M., and Yarkoni, T. (2010). Vive les differences ! individual variation in neural mechanisms of executive control. *Curr. Opin. Neurobiol*, 20:242–250

the spatial arrangement of accept/reject responses is not randomized over trials (Chen, 2014)[96]. This highlights the need for developing approaches that reduce the causal ambiguity of simple brain-behavior mediation analyses.

## 6.A APPENDIX : OLS ESTIMATORS OF PATH COEFFICIENTS

Recall that the second line of Equation 6.2 can be re-written as :

$$\begin{aligned} Y &= Mb + Xc' + \epsilon_Y \\ &= [MX] \begin{bmatrix} b \\ c' \end{bmatrix} + \epsilon_Y \end{aligned} \tag{6.17}$$

The OLS estimators of $b$ and $c'$ path coefficients are thus given by :

$$\begin{aligned} \begin{bmatrix} \hat{b} \\ \hat{c'} \end{bmatrix} &= \left( \begin{bmatrix} M^T \\ X^T \end{bmatrix} [MX] \right)^{-1} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \begin{bmatrix} n & M^T X \\ X^T M n & n \end{bmatrix}^{-1} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \frac{1}{n^2 - M^T X X^T M} \begin{bmatrix} n & -M^T X \\ -X^T M n & n \end{bmatrix} \begin{bmatrix} M^T \\ X^T \end{bmatrix} Y \\ &= \frac{1}{n^2 - M^T X X^T M} \begin{bmatrix} n M^T - M^T X X^T \\ n X^T - X^T M M^T \end{bmatrix} Y \end{aligned} \tag{6.18}$$

where the third line derives from the analytical formulation of 2x2 inverse matrices. Now recall that $M = X\hat{a} + \hat{\epsilon}_M^0$ with $\hat{a} = 1/nM^T X$ . The estimator of path coefficients $b$ and $c'$ thus writes :

$$\begin{aligned} \begin{bmatrix} \hat{b} \\ \hat{c'} \end{bmatrix} &= \frac{1}{n^2 - n^2 a} \begin{bmatrix} n(X\hat{a} + \hat{\epsilon}_M^0)^T - n\hat{a}X^T \\ nX^T - n\hat{a}(X\hat{a} + \hat{\epsilon}_M^0)^T \end{bmatrix} Y \\ &= \frac{1}{n - n\hat{a}^2} \begin{bmatrix} \hat{\epsilon}_M^{0\,T} \\ (1-\hat{a}^2)X^T - \hat{a}\hat{\epsilon}_M^{0\,T} \end{bmatrix} Y \\ &= \begin{bmatrix} \frac{1}{1-\hat{a}^2} \frac{1}{n} \hat{\epsilon}_M^{0\,T} Y \\ \frac{1}{n} X^T Y - \hat{a}\hat{b} \end{bmatrix} \end{aligned} \tag{6.19}$$

This completes the derivation of path coefficients' estimates.

## 6.B APPENDIX : SOBEL'S TEST

In what follows, we summarize the derivation of Sobel's mediation test.

First, recall that, given Equations 6.4 and 6.5, both $\hat{a}$ and $\hat{b}$ follow gaussian distributions, namely : $\hat{a} \sim \mathcal{N}(a, \hat{\sigma}_a^2)$ and $\hat{b} \sim \mathcal{N}(b, \hat{\sigma}_b^2)$. Sobel's approach effectively reduces to a Laplace approximation of the distribution of the product $\hat{a}\hat{b}$ of path coefficients' estimates.

Let $f(\hat{a}, \hat{b}) = \hat{a}\hat{b}$ be the function that maps the pair of path coefficient estimates to their product. One can approximate $f(\hat{a}, \hat{b})$ using a first-order Taylor expansion in the neighborhood of some arbitrary point


93. Poldrack, R. (2015). Is efficiency a useful concept in cognitive neuroscience ? dev. *Cogn. Neurosci*, 11:12–17

94. Ben-Shabat, E., Matyas, T., Pell, G., Brodtmann, A., and Carey, L. (2015). The right supramarginal gyrus is important for proprioception in healthy and stroke-affected participants : A functional mri study. *Front. Neurol*, 6

95. Tunik, E., Lo, O.-Y., and Adamovich, S. (2008). Transcranial magnetic stimulation to the frontal operculum and supramarginal gyrus disrupts planning of outcome-based hand-object interactions. *J. Neurosci*, 28:14422–14427

96. Chen, M.-Y. (2014). The development of bias in perceptual and financial decision-making

$(a_0, b_0)$ :

$$\begin{aligned} f(\hat{a},\hat{b}) &\approx f(a_0,b_0) + \left.\frac{\partial f}{\partial \hat{a}}\right|_{a_0,b_0}(\hat{a}-a_0) + \left.\frac{\partial f}{\partial \hat{b}}\right|_{a_0,b_0}(\hat{b}-b_0) \\ &= a_0 b_0 + b_0(\hat{a}-a_0) + a_0(\hat{b}-b_0) \\ &= a_0\hat{b} + b_0\hat{a} - a_0 b_0 \end{aligned} \tag{6.20}$$

If we choose to use the above Taylor expansion in the neighborhood of the unknown true values of path coefficients $(a_0 = a, b_0 = b)$ , then Equation 6.18 provides us with a Laplace approximation to the first two moments of the bivariate product $\hat{a}\hat{b}$ :

$$\begin{aligned} E[\hat{a}\hat{b}] &\approx ab \\ Var[\hat{a}\hat{b}] &\approx \hat{\sigma}_a^2 b^2 + \hat{\sigma}_b^2 a^2 \end{aligned} \tag{6.21}$$

The Sobel test directly relies on this approximation to form a pseudo z-score $z_{ab}^{(Sobel)}$ for the strength of the indirect path, as follows :

$$z_{ab}^{(Sobel)} = \frac{\hat{a}\hat{b}}{\sqrt{\hat{\sigma}_a^2\hat{b}^2 + \hat{\sigma}_b^2\hat{a}^2}} \tag{6.22}$$

where the unknown path coefficients have been replaced by their OLS estimates. Note that $z_{ab}^{(Sobel)}$ is invariant under arbitrary rescaling of $X$, $Y$ and/or $M$ . Under the null $h_0 : ab = 0$, $z_{ab}^{(Sobel)}$ approximately follows Student's probability density function with appropriate degrees of freedom.

We note that this approximation will be quite tight away from the diagonal lines $\hat{a} = \pm\hat{b}$, where the product $\hat{a}\hat{b}$ will start to behave as a quadratic function. But Sobel's approximation error will grow quicker than estimation errors on path coefficients.

One can also show that Sobel's test statistics is always smaller than conjunctive's test statistics :

$$\begin{aligned} |z_{ab}^{(Sobel)}| &= \frac{1}{\sqrt{\hat{\sigma}_a^2/\hat{a}^2 + \hat{\sigma}_b^2/\hat{b}^2}} \\ &= \frac{1}{\sqrt{1/t_a^2 + 1/t_b^2}} \\ &= \frac{|t_a||t_b|}{\sqrt{t_a^2 + t_b^2}} \\ &= min(|t_a|,|t_b|)\frac{max(|t_a|,|t_b|)}{\sqrt{t_a^2+t_b^2}} \\ &\leq min(|t_a|,|t_b|) \end{aligned} \tag{6.23}$$

where $min(|t_a|, |t_b|)$ is the conjunctive test statistics (cf. Equation 6.9).

## 6.C APPENDIX : DEALING WITH CONTRASTS ON EXPERIMENTAL CONDITIONS

So far, we have only considered simple independent variables . However, a typical experiment includes more than one condition or

factor, and the question of interest might be best framed in terms of mediating the effect of a linear combination of independent variables. In other terms, we want to generalize classical mediation analyses of the sort implied by Equation 6.1 to contrasts of experimental factors.

Without loss of generality, let us consider an experimental design with $n_{cond}$ conditions, which are encoded through a $n \times n_{cond}$ design matrix $\mathbf{X}$. Typically, the entries of $\mathbf{X}$'s columns would be zero everywhere, except at trials that belong to the corresponding condition (where their value would be one). Replacing $X$ with the design matrix $\mathbf{X}$ in Equation 6.2 induces the following two-fold lieanr regression model :

$$\begin{aligned} M &= \mathbf{X}\mathbf{a} + \epsilon_M \\ Y &= Mb + \mathbf{X}\mathbf{c}' + \epsilon_Y \end{aligned} \tag{6.24}$$

where $\mathbf{a}$ and $\mathbf{c}'$ are now $n_{cond} \times 1$ vectors of regression coefficients that encode the condition means. In this context, most experimental questions of interest are framed in terms of contrasts on path coefficients $\mathbf{a}$. So how can one ask whether $M$ mediates the effect of an arbitrary contrast on path coefficients ?

Two cases may arise. In the simplest case, one would deal with single contrasts. Let $\mathbf{W}$ be an arbitrary $n_{cond} \times 1$ vector of contrast weights. For example, in a typical 2×2 factorial design, $x = [1 - 1 - 11]$would be capturing the interaction between the two factors. Single contrasts of this sort do not require any specific adaptation of mediation analyses, because $\mathbf{w}^T\mathbf{a}$ is a scalar, and its OLS estimate has a known fixed-form distribution under the null. In turn, asking whether single contrasts are mediated by $M$ reduce to testing whether $(\mathbf{w}^T\mathbf{a})b \neq 0$, which can be done using either the indirect or conjunctive approaches described above. Slightly more subtle is the case of multiple contrasts, as induced by global null hypotheses tests. For example, let us consider an experimental design with three conditions. In analogy to ANOVA, we wish to test for the mediation of any difference between the conditions. The corresponding contrast of interest $\mathbf{W}$ is now a 3×2 matrix of weights, and $\mathbf{W}^T\mathbf{a}$ becomes a 2×1 vector. Outside the context of brain-behavior mediation analysis, assessing the statistical significance of such a contrast would be performed using an F-test, for which p-values can be derived analytically (Friston et al., 1995). When using the conjunctive approach, this poses no problem, as one would simply compute the p-value of the resulting minimum F-statistics. The indirect approach is more difficult to adapt here. In principle, one would first have to partition the design matrix $\mathbf{X} \leftarrow [\tilde{\mathbf{X}}, \tilde{\mathbf{X}_0}]$ into subspaces respectively spanning the contrast of interest $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$ and the effects of no interest $\tilde{\mathbf{X}_0} = \mathbf{X}(I - \mathbf{W}\mathbf{W}^-)$, where $\mathbf{W}^-$ is the pseudo-inverse of $\mathbf{W}$. Then, one would remove the effects of no interest from both the mediator and the dependant variable. Finally, one would have to test whether any indirect path induced by the ensuing columns of $\tilde{\mathbf{X}}$ is significant. The latter issue is not entirely trivial, but can be solved using the *minimum p-value* approach [97] [98] of conjunction analysis.

97. Friston, K., Holmes, A., Price, C., BÃ¼chel, C., and Worsley, K. (1999). Multisubject fmri studies and conjunction analyses. *NeuroImage*, 10:385–396

98. Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, 25:653–660

## 6.D APPENDIX : GROUP-LEVEL RANDOM-EFFECT ANALYSIS

Let us now consider the specific issue of experiments performed with multiple subjects. For example, let us assume that each subject participates in an experiment consisting of multiple trials, such that Equation 6.2 describes the relationship existing between $X$, $M$ and $Y$ across trials, at the subject-level. We now want to ask whether there is a mediated effect that is consistent across subjects, at the group-level. This calls for mixed-effects analyses, which essentially assume that subject-level path coefficients are sampled from a parent (population) distribution whose mean we wish to infer on. This can be efficiently performed using a summary statistics approach [99] [100], whereby one first estimates subject-level effects (here, $\hat{a}_i$ and $\hat{b}_i$, where $i \in [1, \dots, n]$ is the participant's index, and $n$ is here the number of participants), and then report these for a random-effect analysis at the group-level. Similarly to subject-level analysis, both conjunctive and indirect approaches are possible here. Let $\mu_a$ and $\mu_b$ be the unknown population mean of $a$ and $b$ path coefficients, respectively. At the group-level, the null hypothesis of mediation analysis can be written as follows :

$$\begin{aligned} H_0^{(conjunction)} &: \mu_a = 0 \text{ OR } \mu_b = 0 \\ H_0^{(indirect)} &: \mu_a \mu_b = 0 \end{aligned} \tag{6.25}$$

where $H_0^{(conjunction)} \Leftrightarrow H_0^{(indirect)}$ as before.

The conjunctive approach then simply reduces to testing whether both group-mean estimates $\hat{\mu}_a = 1/n \sum_i \hat{a}_i$ and $\hat{\mu}_b = 1/n \sum_i \hat{b}_i$ differ from zero, which can be tested using the p-value for the minimum statistic. The indirect approach relies on testing whether the group-mean of the indirect effect differs from zero. According to the central limit theorem [101], the distribution of the average product $1/n \sum_i \hat{a}_i \hat{b}_i$ will quickly tend towards a Gaussian distribution. However, any non-zero covariance between path coefficients will bias the inference, because $E[\hat{a}\hat{b}] = E[\hat{a}]E[\hat{b}] + cov[\hat{a}, \hat{b}]$ (Kenny et al., 2003). In other terms, even if the null hypothesis is true, covarying fluctuations in path estimates may significantly differ from zero. This is why the indirect approach should rather rely on testing the product $\hat{\mu}_a \hat{\mu}_b$ of group-mean estimates. This can be done using either parametric (cf. Sobel, Airoian or Goodman test statistics) or non-parametric (cf. M3 bootstrap) approaches.

99. Friston, K., Stephan, K., Lund, T., Morcom, A., and Kiebel, S. (2005b). Mixed-effects and fmri studies. *NeuroImage*, 24:244–252

100. Holmes, A., Friston, K., and Friston, K. (1998). Generalisability, random effects and population inference

101. Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. U. S. A*, 42:43–47

## 6.E APPENDIX : CAUSAL IMPACT OF NEURAL NOISE

The non-trivial impact of neural noise is not a feature of univariate linear brain-behavior mediation models. In fact, one can show that this generalizes to any form of brain-behavior. In what follows, we rely on an information-theoretic framework that was developed for addressing mediation claims, irrespective of the mathematical form that the mediation model may take [102]. The only requirement here,

102. Pearl, J. (2012). The mediation formula : A guide to the assessment of causal pathways in nonlinear models. In *Causality*, pages 151–179. John Wiley Sons, Ltd

is that of a causal cascade from $X$ to $M$ and $Y$, and from $M$ to $Y$ (cf. directed acyclic graph in Figure 6.2, left panel).

Let $IE_{x,x'}(Y)$ be the expected impact of the mediator variable on the behavior, under a virtual change of the manipulation (from $X = x$ to $X = x'$, see Equation 9 in (Pearl, 2012)) :

$$IE_{x,x'}(Y) = \sum_{M} E[Y|X, M]\,(P(M|x') - P(M|x)) \tag{6.26}$$

where $P(M|X)$ is the conditional distribution of the mediator variable. Note that, in Equation 6.26, the causal relationships between $X$, $M$ and $Y$ are implicitly absorbed in conditional distributions. In brief, $IE_{x,x'}(Y)$ measures the strength of the indirect effect of $X$ onto $Y$, i.e. it serves as a summary statistics for significance tests of (possible multivariate and nonlinear) mediated effects.

Now, when there is no neural noise, the mediator variable brings no additional information on the behavior, i.e. $E[Y|X, M] \approx E[Y|X]$. In turn, the mediator's impact $IE_{x,x'}(Y)$ becomes negligible :

$$IE_{x,x'} \approx \; E[Y|X] \left( \sum_{M} \; P(M|x') \; - \; \sum_{M} P(M|x) \right) = 0$$

. In other terms, when the mediator brings no additional information on behavior, it cannot be detected. Conversely, when neural noise dominates, the mediator is effectively independent from the manipulation, i.e. : $P(M|x') \approx P(M|x) \approx P(M)$. It follows that, here again :

$$IE_{x,x'}(Y) \approx \sum_{M} E[Y|X, M]\,(P(M) - P(M)) = 0$$

In other words, when the manipulation brings no or little information on the mediator, no mediation can be detected. In conclusion, mediated effects can only be detected for intermediate neural noise magnitudes, irrespective of the mathematical form of the brain-behavior mediation model.

## 6.F APPENDIX : EQUIVALENCE OF CAUSAL INTERPRETATIONS OF MEDIATION ANALYSIS

In what follows we give a proof of (i) Equation 6.12 in the main text, and (ii) equality of t-statistics of "native" and "swapped" path coefficients.

First one can use the expressions of their OLS estimates to derive

the ratio of the two path coefficients (see Appendix 6.A) :

$$\begin{aligned}\frac{\hat{b}}{\hat{d}} &= \frac{(nM^T - M^TXX^T)Y}{n^2 - M^TXX^TM} \times \frac{n^2 - Y^TXX^TY}{(nY^T - Y^TXX^T)M} \\ &= \frac{1-(Y^TX/n)^2}{1-(M^TX/n)^2} \times \frac{(nM^T - M^TXX^T)Y}{(nY^T - Y^TXX^T)M} \\ &= \frac{1-\hat{\rho}(X,Y)^2}{1-\hat{\rho}(X,M)^2} \times \frac{M^TY/n - M^TX/n \times Y^TX/n}{Y^TM/n - Y^TX/n \times X^TM/n} \\ &= \frac{1-\hat{\rho}(X,Y)^2}{1-\hat{\rho}(X,M)^2}\end{aligned} \tag{6.27}$$

where $\hat{\rho}(\cdot,\cdot)$ is the sample correlation between arbitrary vectors. Now, recall that, in Equations 6.1-6.2, (i) all mediation variables are z-scored and (ii) residual estimates are, by construction, orthogonal to the variable $X$. Therefore :

$$\begin{aligned}1 &= \hat{\rho}(X,Y)^2 + \hat{\sigma}^2_{Y|X} \\ 1 &= \hat{\rho}(X,M)^2 + \hat{\sigma}^2_{M|X}\end{aligned} \tag{6.28}$$

This concludes the demonstration of Equation 6.12 of the main text ($\hat{b} = \hat{d} \times \hat{\sigma}^2_{Y|X}/\hat{\sigma}^2_{M|X}$).

Now let us prove the equality of t-statistics of "native" and "swapped" path coefficients. Using the definition of these test statistics we have :

$$\begin{aligned}t_b =& t_d \\ \Leftrightarrow& \frac{\hat{b}}{\hat{\sigma}^2_{Y|X,M}}\sqrt{n-2} = \frac{\hat{d}}{\hat{\sigma}^2_{M|X,Y}}\sqrt{n-2} \\ \Leftrightarrow& \frac{\hat{b}}{\hat{d}} = \frac{\hat{\sigma}^2_{Y|X,M}}{\hat{\sigma}^2_{M|X,Y}} \\ \Leftrightarrow& \frac{\hat{\sigma}^2_{Y|X}}{\hat{\sigma}^2_{M|Y}} = \frac{\hat{\sigma}^2_{Y|X,M}}{\hat{\sigma}^2_{M|X,Y}}\end{aligned} \tag{6.29}$$

Now recall the iterative decomposition of the determinant of a gram matrix : $det([A,v]^T[A,v]) = det(A^TA) \times (v^Tv - v^TA(A^TA)^{-1}A^Tv)$, where $A$ and $v$ are an arbitrary invertible matrix and a vector with adequate dimensions, respectively (Csató and Opper, 2003). This yields :

$$\begin{aligned}det([X,M,Y]^T[X,M,Y]) &= n \times det([X,M]^T[X,M]) \times \hat{\sigma}^2_{Y|X,M} \\ &= n^2 \times det(X^TX) \times \hat{\sigma}^2_{M|X}\hat{\sigma}^2_{Y|X,M}\end{aligned} \tag{6.30}$$

Similarly, we have :

$$det([X,Y,M]^T[X,Y,M]) = n^2 \times det(X^TX) \times \hat{\sigma}^2_{Y|X}\hat{\sigma}^2_{M|X,Y} \tag{6.31}$$

Lastly, because the order of the matrix's columns leaves the determinant unchanged, Equations 6.30 and 6.31 are identical. This implies that $\hat{\sigma}^2_{Y|X}\hat{\sigma}^2_{M|X,Y} = \hat{\sigma}^2_{M|X}\hat{\sigma}^2_{Y|X,M}$ , which concludes our proof.

# 7

# *Conclusion sur les analyses de médiation cérébrale*

◂ Chapitre 6 Chapitre 8 ▸

Dans ce chapitre j'ai étudié l'application de l'analyse de médiation aux données IRMf. À première vue ce type d'analyse semble permettre une identification des mécanismes neuraux sans avoir à fournir d'hypothèses à leur propos. Dans une certaine mesure, cette promesse est tenue et cette technique fournit des scénarios semi-génératifs du comportement. C'est-à-dire que le comportement est bien généré à partir des données neurales et des facteurs expérimentaux, mais toute l'activité neurale n'est cependant pas entièrement décrite et comporte des fluctuations incontrôlées.

Mes travaux ont également montré les limites d'une approche se privant d'hypothèses sur les mécanismes neuronaux. Les premières étapes perceptuelles et les dernières étapes motrices d'un processus comportementales seront toujours plus difficiles à détecter que des étapes intermédiaires. Une représentation parfaite des entrées, par exemple, ne pourra jamais être identifiée comme médiateur d'un processus comportemental, quel que soit le test utilisé. Cela ne condamne bien entendu pas cette approche, mais pousse à nuancer ses ambitions.

Une analyse de médiation permet bien de caractériser certains aspects d'un processus comportemental. Même si elle ne permet pas de déterminer toute une chaîne de traitement de l'information, elle permet d'identifier l'existence d'un mécanisme dans une région cérébrale et d'utiliser cette caractérisation pour prédire d'autre aspect du comportement, comme l'aversion à la perte dans mon étude. Cependant, pour comprendre le détail de ces processus des hypothèses sur les mécanismes en jeu sont nécessaires.

Le chapitre suivant explore justement l'évaluation de scénarios physiologiques dont les mécanismes contraignent les processus de traitement de l'information déterminant le comportement.

# Troisième partie

# RÉSEAUX DE NEURONES ARTIFICIELS CONTRAINTS PAR DES HYPOTHÈSES BIOLOGIQUES

# 8

# *Introduction aux contraintes biologiques sur les mécanismes computationelles*

◂ Chapitre 7 Chapitre 9 ▸

Nos choix dépendent de la manière dont nous traitons les informations à notre disposition. Depuis les travaux de Von Neumann [1], Savage [2], Kahneman et Tversky [3] de nombreux biais cognitifs ont été identifiés dans nos processus de décision. Les plus connus sont notre perception biaisée des probabilités [4] et notre aversion aux pertes [5]. Les traces neurales de ces biais ont même pu être identifiées grâce à l'IRMf chez des sujets sains [6] [7] et validés par des études sur des patients neurologiques présentant une lésion de l'amygdale [8].

Plus généralement, de plus en plus de travaux neuroscientifiques montrent une association statistique entre les biais cognitifs et l'organisation de l'activité cérébrale [9] [10] [11]. S'il est difficile d'établir la direction causale de cette association par des études IRMf, une théorie récente suggère que les biais cognitifs sont une conséquence de notre organisation cérébrale, et non l'inverse.

Proposée par Stanislas Dehaene et Laurent Cohen en 2007, la théorie du recyclage neuronal suggère que les fonctions cognitives modernes ont évolué en recyclant des systèmes neuronaux ancestraux [12]. Ce serait le cas par exemple du raisonnement mathématique, de la lecture et, dans une certaine mesure, du langage et de l'utilisation d'outils. En 2010, Michael Anderson pousse le raisonnement plus loin et propose qu'une fonction cognitive ne puisse être acquise que s'il existe déjà une « niche neurale » lui permettant de réutiliser d'anciens circuits neuronaux à de nouvelles fins [13]. La théorie du recyclage neuronal suggère ainsi qu'une fonction cognitive s'appuie exclusivement sur des circuits neuronaux préexistants et hérite donc de leurs limitations.

Ce chapitre propose d'étudier et d'identifier ce type de limitations, lorsqu'elles prennent la forme de contraintes biologiques s'appliquant aux traitements neuraux de l'information déterminant le comportement. Plus précisément, je propose une approche permettant de comparer différents types de mécanismes neuronaux altérant les processus de prise de décision, notamment la plasticité hebbienne et la sensibilité adaptative (prédite par la théorie du codage efficace). Ce type de mécanismes garantit une forme de flexibilité nécessaire au recyclage neuronal, mais pouvant induire une forme d'instabilité qui altère les processus de détermination du comportement.

L'approche présentée ici emprunte la logique des analyses model-

1. Neumann, V. and John Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press
2. Savage, L. (1954). *The foundations of statistics*. John Wiley Sons
3. Kahneman, D. and Tversky, A. (1979). Prospect theory : An analysis of decision under risk. *Econometrika*. *Econometrica*, 47(2):263–292
4. On a tendance à surestimer les faibles probabilités et sous-estimer les fortes probabilités.
5. Face à la perspective de gagner ou de perdre une certaine somme d'argent, on a tendance à donner deux fois plus d'importance aux pertes.
6. Hsu, M., Krajbich, I., Zhao, C., and Camerer, C. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience*, 29(7):2231–2237
7. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518
8. De Martino, B., Camerer, C., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792
9. Drugowitsch, J., Wyart, V., Devauchelle, A., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6):1398–1411
10. Polania, R., Woodford, M., and Ruff, C. (2018). Efficient coding of subjective value. *BioRxiv*, 358317
11. Rouault, M., Drugowitsch, J., and Koechlin, E. (2019). Prefrontal mechanisms combining rewards and beliefs in human decision-making. *Nature Communications*, 10(1)
12. Dehaene, S. and Cohen, L. (2007). *Cultural Recycling of Cortical Maps*. Neuron
13. Anderson, M. (2010). Neural reuse : A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33:245–313

based, mais préserve les propriétés de « microscope mathématique » des modèles décrivant le code neural. J'utilise tout d'abord les réseaux de neurones artificiels comme des modèles génératifs du comportement des sujets. Plusieurs types de réseau sont ici étudiés, chacun étant soumis à une contrainte biologique différente. Le profil d'activité de ces réseaux permet alors d'identifier une trace neurale de ces contraintes, qu'il est possible de comparer aux variations multivariées du signal BOLD par une analyse RSA [14] [15].L'intérêt de cette méthode réside dans sa capacité à s'abstraire de nombreux détails d'implémentation (au sens de Marr) d'un système de traitement neural de l'information [16].

Ce chapitre étudie ainsi la combinaison de réseaux de neurones et d'analyses multivariées IRMf pour identifier les contraintes neuronales qui altèrent les processus neuronaux déterminant le comportement.

14. Representational Similiraty Analysis

15. Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November):1–28

16. Kriegeskorte, N. and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 371(1705)

# 9

# *Blaming blunders on the brain : can indifferent choices be driven by range adaptation or synaptic plasticity ?*

◂ Chapitre 8 Chapitre 10 ▸

Ce chapitre présente l'article "*Blaming blunders on the brain : can indifferent choices be driven by range adaptation or synaptic plasticity ?*" que j'ai publié en préimpression sur le site bioRxiv. Il est en anglais et est présenté sans modifications.

Adresse web :
https://www.biorxiv.org/content/10.1101/2020.09.08.287714v1.

Abstract Computational investigations of learning and decision making suggest that systematic deviations to adaptive behavior may be the incidental outcome of biological constraints imposed on neural information processing. In particular, recent studies indicate that range adaptation, i.e., the mechanism by which neurons dynamically tune their output firing properties to match the changing statistics of their inputs, may drive plastic changes in the brain's decision system that induce systematic deviations to rationality. Here, we ask whether behaviorally-relevant neural information processing may be distorted by other incidental, hard-wired, biological constraints, in particular : Hebbian plasticity. One of our main contributions is to propose a simple computational method for identifying (and comparing) the neural signature of such biological mechanisms or constraints. Using ANNs (i.e., artificial neural network models) and RSA (i.e., representational similarity analysis), we compare the neural signatures of two types of hard-wired biological mechanisms/constraints : namely, range adaptation and Hebbian plasticity. We apply the approach to two different open fMRI datasets acquired when people make decisions under risk. In both cases, we show that although peoples' apparent indifferent choices are well explained by biologically-constrained ANNs, choice data alone does not discriminate between range adaptation and Hebbian plasticity. However, RSA shows that neural activity patterns in bilateral Striatum and Amygdala are more compatible with Hebbian plasticity. Finally, the strength of evidence for Hebbian plasticity in these structures predicts inter-individual differences in choice inconsistency.

## 9.1 INTRODUCTION

Why do we overreact to emotional stimuli ? Why are our judgments plagued with errors and biases ? Why do we engage in behaviors whose consequences may be detrimental ? That the brain's biology is to blame for all kinds of cognitive and/or behavioural flaws is not a novel idea [1] [2] [3] [4]. However, providing neuroscientific evidence that a hard-wired biological constraint shapes and/or distorts the way the brain processes information is not an easy task. This is because whether the brain deviates from how it should process a piece of information is virtually unknown. In this work, we show how one may use multivariate analysis of fMRI data to identify the neural signature of incidental, hard-wired, biological constraints on behaviorally-relevant neural information processing.

Over the past two decades, cognitive neuroscience has involved much effort into developing computational means to understand how the brain processes information. In particular, the computational neuroscience of perception, learning, and decision making has now reached a stage of maturity, both in terms of its methods and models and in terms of the reproducibility of the ensuing results. For example, neuroscientific evidence that basal ganglia encode the reward prediction error that enables reinforcement learning (i.e., learning from reward feedbacks) has been found repetitively in monkeys [5] [6] and humans [7] [8] [9]. From a methodological standpoint, this line of study is remarkable for two reasons. First, it highlights the importance of behavioral measurements for understanding how the brain processes information. This shifts the scientific question from identifying how the brain encodes incoming information (e.g., cues and feedbacks) to assessing how it uses this information to produce behavioral responses. Second, its theoretical basis is derived from formal computational models of learning originating from research in the field of artificial intelligence and robotics [10] [11]. This provides a formal reference point for interpreting neural signals in terms of neural computations, i.e., intermediary steps in neural information processing geared towards producing adapted behavioral responses.

Taken in isolation, none of these two aspects is particularly novel. Retrospectively, the focus on brain-behavior relationships is the hallmark of behavioral neuroscience. And computational neuroscience already had enabled deep quantitative insights for understanding the neural code of perceptual and motor systems, providing unprecedented empirical evidence for, e.g., population coding [12], predictive coding [13] [14], or efficient coding [15] [16]. But in combination, these two aspects allow one to understand how brain computations eventually shape non-trivial behavior. This has typically be done in two different ways. On the one hand, one may look for neural evidence of cognitive mechanisms that provide candidate explanations for observed behavioral deviations to normative theories. For example, this approach has placed the putative distortions of prospective loss perceptions that drive irrational risk attitudes on a firm neuroscientific footing [17] [18] [19].


1. Buschman, T., Siegel, M., Roy, J., and Miller, E. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11252–11255
2. Marois, R. and Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, 9(6):296–305
3. Miller, E. and Buschman, T. (2015). Working memory capacity : Limits on the bandwidth of cognition. *Daedalus*, 144(1)
4. Ramsey, N., Jansma, J., Jager, G., Van Raalten, T., and Kahn, R. (2004). Neurophysiological factors in human information processing capacity. *Brain*, 127(3):517–525
5. Fiorillo, C., Tobler, P., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902
6. Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599
7. Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, 31(2):790–795
8. Rn Diedrichsen, J. and Kriegeskorte, N. (2017a). *Representational models : A common framework for understanding encoding, pattern-component, and representational- similarity analysis*. PLOS Computational Biology
9. Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning : A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(7):1297–1310
10. Dayan, P. and Daw, N. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective and Behavioral Neuroscience*, 8(4):429–453
11. Sutton, R. and Barto, A. (1999). Reinforcement learning : An introduction. *Trends in cognitive sciences*, 3(ttps://doi.org/10.1016/s1364-6613(99)01331-5)
12. Averbeck, B., Latham, P., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366
13. Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711
14. Hosoya, T., Baccus, S., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77
15. Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, page 216–234

Critically, this line of work typically also demonstrates the relevance of neural data for understanding inter-individual differences w.r.t. the magnitude of behavioral distortions. For example, it was shown that those people who exhibit a strong optimism bias are those people whose encoding of disappointing prediction errors (in the right frontal gyrus) was the weakest [20]. On the other hand, one may disclose non-trivial behavioral consequences of the computational properties of neural information processing. For example, it was shown that the brain's reliance on efficient coding induced systematic biases in both perceptual and value-based decisions [21 22 23 24 25]. The irony here is that efficient coding is the brain's optimal solution to the problem of building reliable cognitive representations under limited neural resources [26 27]. In brief, this series of work provides evidence for the impact of biological constraints on behaviorally-relevant information processing.

One critical insight here was that efficient coding induces plastic changes in the brain's decision system that was incidental, i.e., they were not instrumental to the decision task [28]. More precisely, the encoding of value in OFC neurons was shown to obey a ubiquitous, hard-wired, biological constraint, namely : range adaptation [29 30 31 32 33]. Range adaptation is the mechanism by which neurons dynamically tune their output firing properties to match the changing statistics of their inputs, hence implementing efficient coding under the constraint of bounded neural activation range [34 35 36]. Although a major breakthrough in decision neuroscience, these studies suffer from two methodological weaknesses. First, they rely on a normative reference model that describes how the brain should process behaviorally-relevant information, whose computational properties are altered by range adaptation. In turn, neuroscientific evidence for range adaptation is mostly indirect because it relies on validating its corollary consequence in terms of value distortions (e.g., divisive normalization), rather than identifying its neural signature (but see (Zimmermann et al., 2018) [37]). Second, other alternative computational mechanisms that may make qualitatively similar predictions are ignored. In particular, one may argue that many forms of plasticity may, in principle, induce dynamic changes in the brain's decision circuits that may eventually be confounded with range adaptation. A ubiquitous and ever-persistent example of this is Hebbian synaptic plasticity [38], which is central to, e.g., development and recovery from injury [39 40 41]. A plethora of electrophysiological studies have established its many variants, including, but not limited to, spike-timing dependent plasticity and long-term potentiation/depression [42 43 44 45].

Critically, Hebbian plasticity does not reduce to range adaptation, and one may reasonably ask which of these two hard-wired mechanisms is the most constraining for behaviorally-relevant neural information processing.

This work is a first step towards solving the two above issues. In brief, we propose a computational method for identifying (and comparing) the neural signature of biological mechanisms or constraints on behaviorally-relevant neural information processing. We bypass the issue of defining a normative reference model for neural information processing by fitting ANNs (i.e., artificial neural network models) to behavioral data, with and without incidental, hard-wired, constraints. Here, we consider two types of hard-wired biological mechanisms : namely, range adaptation and Hebbian plasticity. We then evaluate the evidence for or against biologically-constrained ANNs using a variant of RSA (i.e., representational similarity analysis), because it exploits detailed multivariate information in the data while being robust to nuisance model misspecifications [46 47 48]. We apply the approach to two different open fMRI datasets acquired when people make decisions under risk [49]. In what follows, we describe our methodological approach and evaluate its statistical properties with numerical Monte-Carlo simulations. We then report the results of the ensuing analysis of concurrent behavior and fMRI data. Finally, we discuss our results in light of the existing literature and highlight potential weaknesses and perspectives.


16. Lewicki, M. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363
17. De Martino, B., Camerer, C., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792
18. De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. (2006). Frames, biases and rational decision-making in the human brain. *Science*, 313(5787):684–687
19. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518
20. Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23):941– 945
21. Louie, K. and Glimcher, P. (2012). Efficient coding and the neural representation of value. *Annals of the New York Academy of Sciences*, 1251(1):13–32
22. Polania, R., Woodford, M., and Ruff, C. (2018). Efficient coding of subjective value. *BioRxiv*, 358317
23. Soltani, A., Martino, B., and Camerer, C. (2012). A range-normalization model of context-dependent choice : A new model and evidence. *PLoS Computational Biology*, 8(7)
24. Wei, X. and Stocker, A. (2015). A bayesian observer model constrained by efficient coding can explain "anti-bayesian" percepts. *Nature Neuroscience*, 18(10):1509–1517
25. Zimmermann, J., Glimcher, P., and Louie, K. (2018). Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nature Communications*, 9(1):1–11
26. Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, page 216–234
27. Simoncelli, E. and Olshausen, B. (2001). Natural image statistic and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216
28. Conen, K. and Padoa-Schioppa, C. (2019). Partial adaptation to the value range in the macaque orbitofrontal cortex. *Journal of Neuroscience*, 39(18):3498–3513
29. Burke, C., Baddeley, M., Tobler, P., and Schultz, W. (2016). Partial adaptation of obtained and observed value signals preserves information about gains and losses. *Journal of Neuroscience*, 36(39):10016–10025
30. Cox, K. and Kable, J. (2014). Bold subjective value signals exhibit robust range adaptation. *Journal of Neuroscience*, 34(49):16533–16543

## 9.2 METHODS

### 9.2.1 *Biologically-constrained artificial neural networks for behavioral data*

Artificial Neural Networks or ANNs provide essentially attempt to decompose a possibly complex form of information processing in terms of a combination of very simple computations performed by connected 'units', which are a mathematical abstraction of neurons. Here, we take inspiration from a growing number of studies that use ANNs as descriptive models of neural information processing, whose relative biological realism is to be gauged with neuroimaging data [50 51 52].

We consider behavioral paradigms akin to decision tasks, whereby subjects need to process some (experimentally controlled) behaviourally-relevant information $u = \{u^{(1}, u^{(2)}, \dots, u^{(n_u)}\}$ to provide a response $r$. In what follows, we will focus on a value-based decision-making task, whereby participants have to accept or reject a risky gamble composed of a 50% chance of winning a gain G and a 50% chance of losing L, i.e., $u$ is composed of $n_u = 2$ input features : $u = \{G, L\}$. In brief, we assume that people's behavioral response $y$ is the output of a neural network that processes the input, i.e. : $r \approx g_{ANN}(u, \theta)$, where $\theta$ are unknown ANN parameters and $g_{ANN}(\cdot)$ is the ANN's input-output transformation function. So-called "shallow" ANNs effectively reduce $g_{ANN}(\cdot)$ to a combination of neural units organized in a single hidden layer. Here, we rather rely on ANNs with two hidden layers. As will be more apparent below, this will facilitate the introduction of Hebbian plasticity mechanisms/constraints.

We assume that each input feature $u_t^{(i)}$ is encoded into the activity of neurons $[x_t^{(i,1)}, x_t^{(i,2)}, \dots, x_t^{(i,j)}, \dots, x_t^{(i,n_x)}]$ of its dedicated "input layer", where $n_x$ is the number of input neurons per input. What we mean here is that the neuron $j$ in the input layer $i$ responds to $u_t^{(i)}$ as follows :

$$x_t^{(i,j)} = f_1(u_t^{(i)}, \theta^{(i,j)}) \tag{9.1}$$

where $f_1(\cdot)$ is the activation function of neural units that compose the ANN's input layer. Collectively, the activity vector $[x_t^{(i,j)}]_{j=1,\dots,n_x}$ forms a representation of the input $u_t^{(i)}$ in the form of a population code.

Critically, we consider activation functions that are bounded, i.e., either a sigmoid or a pseudo-gaussian mapping of inputs (see below) :

$$f(u, \theta) = \begin{cases} f_{Gauss}(u, \theta) \triangleq exp\left(-\frac{(u-\mu)^2}{\sigma^2}\right) \\ \text{or} \\ f_{sigmoid}(u, \theta) \triangleq \frac{1}{1+exp(\gamma\frac{\mu-u}{\sigma})} \end{cases} \tag{9.2}$$

where $\gamma \approx 1.5434$ is a scaling constant that we introduce for mathematical convenience (see Appendix 9.A). The parameters $\theta^{(i,j)} = \{\mu^{(i,j)}, \sigma^{(i,j)}\}$ capture the idiosyncratic properties of the neuron $j$ in the input layer $i$ (e.g., its firing rate threshold $\mu^{(i,j)}$ and the pseudo-variance parameter $\sigma^{(i,j)}$). Note that, when inputs $u$ fall too far away


31. Elliott, R., Agnew, Z., and Deakin, J. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, 27(9):2213–2218
32. Kobayashi, S., De Carvalho, O., and Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience*, 30(2):534–544
33. Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014
34. Brenner, N., Bialek, W., and De Ruyter Van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702
35. Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capactiy
36. Wark, B., Lundstrom, B., and Fairhall, A. (2008). Sensory adaptation. *Physiology*, 17(4):423–429
37. Zimmermann, J., Glimcher, P., and Louie, K. (2018). Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nature Communications*, 9(1):1–11
38. Hebb, D. (1949). *The organization of behavior : a neuropsychological theory*. J. Wiley ; Chapman Hall
39. Fox, K. and Stryker, M. (2017). Integrating hebbian and homeostatic plasticity : Introduction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715)
40. Martens, M., Celikel, T., and Tiesinga, P. (2015). A developmental switch for hebbian plasticity. *PLoS Computational Biology*, 11(7):1–19
41. Turrigiano, G. (2017). The dialectic of hebb and homeostasis. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715):4–6
42. Fox, K. and Stryker, M. (2017). Integrating hebbian and homeostatic plasticity : Introduction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715)
43. Lisman, J. (2017). Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes : Long-term potentiation, long-term depression, short-term potentiation and scaling. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715)
44. Shouval, H., Wang, S., and Wittenberg, G. (2010). Spike timing dependent plasticity : A consequence of more fundamental learning rules. *Frontiers in Computational Neuroscience*, 4(July):1–13

from $\mu$ (say outside a $\pm 2\sqrt{2}\sigma$ range), both these activation functions saturate, i.e., they produce non-discriminable outputs (close to 0 or 1). In other words, the pseudo-variance parameter defines the range of inputs over which units incur no information loss. As we will see below, range adaptation effectively tunes these activation functions to minimize information loss.

Then the output of the input layers is passed to the "integration layer" $[z_t^{(1)}, z_t^{(2)}, \dots, z_t^{(k)}, \dots, z_t^{(n_z)}]$ , i.e., the neuron $k$ of the integration layer responds to $[x_t^{(i,j)}]_{j=1,\dots,n_x}^{i=1,\dots,n_u}$ as follows :

$$z_t^{(k)} = f_2\left(\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)} x_t^{(i,j)}, \Phi^{(k)}\right) \tag{9.3}$$

where $C^{(i,j,k)}$ is the connection weight from the neuron $j$ in the input layer $i$ to the neuron $k$ of the integration layer, and $\Phi^{(k)}$ capture idiosyncratic properties of the integration neuron $k$. For simplicity, we restrain our analysis to $n_z = n_x$.

The behavioral response $r_t$ at time or trial $t$ is then read out from the integration layer as follows :

$$r_t \approx f_{sigmoid}\left(\sum_{k=1}^{n} {}_z W^{(k)} z_t^{(k)}, \nu)\right) \tag{9.4}$$

where the $W^(k)$ can be thought of as connection weights to another system that would implement the decision into an action (.e.g., the motor system). Taken together, Equation 9.1-9.2-9.3 define the ANN's input-output transformation function, when no further biological constraint is introduced (see below) :

$$g_{ANN}^{(0)}(u,\vartheta) \triangleq f_{sigmoid}\left(\sum_{k=1}^{n_z} W^{(k)} f_2\left(\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{i,j,k} f_1(u^{(i)}, \theta^{(i,j)}), \Phi^{(k)}\right), \nu\right) \tag{9.5}$$

where $\vartheta$ lumps all ANN parameters together, i.e. $\vartheta \triangleq \{W, C, \theta, \Phi, \nu\}$, and $f_{i\in\{1,2\}}$ are either gaussian or sigmoid. A schematic summary of the ANN's double-layer structure is shown in Figure 9.1 below.

Although, strictly speaking, this ANN includes one form of biological constraint (cf. bounded units' activation functions), we will refer to it as the 'default' or 'non-constrained' ANN. Note that, provided there are enough neurons in input and integration layers, this ANN architecture can capture any value function defined on the multidimensional input space. However, it cannot capture behavioral hysteresis effects, whereby previous decisions may change the network's response to behaviorally-relevant information. This is why we now introduce range adaptation and Hebbian plasticity.

Recall that range adaptation is a mechanism by which neurons maximize the contrast of their output activity over the natural range


45. Zenke, F. and Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715)

46. Diedrichsen, J., Berlot, E., Mur, M., Schütt, H., and Kriegeskorte, N. (2020). Comparing representational geometries using the unbiased distance correlation. Retrieved from

47. Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November):1–28

48. Rn Diedrichsen, J. and Kriegeskorte, N. (2017a). *Representational models : A common framework for understanding encoding, pattern-component, and representational- similarity analysis*. PLOS Computational Biology

50. Güçlü, U. and Gerven, M. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014

51. Kietzmann, T., McClure, P., and Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *BioRxiv*, 133504

52. Kriegeskorte, N. and Golan, T. (2019). Neural network models and deep learning. *Current Biology*, 29(7):231– 236

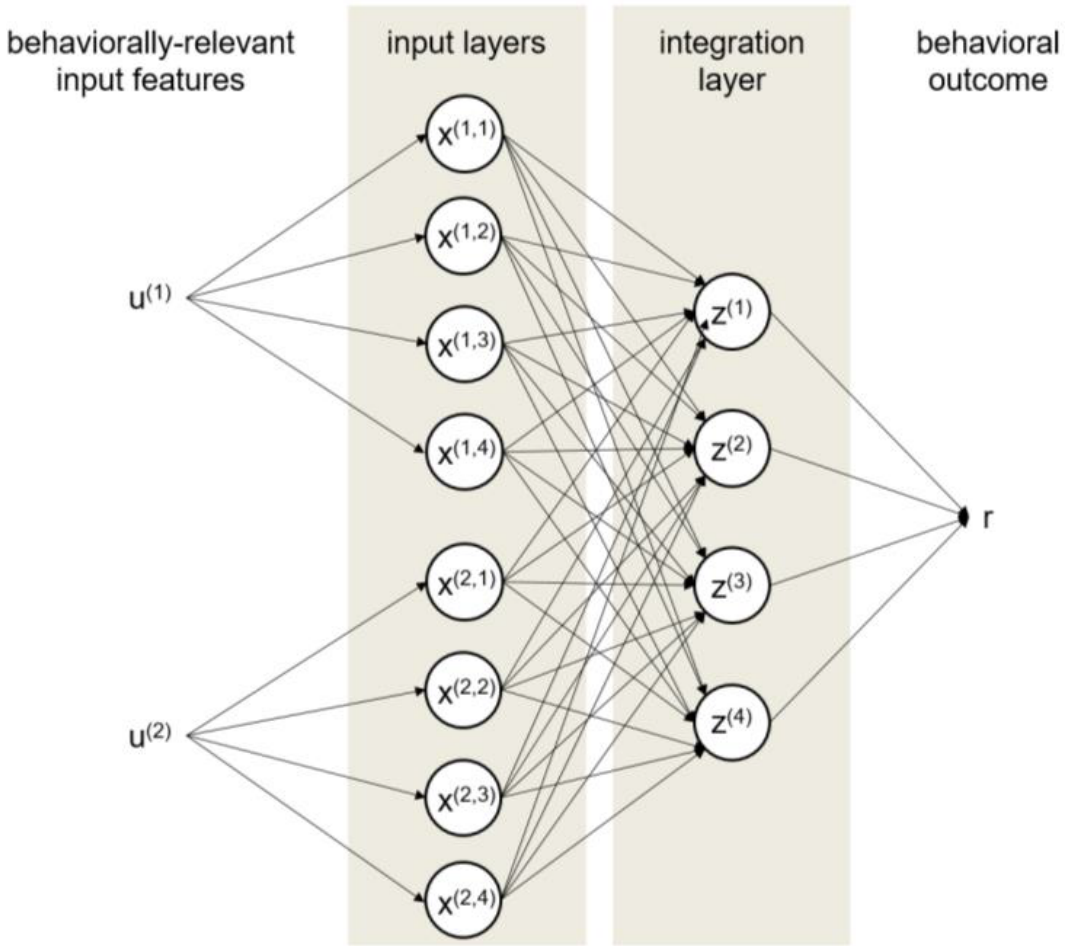



FIGURE 9.1 – *Structure of the 'default' artificial neural network. Behaviorally-relevant input features first enter the 'input' layer, which then sends its multiple outputs to the 'integration' layer. Finally, a behavioral response is produced from the multiple outputs of the 'integration' layer. See the main text for mathematical notations.*

of their inputs. Given that we used sigmoid or pseudo-gaussian activation functions (cf. Equation 9.2), range adaptation adaption then reduces to a learning rule on $f_2$'s pseudo-variance parameters, which are now time-dependent variables and seek to maximize the transmitted information, i.e., the discriminability of the outputs (see Appendix 1 for details) :

$$\sigma_{t+1}^{(k)} = \sigma_t^{(k)} + \alpha_{RA}\left(\left|\mu^{(k)} - \sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)} x_t^{(i,j)}\right| - \sigma_t^{(k)}\right) \tag{9.6}$$

where $\alpha_{RA}$ is the learning rate of range adaptation. Equation 9.6 effectively matches pseudo-variance parameters $\sigma$ with the variability of the recent history of each units' inputs. In turn, units' activation functions are sampled over a range where their output activity does not saturate. Now the two-layers structure of the ANN also enables explicit modeling of Hebbian plasticity. More precisely, the Hebbian adaption rule will strengthen the connection between input and integration units that co-vary. This recapitulates the "fire together, wire together" rule :

$$\begin{aligned} C_t^{(i,j,k)} &= c^{(i,j,k)} f_{sigmoid}(\kappa_t^{(i,j,k)}) \\ \kappa_{t+1}^{(i,j,k)} &= \kappa_t^{(i,j,k)} + \alpha_H(x_t^{(i,j)} z_t^{(k)} - \lambda_H) \end{aligned} \tag{9.7}$$

where $c^{(i,j,k)}$ and $\kappa_t^{(i,j,k)}$ are the static and dynamic components of between-layers connection weights, respectively, $\alpha_H$ is the Hebbian learning rate and $\lambda_H$ is covariance threshold. Equation 9.7 reinforces a connection weight whenever the product of the corresponding units' outputs exceeds the threshold $\lambda_H$.

At the limit when learning rates tend to zero ($\alpha_{RA} \to 0$ or $\alpha_H \to 0$), the constrained ANNs exhibit no plastic change, i.e., they become indistinguishable from the above 'default' ANN. Otherwise, both range adaptation and Hebbian plasticity constraints make the ANN's trial-by-trial response a function of the recent history of inputs to the network. In both cases, learning rates effectively control the amount of plastic changes that modified ANNs will exhibit. Importantly, behavioral distortions and/or neural activity patterns that will be induced with these two types of plastic changes may be different. In other terms, Hebbian plasticity and range adaptation are unlikely to capture similar forms of behavioral and/or neural hysteresis effects. We will comment on the computational properties of Equations 6 and 7 in the Discussion section. Importantly, no normative model exists that can be used as a reference point to set the amount of plastic change that the decision network should exhibit. But one can use observed peoples' behavioral responses to evaluate how much plastic changes the decision network actually does exhibit. Here, we rely on established variational Bayesian model inversion techniques to perform probabilistic parameter estimation [53] [54]. To mitigate the impact of local optima, we use a twofold strategy. First, we concurrently fit the behavior trial series together with its rolling mean and variance (over a sliding temporal window whose width

53. Daunizeau, J. (2017). The variational laplace approach to approximate bayesian inference

54. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the laplace approximation. *NeuroImage*, 34(1):220–234

we set to 5 trials). Second, we use a hierarchical group-level mixed-effects approach that constrains subject-specific parameter estimates with estimated group statistics [55]. The priors on the ANNs' model parameters for the ensuing parametric 'empirical Bayes' approach are summarized in Table 19.1 below.

55. Daunizeau, J. (2019). Variational bayesian modelling of mixed-effects. Icm). Retrieved from

| Parameter | Distributions | Rational |
|---|---|---|
| Pseudo-gaussian mean/ Sigmoid center | $\mu^{(i,j)} \sim \mathcal{N}\left(\frac{1}{n_x+1}, \frac{1}{n_x+1}\right)$ | Homogenous paving of inputs |
| Pseudo-gaussian initial standard deviation | $\sigma_0^{(i,j)} = \lvert\theta\rvert$ with $\theta \sim \mathcal{N}\left(\frac{0.5}{n_x+1}, \frac{0.5}{n_x+1}\right)$ | Overlapping pseudo-gaussian |
| Initial connection weights | $c^{(i,j,k)} \sim \mathcal{N}\left(\frac{1}{n_x}, \frac{1}{n_x}\right)$ | Inputs averaging |
| Range adaptation learning rate | $\alpha_{RA} = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-3, 2)$ | Gradual, stable learning |
| Hebbian-plasticity learning rate | $\alpha_H = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-3, 2)$ | Gradual, stable learning |
| Hebbian plasticity threshold | $\lambda_H = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-1, 1)$ | Comparable to the average product of two bounded units |
| Hebbian initial strength | $\kappa_0^{(i,j,k)} = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(0, 0.5)$ | The middle point between full and null strength |

TABLE 9.1 – *Parameters' priors for biologically-constrained ANNs.*

Note that all our behavioural analyses are performed using the VBA academic freeware [56].

56. Daunizeau, J., Adam, V., and Rigoux, L. (2014). Vba : A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, 10(1)

### 9.2.2 *Assessing the neural signature of candidate biological constraints using RSA*

From a statistical perspective, Equations 6 and 7 provide extra degrees of freedom when fitting the modified ANN to behavioural choices, when compared to the 'no-constraint' ANN. This means that one would expect behavior to be better explained with range adaptation and/or Hebbian constraints, irrespective of whether these constraints are realistic determinants of behavior or not. This is why it is critical to cross-validate behavioral analyses with neural data. This can be done because once fitted to behavioral data our modified ANN models make specific trial-by-trial predictions of neural activity patterns $\{x_t, z_t\}$ that can be compared to multivariate neural signals. Here, we have chosen to rely on a modified representational similarity analysis [57], which possesses the following properties :

1. It is simple (at least from a statistical standpoint).
2. It is robust to assumptions regarding the relationship between modeled and empirical neural time series. In particular, it is not confounded by nonlinearities and/or by dimensionality differences. These, in fact, are known virtues of RSA (Flandin Friston, 2019 ; Rn Diedrichsen Kriegeskorte, 2017) [58] [59].
3. It extracts multivariate information from empirical neural signals that is orthogonal to linear combinations of behaviorally-relevant inputs and behavioral responses. This is necessary (i) to provide

57. Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November):1–28

58. Friston, K. J., Diedrichsen, J., Holmes, E., and Zeidman, P. (2019). Variational representational similarity analysis. *NeuroImage*, 201:115986

59. Rn Diedrichsen, J. and Kriegeskorte, N. (2017b). *Representational models : A common framework for understanding encoding, pattern-component, and representational- similarity analysis*. PLOS Computational Biology

analysis results that are orthogonal to previous mass-univariate analyses, and (ii) to prevent statistical biases towards models that best explain behavioral data.

In brief, RSA consists of evaluating the statistical resemblance between model-based and data-based 'representational dissimilarity matrices' or RDMs, which we derive as follows. Let $Y$ be the $n_Y \times n_t$ multivariate time series of (modeled or empirical) neural activity, where $n_Y$ and $n_t$ are the number of units and trials, respectively. Note that, for model-based RDMs, 'units' mean artificial elementary units in ANNs, whereas for data-based RDMs, 'units' mean either neurons (cf. electrophysiology) or voxels (fMRI). First, we orthogonalize $Y$ with respect to potential confounding sources of between-trial variability, i.e. : $Y \leftarrow Y(I_{n_Y} - X^T(XX^T)^{-1})X)$, where $X$ is the $n_c \times n_t$ confounds matrix. Here, the set of confounds typically include a constant term, behaviorally-relevant inputs $u$, and behavioral responses $r$. Second, we standardize neural time series by zscoring over trials. Now let $D_Y$ be the ensuing $n_t \times n_t$ between-trials Euclidean distance matrix :

$$D_Y = \begin{bmatrix} 0 & D_Y^{2,1} & ... & D_Y^{1,T} \\ D_Y^{2,1} & 0 & & D_Y^{2,T} \\ \vdots & & \ddots & \vdots \\ D_Y T,1 & D_Y^{T,2} & ... & 0 \end{bmatrix} \text{ with } D_Y^{t,t'} = \sqrt{\sum_{i=1}^{n_Y} (Y_t^{(i)} - Y_{t'}^{(i)})^2} \tag{9.8}$$

The matrix element $D_Y^{t,t'}$ thus measures the dissimilarity of neural patterns of activity between trial $t$ and trial $t'$, having removed trial-by-trial variations that can be explained as linear combinations of behaviorally-relevant inputs and behavioral responses. We define the ensuing RDM as the lower-left triangular part of $D_Y$.

In what follows, model-based RDMs are derived using the integration layer of our modified ANNs (i.e. $Y_{ANN} = [z_1, z_2, ..., z_{n_z}]^T$), after having fitted the corresponding model parameters to behavioral responses. Data-based RDMs are derived from the fMRI time series. Here, $Y_{fMRI}$ is obtained by deconvolving BOLD time series from the hemodynamic response function with a Dirac delta or stick basis function set that is time-locked to trial events[60]. RSA then proceeds with the statistical comparison of $D_{Y,ANN}$ and $D_{Y,fMRI}$. In line with recent methodological developments of RSA, we first bin RDMs into 20 quantiles and then compute the Pearson correlation $\rho = corr(RDM_{ANN}, RDM_{fMRI})$ between the binned RDMs. Group-level statistical significance of RDMs' correlations can be assessed using one-sample t-tests on the group mean of Fischer-transformed RDM correlation coefficients $\rho$ (see below). Figure 9.2 below recapitulates the ensuing ANN-RSA approach.

60. Dale, A. (1999). Optimal experimental design for event-related fmri. *Human Brain Mapping*, 8(23):109–114

Note that our ANN-RSA approach does not a priori favor more complex ANNs (i.e., ANNs with more parameters). When fitted to behavioral data, more complex ANNs (i.e., those that include range adaptation or Hebbian plasticity) are expected to yield greater explanatory power. The RDM correlation $\rho$ exhibits no such bias, however. This is because, once fitted to behavioural data, estimated

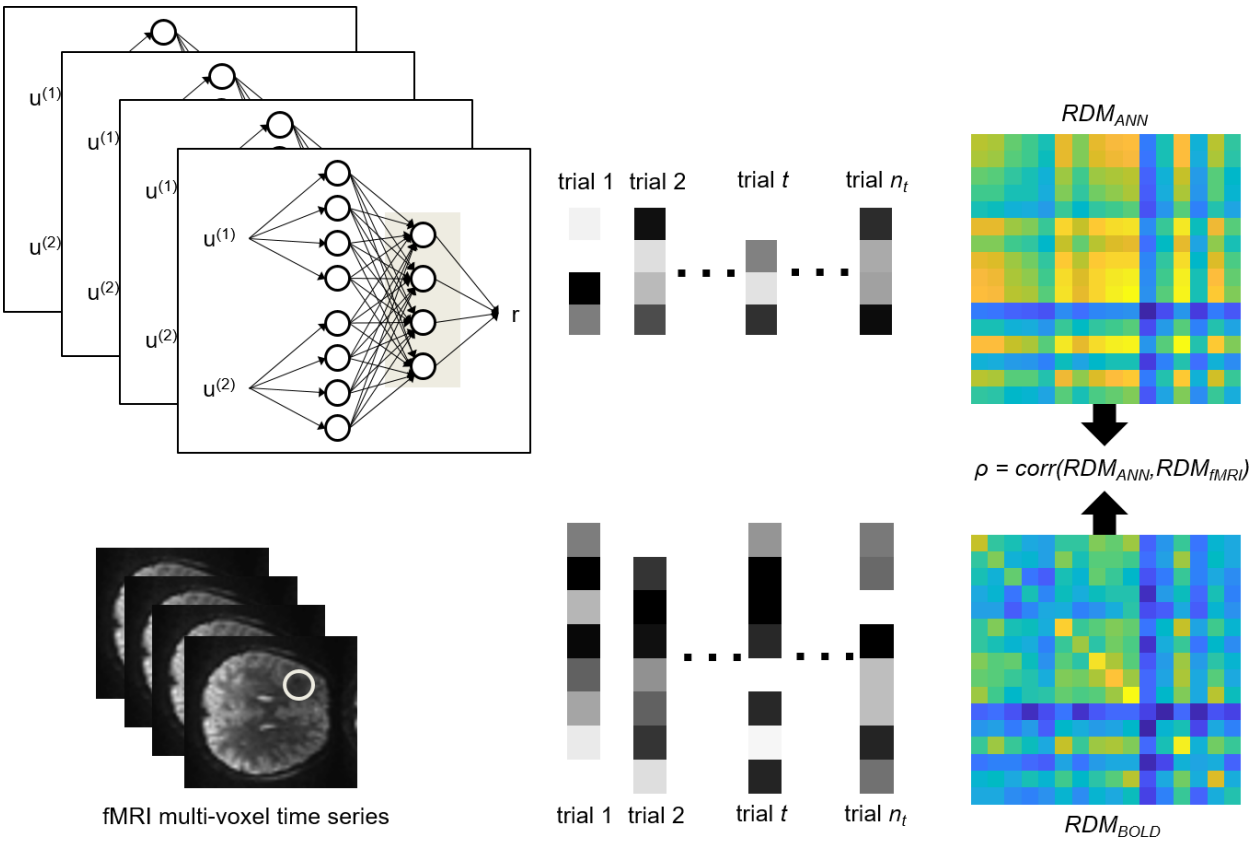



FIGURE 9.2 – *Summary data-analysis pipeline of the ANN-RSA approach. First, trial-by-trial profiles of the ANN's response to behaviourally-relevant inputs (in the integration layer) are estimated. Second, corresponding trial-by-trial multivariate patterns of fMRI activity are extracted in each ROI of interest. Third, corresponding model-based and fMRI-based RDM are derived, whose correlation ρ serves as the RSA summary statistics (which then enters subsequent statistical significance testing).*

ANN activity patterns and their ensuing RDMs have no degree of freedom whatsoever. In particular, this means that default (non-constrained) ANNs may show a greater RDM correlation than ANNs that include range adaptation or Hebbian plasticity. In turn, this enables a simple statistical procedure for comparing candidate models based on group-level comparisons of RDM correlations (see below).

### 9.2.3 *Note on statistical testing and model comparison*

Recall that our model space is factorial, with two orthogonal modeling factors : (i) our factor of interest has three 'levels' : no constraint, range adaptation or Hebbian plasticity, and (ii) our factor of no interest has two 'levels' : sigmoid versus pseudo-gaussian neural activation functions. This means that we will be comparing 2x3=6 models. When assessing the statistical significance of the ensuing model comparison, we will be using a variant of composite null testing. Let $p_m^{m'}$ be the p-value associated with the elementary pairwise comparison of model $m$ and $m'$, whose null hypothesis is $H_0^{(m,m')} : \rho_m \leq \rho_{m'}$, where $\rho_m$ is the corresponding Fisher-transformed RDM correlation ($p_m^{m'}$ can be evaluated using paired t-tests on RDM correlations). For each model $m \in [1, \dots, 6]$, we ask whether its RDM correlation is the highest among the candidate models. This induces the following composite null hypothesis : $H_0^{(m)} : \rho_m \neq \max_{m'} \rho_{m'}$. The maximum

p-value statistics $p_m = \max_{m'} p_m^{m'}$ yields a valid test of the composite null hypothesis, though not necessarily maximally efficient [61]. Because $H_0^{(m)}$ is the conjunction of elementary pairwise null hypotheses $H_0^{(m,m')}$, we refer to this approach as "conjunctive null testing".

One may also want to evaluate the statistical significance of the comparison of RDM correlations across levels of our factor of interest, irrespective of our factor of no interest. The corresponding null hypothesis involves a disjunctive/conjunctive combination of elementary null hypotheses. For example, if one wants to test whether range adaptation has a significantly higher RDM correlation than Hebbian or default (no-constraint) ANNs, the corresponding null hypothesis $H_0^{RA}$ is defined as :

$$H_0^{RA} : \begin{cases} \rho_{RA,Gauss} \neq \max\limits_{m' \notin \{RA,sigmoid\}} \rho_{m'} \\ \text{AND} \\ \rho_{RA,sigmoid} \neq \max\limits_{m' \notin \{RA,Gauss\}} \rho_{m'} \end{cases} \tag{9.9}$$

The following p-value then yields a valid statistical test of $H_0^{(RA)}$ :

$$\widehat{p}_{RA} = 2 \times min \left[ \max_{m' \notin \{RA,sigmoid\}} p_m^{m'}, \max_{m' \notin \{RA,Gauss\}} p_m^{m'} \right] \tag{9.10}$$

By design, the ensuing "disjunctive/conjunctive" approach cannot conclude about the underlying activation functions, i..e. it does not discriminate between sigmoid and pseudo-gaussian functional forms. However, it pools evidence over levels of our factor of no interest, which eventually improves statistical power. This is a frequentist -and simpler- variant of so-called "family inference" in Bayesian model comparison [62], where one marginalizes over modeling factors of no interest, effectively trading statistical power against inference resolution. We will see a direct demonstration of the disjunctive/conjunctive approach below.

61. Wasserman, L. (2004). All of statistics : A concise course in statistical inference brief contents. *Simulation*

62. Penny, W. D., Stephan, K., Daunizeau, J., Rosa, M., Friston, K., Schofield, T., and Leff, A. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3)

### 9.2.4 *fMRI study of risk attitudes : experimental design*

In this work, we compare the neural evidence for candidate biological constraints (range adaptation versus Hebbian plasticity) on behaviorally-relevant neural information processing using a re-analysis of the NARPS dataset [63], openly available on openneuro.org [64]. This dataset includes two studies, each of which is composed of a group of 54 participants who make a series of decisions made of 256 risky gambles. On each trial, a gamble was presented, entailing a 50/50 chance of gaining an amount G of money or losing an amount L. As in (Tom et al., 2007) [65], participants were asked to evaluate whether or not they would like to play each of the gambles presented to them (strongly accept, weakly accept, weakly reject or strongly reject). They were told that, at the end of the experiment, four trials would be selected at random : for those trials in which they had accepted the corresponding gamble, the outcome would be decided with a coin toss, and for the other ones -if any-, the gamble would

63. Botvinik-nezer, R., Holzmeister, F., Camerer, C., and Johannesson, M. (2019). Variability in the analysis of a single neuroimaging dataset by many teams

64. Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., and Milham, M. (2013). Toward open sharing of task-based fmri data : the openfmri project. *Frontiers in Neuroinformatics*, 7(July):12

65. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518

not be played. In the first study (hereafter : "equal range" group), participants decided on gambles made of gain and loss levels that were sampled from the same range (G and L varied between 5 and 20 $). In the second study (hereafter : the "equal indifference" group), gain levels scaled to double the loss levels (L varied between 5 and 20$, and G varied between 10 and 40$). In both studies, all 256 possible combinations of gains and losses were presented across trials, which were separated by 7 seconds on average (min 6, max 10).

MRI scanning was performed on a 3T Siemens Prisma scanner. High-resolution T1w structural images were acquired using a magnetization prepared rapid gradient echo (MPRAGE) pulse sequence with the following parameters : TR = 2530 ms, TE = 2.99 ms, FA = 7, FOV = 224 × 224 mm, resolution = 1 × 1 × 1 mm. Whole-brain fMRI data were acquired using echo-planar imaging with multi-band acceleration factor of 4 and parallel imaging factor (iPAT) of 2, TR = 1000 ms, TE = 30 ms, flip angle = 68 degrees, in-plane resolution of 2X2 mm 30 degrees of the anterior commissure-posterior commissure line to reduce the frontal signal dropout, with a slice thickness of 2 mm and a gap of 0.4 mm between slices to cover the entire brain. See https://www.narps.info/analysis.htmlprotocol for more details. Data preprocessing included standard realignment and movement correction steps. Note that we excluded 5 participants from the 'equal-range' group because the misalignment between functional and anatomical scans could not be corrected. No spatial smoothing was applied.

Previous mass-univariate analyses of these datasets, including a recent study of the analysis variability among multiple research groups [66], provided evidence for the implication of multiple brain systems in response to either gains and/or losses, in particular : the ventromedial prefrontal cortex or vmPFC, the dorsolateral prefrontal cortex or dlPFC, the anterior cingulate cortex or ACC, the posterior cingulate cortex or PCC, the Amygdala, the Striatum and the Insula. Given the anatomo-functional variability of these regions, we opted for a multiple ROI analysis. Using the NeuroQuery website [67], we selected spatial maps based on the following 12 terms : vmPFC, dlPFC, ACC, dACC, PCC, Amygdala, Striatum, and Insula. We also included primary motor and primary visual cortices, which serve as sensory/motor control regions. Then we took the 2000-th strongest voxels, excluded those that belonged to clusters smaller than 200 voxels, smooth the resulting map, filter out white matter overlaps, and kept the 200 strongest voxels of each remaining clusters. This procedure yielded 18 approximately spherical ROIs spanning both hemispheres, which are shown in Figure 9.3 below.

In each ROI, we regressed trial-by-trial activations with SPM through a GLM that included one stick regressor for each trial (at the time of the gamble presentation onset), which was convolved with the canonical HRF. To account for variations in hemodynamic delays, we added the basis function set induced by the HRF temporal derivative [68]. To correct for movement artifacts, we also included the

66. Botvinik-nezer, R., Holzmeister, F., Camerer, C., and Johannesson, M. (2019). Variability in the analysis of a single neuroimaging dataset by many teams

67. Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., and Varoquaux, G. (2020). Neuroquery : comprehensive meta-analysis of human brain mapping. Retrieved from

68. Hopfinger, J., Büchel, C., Holmes, A., and Friston, K. (2000). A study of analysis parameters that influence the sensitivity of event- related fmri analyses. *NeuroImage*, 11(4):326–333

six head movement regressors and their squared values. We then extracted the 256 trial-wise regression coefficients in each voxel of each ROI. Finally, we orthogonalized the resulting fMRI trial series w.r.t. gains, losses, and choices, zscored them and computed the 18 ROI-specific RDMs.

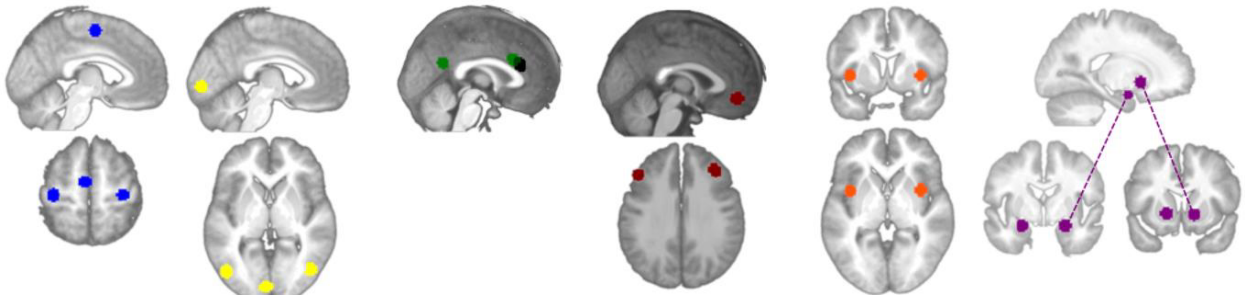


FIGURE 9.3 – *Regions of interest. Control ROIs : Motor left, median, and right (blue), Visual left, median, right (yellow). ROIs of interest : PCC, ACC and dACC (green), vmPFC and dlPFC left and right (red), Insula left and right (orange), Amygdala left and right, and Striatum left and right (purple).*

## 9.3 RESULTS

### 9.3.1 *Assessing expected model confusion using numerical Monte-Carlo simulations*

Prior to presenting our fMRI analyses, we ought to provide evidence that our combined ANN-RSA approach exhibits the statistical robustness that is required for a reliable interpretation of results. In particular, one may ask whether the approach is robust to modeling assumptions regarding (i) the (necessarily underestimated) dimensionality of ANNs that process behaviorally-relevant information, and (ii) the form of units' activation functions (cf. sigmoid versus pseudo-gaussian). More precisely, we ask whether the approach discriminates between the three candidate biological mechanisms of interest (range adaptation, Hebbian plasticity, and 'default'), even when the data are generated with higher-dimensional ANNs. We thus performed a series of Monte-Carlo simulations that recapitulates the design of the fMRI experiment.

We considered a decision task that requires the integration of two inputs $u = \{u^{[1]}, u^{(2)}\}$ that vary randomly across 256 trials. We simulated six series of datasets, corresponding to the 2x3=6 alternative modified ANN models described above. Each dataset was composed of 20 virtual subjects whose trial-by-trial behavior and neural responses were generated under an ANN with sets of either $n_x$=20, 30, or 50 neural units. We allowed for inter-individual variability, derived from sampling ANN parameters under their respective prior probability density functions (cf. Table 9.1). Each simulated dataset was then analyzed using the ANN-RSA approach described above. In brief, each behavioral trial series was fitted with the 2x3 candidate ANNs, and the resulting estimated neural activity profiles were compared to simulated neural activity profiles using our modified RSA. Importantly, fitted ANNs contained smaller sets of $n_x$ =10 units. For each dataset,

we then compared models using conjunctive null testing. We repeat this procedure 50 times and keep track of all positive tests (with a 5% significance level). The upper panel of Figure 9.4 shows the frequency of positive conjunctive testing for all candidate models for each type of simulated data.

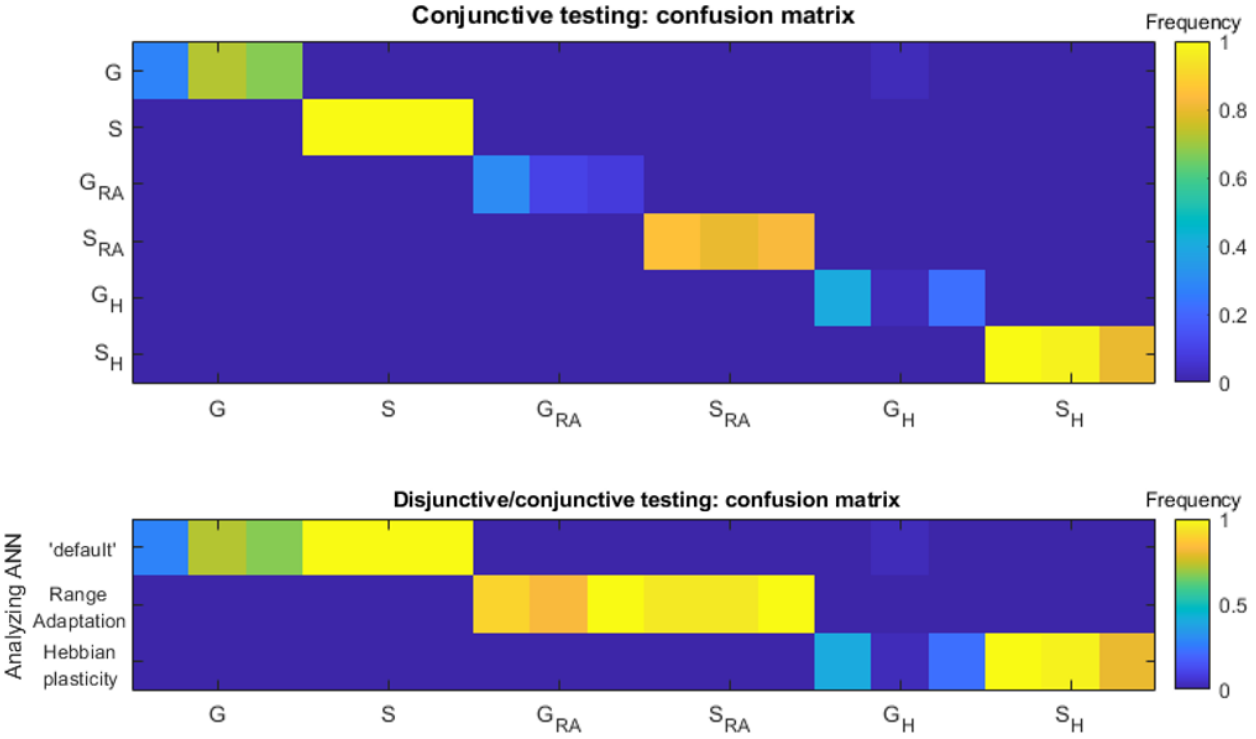



FIGURE 9.4 – *Robustness of the ANN-RSA approach : Monte-Carlo simulations. In what follows, so-called "generative" ANNs were used to simulate data. They can be of 2x3=6 sorts : pseudo-Gaussian/sigmoid 'default' ANNs, pseudo-Gaussian/sigmoid range adaptation ANNs, and pseudo-Gaussian/sigmoid Hebbian ANNs. Each of these sorts of ANNs had three possible dimensions with sets of $n_x$ =20, 30, or 50 units. In contrast, "analyzing" ANNs only included sets of $n_x$ =10 units. Upper panel : confusion matrix of the conjunctive testing approach. The rate at which each "analyzing" ANN (y-axis) exhibits significantly higher RDM correlations than other models, for each "generative" ANNs (x-axis) is color-coded. The three alternative dimensions of "generative" ANNs are presented side to side, from left to right. Lower panel : confusion matrix of the disjunctive/conjunctive approach. Same format, except that the y-axis now shows candidate mechanisms.*

First, note that the conjunctive approach exhibits almost no model confusion. More precisely, the maximum frequency of a model selection error is about 10% (generative ANN = pseudo-gaussian ANN with Hebbian plasticity and 30 units, analyzing ANN= pseudo-gaussian ANN). However, its statistical power is variable (from about 92% ±2% on average for all sigmoid ANNs to about 31% ±25% on average for all pseudo-gaussian ANNs). In other words, the conjunctive testing approach may be too conservative in detecting the correct ANN. Second, the dimensionality of generative ANNs seems to have almost no impact on statistical power. In other words, the relatively small dimensionality of analyzing ANNs (when compared to generative ANNs) does not seem to impair the method's ability to detect the correct underlying mechanism.

Now the lower panel of Figure 9.4 shows the frequency of positive disjunctive/conjunctive testing for the three types of biological mechanisms (no constraint, range adaptation, or Hebbian plasticity) for each type of simulated data. One can see that model confusion is

similar to the conjunctive approach above. However, statistical power is much improved, in particular for detecting range adaptation (94% ±7% on average). Here again, the dimensionality of generative ANNs seems to have no impact on statistical power.

In conclusion, the ANN-RSA approach is robust to violations of modeling and statistical assumptions, including the low dimensionality of analyzing ANNs or the distribution of test statistics. In particular, this implies that, if a candidate mechanism eventually reaches statistical significance using the disjunctive/conjunctive approach, then we can safely infer that it is a more likely explanation of fMRI activity patterns than other candidate mechanisms.

### 9.3.2 *Behavioural analyses*

Each participant's choice sequence data were fitted with the six candidates ANNs, as well as with a simple logistic model. We used sets of $n_x = 4$ units and normalized the gain and loss levels by their averaged sum before feeding them to the input layer. The latter logistic model is the typical agnostic modeling choice in decision paradigms of this kind and was used to measure loss aversion in a previous study relying on the same behavioral design [69]. Here, it will serve as a reference model for evaluating the predictive power of ANNs. Each group was fitted independently through the VBA empirical Bayes procedure. All summary statistics of these behavioural analyses are provided in Tables 1 and 2 of the Appendix. Figure 9.5 below summarizes the fit accuracy of the seven models for the 'equal range' group.

69. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518

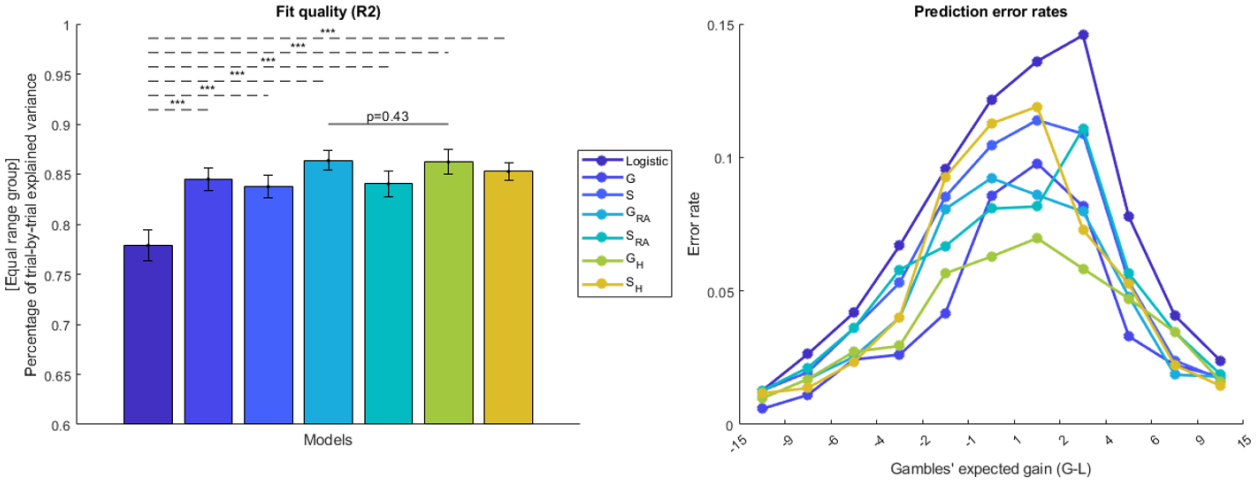



FIGURE 9.5 – *Behavioral results : 'equal range' group. Left panel : mean percentage of variance explained in trial-by-trial choices ± one standard error of the mean (y-axis) is shown for each candidate model (x-axis : from left to right : logistic reference model, pseudo-Gaussian default ANN, sigmoid default ANN, pseudo-Gaussian range adaptation ANN, sigmoid range adaptation ANN, pseudo-Gaussian Hebbian ANN and sigmoid Hebbian ANN). Right panel : The average rate of prediction error (y-axis) is plotted as a function of gambles' expected gain (i.e., G-L, x-axis) for each candidate model (same color code as left panel). Note that the indifference point (maximal prediction error) seems to be biased towards positive expected gains.*

First, one can see that all candidate ANNs perform much better than the simple (reference) logistic model. In fact, they all exhibit a significantly higher percentage of explained variance (all $p < 10^{-5}$).

It turns out that most of the fit improvement lies around the indifference point, where gains and losses balance out (cf. right panel of Figure 9.4). Around that point (i.e., within the [-1,4] interval of expected utility), the logistic reference model necessarily makes unreliable predictions and yields an average error rate of about 12.2% to 14.6%. In comparison, ANNs seem to be able to reduce the apparent randomness in participants' choices, even around the indifference point. This is clearly the case for the model that achieves the lowest average error rate (about 6.3% to 7.0%) : namely : the 'pseudo-gaussian Hebbian' ANN. A likely explanation here is that Hebbian plasticity may effectively change, in a deterministic but nonlinear manner, the network response to repetitions of -otherwise indifferent- gambles. In turn, seemingly random choices may be, at least partially, predicted from the history of past network inputs. This may be taken as evidence against the range adaptation mechanism, which exploits qualitatively similar history-dependent effects to find predictors of peoples' choices around the indifference point. However, it is difficult to conclude from behavioral data alone, because there is no strong statistical evidence that the 'pseudo-gaussian Hebbian' ANN has better explanatory power than the 'pseudo-gaussian range adaptation' ANN, which is the next best model in terms of behavioral fit accuracy (average R2 difference = 0.1% ± 5.6%, p=0.43). Figure 9.6 below presents the results of the same analysis for the 'equal indifference' group.

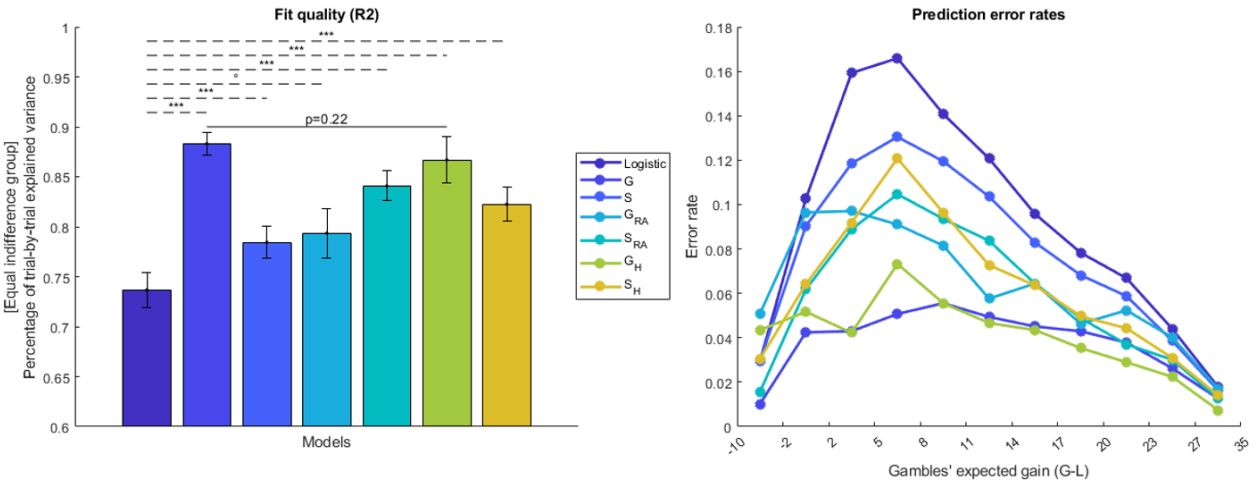



FIGURE 9.6 – *Behavioral results : 'equal indifference' group. Same format as Figure 9.5*

In brief, the same observations can be made, i.e., the behavioral analysis replicates on this second study. In particular, here again, the 'pseudo-gaussian Hebbian' ANN achieves an average error rate of about 6.4% to 9.2% around the indifference point but shows no significant difference in explanatory power with the next best model (average R2 difference = 1.6% ± 15.1%, p=0.22).

At this stage, one would conclude that although biologically-constrained ANNs seem to provide clear improvements over simple statistical behavioural models, behavioral data alone does not clearly

discriminate between candidate underlying biological mechanisms/constraints.

### 9.3.3 *FMRI analyses*

We now aim at identifying the neural signature of candidate biological mechanisms/constraints that may determine people's choice sequences. To begin with, we simply ask whether any candidate model actually explain multivariate fMRI time series in any ROI that we included in our analysis. Figure 9.7 below summarizes the ANN-RSA analysis, in terms of the group-average RDM correlations $\rho$ for each pair of candidate model and ROI ('equal range' group). Table 3 in the Appendix provides the ensuing p-value of RDM correlations' group-level statistical significance ($H_0 : \rho \leq 0$ , one-sided t-test). Note that instead of using units activity, we computed the RDM of the logistic model from the gain and loss levels weighted by the regression coefficients, and orthogonalized from the subject's choices only.

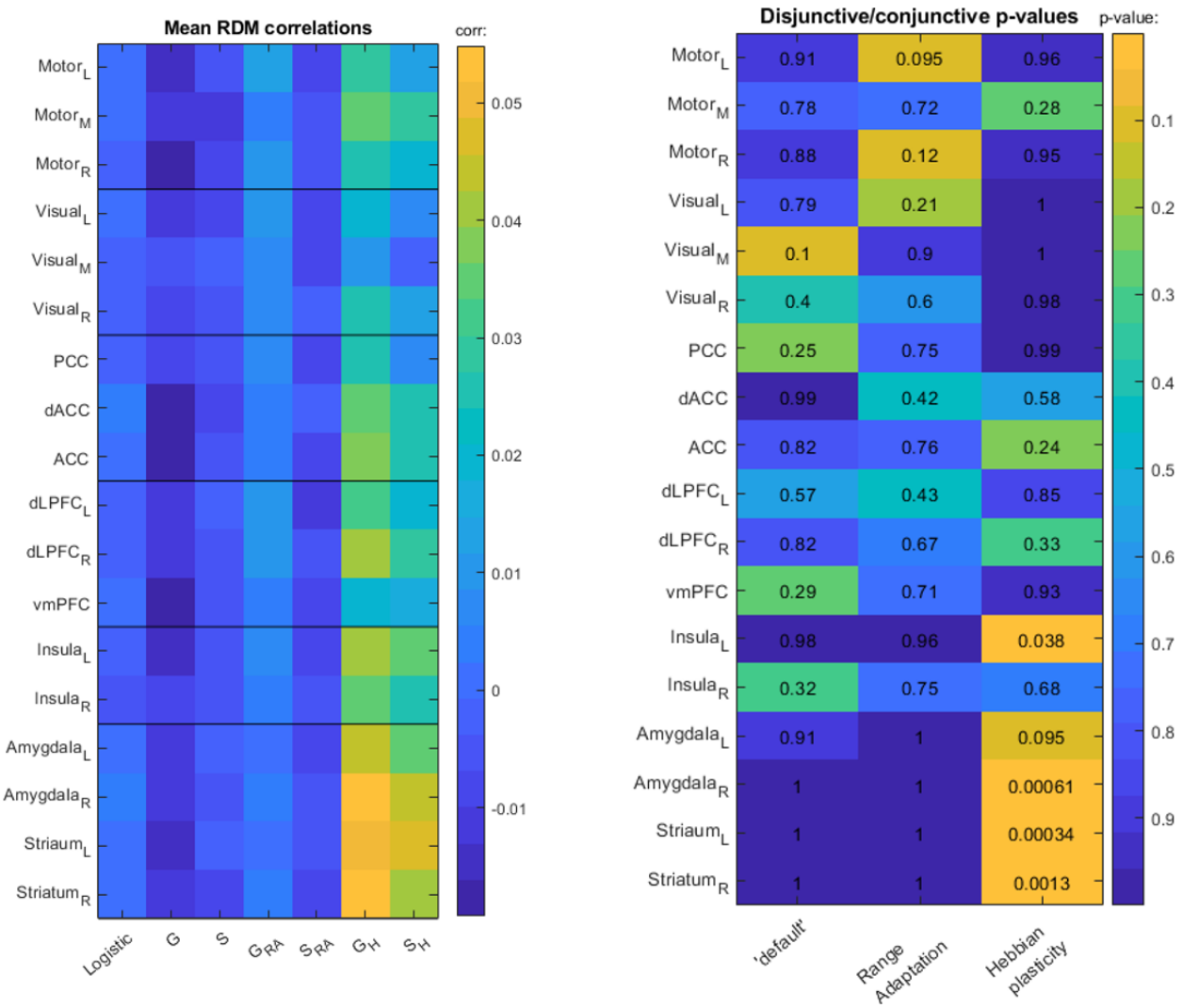



FIGURE 9.7 – *FMRI results : 'equal range' group. Left : group means RDM correlations are shown for each candidate model (x-axis, from left to right : logistic reference model, pseudo-Gaussian default ANN, sigmoid default ANN, pseudo-Gaussian range adaptation ANN, sigmoid range adaptation ANN, pseudo-Gaussian Hebbian ANN, and sigmoid Hebbian ANN) and each ROI (y-axis, from top to bottom : left motor, medial motor, right motor, left visual, medial visual, right visual, PCC, dorsal ACC, ACC, left DLPFC, right DLPFC, vmPFC, left Insula, right Insula, left Amygdala, right Amygdala, left ventral Striatum, right ventral Striatum). Right : group-level p-values of the disjunctive/conjunctive approach to comparing candidate mechanisms are shown for each mechanism (x-axis, for left to right : 'default', range adaptation, and Hebbian plasticity) and each ROI (y-axis, same order as left panel).*

One can see that non-Hebbian models exhibit very small RDM correlations when compared to Hebbian models. Also, the RDM correlations of all models (including Hebbian models) are very weak in control (visual and motor) ROIs. More precisely, no model reaches statistical significance in control regions when correcting for multiple comparisons (all $p > 0.0008$, Bonferroni-corrected threshold$= 0.00046$). In fact, only RDM correlations of Hebbian ANNs reach statistical significance, and only in right DLPFC (pseudo-Gaussian : $p = 0.0004$, sigmoid : $p = 0.0004$), left insula (pseudo-Gaussian : $p = 0.0004$, sigmoid : $p < 10^{-4}$), left amygdala (pseudo-Gaussian trend : $p = 0.0006$, sigmoid : $p = 0.0001$), right amygdala (pseudo-Gaussian : $p < 10^{-4}$, sigmoid : $p < 10^{-4}$), left striatum (pseudo-Gaussian : $p < 10^{-4}$, sigmoid : $p < 10^{-4}$) and right striatum

(pseudo-Gaussian : $p = 0.0001$, sigmoid : $p = 0.0002$).

We then compared Hebbian plasticity to other biological mechanisms of interest using disjunctive/conjunctive testing, whose ensuing p-values are shown in Figure 9.7(right panel Bonferroni-corrected threshold= 0.0028). We found that the comparison of RDM correlations reached statistical significance in bilateral Striatum (left Striatum : $p = 0.0003$, right Striatum : $p = 0.001$) and in the right Amygdala ($p = 0.0006$). In control ROIs, no comparison of RDM correlations achieves statistical significance (all $p > 0.1$, uncorrected). Furthermore, the RDM correlations of range adaptation are never significantly higher than those of other models (all $p > 0.095$, uncorrected). Figure 9.8 below summarizes the results of the same analysis for the 'equal indifference' group (Table 4 in the Appendix provides the ensuing p-value of RDM correlations).

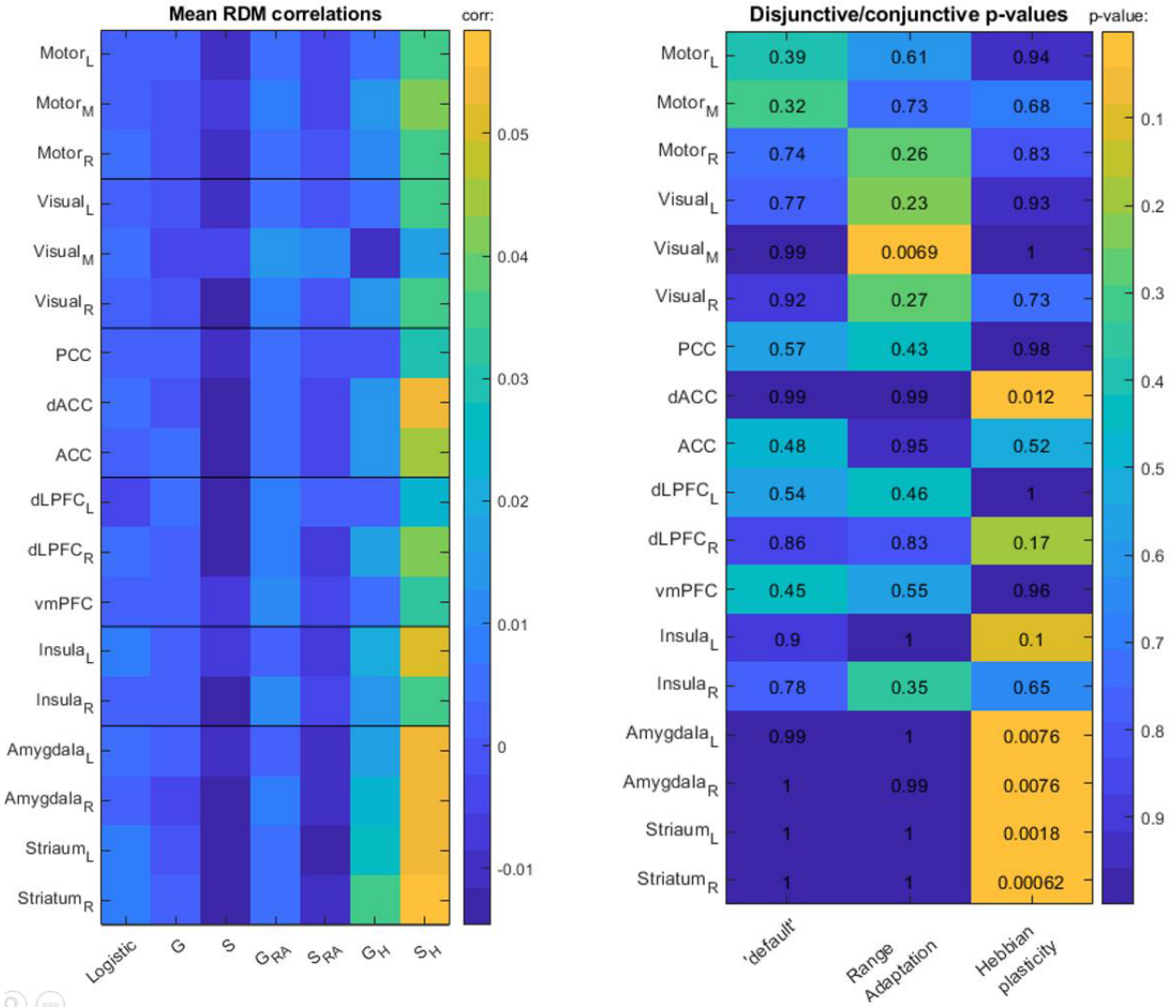



FIGURE 9.8 – *FMRI results : 'equal indifference' group. Same format as Figure 9.7.*

In brief, results remarkably replicate the 'equal range' study. Here again, the RDM Hebbian plasticity reaches statistical significance in the left Striatum (pseudo-Gaussian : $p = 0.0004$, sigmoid :

$p < 10^{-4}$), right Striatum (pseudo-Gaussian : $p = 0.0001$, sigmoid : $p < 10^{-4}$), left amygdala (pseudo-Gaussian trend : $p = 0.002$, sigmoid : $p < 10^{-4}$) and right Amygdala (pseudo-Gaussian : $p = 0.0002$, sigmoid : $p < 10^{-4}$). We note that here, the RDM correlations of sigmoid-Hebbian ANNs reach statistical significance in all other ROIs except in the medial visual cortex (all $p < 10^{-4}$). Notably, the RDM correlations of Hebbian ANNs are only significantly higher than other mechanisms of interest in bilateral Striatum (left Striatum : $p = 0.0018$, right Striatum : $p = 0.0006$). However, there is a trend in bilateral Amygdala (left/right Amygdala : $p = 0.0076$). In control ROIs, no model comparison achieves statistical significance, and the RDM correlations of range adaptation are never statistically higher than those of other mechanisms.

At this stage, one may safely conclude that Hebbian plasticity is a more likely explanation for fMRI activity patterns during risky decisions than range adaptation (or the default, non-constrained, biological scenario). But is Hebbian plasticity impairing or enabling adaptive behavior ? Numerical simulations on fitted Hebbian ANNs show that reducing Hebbian learning rates $\alpha_H$ (keeping all other estimated parameters the same) altered the decisions' sensitivity to small gains and high losses, effectively increasing loss aversion. But computational investigations of this sort cannot tell us whether and how people's behavior change when their brain activity displays more Hebbian-ness, i.e., when it becomes more similar to predictions from Hebbian ANNs. We thus ask whether inter-individual differences in Hebbian-ness may explain inter-individual differences in behavior, in particular : choice inconsistency. We define the Hebbian-ness of fMRI activity patterns in terms of the increase in neural evidence for the Hebbian ANN when compared to the default (non-constrained) ANN. Let $R^2_m$ be the percentage of explained variance in the fMRI RDM using the model $m$ (in each ROI). We then measure Hebbian-ness using the following pseudo F-score : $R^2_{Hebb} - R^2_{default}$. We define choice inconsistency in terms of the number of choices that contradict the logistic reference model, once it has been fitted to behavioral data. This effectively measures the rate of decisions, close to a subject's subjective indifference point, that contradicts its average preference. We then regress choice inconsistency against Hebbian-ness in bilateral Striatum and Amygdala concurrently (independently for both sigmoid and pseudo-gaussian ANNs). Figure 9.9 below summarizes this analysis for both groups of participants.

One can see that, when using pseudo-gaussian ANNs, Hebbian-ness does not predict inter-individual differences in choice inconsistency ('equal range' group : $p = 0.22$, 'equal indifference' group : $p = 0.37$, omnibus F-test). However, when using sigmoid ANNs, inter-individual differences in choice inconsistency can be predicted from fMRI measures of Hebbian-ness ('equal range' group : $p = 0.044$, 'equal indifference' group : $p = 0.012$, omnibus F-test). Now whether Hebbian-ness facilitates or hinders choice consistency seems to depend upon where in the brain it is measured. More precisely, increasing

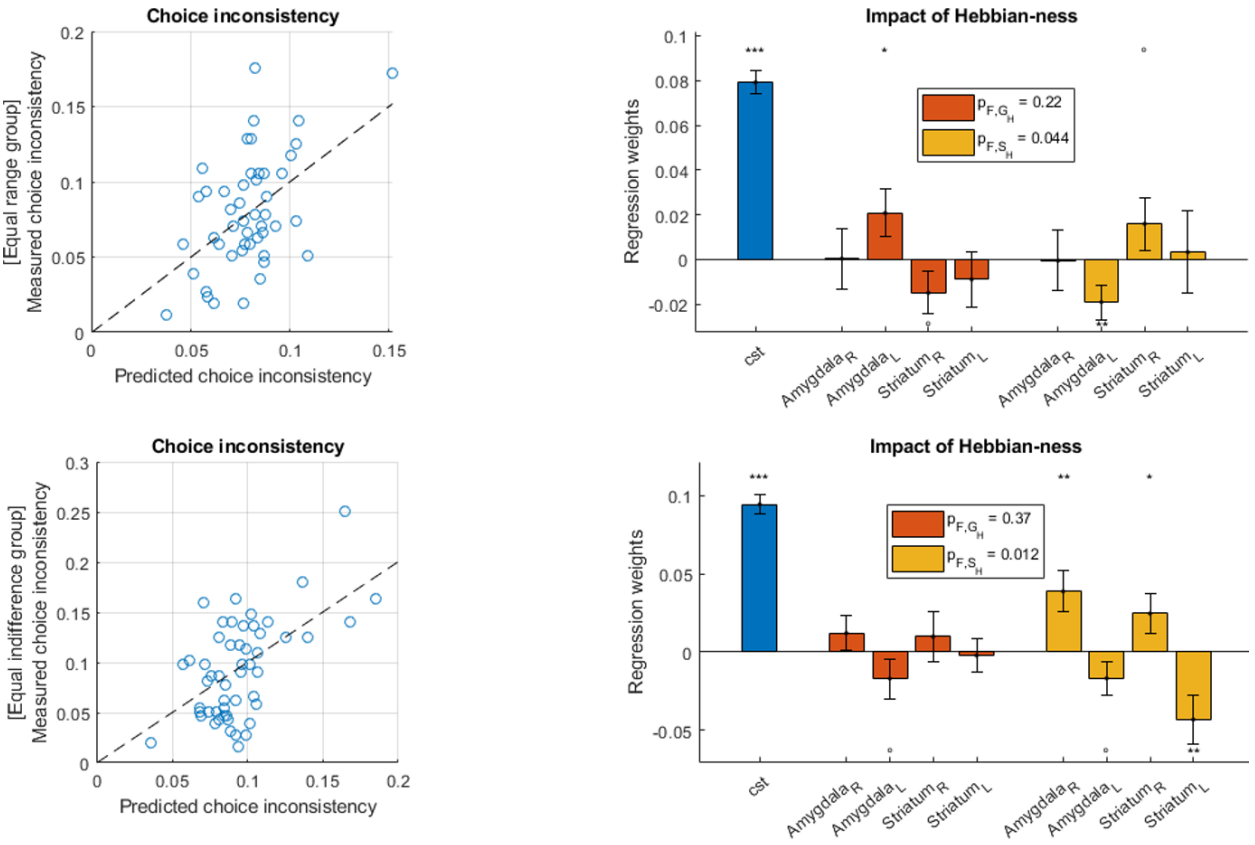



FIGURE 9.9 – *Analysis of inter-individual variability. Upper-left panel : measured (x-axis) and predicted (y-axis) rates of choice inconsistency are plotted against each other for the 'equal range' group (each dot is a participant). Lower-left panel : same as above, for the 'equal indifference' group. Upper-right panel : the normalized regression weight estimates (y-axis) are shown for each corresponding ROI (x-axis, from left to right : left Amygdala, right Amygdala, left ventral Striatum, right ventral Striatum), for both pseudo-gaussian (red) and sigmoid (yellow) Hebbian ANNs. Lower-right panel : same as above, for the 'equal indifference' group.*

Hebbian-ness in the left amygdala decreases choice inconsistency ('equal range' group : $p = 0.019$, 'equal indifference' group trend : $p = 0.057$), whereas (right-)striatal Hebbian-ness increases it ('equal range' group trend : $p = 0.18$, 'equal indifference' group :$p = 0.026$). We note that Hebbian-ness in the right Amygdala and left Striatum does not seem to have a robust effect on choice inconsistency, since statistical significance is reached only for the 'equal indifference' group (right Amygdala : $p = 0025$, left Striatum : $p = 0.0036$), but not for the 'equal range' group (right amygdala : $p = 0.97$, left Striatum : $p = 0.85$).

## 9.4 DISCUSSION

In this work, we identify the neural signature of candidate biological constraints and/or mechanisms that may shape or distort neural information processing. Rather than using normative models of behavior, we quantify the (potentially idiosyncratic) impact of biological constraints by fitting constrained ANNs to people's behavioral responses. We then use RSA to compare the estimated neural activity profiles to multivariate fMRI signals. Using numerical Monte-Carlo simulations, we demonstrate that the ensuing ANN-RSA approach is robust to modeling and statistical assumptions of no interest. We then show, on two independent fMRI studies, that (i)

seemingly indifferent choices in risky gambles are partially determined by range adaptation and/or Hebbian plasticity, (ii) multivariate activity in Striatum and Amygdala during choice is better explained by Hebbian plasticity than with range adaptation, and (iii) the Hebbian-ness of striatum and amygdala activity profiles predicts inter-individual differences in choice inconsistency. From a methodological standpoint, our main contribution is to show how to quantify the neural evidence for or against incidental, hard-wired, biological constraints on behaviorally-relevant information processing. With this aim, we retain the simplicity of established 'model-based' fMRI approaches [70 71], which proceed by cross-validating the identification of hidden computational determinants of behavior with neural data. In addition, we leverage the flexibility of ANNs and RSA to extend the breadth of empirical questions that can be addressed using dual computational/behavioural means.

In particular, this enables us to quantify the statistical evidence for neurophysiological mechanisms that are difficult –if not impossible- to include in computational models that are defined at Marr's algorithmic level [72], e.g., normative models of behavior (as derived from, e.g., learning or decision theories) and/or cognitive extensions thereof. Hebbian plasticity is a paradigmatic example of what we mean here. Recall that it was initially proposed as an explanation -at the neural or Marr's implementational level- for learning, memory, and sensory adaptation [73]. Since then, Hebbian-like synaptic plasticity that serves well-defined computational purposes of this sort has been superseded by theoretical frameworks that transcend the three Marr's analysis levels, e.g., the "Bayesian brain" hypothesis [74 75 76]. But hard-wired biological mechanisms of this sort may not always be instrumental to the cognitive process of interest. In turn, it may be challenging to account for incidental biological disturbances of neural information processing, when described at the algorithmic level. A possibility here is to conceive of these disturbances as some form of random noise that perturbs cognitive computations [77]. That these stochastic scenarios remain agnostic about the underlying (most likely hard-wired and deterministic) biological processes is both their strength and their weakness.

Of course, the field has been using neural network models of behavior for decades [78 79 80 81 82]. However, existing models are typically difficult to generalize beyond the empirical frame within which they have been derived. This is because model-based predictions typically rely on many assumptions that are specific to the neural circuit and/or the cognitive process of interest. In contrast, we take inspiration from recent theoretical work promoting the advantages of pairing ANNs with RSA [83], and search for neural evidence buried in multivariate patterns of brain activity while marginalizing over modeling assumptions of no interest. The aim here is to keep the modeling simple and protect the ensuing statistical inference from quantitative assumptions that have no theoretical or empirical support (cf., e.g., ANN dimensionality and/or sigmoid versus pseudo-gaussian


70. Borst, J., Taatgen, N., and Van Rijn, H. (2011). Using a symbolic process model as input for model-based fmri analysis : Locating the neural correlates of problem state replacements. *NeuroImage*, 58(1):137–147

71. O'Doherty, J., Hampton, A., and Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53

72. McClamrock, R. (1991). Marr ' s three levels : A re-evaluation. In *Minds and Machines*, page 185–196

73. Hebb, D. (1949). *The organization of behavior : a neuropsychological theory*. J. Wiley ; Chapman Hall

74. Aitchison, L. and Lengyel, M. (2017). With or without you : predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227

75. Doya, K., Ishii, S., Pouget, A., and Rao, R. (XXXX). Bayesian brain : Probabilistic approaches to neural coding

76. Friston, K. (2012). The history of the future of the bayesian brain. *NeuroImage*, 62(2):1230–1233

77. Drugowitsch, J., Wyart, V., Devauchelle, A., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6):1398–1411

78. Deco, G., Rolls, E., Albantakis, L., and Romo, R. (2013). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology*, 103:194–213

79. Frank, M. (2006). Hold your horses : A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, 19(8):1120–1136

80. Jocham, G., Hunt, L., Near, J., and Behrens, T. (2014). A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nature Neuroscience*, 15(7):960–961

81. Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221

82. Xiaojing, W. (2008). Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234

83. Kriegeskorte, N. and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 371(1705)

activation functions). Although the numerical simulations we present here tend to validate our statistical treatment, we think that this kind of problem is more flexibly solved using the so-called 'family inference' in the context of Bayesian model comparison [84]. In brief, the family inference is an optimal method for pooling statistical evidence over modeling factors of no interest and has proven both specific and sensitive in the context of large model spaces [85]. This would be most likely needed when extending the set of candidate biological constraints and/or when studying their interactions (see below).

A related point is the issue of defining which data feature(s) is eventually compared to model predictions. By construction, RSA assumes that candidate scenarios can be faithfully evaluated in terms of their ability to predict the trial-to-trial similarity/dissimilarity of multivariate (fMRI) patterns of neural activity. At the very least, this discards potentially relevant information, e.g., voxels' spatial location and peri-stimulus dynamics are lost [86]. Whether and how one may improve the statistical efficiency and robustness of RSA are unresolved issues [87 88 89]. In our context, this has two practical consequences. First, we used control sensory and motor ROIs to demonstrate the anatomo-functional specificity of our inference. Problematic here is the fact that we relied on negative results (in control ROIs), which may follow from the limited statistical efficiency of RSA. In the context of classical mass-univariate approaches, the issue of comparing different brain regions is known to be bound to many intricate confounds [90]. How these interact with the statistical properties of RSA is virtually unknown. Second, one may question the way we defined our set of confounds when deriving the ANN and fMRI RDMs. More precisely, we removed trial-by-trial variations that can be explained by linear combinations of inputs and outputs. This is important if one is to (i) draw inferences that are orthogonal to linear univariate event-related fMRI analyses, and (ii) prevent a bias towards models that fit behavior best. Note that the latter issue is critical for our definition of Hebbian-ness, whose inter-individual variations may otherwise be driven by statistical artifacts that grow with behavioral atypicality. The obvious cost of this conservative strategy is in terms of information loss. Although our results are qualitatively unchanged when excluding inputs and outputs from the set of confounds (not shown), this may not always be the case. In our opinion, addressing these sorts of issues may require the development of more sophisticated computational approaches that can treat behavioral and neural data in a statistically symmetrical manner [91 92]. We intend to pursue this type of approach in subsequent publications. At this point, and given the above limitations, we acknowledge that our neuroscientific claim is quite modest. In brief, our results support Hebbian plasticity as a valid alternative to range adaptation in the context of risky gambles. That we eventually identify the Striatum and the Amygdala to be specifically involved in this context is well aligned with the existing literature. On the one hand, the ventral Striatum is known to encode value and risk [93], and the tendency to opt for a risky choice increases


84. Penny, W. D., Stephan, K., Daunizeau, J., Rosa, M., Friston, K., Schofield, T., and Leff, A. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3)

85. Penny, W. D. and Ridgway, G. (2013). Efficient posterior probability mapping using savage-dickey ratios. *PLoS ONE*, 8(3)

86. Kriegeskorte, N. and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 371(1705)

87. Diedrichsen, J., Berlot, E., Mur, M., Schütt, H., and Kriegeskorte, N. (2020). Comparing representational geometries using the unbiased distance correlation. Retrieved from

88. Flandin, G. and Friston, K. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, 40(7):2052–2054

89. Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (XXXX). Reliability of dissimilarity measures for multi-voxel pattern analysis. Retrieved from

90. Henson, R. (2006). Forward inference using functional neuroimaging : Dissociations versus associations. *Trends in Cognitive Sciences*, 10(2):64–69

91. Lohmann, G., Müller, K., and Turner, R. (2013). Response to commentaries on our paper : Critical comments on dynamic causal modelling. *NeuroImage*, 75:279–281

92. Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221

93. Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C., Tobler, P., and Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 363(1511):3801–3811

with the magnitude of the striatal response to risk [94 95]. In fact, the same experimental protocol as we use here ('equal indifference' range) already served to demonstrate that the differential striatal responses to losses and gains drive inter-individual variations in loss aversion [96]. On the other hand, it was also shown that the prospect of a possible loss might activate Amygdala, which would trigger a cautionary brake on behavior that facilitates loss aversion [97 98]. How ventral Striatum and Amygdala eventually interact with each other to determine loss aversion is unknown, and the present study does not resolve this debate. In line with recent studies of hysteretic effects in the brain's decision system [99 100 101], we rather focus on seemingly indifferent and/or inconsistent choices, which remain otherwise unexplained. The present results illustrate how neuroimaging can be used to directly test whether candidate hard-wired, incidental, biological constraints may impact on behavior : in this case, the hysteretic effects of range adaptation and/or Hebbian plasticity. Hebbian plasticity, but not range adaptation, was observed in both brain systems that were previously shown to regulate loss aversion. Retrospectively, however, many other candidate mechanisms may, in principle, explain such hysteretic effects, e.g., homeostatic plasticity [102 103 104 105]. To what extent seemingly indifferent and/or inconsistent choices may eventually be explained away with these and/or similar biological constraints is an open and challenging issue.


94. Christopoulos, G., Tobler, P., Bossaerts, P., Dolan, R., and Schultz, W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience*, 29(40):12574–12583

95. Kuhnen, C. and Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, 47(5):763–770

96. Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518

97. De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. (2006). Frames, biases and rational decision-making in the human brain. *Science*, 313(5787):684–687

98. De Martino, B., Camerer, C., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792

99. Conen, K. and Padoa-Schioppa, C. (2019). Partial adaptation to the value range in the macaque orbitofrontal cortex. *Journal of Neuroscience*, 39(18):3498–3513

100. Antonio, R. and Clithero, J. (2012). Value normalization in decision making : theory and evidence. *Current Opinion in Neurobiology*, 22(6):970–981

101. Soltani, A., Martino, B., and Camerer, C. (2012). A range-normalization model of context-dependent choice : A new model and evidence. *PLoS Computational Biology*, 8(7)

102. Fox, K. and Stryker, M. (2017). Integrating hebbian and homeostatic plasticity : Introduction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715)

103. Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134:17–35

104. Toyoizumi, T., Kaneko, M., Stryker, M., and Miller, K. (2015). Modeling the dynamic interaction of hebbian and homeostatic plasticity

105. Turrigiano, G. (2017). The dialectic of hebb and homeostasis. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715):4–6

## 9.A APPENDIX : RANGE ADAPTATION

In what follows we provide the mathematical derivation of equation 9.6 of the main text. Let $\sigma^*$ be the value of the pseudo-variance parameter $\sigma$ that maximizes the derivative of a given ANN unit activation function w.r.t. to its inputs, i.e. :

$$\sigma^* = arg\max_{\sigma} \left| \frac{\partial f_.}{\partial u}\Big|_{u,\sigma} \right| \tag{9.11}$$

where $f_.$ is the activation function of neural units in the ANN (cf. Equation 9.2), and we have dropped unit indices for mathematical convenience.

Range adaptation proceeds by modifying the pseudo-variance parameter in the direction of $\sigma^*$, with a step size that is controlled by the learning rate $\alpha_{RA}$, i.e. :

$$\sigma_{t+1} = \sigma_t + \alpha_{RA} \times (\sigma^* - \sigma_t) \tag{9.12}$$

Setting $\alpha_{RA} < 1$ ensures that the pseudo-variance parameter integrates the history of past inputs when adapting its range of activation. Let us first focus on pseudo-gaussian activation functions. Without loss of generality, we will drop the time index and use a centred input $\tilde{u} = u - \mu$. The first derivative of the activation function is given by :

$$\frac{\partial f_{Gauss}}{\partial u}\Big|_{\tilde{u},\mu,\sigma} = \frac{2\tilde{u}}{\sigma^2} exp\left(-\frac{\tilde{u}}{\sigma^2}\right) \tag{9.13}$$

Range adaptation proceeds by maximizing Equation A3 with respect to , which reduces to finding the zero of the mixed partial derivative of $f_{Gauss}$ :

$$\frac{\partial^2 f_{Gauss}}{\partial\sigma\partial\tilde{u}}\Big|_{\tilde{u},\mu,\sigma} = (\sigma^2 - \tilde{u}^2)\frac{4\tilde{u}}{\sigma^5} exp\left(-\frac{\tilde{u}^2}{\sigma^2}\right) \tag{9.14}$$

Since $\sigma$ is positive, the maximum of Equation A3 is simply given by $\sigma^* = |\tilde{u}| = |u - \mu|$. Inserting the expression for into Equation A2 then provides the following learning rule :

$$\sigma_{t+1} = \sigma_t + \alpha_{RA} \times (|\mu - u_t| - \sigma_t) \tag{9.15}$$

Noting the input to the integration layer is given by $\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)} x_t^{(i,j)}$ eventually yields Equation 9.6 of the main text.

Let us now focus on sigmoid activation functions. In this case, we use the following change of variable : $\tilde{u} = (u - \mu) * \gamma$, where $\gamma$ is the arbitrary scaling factor of the sigmoid activation function (cf. Equation 9.2). We will see that it is possible to set such that the range adaptation learning rule is identical for both pseudo-gaussian and sigmoid activation functions. It is trivial to show that the first and mixed partial derivative of the sigmoid activation function are

given by :

$$\begin{aligned}\left.\frac{\partial f_{sigmoid}}{\partial \tilde{u}}\right|_{\tilde{u},\mu,\sigma} &= \frac{1}{\sigma}\frac{e^{-\tilde{u}/\sigma}}{\left(e^{-\tilde{u}/\sigma}+1\right)^2} \\ \left.\frac{\partial f_{sigmoid}}{\partial \tilde{u}}\right|_{\tilde{u},\mu,\sigma} &= e^{-\tilde{u}/\sigma}\frac{-(\tilde{u}/\sigma+1)e^{-\tilde{u}/\sigma}+(\tilde{u}/\sigma-1)}{\sigma^2\left(e^{-\tilde{u}/\sigma}+1\right)^3}\end{aligned} \tag{9.16}$$

Finding the zero of the mixed partial derivative reduces to solving $(y-1)-e^{-y}(y+1)$ with $y=\tilde{u}/\sigma$. There is no analytical closed form solution to this equation, but a numerical approach yields $y^* \approx \pm 1.5434$. Since $\sigma$ is positive, the solution is simply given by $\sigma^* = |\tilde{u}/y^*| = |u-\mu| \times \gamma/1.5434$. Setting $\gamma = 1.5434$ then simplifies the solution to $\sigma^* = |u-\mu|$, which thus provides the same learning rule as Equation A3 above.

## 9.B APPENDIX : FMRI RESULTS STATISTICS

In what follows, we provide summary statistics of our ANN-based behavioural and fMRI analyses. Table 1 below gives the mean percentage of explained behavioral variance (R2) and it standard deviation (across participants) for each model, for both groups. Table 2 below gives the mean R2 difference between each ANN model and the reference logistic model, its standard deviation, and the resulting p-value ($H_0$ : no $R^2$ difference, dof=53), for both groups. Table 3 below gives the p-value of RDM correlations for each model and each ROI ($H_0 : \rho \leq 0$, one-sided t-test, dof=53), in the 'equal' range' group. Table 4 below gives the p-value of RDM correlations for each model and each ROI ($H_0 : \rho \leq 0$, one-sided t-test, dof=53), in the 'equal indifference' group. In all tables, statistical significance is highlighted in green (with the appropriate threshold correction for multiple comparisons).

| | 'equal range' | | 'equal indifference' | |
|---|---|---|---|---|
| Model | mean | std | mean | std |
| logistic | 0.7789 | 0.1049 | 0.7367 | 0.1273 |
| G-ANN | 0.8451 | 0.0784 | 0.8831 | 0.0866 |
| S-ANN | 0.8374 | 0.0793 | 0.7845 | 0.1166 |
| G-RA-ANN | 0.8636 | 0.0682 | 0.7937 | 0.1822 |
| S-RA-ANN | 0.8402 | 0.0878 | 0.8409 | 0.1094 |
| G-H-ANN | 0.8621 | 0.085 | 0.867 | 0.1673 |
| S-H-ANN | 0.8527 | 0.0623 | 0.8225 | 0.127 |

TABLE 9.2 – *Mean $R^2$ and its standard deviation for each model, for both groups.*

| | 'equal range' | | | 'equal indifference' | | |
|---|---|---|---|---|---|---|
| Model | mean | std | p-value | mean | std | p-value |
| G-ANN | 0.0662 | 0.0939 | 5.00E-06 | 0.1463 | 0.0748 | 3.00E-20 |
| S-ANN | 0.0586 | 0.0446 | 1.00E-12 | 0.0477 | 0.0311 | 6.00E-16 |
| G-RA-ANN | 0.0847 | 0.0725 | 6.00E-11 | 0.0569 | 0.1884 | 1.53E-02 |
| S-RA-ANN | 0.0614 | 0.0363 | 4.00E-16 | 0.1041 | 0.0669 | 4.00E-16 |
| G-H-ANN | 0.0832 | 0.0601 | 3.00E-13 | 0.1302 | 0.167 | 3.00E-07 |
| S-H-ANN | 0.0738 | 0.0841 | 8.00E-08 | 0.0857 | 0.119 | 2.00E-06 |

TABLE *9.3 – Mean $R^2$ difference and its standard deviation for each ANN model, for both groups.*

| | G-ANN | S-ANN | G-RA-ANN | S-RA-ANN | G-H-ANN | S-H-ANN |
|---|---|---|---|---|---|---|
| $\text{Motor}_L$ | 0.9985 | 0.9684 | 0.0056 | 0.9532 | 0.0096 | 0.0455 |
| $\text{Motor}_M$ | 0.9909 | 0.998 | 0.1474 | 0.911 | 0.0022 | 0.0008 |
| $\text{Motor}_R$ | 0.9994 | 0.9558 | 0.0335 | 0.7906 | 0.0113 | 0.0195 |
| $\text{Visual}_L$ | 0.9978 | 0.9768 | 0.01 | 0.9658 | 0.0684 | 0.2107 |
| $\text{Visual}_M$ | 0.9146 | 0.7227 | 0.0762 | 0.9782 | 0.184 | 0.5914 |
| $\text{Visual}_R$ | 0.9458 | 0.8604 | 0.0707 | 0.8023 | 0.0226 | 0.0626 |
| PCC | 0.9737 | 0.9226 | 0.0644 | 0.9616 | 0.0101 | 0.1186 |
| dACC | 0.9999 | 0.9868 | 0.1332 | 0.7359 | 0.0017 | 0.0008 |
| ACC | 0.9996 | 0.9487 | 0.2128 | 0.9518 | 0.0014 | 0.0037 |
| $\text{dLPFC}_L$ | 0.9856 | 0.7515 | 0.0271 | 0.9953 | 0.004 | 0.0082 |
| $\text{dLPFC}_R$ | 0.9887 | 0.7984 | 0.0034 | 0.7863 | 0.0004 | 0.0004 |
| vmPFC | 0.9999 | 0.8739 | 0.1146 | 0.9821 | 0.0407 | 0.0535 |
| $\text{Insula}_L$ | 0.9965 | 0.9485 | 0.0507 | 0.9592 | 0.0004 | 3.57E-05 |
| $\text{Insula}_R$ | 0.9642 | 0.9376 | 0.1517 | 0.9201 | 0.0052 | 0.0067 |
| $\text{Amygdala}_L$ | 0.9889 | 0.5912 | 0.4413 | 0.983 | 0.0006 | 0.0001 |
| $\text{Amygdala}_R$ | 0.9773 | 0.9464 | 0.2797 | 0.9175 | 1.51E-05 | 9.45E-07 |
| $\text{Striatum}_L$ | 0.9986 | 0.716 | 0.3392 | 0.8637 | 1.86E-05 | 3.13E-06 |
| $\text{Striatum}_R$ | 0.9956 | 0.9768 | 0.222 | 0.9845 | 0.0001 | 0.0002 |

TABLE *9.4 – P-value of RDM correlations for each model and each ROI ('equal' range').*

| | G-ANN | S-ANN | G-RA-ANN | S-RA-ANN | G-H-ANN | S-H-ANN |
|---|---|---|---|---|---|---|
| $\text{Motor}_L$ | 0.4258 | 0.9914 | 0.0784 | 0.7568 | 0.1149 | 1.88E-06 |
| $\text{Motor}_M$ | 0.5736 | 0.9907 | 0.0324 | 0.839 | 0.0066 | 1.94E-06 |
| $\text{Motor}_R$ | 0.6155 | 0.9975 | 0.0559 | 0.6253 | 0.0336 | 7.20E-07 |
| $\text{Visual}_L$ | 0.6878 | 0.9954 | 0.153 | 0.5337 | 0.2292 | 2.41E-05 |
| $\text{Visual}_M$ | 0.7691 | 0.9556 | 0.0014 | 0.0344 | 0.9672 | 3.22E-03 |
| $\text{Visual}_R$ | 0.5845 | 0.9997 | 0.0249 | 0.5964 | 0.0202 | 1.22E-06 |
| PCC | 0.2997 | 0.9995 | 0.1245 | 0.6836 | 0.4924 | 9.47E-05 |
| dACC | 0.455 | 0.9998 | 0.1214 | 0.8494 | 0.0097 | 1.00E-07 |
| ACC | 0.0753 | 0.9993 | 0.1022 | 0.8217 | 0.0026 | 6.30E-07 |
| $\text{dLPFC}_L$ | 0.051 | 0.9995 | 0.0146 | 0.4022 | 0.4083 | 1.94E-04 |
| $\text{dLPFC}_R$ | 0.3554 | 0.9998 | 0.0249 | 0.9613 | 0.0037 | 8.10E-07 |
| vmPFC | 0.3914 | 0.9742 | 0.0148 | 0.7574 | 0.1307 | 6.89E-05 |
| $\text{Insula}_L$ | 0.2355 | 0.9684 | 0.3209 | 0.964 | 0.0009 | 1.00E-08 |
| $\text{Insula}_R$ | 0.3882 | 0.9999 | 0.0135 | 0.8121 | 0.0171 | 3.71E-05 |
| $\text{Amygdala}_L$ | 0.4186 | 0.9998 | 0.3852 | 0.9786 | 0.0026 | 2.00E-08 |
| $\text{Amygdala}_R$ | 0.7117 | 0.9987 | 0.0353 | 0.9971 | 2.00E-04 | 3.00E-08 |
| $\text{Striatum}_L$ | 0.5835 | 0.9997 | 0.1023 | 0.9974 | 2.00E-04 | 4.00E-08 |
| $\text{Striatum}_R$ | 0.3378 | 0.9998 | 0.0955 | 0.9782 | 0.0001 | 3.30E-07 |

TABLE *9.5 – P-value of RDM correlations for each model and each ROI ('equal' indifference)*

# 10

# *Conclusion sur les réseaux de neurones contraints*

◂ Chapitre 9 Chapitre 11 ▸

Dans ce chapitre, j'ai étudié une technique permettant de comparer différentes contraintes altérant les mécanismes neuronaux des processus de décision.

Tout comme les travaux de Rao et Ballard (voir chapitre 2), ou plus récemment de Josh McDermott [1], de Bernard Balleine [2] de Xia-Jing Wang [3] ou encore de Tim Behrens [4], ces travaux proposent de décrire les processus de traitement cérébral de l'information à l'aide de réseaux de neurones artificiels, puis de comparer leurs propriétés à celles de véritables systèmes neuronaux. Cependant, mon approche ne s'accorde pas rigoureusement avec la méthode des niveaux de Marr, telle qu'elle est typiquement utilisée pour l'étude de la prise de décision.

D'une part, je n'utilise pas de modèle normatif de la décision. Mes réseaux de neurones artificiels ne cherchent pas à prendre une décision de manière rationnelle : leur objectif est de reproduire fidèlement le comportement des sujets. Certes, cela revient au même si les sujets effectuent rationnellement la tâche qui leur est imposée, cependant, dès lors que les sujets dévient de ce comportement idéal [5], l'ajustement des modèles s'y adaptera et reproduira ces déviations. D'une certaine manière, on pourrait dire que le réseau adopte non pas le but de la tâche, mais le but implicite de chaque sujet, celui qui sous-tend ses choix.

D'autre part, la relation entre les niveaux deux et trois de Marr de mes modèles de génération du comportement est ambiguë. Il ne s'agit pas ici de déterminer des représentations sans tenir compte de leur substrat (niveau deux), ni de caractériser les propriétés biologiques du système indépendamment de sa fonction (niveau trois). Je cherche précisément à modéliser l'influence du substrat sur la fonction : les niveaux deux et trois de Marr s'entrecroisent. Ironiquement, il s'agit là du 4e niveau initialement proposé par Marr et Poggio [6], celui qui décrit la dépendance entre un algorithme et les opérations autorisées par son substrat.

Vue sous cet angle, mon approche souligne la dépendance mutuelle des niveaux deux et trois de Marr : bien que tous les réseaux soient « ajustés au même but » [7], ils aboutissent à des algorithmes différents [8], selon les contraintes physiologiques qui leur sont imposées. En d'autres termes, il existe une forme de dégénérescence comportementale qui s'exprime sur le plan algorithmique, mais qu'il est, en principe, possible de lever en tenant compte des contraintes

1. Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644

2. Dezfouli, A., Morris, R., Ramos, F. T., Dayan, P., and Balleine, B. (2018). Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models. In *Advances in Neural Information Processing Systems*, pages 4228–4237

3. Song, H. F., Yang, G. R., and Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife*, 6:e21492

4. Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2019b). The tolman-eichenbaum machine : Unifying space and relational memory through generalisation in the hippocampal formation. *BioRxiv*, page 770495

5. Dans le cas de la tâche précédente, la réponse rationnelle sur le plan économique est d'accepter la loterie dès lors que les gains excèdent les pertes.

6. Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry

7. Dans l'étude précédente, tous les réseaux contraints (par plasticité hebbienne ou sensibilité adaptative) sont ajustées aux mêmes données, et ils expliquent le comportement avec le même degré de fiabilité. En ce sens, ils sont « ajustés au même but ».

8. Bien qu'ils ne soient pas explicitement exhibés, les algorithmes de chaque réseau diffère nécessairement, puisque leurs représentations internes diffèrent.

biologiques qui s'expriment sur le système.

D'autre part, elle suggère que le comportement des sujets diffère selon l'algorithme qu'ils emploient, et par conséquent que le but implicite de leurs actions diffère selon la force des contraintes s'appliquant à leurs processus décisionnels neuronaux.

# Quatrième partie

# CONCLUSION

# 11

# *Comparaison de l'analyse de médiation et des RNA contraints*

◂ Chapitre 10

Synopsis Dans les chapitres précédents, j'ai présenté deux techniques permettant d'étudier les étapes d'un processus neural de traitement de l'information impliqué dans la détermination du comportement. Dans ce chapitre final, je reviens sur leurs intérêts et limitations, leurs similarités et différences, ainsi que sur leurs combinaisons et extensions potentielles. Enfin, je conclus sur une réflexion plus générale concernant l'apport de ce travail de thèse à ma formation scientifique.

## 11.1 APPORTS MÉTHODOLOGIQUES

Dans l'étude présentée dans au chapitre 6, j'ai établi que, dans une analyse massivement univariée de données d'IRMf, seul un test conjonctif pouvait détecter efficacement des médiateurs neuronaux. J'ai également montré qu'une telle analyse était peu sensible aux premières et dernières étapes d'un système séquentiel de traitement de l'information. Dans l'étude présentée dans au chapitre 9, j'ai proposé un cadre pour identifier les contraintes biologiques qui s'expriment sur les processus neuronaux de traitement de l'information. J'ai également démontré que, bien que ces contraintes ne soient pas discriminables sur le plan comportemental, il est possible d'en retrouver la trace neurale à l'aide d'analyses IRMf multivariées.

## 11.2 APPORTS COGNITIFS

Dans l'étude du chapitre 6, j'ai identifié dix aires cérébrales susceptibles d'effectuer la médiation de l'effet des gains sur le rejet ou l'acceptation d'un pari. Parmi eux le cortex préfrontal dorsolatéral postérieur gauche et également associé au degré d'aversion à la perte des sujets. Cela suggère que le profil de médiation du système cérébral impliqué permet de caractériser certains aspects idiosyncratiques du comportement. Dans l'étude du chapitre 9, j'ai montré que les décisions apparemment aléatoires d'un sujet [1] peuvent être expliquées par une forme de plasticité hebbienne altérant l'organisation des systèmes cérébraux impliqués dans la prise de décision (en particulier : l'amygdale et le striatum, de façon bilatérale). Enfin, j'ai montré que l'incohérence du comportement d'un sujet est d'autant plus importante que la plasticité hebbienne est plus marquée dans son amygdale droite, et plus faible dans son striatum gauche. Cela suggère que le profil « d'hebbianité » d'un sujet permet de caractériser certains aspects idiosyncratiques de son comportement.

1. C'est à dire les décisions qui ne sont pas expliquées par un modèle logistique prenant en compte les gain set les pertes. Typiquement, il s'agit des choix effectués autour du point d'indifférence.

## 11.3 LIMITES

Bien que ces interprétations soient séduisantes, elles doivent être nuancées par les limites méthodologiques des approches. Tout d'abord, si les analyses mises en œuvre ici suggèrent l'implication de telle ou telle région cérébrale dans la génération du comportement, elles ne peuvent pas l'établir de manière irréfutable. Pour cela, une analyse causale par perturbation est nécessaire, c'est-à-dire en manipulant directement l'activité d'une région ((Smith et al., 2011) propose cependant d'autres types de méthodes).

De plus, ces deux approches étudient les éléments du système cérébral impliqué indépendamment les uns des autres. En principe, il n'est donc pas possible de comprendre l'implication spécifique d'un élément du système dans la chaîne de traitement de l'information.

Ensuite, ces analyses encourent toujours un risque de biais statis-

tique induit par la forme mathématique des modèles. Par exemple, l'étude du Chapitre 3 emploie exclusivement des modèles linéaires. Si les données neurales présentent un format différent, cela peut induire des artefacts dont l'impact reste à déterminer. De même, les réseaux de neurones artificiels contraints du chapitre 9 n'emploient que des fonctions d'activation gaussiennes ou sigmoïdales. Cependant, d'autres formes de sensibilité neuronale ont été suggérées par le passé : fonctions Gabor [2], cellules simples, complexes [3], hypercomplexes [4], cellules de lieux, cellules de grilles [5], etc. Le degré de confusion que ces formes induiraient dans l'analyse des contraintes biologique est pour l'instant inconnu.

2. Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*

3. Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025

4. Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex : A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87

5. Whittington, J., Muller, T., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. (2019a). The tolman-eichenbaum machine : Unifying space and relational memory through generalisation in the hippocampal formation

## 11.4 SIMILARITÉ

Les deux approches présentées dans cette thèse proposent d'une part d'identifier de nouveaux mécanismes décisionnels, et d'autre part de valider ces mécanismes par des mesures comportementales.

Je note que les deux méthodes offrent une grande flexibilité dans les applications qui peuvent en être faites. La seule contrainte d'une analyse de médiation est de combiner linéairement le médiateur neural au comportement. De leur côté, les réseaux de neurones artificiels sont des « approximateurs universels » [6] et bénéficient de nombreux paramètres ajustables. Dès lors, ces modèles peuvent capturer des variations inattendues du comportement [7], et permettent d'explorer des mécanismes échappant à des modèles moins agnostiques. Bien entendu, le danger de ce type de modélisation réside dans un surajustement des modèles aux bruits des données expérimentales : « *l'over-fitting* ». Un modèle sera surajusté s'il décrit bien les données pour une raison purement statistique, d'une manière indépendante des mécanismes qui déterminent le comportement de chaque sujet. Ici, le problème qu'induirait un surajustement du modèle est double. D'une part, il entraînerait une erreur de généralisation des prédictions comportementales. D'autre part, il pourrait impliquer une confusion sur l'identification des mécanismes neuraux sous-jacents à la décision. Cela dit, le risque de surajustement est limité, dès lors que les résultats de ces analyses peuvent être utilisés pour prédire une caractéristique du comportement orthogonale à l'information utilisée pour les obtenir.

6. C'est-à-dire que n'importe quelle fonction pourra toujours être décrite par un réseau de neurones artificiels, s'il possède suffisamment d'unités.

7. Des variations indépendantes des facteurs expérimentaux, ou répondant à des traitements non-linéaires.

C'est la raison pour laquelle j'ai procédé en deux étapes, dans le contexte des applications de ces techniques à la prise de décision risquée. Dans un premier temps, j'ai cherché à tirer des conclusions valides au niveau du groupe de participants. En général, ces conclusions peuvent effectivement être sujettes au risque de surajustement. Dans un deuxième temps, j'ai utilisé les résultats des analyses pour extraire des marqueurs biologiques des processus de détermination du comportement [8] permettant de prédire les différences inter-individuelles. Ainsi, si le profil de médiation ou « l'hebbianité » d'une aire cérébrale permettent de prédire l'aversion à la perte et/ou l'incohérence comportementale d'un sujet, cela valide les résultats

8. La force de médiation des différentes régions ou leur hébbianité.

obtenus au niveau du groupe. À ce stade, il faut cependant noter que cette approche n'est qu'approximative et nécessiterait une formalisation plus poussée. Une compréhension plus globale de la détermination du comportement requiert un modèle associant, à la fois une description des mécanismes internes des éléments cérébraux impliqués, et une prise en compte de leur interdépendance.

## 11.5 DIFFÉRENCES

Rétrospectivement, ces deux approches traitent le problème de l'identification des déterminants biologiques du comportement de manière relativement distincte.

L'analyse de médiation identifie les aires cérébrales dont l'activité permet d'améliorer la prédiction du comportement. La détection d'un effet indirect significatif suggère que la région correspondante opère une transformation de l'information entrante (stimuli, instructions) qui détermine le comportement, mais dont la nature n'est pas informée (ou même confirmée) par l'analyse. L'analyse de médiation repose sur un modèle simpliste de cette transformation (l'ajout d'un bruit neural), qui peut, en principe, absorber l'influence de facteurs ignorés, ou d'aspects ignorés des facteurs expérimentaux [9].

L'approche en réseaux de neurones contraints, pour sa part, permet de comparer différents mécanismes de traitement de l'information grâce aux données neuronales et d'identifier le plus probable. Ce faisant, elle exploite les données neuronales pour spécifier la nature de la transformation de l'information opérée par les aires cérébrales impliquées. Cependant, si elle permet d'explorer certaines formes non linéaires de cette transformation, elle ne permet pas de découvrir des effets non modélisés. Par ailleurs, cette analyse exploite, par construction, des variations du signal BOLD orthogonales à celles qui importent à l'analyse de médiation [10].

Ces différences expliquent en partie les écarts de résultats entre les deux études : l'amygdale et le striatum ne sont pas détectés comme des médiateurs de l'effet des gains ou des pertes sur le choix des sujets. Certes, elles exhibent un profil partiellement similaire à un réseau de neurones hebbien, et ce dernier offre une meilleure description du comportement qu'un modèle logistique. Cependant, cette similarité est relativement faible [11], et elle n'implique pas que l'activité de ces régions permet de prédire le comportement avec la même performance que le réseau de neurones artificiels.

9. Puisque le « bruit neural » capture n'importe quelle déviation de l'équation d'encodage de l'analyse de médiation.

10. L'analyse multivariée RSA du signal BOLD est précédée d'une étape de pré-traitement qui projette les données (et les prédictions du modèle) dans l'espace nul des entrées et sorties comportementales.

11. Les corrélations sont de l'ordre de 6%

## 11.6 COMPLÉMENTARITÉ

Les deux approches peuvent également être combinées pour affiner l'identification des déterminants du comportement. Par exemple, le bruit neural instrumental [12] révélé par l'analyse de médiation peut être comparé aux profils d'activité des réseaux de neurones contraints. L'aspect exploratoire de l'analyse de médiation permettrait ainsi de

12. La partie du médiateur orthogonale aux facteurs expérimentaux mais associée au comportement.

détecter les éléments du système impliqué dans la détermination du comportement, dont les mécanismes de traitement de l'information pourraient être alors identifiés via l'approche en réseaux de neurones. D'une certaine manière, on remplacerait ainsi une description stochastique et agnostique de la transformation des entrées du système en un traitement non linéaire dont on pourrait interpréter les propriétés (potentiellement contraintes par la biologie du système). Inversement, on peut appliquer une analyse de médiation aux représentations internes d'un réseau de neurones ajusté sur le comportement. Cela permettrait d'identifier, soit une étape de traitement supplémentaire, soit une nouvelle source de variance (potentiellement alimentée par l'activité d'autres éléments du système). Il serait par exemple intéressant de vérifier si l'amygdale et le striatum peuvent être identifiés comme des médiateurs entre la première couche d'un des réseaux hebbiens du chapitre 9 et le comportement des sujets.

## 11.7 PERSPECTIVES

Chaque approche peut également être étendue pour couvrir plusieurs régions cérébrales et modéliser leurs interactions. Pour l'analyse de médiation, cela revient à ajouter un ou plusieurs nœuds supplémentaires entre les facteurs expérimentaux et le comportement. L'ajout d'étapes intermédiaires dans le graphe de médiation permettra notamment de nuancer le compromis entrée-sortie (voir chapitre 6), et ainsi révéler une plus grande partie des processus décisionnels. Pour les réseaux de neurones contraints, cette extension revient à combiner plusieurs réseaux de neurones anatomiquement ségrégés (à la manière d'un bDCM) pour former un graphe de modules interagissant entre eux. L'architecture globale du réseau sera alors critique pour interpréter le rôle de chaque composant. Cela dit, ce type d'extension soulève une difficulté de principe. Dans le chapitre 9, la deuxième couche intégrait les représentations des gains et des pertes pour déterminer le choix : elle implémentait un processus de décision. Or si plusieurs modules interagissent entre eux pour déterminer un ou plusieurs aspects du comportement, leur contribution respective devient plus difficile à identifier[13]. Ce problème peut être partiellement contourné en s'assurant que l'architecture du modèle respecte certaines contraintes issues de scénarios neurocognitifs prédéfinis. Il devient alors possible de contrôler précisément les interactions des modules et d'y assigner distinctement la détermination d'aspects spécifiques du traitement de l'information : évaluation des options, comparaison des valeurs, estimation de la confiance dans le choix, etc. Une fois le modèle ajusté grâce aux données comportementales et IRMf, il est alors possible d'effectuer une analyse systématique de l'impact de lésions virtuelles sur le modèle, dont les prédictions pourraient être validées expérimentalement à l'aide de méthodes causales par perturbation et/ou par l'étude de patients cérébro-lésés.

À ce stade, il faut noter que la mise en pratique de ce genre d'extension est difficilement compatible avec une approche purement

13. Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2018). A benchmark for interpretability methods in deep neural networks. NeurIPS). Retrieved from

exploratoire. En premier lieu, le problème des comparaisons multiples pourrait s'avérer prohibitif. Dans un modèle de médiation, chaque nœud du graphe induit au minimum deux connexions à tester. Dans un modèle de réseaux de neurones contenant plusieurs modules, il faut comparer chaque module à chaque région candidate, puis vérifier que les relations inter-régionales correspondent également aux relations inter-modules. Enfin, sans hypothèse préalable sur les régions cérébrales impliquées dans la prise de décision, il faut tester toutes les combinaisons possibles. Or, ces dernières croient exponentiellement en fonction du nombre de régions étudiées.

On peut néanmoins contourner ces difficultés de plusieurs manières : tout d'abord, en limitant le nombre de régions explorées et leurs interactions à l'aide de données supplémentaires. Par exemple, en sélectionnant uniquement les régions associées à une tâche cognitive donnée [14], et en ne considérant que les connexions anatomiques liées aux faisceaux de substance blanche [15]. Ensuite, au lieu d'évaluer indépendamment chaque élément d'un graphe, on peut extraire un score global représentant l'adéquation des données IRMf à la structure du graphe [16]. Il suffit alors de comparer les scores de plusieurs graphes concurrents pour déterminer le plus probable [17].

14. En effectuant une revue de la littérature, ou grâce à des sites comme https://neurosynth.org ou https://neuroquery.org.

15. On pourra utilisé des données de tractographies (voir chapitre 2).

16. Par exemple via le pourcentage de variance expliquée dans les données, ou d'autres mesures plus avancées telles que la probabilité (bayésienne) du modèle, i.e. de la structure du graphe et ses mécanismes.

17. On pourra également s'appuyer sur une méthode bayésienne de sélection de modèles, comme c'est l'usage pour les analyses de type DCM.

## 11.8 Conclusion scientifique

Que retenir de tous les points de vues que j'ai présentés, de toutes les études dont j'ai parlé et des méthodes que j'ai élaborés ? En premier lieu, et malgré la complexité du fonctionnement du cerveau, il faut retenir que l'on possède aujourd'hui de nombreux moyens nous permettant de l'étudier. Au cours de cette thèse j'ai essayé de montrer que les limitations actuelles tenaient plus de l'usage des méthodes à notre disposition que de limitations techniques ou technologiques. La modélisation de mécanismes cognitifs, les études d'encodages et de décodages, la modélisation du code neural, l'étude des propriétés statistiques d'une méthode, toutes ces approches sont relativement simples à mettre en œuvre [18]. Elles sont avant tout limitées par leurs hypothèses, et celles-ci sont constamment allégées.

18. Car elles ne nécessitent pas de lourds investissements financiers, mais plutôt des investissements académiques et humains.

En second lieu, il existe un point sur lequel j'aimerai revenir, celui du traitement local de l'information et des analyses systémiques. Mes travaux ont montré qu'on pouvait identifier, relativement efficacement, des déterminants neuronaux du comportement, et ceci par des méthodes génériques. Cependant, mes travaux, comme bien d'autres, font l'hypothèse d'un traitement purement local de l'information [19]. C'est-à-dire qu'ils s'appuient directement sur les niveaux des facteurs expérimentaux et pour produire le comportement observé. Or, il est bien improbable qu'une région comme le cortex préfrontal soit directement connectée aux capteurs sensoriels et aux muscles de la main. Le traitement cérébral de l'information s'effectue vraisemblablement de manière graduelle, sans sauter directement de l'environnement à une région cérébrale, puis à un muscle. L'information que reçoit une région cérébrale est celle émise par une autre région, dans son

19. Que ce soit par des analyses univariées ou multivariées.

langage propre, et après avoir déjà subi un traitement plus ou moins complexe par rapport aux stimuli sensoriels bruts. Certes, certaines manipulations expérimentales peuvent être remarquablement bien alignées aux mécanismes cérébraux et aux représentations neurales [20], mais cette correspondance est rarement formalisée. Rares sont les modèles qui proposent des traitements neuronaux complets et plausibles de la génération du comportement. Pourtant, de tels modèles permettraient d'affiner considérablement nos scénarios neurocognitifs, ainsi que leurs applications cliniques. Cela nous permettrait d'associer un mécanisme neuro-comportemental à chaque région, et de définir ce mécanisme à partir des régions précédant et suivant la région étudiée dans les chaînes cérébrales de traitement de l'information. Dans mes recherches futures, j'ai bon espoir d'intégrer cette perspective et d'élaborer de nouvelles méthodes génériques d'analyse de réseaux neuro-comportementaux.

20. Les représentations de la valeur dans le vmPFC par exemple.

## 11.9 MOT DE LA FIN

Au cours de cette thèse, j'ai appris énormément sur l'étude du cerveau, de la cognition en général et de la prise de décision en particulier. Au-delà de la quantification des relations cerveaux-comportement, je cherchais à comprendre comment étudier l'intelligence humaine. J'ai découvert qu'il n'existait pas une manière formelle d'aborder le problème, mais plutôt mille et une manières, chacune offrant un éclairage différent, en des termes sociaux, économiques, évolutionnaires, perceptuels, chimiques, statistiques, etc. Chacune possède ses forces, ses faiblesses et, surtout, ses biais conceptuels et méthodologiques. Laissant de côté l'attrait d'un formalisme unique permettant de décrire chaque aspect de l'intelligence, j'ai appris à apprécier la relativité scientifique des différents domaines des neurosciences cognitives et, plus généralement, des sciences expérimentales. En conclusion, l'étude des mécanismes cognitifs dépend du niveau de compréhension que l'on recherche [21].

Dans mon cas, il s'agit de modèles computationnels empreints de contraintes biologiques et d'interprétations comportementales. J'y vois un compromis intéressant entre la modélisation mathématique et les neurosciences cognitives. La première apporte sa rigueur quantitative et permet d'explorer de nombreux scénarios de mécanismes, la seconde permet l'interprétation des modèles et ancre la réflexion dans des données concrètes. Ce travail de thèse m'a permis d'élaborer deux méthodes combinant ces aspects et m'a donné l'opportunité d'évaluer empiriquement [22] leurs intérêts respectifs. Il m'a également permis d'étudier de très (trop ?) nombreux aspects des neurosciences cognitives, tous n'ayant malheureusement pas trouvé leur place dans cette thèse. Enfin, ces quatre années m'ont progressivement amené à cerner les limitations des approches actuelles des méthodes d'étude des relations entre cerveau et comportement. Plutôt qu'avoir répondu de manière définitive à une question, j'ai découvert un océan d'interrogations subsidiaires qu'il me tarde d'étudier en détail.

21. Chirimuuta, M. (2020). *Prediction versus understanding in computationally enhanced neuroscience*. Synthese

22. Par différentes modélisations, simulations et analyses IRMf et comportementales.

# *References*


Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, 31(2):790–795.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Curr. Opin. Neurobiol*, 12:169–177.

Aerts, H., Fias, W., Caeyenberghs, K., and Marinazzo, D. (2016). Brain networks under attack : robustness properties and the impact of lesions. *Brain J. Neurol*.

Aitchison, L. and Lengyel, M. (2017). With or without you : predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227.

Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Comput. Biol*, 5:1000408.

Anderson, M. (2010). Neural reuse : A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33:245–313.

Antonio, R. and Clithero, J. (2012). Value normalization in decision making : theory and evidence. *Current Opinion in Neurobiology*, 22(6):970–981.

Atlas, L., Bolger, N., Lindquist, M., and Wager, T. (2010). Brain mediators of predictive cue effects on perceived pain. *J. Neurosci. Off. J. Soc. Neurosci*, 30:12964–12977.

Atlas, L., Lindquist, M., Bolger, N., and Wager, T. (2014). Brain mediators of the effects of noxious heat on pain. *Pain*, 155:1632–1648.

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

Averbeck, B., Latham, P., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366.

Bandettini, P., Wong, E., Hinks, R., Tikofsky, R., and Hyde, J. (1992). Time course epi during task activation. *Magnetic Resonance in Medicine*, 25:390–397.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, page 216–234.

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research : conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol*, 51:1173–1182.

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4).

Bays, P. (2014). Noise in neural populations accounts for errors in working memory. *J. Neurosci*, 34:3632–3645.

Ben-Shabat, E., Matyas, T., Pell, G., Brodtmann, A., and Carey, L. (2015). The right supramarginal gyrus is important for proprioception in healthy and stroke-affected participants : A functional mri study. *Front. Neurol*, 6.

Borst, J., Taatgen, N., and Van Rijn, H. (2011). Using a symbolic process model as input for model-based fmri analysis : Locating the neural correlates of problem state replacements. *NeuroImage*, 58(1):137–147.

Botvinik-nezer, R., Holzmeister, F., Camerer, C., and Johannesson, M. (2019). Variability in the analysis of a single neuroimaging dataset by many teams.

Braver, T., Cole, M., and Yarkoni, T. (2010). Vive les differences ! individual variation in neural mechanisms of executive control. *Curr. Opin. Neurobiol*, 20:242–250.

Brenner, N., Bialek, W., and De Ruyter Van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702.

Brette, R. (2019). Is coding a relevant metaphor for the brain ? behavioral and brain sciences.

Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins de La Société d'anthropologie de Paris*, 6(1):377–393.

Brochard, J. and Daunizeau, J. (2020). Blaming blunders on the brain : can indifferent choices be driven by range adaptation or synaptic plasticity ? *BioRxiv*.

Burke, C., Baddeley, M., Tobler, P., and Schultz, W. (2016). Partial adaptation of obtained and observed value signals preserves information about gains and losses. *Journal of Neuroscience*, 36(39):10016–10025.

Buschman, T., Siegel, M., Roy, J., and Miller, E. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11252–11255.

Caton, R. (1875). The electric currents of the brain. *British Medical Journal*, 2(765).

Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., and Liu, T. (2014). Survey of encoding and decoding of visual stimulus via fmri : An image analysis perspective. *Brain Imaging and Behavior*, 8(1):7–23.

Chen, M.-Y. (2014). The development of bias in perceptual and financial decision-making.

Chen, O., Crainiceanu, C., Ogburn, E., Caffo, B., Wager, T., and Lindquist, M. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostat. Oxf. Engl*, 19:121–136.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136.

Chib, V., Rangel, A., Shimojo, S., and O'Doherty, J. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, 29(39):12315–12320.

Chirimuuta, M. (2020). *Prediction versus understanding in computationally enhanced neuroscience*. Synthese.

Christopoulos, G., Tobler, P., Bossaerts, P., Dolan, R., and Schultz, W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience*, 29(40):12574–12583.

Cohen, D. (1966). Magnetoencephalography : Evidence of mapetic fields produced by alpha-rhylhm currents. *Science*, 161.

Conen, K. and Padoa-Schioppa, C. (2019). Partial adaptation to the value range in the macaque orbitofrontal cortex. *Journal of Neuroscience*, 39(18):3498–3513.

Cox, K. and Kable, J. (2014). Bold subjective value signals exhibit robust range adaptation. *Journal of Neuroscience*, 34(49):16533–16543.

Dale, A. (1999). Optimal experimental design for event-related fmri. *Human Brain Mapping*, 8(23):109–114.

Daunizeau, J. (2017). The variational laplace approach to approximate bayesian inference.

Daunizeau, J. (2019). Variational bayesian modelling of mixed-effects. Icm). Retrieved from.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). Vba : A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, 10(1).

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., and Depaulis, A. (2008). Identifying neural drivers with functional mri : An electrophysiological validation. *PLoS Biol*, 6:315.

Dayan, P. and Daw, N. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive , Affective and Behavioral Neuroscience*, 8(4):429–453.

De Martino, B., Camerer, C., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792.

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. (2006). Frames, biases and rational decision-making in the human brain. *Science*, 313(5787):684–687.

de Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain ? *Psychonomic Bulletin  Review*, page 1–14.

Deco, G., Rolls, E., Albantakis, L., and Romo, R. (2013). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology*, 103:194–213.

Dehaene, S. and Cohen, L. (2007). *Cultural Recycling of Cortical Maps*. Neuron.

Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on granger causality analysis of fmri. *NeuroImage*, 52:884–896.

Dezfouli, A., Morris, R., Ramos, F. T., Dayan, P., and Balleine, B. (2018). Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models. In *Advances in Neural Information Processing Systems*, pages 4228–4237.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H., and Kriegeskorte, N. (2020). Comparing representational geometries using the unbiased distance correlation. Retrieved from.

Dinstein, I., Heeger, D., and Behrmann, M. (2015). Neural variability : friend or foe ? trends cogn. *Sci*, 19:322–328.

Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., and Varoquaux, G. (2020). Neuroquery : comprehensive meta-analysis of human brain mapping. Retrieved from.

Doi, E. and Lewicki, M. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Comput*, 23:2498–2510.

Donders, F. (1969). On the speed of mental processes. *Acta Psychologica*, 30(C):412–431.

Doya, K., Ishii, S., Pouget, A., and Rao, R. (XXXX). Bayesian brain : Probabilistic approaches to neural coding.

Drugowitsch, J., Wyart, V., Devauchelle, A., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6):1398–1411.

Eklund, A., Nichols, T., and Knutsson, H. (2016). Cluster failure : Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413.

Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., and Hariri, A. (2020). What is the test-retest reliability of common task-functional mri measures ? new empirical evidence and a meta-analysis. *Psychological Science*, 31(7):792–806.

Elliott, R., Agnew, Z., and Deakin, J. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, 27(9):2213–2218.

Evans, A., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., and Bub, D. (1992). Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage*, 1(1):43–53.

Fairchild, A. J. and MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, 10(2):87–99.

Faisal, A., Selen, L., and Wolpert, D. (2008). Noise in the nervous system. *Nat. Rev. Neurosci*, 9:292–303.

Ferster, D. (1996). Is neural noise just a nuisance ? *Science*, 273:1812–1812.

Fiorillo, C., Tobler, P., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902.

Flandin, G. and Friston, K. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, 40(7):2052–2054.

Forstmann, B., Dutilh, G., Brown, S., Neumann, J., Cramon, D., Ridderinkhof, K., and Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proc. Natl. Acad. Sci*, 105:17538–17542.

Fox, K. and Stryker, M. (2017). Integrating hebbian and homeostatic plasticity : Introduction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715).

Frank, M. (2006). Hold your horses : A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, 19(8):1120–1136.

Frank, M. J., Samanta, J., Moustafa, A. A., and Sherman, S. J. (2007). Hold your horses : impulsivity, deep brain stimulation, and medication in parkinsonism. *science*, 318(5854):1309–1312.

Friston, K. (2011). Functional and effective connectivity : A review. *Brain Connect*, 1:13–36.

Friston, K. (2012). The history of the future of the bayesian brain. *NeuroImage*, 62(2):1230–1233.

Friston, K., Holmes, A., Price, C., BÃ¼chel, C., and Worsley, K. (1999). Multisubject fmri studies and conjunction analyses. *NeuroImage*, 10:385–396.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the laplace approximation. *NeuroImage*, 34(1):220–234.

Friston, K., Penny, W., and Glaser, D. (2005a). Conjunction revisited. *NeuroImage*, 25:661–667.

Friston, K., Stephan, K., Lund, T., Morcom, A., and Kiebel, S. (2005b). Mixed-effects and fmri studies. *NeuroImage*, 24:244–252.

Friston, K. J., Diedrichsen, J., Holmes, E., and Zeidman, P. (2019). Variational representational similarity analysis. *NeuroImage*, 201:115986.

Galton, F. (1888). Co-relations and their measurement , chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45:135–145.

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning : A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(7):1297–1310.

Gbadeyan, O., McMahon, K., Steinhauser, M., and Meinzer, M. (2016). Stimulation of dorsolateral prefrontal cortex enhances adaptive cognitive control : A high-definition transcranial direct current stimulation study. *J. Neurosci*, 36:12530–12536.

Georgopoulos, A., Schwartz, A., and Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419.

Geuter, S., Losin, E., Roy, M., Atlas, L., Schmidt, L., Krishnan, A., Koban, L., Wager, T., and Lindquist, M. (2018). Multiple brain networks mediating stimulus-pain relationships in humans. *BioRxiv*, 298927.

Ghuman, A. and Martin, A. (2019). Dynamic neural representations : An inferential challenge for fmri. *Trends in Cognitive Sciences*, 23(7):534–536.

Gitelman, D., Penny, W., Ashburner, J., and Friston, K. (2003). Modeling regional and psychophysiologic interactions in fmri : the importance of hemodynamic deconvolution. *NeuroImage*, 19:200–207.

Glasser, M., Coalson, T., Robinson, E., Hacker, C., Harwell, J., Yacoub, E., and Van Essen, D. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, page 1–11.

Gläscher, J., Hampton, A., and O'Doherty, J. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19(2):483–495.

Gläscher, J. and O'Doherty, J. (2010). Model-based approaches to neuroimaging : Combining reinforcement learning theory with fmri data. *Wiley Interdisciplinary Reviews : Cognitive Science*, 1(4):501–510.

Guest, O. and Love, B. (2016). What the success of brain imaging implies about the neural code. *BioRxiv*, 1:071076.

Güçlü, U. and Gerven, M. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., Van Lier, R., and Van Gerven, M. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, page 4247–4258.

Haas, L. and Caton, R. (2003). Hans berger (1873-1941. *Journal of Neurology, Neurosurgery, and Psychiatry*, 74(1):9.

Hampton, A. and O'Doherty, J. (2007). Decoding the neural substrates of reward-related decision making with functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4):1377–1382.

Harlow, J. (1868). *Passage of an iron bar through the head*. Massachusetts Medical Society.

He, Y. and Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Curr. Opin. Neurol*, 23:341–350.

Hebb, D. (1949). *The organization of behavior : a neuropsychological theory*. J. Wiley ; Chapman Hall.

Henson, R. (2006). Forward inference using functional neuroimaging : Dissociations versus associations. *Trends in Cognitive Sciences*, 10(2):64–69.

Hoang, L. (2018). La formule du savoir.

Holmes, A., Friston, K., and Friston, K. (1998). Generalisability, random effects and population inference.

Hong, S. and Rebec, G. (2012). A new perspective on behavioral inconsistency and neural noise in aging : compensatory speeding of neural communication. *Front. Aging Neurosci*, 4.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2018). A benchmark for interpretability methods in deep neural networks. NeurIPS). Retrieved from.

Hopfinger, J., Büchel, C., Holmes, A., and Friston, K. (2000). A study of analysis parameters that influence the sensitivity of event- related fmri analyses. *NeuroImage*, 11(4):326–333.

Horn, J., Irimia, A., Torgerson, C., Chambers, M., Kikinis, R., and Toga, A. (2012). Mapping connectivity damage in the case of phineas gage. *PLoS ONE*, 7(5).

Hosoya, T., Baccus, S., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77.

Hsu, M., Krajbich, I., Zhao, C., and Camerer, C. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience*, 29(7):2231–2237.

Huffman, C. (2017). *Alcmaeon*. The Stanford Encyclopedia of Philosophy.

Jarius, S. and Wildemann, B. (2017). Pavlov's reflex before pavlov : Early accounts from the english, french and german classic literature. *European Neurology*, 77(5–6):322–326.

Jocham, G., Hunt, L., Near, J., and Behrens, T. (2014). A mechanism for value-guided choice based on the excitation- inhibition balance in prefrontal cortex. *Nature Neuroscience*, 15(7):960–961.

Kahneman, D. and Tversky, A. (1979). Prospect theory : An analysis of decision under risk. *Econometrika. Econometrica*, 47(2):263–292.

Kass, R. and Raftery, A. (1995). Bayes factors. *J. Am. Stat. Assoc*, 90:773–795.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.

Kietzmann, T., McClure, P., and Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *BioRxiv*, 133504.

Kim, H., Shimojo, S., and O'Doherty, J. (2006). Is avoiding an aversive outcome rewarding ? neural substrates of avoidance learning in the human brain. *PLoS Biology*, 4(8):1453–1461.

King, J. and Dehaene, S. (2014). Characterizing the dynamics of mental representations : The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

Kisbu-Sakarya, Y., MacKinnon, D. P., and Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate behavioral research*, 49(3):261–268.

Koban, L., Jepma, M., LÃ³pez-SolÃ , M., and Wager, T. (2019). Different brain networks mediate the effects of social and conditioned expectations on pain. *Nat. Commun*, 10:4096.

Koban, L., Kross, E., Woo, C.-W., Ruzic, L., and Wager, T. (2017). Frontal-brainstem pathways mediating placebo effects on social rejection. *J. Neurosci*, 37:3621–3631.

Kobayashi, S., De Carvalho, O., and Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience*, 30(2):534–544.

Krakauer, J., Ghazanfar, A., Gomez-Marin, A., MacIver, M., and Poeppel, D. (2017). Neuroscience needs behavior : Correcting a reductionist bias. *Neuron*.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November):1–28.

Kriegeskorte, N. and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 371(1705).

Kriegeskorte, N. and Golan, T. (2019). Neural network models and deep learning. *Current Biology*, 29(7):231– 236.

Kuhnen, C. and Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, 47(5):763–770.

KWONG, K., BELLIVEAU, J., CHESLER, D., GOLDBERG, I., WEISSKOFF, R., PONCELET, B., and ROSEN, B. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capactiy.

Levy, B. and Wagner, A. (2011). Cognitive control and right ventrolateral prefrontal cortex : reflexive reorienting, motor inhibition, and action updating. *Ann. N. Y. Acad. Sci*, pages 40–62.

Lewicki, M. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363.

Liao, C., Worsley, K., Poline, J.-B., Aston, J., Duncan, G., and Evans, A. (2002). Estimating the delay of the fmri response. *NeuroImage*, 16:593–606.

Lindquist, M. (2012a). Functional causal mediation analysis with an application to brain connectivity. *J. Am. Stat. Assoc*, 107:1297–1309.

Lindquist, M. and Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosom. Med*, 77:114.

Lindquist, M. A. (2012b). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309.

Lisman, J. (2017). Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes : Long-term potentiation, long-term depression, short-term potentiation and scaling. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715).

Liu, C. and Aitkin, M. (2008). Bayes factors : Prior sensitivity and model generalizability. *J. Math. Psychol*, 52:362–375.

Liu, J., Harris, A., and Kanwisher, N. (2010). *Perception of Face Parts and Face Configurations : An fMRI Study*, volume 22. Liu, J., Harris, A., Kanwisher, N.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157.

Lohmann, G., Müller, K., and Turner, R. (2013). Response to commentaries on our paper : Critical comments on dynamic causal modelling. *NeuroImage*, 75:279–281.

Louie, K. and Glimcher, P. (2012). Efficient coding and the neural representation of value. *Annals of the New York Academy of Sciences*, 1251(1):13–32.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

MacKinnon, D., Fairchild, A., and Fritz, M. (2007). Mediation analysis. *Annu. Rev. Psychol*, 58:593.

MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods*, 7:83–104.

MacKinnon, D., Lockwood, C., and Williams, J. (2004). Confidence limits for the indirect effect : Distribution of the product and resampling methods. *Multivar. Behav. Res*, 39:99.

Madsen, K. (1988). A history of psychology in metascientific perspective. In *Advances in Psychology*, volume 53, page 193–257. Retrieved from.

Makovac, E., Meeten, F., Watson, D., Garfinkel, S., Critchley, H., and Ottaviani, C. (2016). Neurostructural abnormalities associated with axes of emotion dysregulation in generalized anxiety. *NeuroImage Clin*, 10:172–181.

Marois, R. and Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, 9(6):296–305.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*.

Marrelec, G., Daunizeau, J., Pelegrini-Issac, M., Doyon, J., and Benali, H. (2005). Conditional correlation as a measure of mediated interactivity in fmri and meg/eeg. *IEEE Trans. Signal Process*, 53:3503–3516.

Martens, M., Celikel, T., and Tiesinga, P. (2015). A developmental switch for hebbian plasticity. *PLoS Computational Biology*, 11(7):1–19.

Martin, C., Martindale, J., Berwick, J., and Mayhew, J. (2006). Investigating neural-hemodynamic coupling and the hemodynamic response function in the awake rat. *NeuroImage*, 32:33–48.

Mather, M., Cacioppo, J., and Kanwisher, N. (2008). How fmri can inform cognitive theories. *Bone*, 23(1):1–7.

McClamrock, R. (1991). Marr ' s three levels : A re-evaluation. In *Minds and Machines*, page 185–196.

McDonnell, M. and Ward, L. (2011). The benefits of noise in neural systems : bridging theory and experiment. *Nat. Rev. Neurosci*, 12:415–426.

McGill, W. (1954). Multivariate information transmission. *Psychometrika*, 19:97–116.

Miller, E. and Buschman, T. (2015). Working memory capacity : Limits on the bandwidth of cognition. *Daedalus*, 144(1).

Moran, P. P. (1970). On asymptotically optimal tests of composite hypotheses. *Biometrika*, 57:47–55.

Nachev, P., Wydell, H., O'Neill, K., Husain, M., and Kennard, C. (2007). The role of the pre-supplementary motor area in the control of action. *Neuroimage*, 36:155– 163.

Naselaris, T., Kay, K., Nishimoto, S., and Gallant, J. (2011). Encoding and decoding in fmri. *NeuroImage*, 56:400–410.

Nee, D. and D'Esposito, M. (2017). Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *ELife*, 6:28040.

Neumann, V. and John Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, 25:653–660.

Nieder, A. and Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, 32:185–208.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.

Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation : Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5951–5955.

Olivi, E. (2011). *Coupling of numerical methods for the forward problem in Magneto- and Electro-Encephalography*. Nice - Sophia Antipolis.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*.

O'Doherty, J., Hampton, A., and Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014.

Palestro, J., Bahg, G., Sederberg, P., Lu, Z.-L., Steyvers, M., and Turner, B. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *J. Math. Psychol*, 84:20–48.

Pauling, L. and Coryell, C. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, 22(4):210–216.

Pavlov, I. (1903). The experimental psychology and psychopathology of animals. In *The 14th International Medical Congress*, page 23–30.

Pearl, J. (2012). The mediation formula : A guide to the assessment of causal pathways in nonlinear models. In *Causality*, pages 151–179. John Wiley  Sons, Ltd.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., and Gramfort, A. (2015). Data-driven hrf estimation for encoding and decoding models. *NeuroImage*, 104.

Penny, W. D. and Ridgway, G. (2013). Efficient posterior probability mapping using savage-dickey ratios. *PLoS ONE*, 8(3).

Penny, W. D., Stephan, K., Daunizeau, J., Rosa, M., Friston, K., Schofield, T., and Leff, A. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3).

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134:17–35.

Polania, R., Woodford, M., and Ruff, C. (2018). Efficient coding of subjective value. *BioRxiv*, 358317.

Poldrack, R. (2015). Is efficiency a useful concept in cognitive neuroscience ? dev. *Cogn. Neurosci*, 11:12–17.

Poldrack, R. (2018). The new mind readers : What neuroimaging can and cannot reveal about our thoughts.

Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., and Milham, M. (2013). Toward open sharing of task-based fmri data : the openfmri project. *Frontiers in Neuroinformatics*, 7(July):12.

Poldrack, R. and Yarkoni, T. (2016). From brain maps to cognitive ontologies : Informatics and the search for mental structure. *Annual Review of Psychology*, 67(1):587–612.

Popper, K. R. (1989). *Logik der Forschung*, volume 9. JCB Mohr Tübingen.

Posner, M., Petersen, S., Fox, P., and Raichle, M. (1988). Localization of cognitive operations in the human brain. *Science*, 240(4859):1627–1631.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132.

Power, J., Cohen, A., Nelson, S., Wig, G., AnneBarnes, K., Church, J., and Petersen, S. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Preacher, K. (2015). Advances in mediation analysis : A survey and synthesis of new developments. *Annu. Rev. Psychol*, 66:825–852.

Ramsey, N., Jansma, J., Jager, G., Van Raalten, T., and Kahn, R. (2004). Neurophysiological factors in human information processing capacity. *Brain*, 127(3):517–525.

Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex : A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Rigoux, L. and Daunizeau, J. (2015). Dynamic causal modelling of brain–behaviour relationships. *NeuroImage*, 117:202–221.

Ritchie, J., Kaplan, D., and Klein, C. (2019). Decoding the brain : Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607.

Rn Diedrichsen, J. and Kriegeskorte, N. (2017a). *Representational models : A common framework for understanding encoding, pattern-component, and representational- similarity analysis*. PLOS Computational Biology.

Rn Diedrichsen, J. and Kriegeskorte, N. (2017b). *Representational models : A common framework for understanding encoding, pattern-component, and representational- similarity analysis*. PLOS Computational Biology.

Robbins, T. (2011). Cognition : The ultimate brain function. *Neuropsychopharmacology*, 36:1–2.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. U. S. A*, 42:43–47.

Rouault, M., Drugowitsch, J., and Koechlin, E. (2019). Prefrontal mechanisms combining rewards and beliefs in human decision-making. *Nature Communications*, 10(1).

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity : Uses and interpretations. *NeuroImage*, 52:1059–1069.

Sandrone, S., Bacigaluppi, M., Galloni, M., Cappa, S., Moro, A., Catani, M., and Martino, G. (2014). Weighing brain activity with the balance : Angelo mosso's original manuscripts come to light. *Brain*, 137(2):621–633.

Savage, L. (1954). *The foundations of statistics*. John Wiley Sons.

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C., Tobler, P., and Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 363(1511):3801–3811.

Senden, M., Emmerling, T., Hoof, R., Frost, M., and Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, 224(3):1167–1183.

Seth, A., Chorley, P., and Barnett, L. (2013). Granger causality analysis of fmri bold signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–555.

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

Shadlen, M. and Newsome, W. (1994). Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol*, 4:569–579.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656.

Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23):941– 945.

Shouval, H., Wang, S., and Wittenberg, G. (2010). Spike timing dependent plasticity : A consequence of more fundamental learning rules. *Frontiers in Computational Neuroscience*, 4(July):1–13.

Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies : new procedures and recommendations. *Psychological methods*, 7(4):422.

Silani, G., Lamm, C., Ruff, C., and Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *J. Neurosci*, 33:15466–15476.

Silva, F. (2013). Eeg and meg : Relevance to neuroscience. *Neuron*, 80(5):1112–1128.

Simoncelli, E. and Olshausen, B. (2001). Natural image statistic and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., and Woolrich, M. (2011). Network modelling methods for fmri. *NeuroImage*, 54:875–891.

Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol*, 13:290–312.

Soltani, A., Martino, B., and Camerer, C. (2012). A range-normalization model of context-dependent choice : A new model and evidence. *PLoS Computational Biology*, 8(7).

Song, H. F., Yang, G. R., and Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife*, 6:e21492.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543–545.

Soutschek, A. and Tobler, P. (2020). *Causal role of lateral prefrontal cortex in mental effort and fatigue*. Hum. Brain Mapp.

Sporns, O. (2013). Making sense of brain network data. *Nat. Methods*, 10:491–493.

Stein, R., Gossen, E., and Jones, K. (2005). Neuronal variability : noise or part of the signal? *Nat. Rev. Neurosci*, 6:389–397.

Sutton, R. and Barto, A. (1999). Reinforcement learning : An introduction. *Trends in cognitive sciences*, 3(ttps://doi.org/10.1016/s1364-6613(99)01331-5).

Talairach, J. and Tournoux, P. (1988). *Co–Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System : An Approach to Cerebral Imaging*. Thieme Medical Publishers.

Ter Pogossian, M., Phelps, M., Hoffman, E., and Mullani, N. (1975). A positron emission transaxial tomograph for nuclear imaging (pett. *Radiology*, 114(1):89–98.

Thorndike, E. (1898). *Animal intelligence : an experimental study of the associative processes in animals*. Macmillan, New York.

Tom, S., Fox, C., Trepel, C., and Poldrack, R. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518.

Toyoizumi, T., Kaneko, M., Stryker, M., and Miller, K. (2015). Modeling the dynamic interaction of hebbian and homeostatic plasticity.

Trebaul, L., Deman, P., Tuyisenge, V., Jedynak, M., Hugues, E., Rudrauf, D., and David, O. (2018). Probabilistic functional tractography of the human cortex revisited. *NeuroImage*, 181(January):414–429.

Tu, Y., Tan, A., Bai, Y., Sam Hung, Y., and Zhang, Z. (2016). Decoding subjective intensity of nociceptive pain from pre-stimulus and post-stimulus brain activities. *Frontiers in Computational Neuroscience*, 10(APR):1–11.

Tunik, E., Lo, O.-Y., and Adamovich, S. (2008). Transcranial magnetic stimulation to the frontal operculum and supramarginal gyrus disrupts planning of outcome-based hand-object interactions. *J. Neurosci*, 28:14422–14427.

Turner, B., Palestro, J., MiletiÄ‡, S., and Forstmann, B. (2019a). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neurosci. Biobehav. Rev*, 102:327–336.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206.

Turner, B. M., Palestro, J. J., Miletić, S., and Forstmann, B. U. (2019b). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews*, 102:327–336.

Turrigiano, G. (2017). The dialectic of hebb and homeostasis. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715):4–6.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289.

Vaidya, A., Pujara, M., Petrides, M., Murray, E., and Fellows, L. (2019). Lesion studies in contemporary neuroscience. In *Trends in Cognitive Sciences*, volume 23.

van Kesteren, E.-J. and Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling : A Multidisciplinary Journal*, 26(5):710–723.

Vanrullen, R. and Reddy, L. (XXXX). Reconstructing faces from fmri patterns using deep generative neural networks, 31052.

Vinckier, F., Rigoux, L., Oudiette, D., and Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature communications*, 9(1):1–12.

Wager, T. (2008). canlab/ m3 mediationtoolbox (cognitive and affective neuroscience laboratory.

Wager, T., Ast, V., Hughes, B., Davidson, M., Lindquist, M., and Ochsner, K. (2009a). Brain mediators of cardiovascular responses to social threat, part ii : Prefrontal-subcortical pathways and relationship with anxiety. *NeuroImage*, 47:836–851.

Wager, T., Atlas, L., Lindquist, M., Roy, M., Woo, C., and Kross, E. (2013). An fmri-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397.

Wager, T., Davidson, M., Hughes, B., Lindquist, M., and Ochsner, K. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, 59:1037–1050.

Wager, T., Waugh, C., Lindquist, M., Noll, D., Fredrickson, B., and Taylor, S. (2009b). Brain mediators of cardiovascular responses to social threat : part i : Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. *NeuroImage*, 47:821–835.

Wagner, R. and R, A. (1972). A theory of classical conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, 21(6):64–99.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (XXXX). Reliability of dissimilarity measures for multi-voxel pattern analysis. Retrieved from.

Wark, B., Lundstrom, B., and Fairhall, A. (2008). Sensory adaptation. *Physiology*, 17(4):423–429.

Wasserman, L. (2004). All of statistics : A concise course in statistical inference brief contents. *Simulation*.

Wei, X. and Stocker, A. (2015). A bayesian observer model constrained by efficient coding can explain "anti-bayesian" percepts. *Nature Neuroscience*, 18(10):1509–1517.

Wernicke, C. (1970). The aphasic symptom-complex. a psychological study on an anatomical basis. *Archives of Neurology*, 22(3):280–282.

Whittington, J., Muller, T., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. (2019a). The tolman-eichenbaum machine : Unifying space and relational memory through generalisation in the hippocampal formation.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2019b). The tolman-eichenbaum machine : Unifying space and relational memory through generalisation in the hippocampal formation. *BioRxiv*, page 770495.

Woo, C.-W., Roy, M., Buhle, J., and Wager, T. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLOS Biol*, 13:1002036.

Wood, G., Nuerk, H., Sturm, D., and Willmes, K. (2008). Using parametric regressors to disentangle properties of multi-feature processes. *Behavioral and Brain Functions*, 4(i):1–12.

Worsley, K. and Friston, K. (1995). Analysis of fmri time-series revisited–again. *NeuroImage*, 2:173–181.

Xiaojing, W. (2008). Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234.

Yamamoto, D., Woo, C.-W., Wager, T., Regner, M., and Tanabe, J. (2015). Influence of dorsolateral prefrontal cortex and ventral striatum on risk avoidance in addiction : a mediation analysis. *Drug Alcohol Depend*, 149:10–17.

Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., and Wager, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670.

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural computation*, 10(2):403–430.

Zenke, F. and Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1715).

Zhang, S., Mano, H., Lee, M., Yoshida, W., Robbins, T., Kawato, M., and Seymour, B. (2017). The control of tonic pain by active relief learning. *BioRxiv*, 222653.

Zhao, Y. and Luo, X. (2017). Granger mediation analysis of multiple time series with an application to fmri.

Zimmermann, J., Glimcher, P., and Louie, K. (2018). Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nature Communications*, 9(1):1–11.